

Samples of My Publications

(To view a publication, please click its title)

Jianming Liang

- [M. M. Rahman Siddiquee, Zongwei Zhou, Ruibin Feng, Nima Tajbakhsh, Michael Gotway, Yoshua Bengio, and Jianming Liang. Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization. \[17 pages with supplementary materials\] \(submitted to CVPR-2019, a top conference in computer vision and machine learning\)](#)

Introduce a new concept called fixed-point translation in generative adversarial networks (GAN), result in a new GAN called Fixed-Point GAN, and lead to a novel method for disease detection and localization using only image-level annotation. Our Fixed-Point GAN outperforms the state of the art in image-to-image translation and in disease detection and localization by a large margin (this work was initiated when I was on sabbatical at Mila – Quebec Artificial Intelligence Institute led by Dr. Yoshua Bengio).

- [N. Tajbakhsh, J. Y. Shin, S. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*. 35\(5\):1299-312, 2016. \[16 pages with supplementary materials\] \(Impact Factor: 4.27; the best journal in medical imaging\)](#)

Systematically demonstrate the capability of transfer learning (fine-tuning) from natural images to biomedical images across diseases, imaging modalities, and medical specialties for the first time. This paper has received a great attention in the field and been consistently listed as one of the most popular papers of this top journal.

- [Z. Zhou, J. Shin, S. Gurudu, M. B. Gotway, and J. Liang. Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition \(CVPR'17\)*, Pages 7340–51. \[12 pages with supplementary materials\] \(a top conference in computer vision and pattern recognition; one of only five papers in biomedical imaging accepted by CVPR-2017\)](#)

Dramatically cut annotation cost further via a novel integration of active learning and transfer learning (fine-tuning) across diseases, imaging modalities, and medical specialties.

- [Z. Zhou, J. Shin, R. Feng, R. T. Hurst, C. B. Kendall, and J. Liang. Integrating Active Learning and Transfer Learning for Carotid Intima-Media Thickness \(CIMT\) Video Interpretation. *Journal of Digital Imaging* 2018. \[10 pages\] \(Impact Factor: 1.536; the best journal in imaging informatics\)](#)

Dramatically reduce annotation efforts for CIMT via an original concept of annotation unit and a new multi-class active fine-tuning.

- [Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *The Fourth Workshop on Deep Learning in Medical Image Analysis \(DLMIA 2018\)*. \[8 pages\] \(the best workshop focused on deep learning for medical image analysis\)](#)

Innovate skip connections, a key component of U-Net and FCNs, to boost segmentation accuracy substantially.

- [N. Tajbakhsh, M. Gotway, and J. Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. *Medical image computing and computer-assisted intervention, Lecture Notes in Computer Science*, 9350:62–9, 2015. \[8 pages\] \(Selected as a **Finalist for a Young Scientist Award** at MICCAI 2015, one of the two most prestigious conferences in biomedical image analysis\)](#)

Significantly boost deep learning performance via a novel image presentation for computer-aided pulmonary embolism detection in CT scans (our system is ranked #1 in the CAD PE competition).

- [J. Y. Shin, N. Tajbakhsh, R. T. Hurst, C. B. Kendall, and J. Liang. Automating carotid intima-media thickness video interpretation with convolutional neural networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition \(CVPR'16\)*, Pages 2526-35. \[14 pages with supplementary materials\] \(a top conference in computer vision and pattern recognition; one of only six papers in biomedical imaging accepted by CVPR-2016\)](#)

Fully automate the tedious CIMT (carotid intima-media thickness) video interpretation for the first time by hybridizing deep learning with novel pre- and post-processing.

- N. Tajbakhsh, S. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using context-aware shape features. *IEEE Transactions on Medical Imaging*, 35(2):630-44, 2016. [17 pages with supplementary materials] (Impact Factor: 4.27; the best journal in medical imaging)

Automatically detect polyps in colonoscopy using novel context-aware shape features.

- N. Tajbakhsh, S. Gurudu, and J. Liang. A comprehensive computer-aided polyp detection system for colonoscopy videos. *Information processing in medical imaging (IPMI-2015)*, Lecture Notes in Computer Science, 9123:327–38, 2015. [12 pages] (one of the two most prestigious conferences in medical image analysis)

Dramatically enhance polyp detection performance through a novel integration of deep learning.

- J. Bernal, N. Tajbakhsh, ... S. Gurudu, ... J. Liang, A. Histace. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Transactions on Medical Imaging*. 36(6):1231-49, 2017. [19 pages] (The idea polyp detection competition was conceived by me, and the ASU team shared first authorship with J. Bernal from Spain and the senior authorship with A. Histace from France) (Impact Factor: 4.27; the best journal in medical imaging)

Our system for polyp detection is ranked #1 in this competition.

- J. Liang, T. McInerney, and D. Terzopoulos. United snakes. *Medical Image Analysis*, 10(2):215-233, 2006. [19 pages] (Impact Factor; 5.356; the best journal in medical image analysis).

United Snakes” is a general-purpose, interactive image segmentation and analysis tool with many applications. It unifies the most popular snakes (active contours) variants within a comprehensive finite element framework and, via a hard constraint mechanism, further combines snakes with a complementary technique known as livewire. United Snakes improves the efficiency, accuracy, and reproducibility of interactive image segmentation, offering more flexible control while reducing the need for user interaction. This article captured the No. 1 spot in the *Medical Image Analysis* journal’s top 25 hottest papers, and it has also been frequently used by university professors in their courses taught at the graduate and advanced undergraduate levels.

Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization

Md Mahfuzur Rahman Siddiquee¹, Zongwei Zhou^{1,3}, Ruibin Feng¹, Nima Tajbakhsh¹

Michael B. Gotway², Yoshua Bengio³, and Jianming Liang^{1,3}

¹Arizona State University, Scottsdale, AZ 29590 USA; ²Mayo Clinic Arizona, Scottsdale, AZ 29590 USA

³Mila – Quebec Artificial Intelligence Institute, Montreal, QC H2S 3H1 Canada

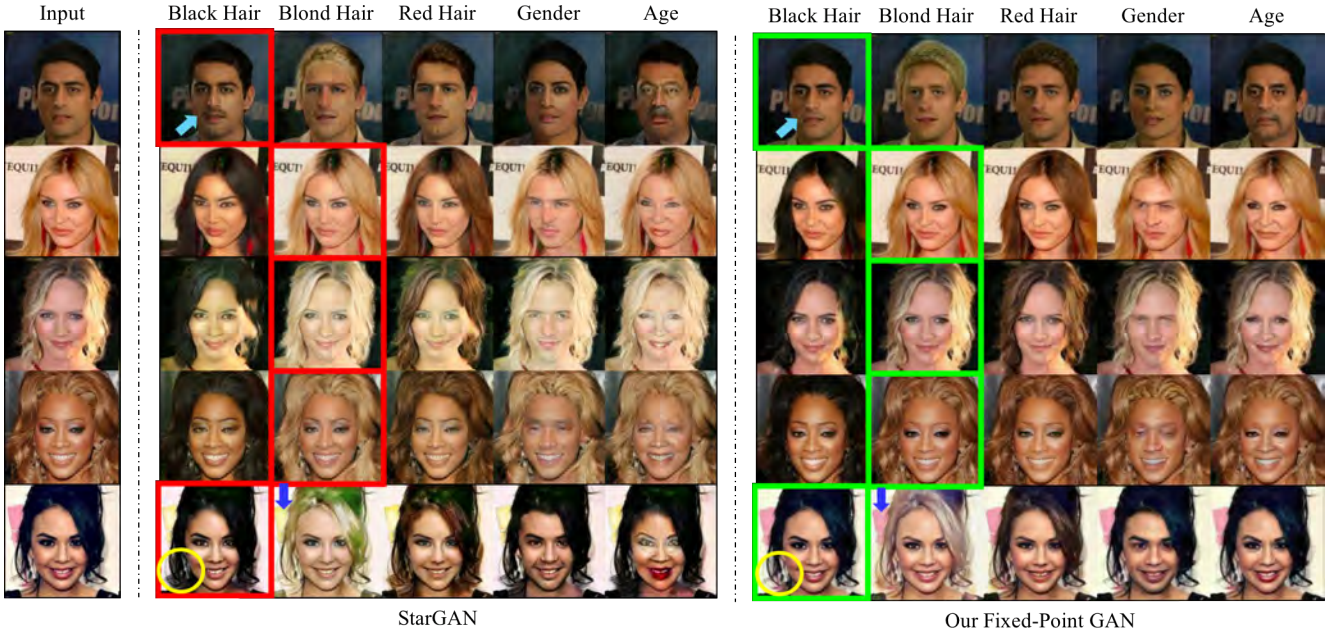


Fig. 1: Comparing our Fixed-Point GAN with StarGAN [5], the current state of the art in multi-domain image-to-image translation, by translating images into five domains. Combining the domains may yield a same-domain (e.g., black to black hair) or cross-domain (e.g., black to blond hair) translation. For clarity, same-domain translations are framed in red for StarGAN and in green for Fixed-Point GAN. As illustrated, during cross-domain translations, and especially during same-domain translations, StarGAN generates artifacts: introducing a mustache (Row 1, Col. 2; light blue arrow), changing the face colors (Rows 2–5, Cols. 2–6), adding more hair (Row 5, Col. 2; yellow circle), and altering the background (Row 5, Col. 3; blue arrow). Our Fixed-Point GAN overcomes these drawbacks via fixed-point translation learning (see Sec. 3) and leads to a framework for disease detection and localization with image-level annotation (see Fig. 2).

Abstract

Generative adversarial networks (GANs) have brought about a revolution in image-to-image translation. Now, can we train a GAN to remove an object, if present, from an image while otherwise preserving the image? Specifically, can a GAN “heal” a patient virtually by turning his image, diseased or healthy, into a healthy one, so that diseased regions could be revealed by subtracting those two images? Such a task requires a GAN to identify a minimal subset of target pixels for domain translation, an ability that we call fixed-point translation and that no GAN is equipped with yet. Therefore, we propose a new GAN, called Fixed-Point GAN, trained by (1) supervising same-domain trans-

lation through a conditional identity loss, and (2) regularizing cross-domain translation through a modified loss of adversarial, domain classification, and cycle consistency. Based on fixed-point translation, we further derive a novel framework for disease detection and localization using only image-level annotation. Qualitative and quantitative evaluations demonstrate that the proposed method outperforms StarGAN [5] in image-to-image translation and is superior in disease detection and localization by a large margin. Our method has the potential to exert significant clinical impact on computer-aided diagnosis in medical imaging, since training our Fixed-Point GAN requires only image-level annotation, which could be achieved by analyzing radiological reports via natural language processing.

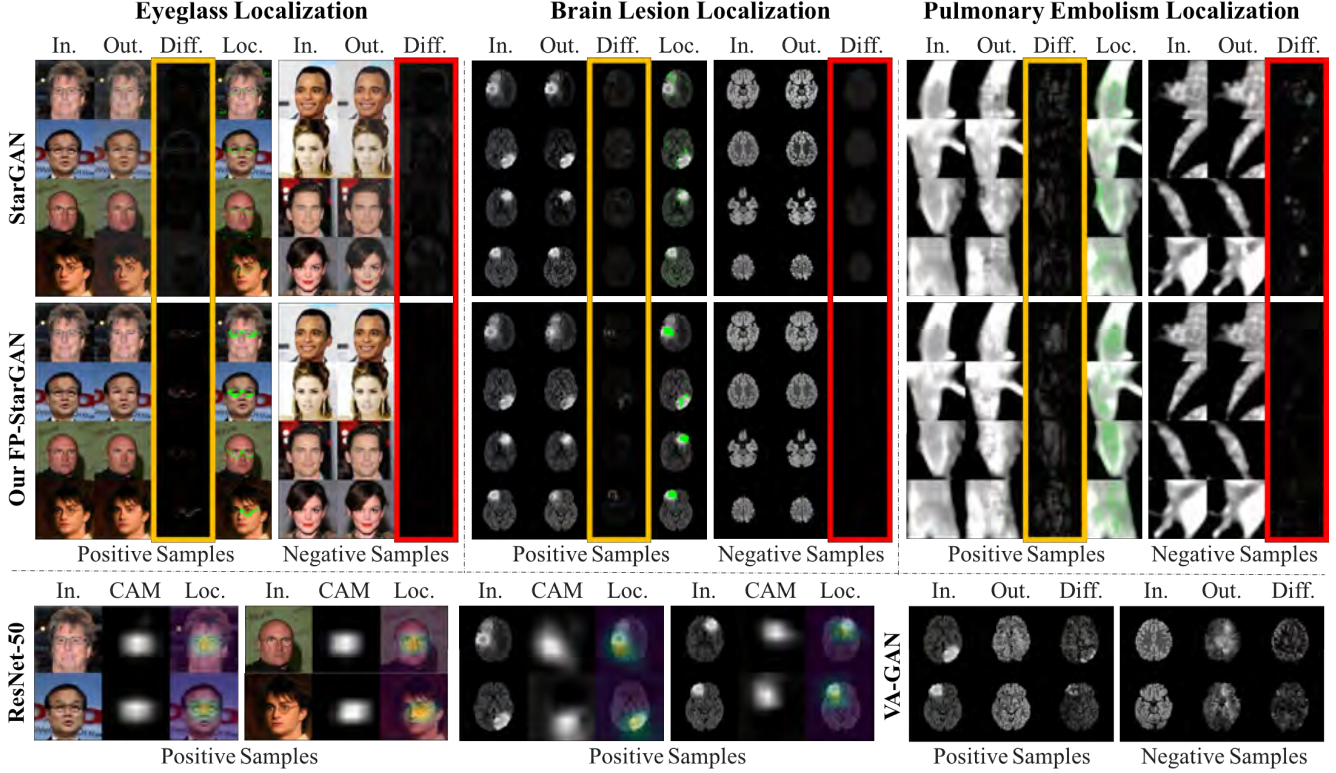


Fig. 2: [Better viewed on-line in color and zoomed in for details] Comparing Fixed-Point GAN with StarGAN for detecting and localizing eyeglasses and diseases with image-level annotation. Taking detecting diseases as an example, our idea is to translate any image, diseased or healthy, into a healthy image, so that diseased regions could be revealed by subtracting those two images. By fixed-point translation learning, our Fixed-Point GAN aims to preserve healthy images during the translation, thereby few differences between the generated (healthy) images and the original (healthy) images are observed in the difference maps (columns framed in red). For diseased images, thanks to the transformation learning from diseased images to healthy ones, disease locations are revealed in the difference maps (columns framed in yellow). For comparison, the localized diseased regions are superimposed on the original images (Loc. Columns), showing Fixed-Point GAN is more precise than CAMs [21] and VA-GAN [2] in localizing eyeglasses and diseases (bottom row; detailed in Sec. 5).

1. Introduction

Generative adversarial networks (GANs) [6] have proven to be powerful in image-to-image translation, such as changing the hair color, facial expression, and makeup of a person [5, 3], and converting MRI scans to CT scans [19]. Now, we start seeking to answer a generic question: *Can GANs remove an object, if present, from an image while otherwise preserving the image content?* Specifically, can we train a GAN to remove eyeglasses from any image of face with eyeglasses while maintaining unchanged those without eyeglasses? Or, can a GAN “heal” a patient on his medical image virtually¹? Such a task appears simple, but it actually demands the following four stringent requirements:

- **Req. 1:** The GAN must handle unpaired images. It may be too arduous to collect a perfect pair of photos of the same person with and without eyeglasses, and it would be too late to acquire a healthy image of a sick patient.

¹Virtual healing (see Fig. 6 in Appendix) turns an image (diseased or healthy) into a healthy image, thereby subtracting the two images reveals diseased regions.

- **Req. 2:** The GAN must require no source domain label when translating an image into a target domain (*i.e.*, source-domain-independent translation). For instance, a GAN trained for virtual healing aims to turn any image, with unknown health status, into a healthy one.
- **Req. 3:** The GAN must conduct an identity transformation for same-domain translation. For virtual healing, the GAN should leave a healthy image intact, injecting neither artifacts nor new information into the image.
- **Req. 4:** The GAN must perform a minimal image transformation for cross-domain translation. Changes should be applied to only the image attributes directly relevant to the translation task, with no impact on unrelated attributes. For instance, removing eyeglasses should not affect the remainder of the image (*e.g.*, the hair, face color, and background), or removing diseases from a diseased image should not impact the region of the image labeled as normal.

Currently, no single image-to-image translation method can meet all aforementioned requirements. The conventional

GANs for image-to-image translation [9], although successful, require paired images. CycleGAN [23] mitigates this limitation through cycle consistency, but it still requires two dedicated generators for each pair of image domains. Facing a scalability issue due to dedicated generators, CycleGAN also fails to support source-domain-independent translation: selecting the suitable generator requires labels for both the source and target domains. StarGAN [5] overcomes both limitations by learning one single generator for all domain pairs of interest. However, StarGAN has its own shortcomings. First, StarGAN tends to make unnecessary changes during cross-domain translation. As illustrated in Fig. 1, StarGAN tends to alter the face color while the goal of domain translation is to change the gender, age, or hair color in images from the CelebFaces dataset [15]. Second, StarGAN fails to competently handle same-domain translation. Referring to examples framed with the red boxes in Fig. 1, StarGAN needlessly adds a mustache to the face in Row 1, and unnecessarily alters the hair color in Rows 2–5, where an identity transformation is simply desired. These shortcomings may be acceptable for image-to-image translation in natural images, but in sensitive domains, such as medical imaging, they may lead to grievous consequences—unnecessary changes and artifacts may result in misdiagnosis. Furthermore, overcoming the above limitations is essential for adapting GANs for object/disease detection, localization, segmentation—and removal.

Therefore, we propose a novel GAN. We call it Fixed-Point GAN for its new fixed-point² translation ability, which allows the GAN to identify a minimal subset of pixels for domain translation. To achieve this ability, we devise a new training scheme to promote the fixed-point translation during training (Fig. 3-3) by (1) supervising same-domain translation through an additional conditional identity loss (Fig. 3-3B), and (2) regularizing cross-domain translation through a modified loss of adversarial (Fig. 3-3A), domain classification (Fig. 3-3A), and cycle consistency (Fig. 3-3C). Owing to its fixed-point translation ability, Fixed-Point GAN performs a minimal transformation for cross-domain translation and strives for an identity transformation for same-domain translation. Consequently, Fixed-Point GAN not only achieves better image-to-image translation for natural images but also offers a novel framework for disease detection and localization with only image-level annotation. Our experiments demonstrate that Fixed-Point GAN significantly outperforms StarGAN over multiple datasets for the tasks of image-to-image translation, disease detection, and disease localization. Formally, we make the following contributions:

1. We introduce a new concept: fixed-point translation,

²Mathematically, x is a fixed point of function $f(\cdot)$ if $f(x) = x$. We borrow the term to describe the pixels to be preserved when applying the GAN translation function.

leading to a new GAN: Fixed-Point GAN.

2. We devise a new scheme to train fixed-point translation by supervising same-domain translation and regularizing cross-domain translation.
3. We show that Fixed-Point GAN outperforms the original StarGAN for the task of image-to-image translation in both natural and medical images.
4. We derive a novel method for disease detection and localization using image-level annotation based on fixed-point translation learning.
5. We demonstrate that our disease detection and localization method based on Fixed-Point GAN is superior to its counterpart based on StarGAN.

Our method has the potential to exert important clinical impact on computer-aided diagnosis in medical imaging, because training our Fixed-Point GAN requires only image-level annotation. Obtaining image-level annotation is much easier than lesion-level annotation, as a large number of diseased and healthy images can be collected from the picture archiving and communication systems, and labeled at the image level by analyzing their radiological reports with NLP. With the availability of large databases of medical images and their corresponding radiological reports, we envision not only that Fixed-Point GAN will detect and localize diseases more accurately, but also that it may eventually be able to “cure”¹, thus segment diseases in the future.

2. Related Work

The literature of GANs [6] for image-to-image translation is wide and deep [9, 23, 10, 24, 14, 20, 5, 12]; therefore we limit our discussion to only the most relevant works. CycleGAN [23] has made a breakthrough in *unpaired* image-to-image translation via cycle consistency. The cycle consistency has proven to be effective in preserving object shapes in translated images, but it may not preserve other image attributes, such as color; therefore, when converting Monet’s painting to photos (a cross-domain translation), Zhu *et al.* [23] imposes an extra identity loss to preserve the colors of input images. However, the identity loss cannot be used in cross-domain translation in general, as it would limit the transformation power. For instance, it would make it impossible to translate black hair to blond hair. Therefore, unlike CycleGAN, we conditionally incorporate the identity loss only during fixed-point translation learning for same-domain translations. More severely, during inference, CycleGAN requires that the source domain be provided, thereby violating our Req. 2 as discussed in Sec. 1 and rendering it unsuitable for our purpose. StarGAN [5] empowers a single generator with a capability for *multi-domain* image-to-image translation, and does not require the source domain of the input image at inference time. However, StarGAN has its own shortcomings, which

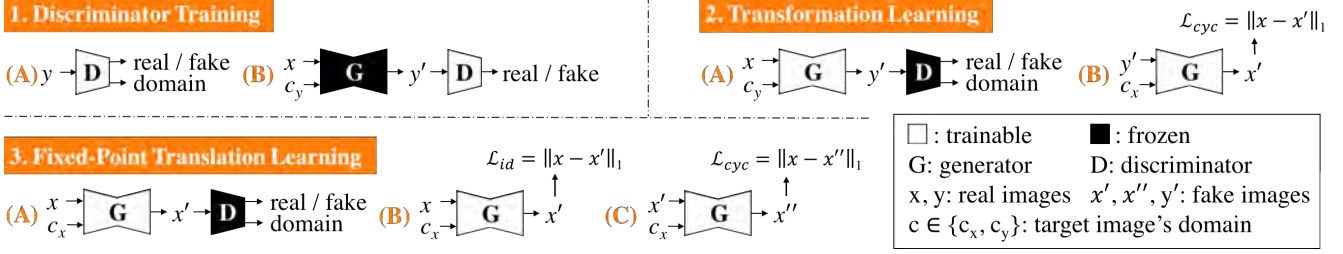


Fig. 3: Fixed-Point GAN training scheme. Like StarGAN, our discriminator learns to distinguish real/fake images (1A) and classify the domains of input images (1B). However, unlike StarGAN, our generator learns to perform not only cross-domain translations via transformation learning (2A–B), but also same-domain translations via fixed-point translation learning (3A–C), which is essential for mitigating the limitations of StarGAN (Fig. 1) and realizing disease detection and localization using only image-level annotation (Fig. 2).

violate Reqs. 3 and 4 as discussed in Sec. 1. Our Fixed-Point GAN overcomes StarGAN’s shortcomings, not only improving image-to-image translation but also opening the door to an innovative use of the generator as disease detector and localizer.

The literature regarding **GANs for disease detection and localization in medical images** is sparse [4, 2]. As far as we know, we are among the first to exploit GANs trained in image-to-image translation for disease detection and localization with image-level annotation. Chen *et al.* [4] train an autoencoder adversarially on only *healthy* brain images to detect brain *lesions*. The results reported by the authors show that the reconstructed brain image from an *abnormal* case cannot preserve its original anatomical structure. This can be explained because, essentially, their method learns only *one* translation: healthy to healthy. By contrast, our Fixed-Point GAN learns *all* 4 translations of diseased and healthy domains. Therefore, when translating an image (diseased or healthy) to a healthy version, Fixed-Point GAN preserves brain anatomy and leaves the fine details of normal brain structures intact, yielding cleaner difference maps (see Sec. 5.2 for details). Baumgartner *et al.* [2] developed VA-GAN to learn the difference between a healthy brain and one affected by Alzheimer’s disease. Although unpaired, VA-GAN requires that all images be registered. Without registration, it cannot preserve the brain structure as shown at the bottom right in Fig. 2. Furthermore, it requires the source-domain label at inference time, thus violating our Req. 2 as listed in Sec. 1.

3. Method

In the following, we present a high-level overview of Fixed-Point GAN, followed by a detailed mathematical description of each individual loss function.

Like StarGAN, our discriminator is trained to classify an image as real/fake and its associated domain (Fig. 3-1). In our new training scheme, the generator learns both cross- and same-domain translation, which differs from StarGAN, wherein the generator only learns the former. Mathemati-

cally, for any input x from domain c_x and target domain c_y , the generator in StarGAN learns to perform cross-domain translation ($c_x \neq c_y$), $G(x, c_y) \rightarrow y'$, where y' is the image in domain c_y . The generator in Fixed-Point GAN, in addition to learning the cross-domain translation, learns to perform the same-domain translation as $G(x, c_x) \rightarrow x'$.

Our new fixed-point translation learning (Fig. 3-3) not only enables same-domain translation but also regularizes cross-domain translation (Fig. 3-2) by encouraging the generator to find a minimal transformation function, thereby penalizing changes unrelated to the present domain translation task. Trained for only cross-domain image translation, StarGAN cannot benefit from such regularization, resulting in many artifacts as illustrated in Fig. 1. Consequently, our new training scheme offers three advantages: (1) reinforced same-domain translation, (2) regularized cross-domain translation, and (3) source-domain-independent translation. To realize these advantages, we define the loss functions of Fixed-Point GAN as follows:

Adversarial Loss. In the proposed method, the generator learns the cross- and same-domain translations. To ensure the generated images look realistic in both scenarios, the adversarial loss is updated as follows:

$$\mathcal{L}_{adv} = \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c} [\log(1 - D_{real/fake}(G(x, c)))] + \mathbb{E}_x [\log D_{real/fake}(x)] \quad (1)$$

Domain Classification Loss. The adversarial loss ensures the generated images look realistic, but it cannot guarantee domain correctness. As a result, the discriminator is trained with an additional domain classification loss, which forces the generated images to be of the correct domain. The domain classification loss for the discriminator is identical to that of StarGAN,

$$\mathcal{L}_{domain}^r = \mathbb{E}_{x,c_x} [-\log D_{domain}(c_x|x)] \quad (2)$$

but we have updated the domain classification loss for the generator to account for both same- and cross-domain trans-

lations, ensuring that the generated image is from the correct domain in both scenarios:

$$\mathcal{L}_{domain}^f = \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c} [-\log D_{domain}(c|G(x, c))] \quad (3)$$

Cycle Consistency Loss. Optimizing the generator with only the adversarial loss has multiple possible but random solutions. The additional *cycle consistency loss* (Eq. 4) helps the generator to learn a transformation that can preserve enough input information, such that the generated image can be translated back to original domain. Our modified *cycle consistency loss* ensures that both cross- and same-domain translations are cycle consistent.

$$\mathcal{L}_{cyc} = \mathbb{E}_{x, c_x, c_y} [\|G(G(x, c_y), c_x) - x\|_1] + \mathbb{E}_{x, c_x} [\|G(G(x, c_x), c_x) - x\|_1] \quad (4)$$

Conditional Identity Loss. StarGAN [5] during training focuses on translating the input image to different target domains. This strategy cannot penalize the generator when it changes aspects of the input that are irrelevant to match target domains (Fig. 1). Beside learning a translation to different domains, we force the generator using the *conditional identity loss* (Eq. 5) to preserve the domain identity while translating the image to the source domain. This also helps the generator learn a minimal transformation for translating the input image to the target domain.

$$\mathcal{L}_{id} = \begin{cases} 0, & c = c_y \\ \mathbb{E}_{x,c} [\|G(x, c) - x\|_1], & c = c_x \end{cases} \quad (5)$$

Full Objective. Combining all losses, the final full objective function for discriminator and generator can be described by Eq. 6 and Eq. 7, respectively.

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{domain} \mathcal{L}_{domain}^r \quad (6)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{domain} \mathcal{L}_{domain}^f + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_{id} \mathcal{L}_{id} \quad (7)$$

where λ_{domain} , λ_{cyc} , and λ_{id} determine the relative importance of the *domain classification loss*, *cycle consistency loss*, and *conditional identity loss*, respectively. Tab. 1 summarize the loss functions of Fixed-Point GAN.

4. Implementation

We revise adversarial loss (Eq. 1) based on the Wasserstein GAN [1] objective by adding gradient penalty [7] to stabilize the training. The gradient penalty coefficient is set to 10 for all experiments. Values for λ_{domain} , λ_{cyc} is used 1 and 10 respectively for all experiments. λ_{id} is set to 10 for CelebA, 0.1 for BRATS 2013, and 1 for PE dataset. 200K iteration is found to be sufficient for CelebA and PE dataset where BRAT 2013 requires 300K iteration for generating good quality images. For a fair comparison, we use

Loss	Definition
Eq. 1 \mathcal{L}_{adv}	$= \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c} [\log(1 - D_{r/f}(G(x, c)))] + \mathbb{E}_x [\log D_{r/f}(x)]$
Eq. 2 \mathcal{L}_{domain}^r	$= \mathbb{E}_{x, c_x} [\log D_{domain}(c_x x)]$
Eq. 3 \mathcal{L}_{domain}^f	$= \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c} [-\log D_{domain}(c G(x, c))]$
Eq. 4 \mathcal{L}_{cyc}	$= \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x, c_x, c} [\ G(G(x, c), c_x) - x\ _1]$
Eq. 5 \mathcal{L}_{id}	$= \mathbb{E}_{x,c} [\ G(x, c) - x\ _1] \text{ if } c = c_x; 0 \text{ otherwise}$
Eq. 6 \mathcal{L}_D	$= -\mathcal{L}_{adv} + \lambda_{domain} \mathcal{L}_{domain}^r$
Eq. 7 \mathcal{L}_G	$= \mathcal{L}_{adv} + \lambda_{domain} \mathcal{L}_{domain}^f + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_{id} \mathcal{L}_{id}$

Tab. 1: Loss functions in Fixed-Point GAN. Terms inherited from StarGAN are in black, while highlighted in blue are our modifications to mitigate StarGAN’s limitations (Fig. 1) and realize disease detection and localization using image-level annotation (Fig. 2).

the same generator and discriminator architectures as the public implementation of StarGAN. All models are trained using the Adam optimizer with learning rate $1e^{-4}$ for both generator and discriminator across all experiments.

Following [2], we slightly change the architecture of the generator to predict a residual (delta) map rather than the desired image directly. Specifically, the generator’s output is computed by adding the delta map to the input image followed by the application of a *tanh* activation function, $\tanh(G(x, c) + x)$. Our ablation study, summarized in Tab. 2, shows the disease detection and localization performance of StarGAN (baseline approach), and the incremental performance improvement using the delta map learning, fixed-point translation learning, and the two approaches combined. We find that the major improvement over StarGAN comes from the fixed-point translation learning, but the combined approach, for most cases, provides enhanced performance compared to each individual approach. We therefore use the combination of delta map learning and fixed-point translation learning in our proposed Fixed-Point GAN, noting that the major improvement over StarGAN is due to the proposed fixed-point translation learning scheme. The implementation will be made publicly available at http://github.com/****/****

5. Applications

5.1. Multi-Domain Image-to-Image Translation

Dataset. To compare the proposed Fixed-Point GAN with StarGAN [5] (the current state of the art), we use CelebFaces Attributes (CelebA) dataset [15]. This dataset is composed of a total of 202,599 face images of various celebrities. Each face image has 40 different attributes, following StarGAN’s public implementation [5] we adopt 5 domains (black hair, blond hair, brown hair, male, and young) in our experiments and preprocess the images by cropping the original 178×218 images into 178×178 and then re-scaling to 128×128 . We use a random subset of 2,000 samples for testing and the rest for training.

Dataset	Image-Level Detection (AUC)				Lesion-Level Localization Sensitivity		
	StarGAN	w/ Delta	w/ Fixed-Point Translation	w/ Both	StarGAN	w/ Fixed-Point Translation	w/ Both
BRATS	0.4611	0.5246	0.9980	0.9831	0.2500	0.8376	0.8449
PE	0.8832	0.8603	0.9216	0.9668	0.5833	0.7778	0.8056

Tab. 2: Ablation study of the generator’s configuration on brain lesion (BRATS2013) and pulmonary embolism (PE) detection. Selected combinations are in bold. The empirical results show that the performance gain is largely due to fixed-point translation learning—the contribution of delta map is minor and application dependent (detailed in Sec. 4)

StarGAN	Our Fixed-Point GAN	Autoencoder
2.40 ± 1.24	0.36 ± 0.35	0.11 ± 0.09

Tab. 3: Image-level L_1 distance comparison for same-domain translation. The ideal L_1 distance for same-domain translation should be zero, however, deep autoencoder networks have inevitable reconstruction error. With our fixed-point translation learning, we try to minimize unnecessary image manipulations for both cross- and same-domain translations. We show that Fixed-Point GAN minimizes the same-domain translation difference map dramatically and is very close to the underlying generator’s reconstruction lower bound as opposed to StarGAN.

Method and Evaluation. We evaluate the cross-domain image translation qualitatively by changing one attribute (such as hair color, gender, or age) from the source domain at a time. This step-wise evaluation facilitates tracking changes to image content. We further evaluate the same-domain image translation both qualitatively and quantitatively by measuring image-level L_1 distance between the input and translated images.

Results. Fig. 1 presents a qualitative comparison between StarGAN and Fixed-Point GAN for multi-domain image-to-image translation. For the cross-domain image translation, StarGAN tends to make unnecessary changes, such as altering the face color when the goal of translation is to change the gender, age, or hair color (Rows 2–5 in Fig. 1). Fixed-Point GAN, however, preserves the face color while successfully translating the images to the target domains. Furthermore, Fixed-Point GAN preserves the image background (marked with a blue arrow in Row 5 of Fig. 1), but StarGAN fails to do so.

For the same-domain image translation, the superiority of Fixed-Point GAN over StarGAN is even more striking. As shown in Fig. 1, Fixed-Point GAN effectively keeps the image content intact (images outlined in green) while StarGAN undesirably changes the image content (images outlined in red). For instance, the input image in the fourth row of Fig. 1 is from the domains of blond hair, female, and young. The same domain translation with StarGAN results in an image in which the hair and face colors are significantly altered. Although this color is closer to the average blond hair color in the dataset, it is far from that in the input image. Fixed-Point GAN with fixed-point translation ability handles this problem properly. Further qualitative

comparisons between StarGAN and Fixed-Point GAN are provided in the supplementary material.

Tab. 3 presents a quantitative comparison between StarGAN and Fixed-Point GAN for the task of same-domain image translation. We use the image-level L_1 distance between the input and generated images as the performance metric. To further gain insights into the comparison, we have included a dedicated autoencoder model that has the same architecture as the generator used in StarGAN and Fixed-Point GAN. As seen, the dedicated autoencoder has an image-level L_1 reconstruction error of 0.11 ± 0.09 , which can be deemed as a technical lower bound for the reconstruction error. The proposed Fixed-Point GAN dramatically reduces the reconstruction error of StarGAN from 2.40 ± 1.24 to 0.36 ± 0.35 . Our quantitative comparisons are in-line with the qualitative results shown in Fig. 1.

5.2. Brain Lesion Detection and Localization with Image-Level Annotation

Dataset. We extend Fixed-Point GAN from an image-to-image translation method to a brain lesion detection and localization method where only image-level annotation are available. As a proof of concept, we use the BRATS 2013 dataset [16, 11]. The main dataset is divided into real and synthetic images. We use the latter in our experiments since the images show more fine details compared to the former; thus, they are more sensitive to small changes. The synthetic images are further divided into two categories: (1) high-grade gliomas (HG), and (2) low-grade gliomas (LG). Each of these categories has 25 patients and for each patient, FLAIR, T1, T2, and post-Gadolinium T1 Magnetic Resonance (MR) images are available. To ease the analysis, we keep the input features consistent by taking only one type (FLAIR) of MR images for all patients from both HG and LG categories, resulting in a total of 9,050 MR slices. We further preprocess the dataset by removing all slices which are blank or have very little brain information. Finally, we randomly select 40 patients including 5,827 slices for training and 10 patients with 1,461 slices for testing. During training, we set aside one batch of the random samples from training dataset for validation. We pad the slices to 300×300 and then center crop to 256×256 ensuring the brain structures appear in the center of the image. Each pixel of MR slices are annotated with one of the five possi-

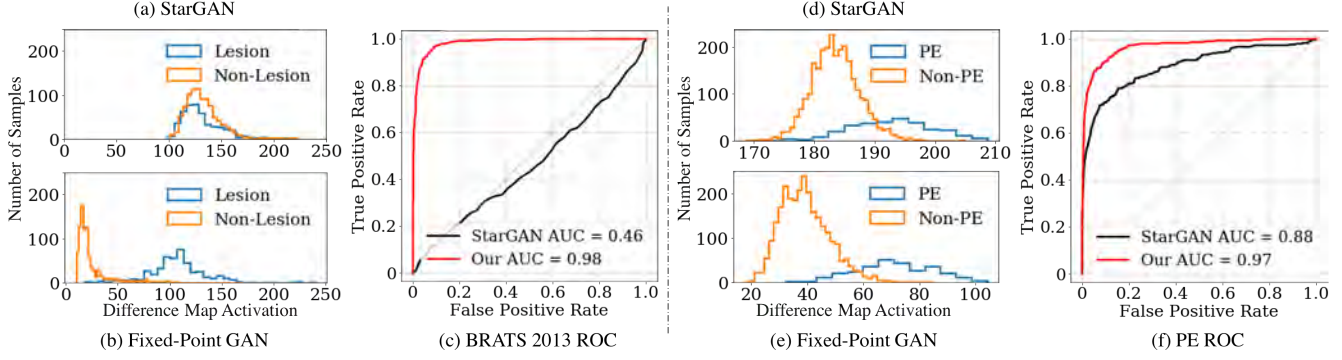


Fig. 4: Performance comparisons of our Fixed-Point GAN with StarGAN on BRATS 2013 (a–c) and Pulmonary Embolism (PE) dataset (d–f). We take the maximum value from the difference map and plot histograms (a–b) and (d–e) across all the test images. Comparing (a) with (b) and (d) with (e) reveals that our Fixed-Point GAN is better than StarGAN in separability between healthy and diseased images, naturally yielding a significant performance gain for image-level disease detection as shown in ROC curves (c) and (f).

ble labels: 1 for non-brain, non-tumor, necrosis, cyst, hemorrhage, 2 for surrounding edema, 3 for non-enhancing tumor, 4 for enhancing tumor core, and 0 for everything else. We group the MR slices with pixels annotated as only label 0 as the healthy domain and the others are associated with the diseased domain.

Method and Evaluation. For training we use only image-level annotation (healthy/diseased). The model is trained for the cross-domain translation (diseased images to healthy images and vice versa) as well as the same-domain translation using the proposed method. At inference time, we focus on translating any images to the healthy domain. The desired behaviour for the GAN is to translate diseased images to healthy ones while keeping healthy images intact. Having translated the images to the healthy domain, we detect the presence and location of lesion in the difference image by subtracting the translated healthy image from the input image. We refer to the resultant difference image as the *difference map*.

We evaluate the difference map at two different levels: (1) image-level disease detection and (2) lesion-level localization. For image-level detection, we take the maximum value across all pixels in the difference map as the *detection score*. We then perform receiver operating characteristics (ROC) curves using the resultant detection scores. For the lesion-level localization task, we first binarize the difference maps using color quantization followed by clustering the foreground pixels into the connected components. Each connected component with an area larger than ten pixels is considered as a lesion candidate. A lesion is considered “detected” if the centroid of at least a lesion candidate falls inside the lesion ground truth.

Results. Fig. 4a–b shows the histograms of maximum values of the difference map from StarGAN (Fig. 4a) and the proposed method (Fig. 4b) for the diseased and healthy images. From Fig. 4a it is clear that the distributions of healthy and diseased images are not distinguishable, suggesting that

StarGAN manipulates both diseased and healthy images in a similar manner, irrespective of their domains. As a result, the difference maps from StarGAN cannot be used for disease detection. On the other hand, with the proposed Fixed-Point GAN, the distributions of the detection scores for the healthy and diseases images have far less overlap, suggesting that the generator of Fixed-Point GAN manipulates the images differently depending on their source domains. Fig. 4c compares ROC curves of both methods for image-level lesion detection. Fixed-Point GAN achieves an area under the curve (AUC) score of 0.9831, which is far superior to StarGAN with an AUC of 0.4611. We further compare StarGAN and our proposed method for the lesion-level localization problem. For StarGAN, the lesion detection sensitivity is only 0.1358 while the proposed Fixed-Point GAN successfully achieves a sensitivity of 0.8122 at 1 false positive per image (Fig. 5a). Qualitative comparison between StarGAN and Fixed-Point GAN for brain lesion detection and localization is provided in Fig. 2. More examples are available in the supplementary material.

5.3. Pulmonary Embolism Detection and Localization with Image-Level Annotation

Dataset. Pulmonary embolism (PE) is a blood clot that travels from a lower extremity source to the lung, where it causes blockage of the pulmonary arteries. It is a major national health problem, but computer-aided PE detection and localization can improve diagnostic capabilities of radiologists for the detection of this disorder, leading to earlier and effective therapy for this potentially deadly disorder. We utilize a database consisting of 121 computed tomography pulmonary angiography (CTPA) scans with a total of 326 emboli. The dataset is organized exactly in the same manner as suggested in [22, 18]. A candidate generator [13] is first applied to generate a set of PE candidates, and then by comparing against the ground truth, they are labeled as PE and non-PE. Finally, a 2D patch of size 15×15 mm is

extracted around each PE candidate according to a vessel-aligned image representation [17]. As a result, PE appears at the center of the PE images. The extracted images are rescaled to 128×128 . The dataset is divided at the patient-level into a training set with 434 PE images (199 unique PEs) and 3,406 non-PE images, and a test set with 253 PE images (127 unique PEs) and 2,162 non-PE images. To enrich the training set, rotation-based data augmentation is applied for both PE and non-PE images.

Method and Evaluation. As with brain lesion detection and localization (Sec. 5.2), we use only image-level annotations during training. At inference time, we always remove PE from the input image (*i.e.* translating both PE and non-PE images into the non-PE domain) irrespective of whether PE is present in the input image. We follow the same procedure described in Sec. 5.2 to generate the difference maps, detection scores, and ROC curves. Note that, since each PE image has an embolus in its center, an embolus is considered as “detected” if the corresponding PE image is correctly classified; otherwise, “missed”. As such, unlike Sec. 5.2, we do not pursue a connected component analysis for PE localization.

Results. Our proposed Fixed-Point GAN outperforms StarGAN in both PE image-level detection and lesion-level localization. Fig. 4d–e show the distributions of the detection scores from StarGAN and Fixed-Point GAN. Despite the similar shapes, the overlapping area between the PE and non-PE distributions of Fixed-Point GAN is much smaller than that of StarGAN. The superiority of Fixed-Point GAN is better reflected in Fig. 4f, where Fixed-Point GAN achieves an AUC of 0.9668 while StarGAN’s AUC is only 0.8832 for image-level PE detection. Fixed-Point GAN also outperforms StarGAN in lesion-level localization by achieving a sensitivity of 0.8056 with 2.94 false positives whereas StarGAN achieves a sensitivity of 0.4167 at 2.95 false positives per volume. The qualitative comparison for PE removal between StarGAN and Fixed-Point GAN is given in Fig. 2. For further analysis, Free-Response ROC curves are provided in Fig. 5b.

6. Discussions

For detecting diseases, one could train an image classifier using weak image-level annotation, and then use the class activation map (CAM) [21] for localization. For comparison, we trained a ResNet-50 [8], with ~ 25 M parameters, to detect and localize eyeglasses and brain lesions. As shown in the bottom panel of Fig. 2, the resulting CAMs have low precision, covering both the regions containing the eyeglasses/lesions and the unrelated background regions. Although the low-precision CAMs may be suitable for image-level brain lesion detection, they impede precise localization of the brain lesions. Quantitatively, ResNet-50 and Fixed-Point GAN achieve an AUC score of 0.9846

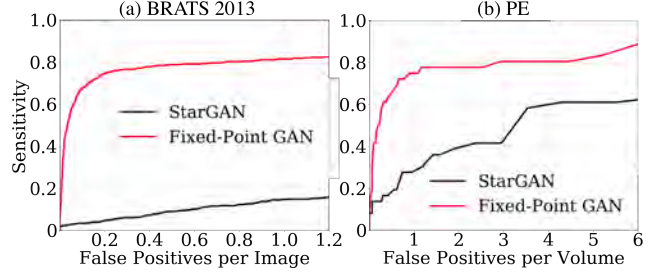


Fig. 5: FROCs of Fixed-Point GAN and StarGAN for lesion-level localization on (a) BRATS 2013 and (b) PE dataset.

and 0.9831 respectively for image-level brain lesion detection, but the performance gap widens for brain lesion localization with ResNet-50 and Fixed-Point GAN, respectively yielding a mean of IOU (intersection over union) of 0.1855 ± 0.0560 and 0.3483 ± 0.2420 . Our results suggest that Fixed-Point GAN is more suitable than CAMs for object localization with image-level annotation.

In Fig. 4c, we show that StarGAN performs poorly for image-level brain lesion detection, because StarGAN is designed to perform general-purpose image translations, rather than an image translation suitable for the task of disease detection. Owing to our new training scheme, Fixed-Point GAN can achieve precise detection.

Referring to Tab. 2, we observe that StarGAN performs far better for PE than brain lesion detection. We believe this is because brain lesions can appear anywhere in the input images, whereas PE always appears in the center of the input images, resulting in a less challenging problem for StarGAN to solve. Nonetheless, Fixed-Point GAN outperforms StarGAN for PE detection, achieving an AUC score of 0.9668 compared to 0.8832 by StarGAN.

Referring to Fig. 2, we further observe that neither StarGAN nor Fixed-Point GAN can completely remove large objects, like sunglasses or brain lesions, from the images. Nevertheless, for image-level detection and lesion-level localization, it is sufficient to remove the objects partially, but precise lesion-level segmentation using image-to-image translation network requires complete removal of the object. We will study this in our future work.

7. Conclusion

In this paper, we have introduced a new concept called fixed-point translation, and developed a new GAN called Fixed-Point GAN. Our comprehensive evaluation demonstrates that our Fixed-Point GAN outperforms StarGAN [5] in image-to-image translation and is superior in disease detection and localization by a large margin. This performance of Fixed-Point GAN is attributed to our new training scheme, realized by supervising same-domain translation and regularizing cross-domain translation.

References

- [1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 5
- [2] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu. Visual feature attribution using wasserstein gans. In *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2017. 2, 4, 5
- [3] H. Chang, J. Lu, F. Yu, and A. Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [4] X. Chen and E. Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018. 4
- [5] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. The implementation is publicly available at <https://github.com/yunjey/StarGAN>. 1, 2, 3, 5, 8
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 3
- [7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 5
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8
- [9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 3
- [10] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017. 3
- [11] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler. The virtual skeleton database: An open access repository for biomedical research and collaboration. *J Med Internet Res*, 15(11):e245, Nov 2013. 6
- [12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017. 3
- [13] J. Liang and J. Bi. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in ct pulmonary angiography. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 630–641. Springer, 2007. 7, 14
- [14] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 3
- [15] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 3, 5
- [16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015. 6
- [17] N. Tajbakhsh, M. B. Gotway, and J. Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 62–69. Springer, 2015. 8
- [18] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 7
- [19] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017. 2
- [20] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017. 3
- [21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 8
- [22] Z. Zhou, J. Y. Shin, L. Zhang, S. R. Gurudu, M. B. Gotway, and J. Liang. Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In *CVPR*, pages 4761–4772, 2017. 7
- [23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 3
- [24] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017. 3

Appendix

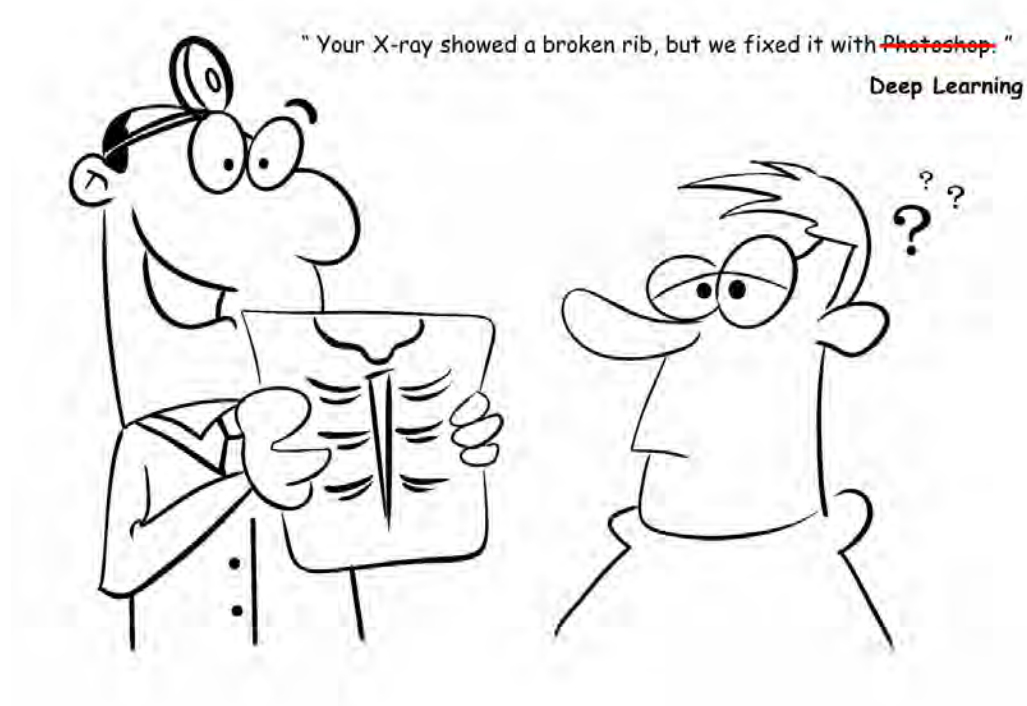


Fig. 6: Now, all humor aside, if a GAN can remove diseases *completely* from images, it will offer a *perfect* method for segmenting *all* diseases, an ambitious goal that has yet to achieve. Nevertheless, our method is still of great clinical significance in computer-aided diagnosis for medical imaging, because it can be used for disease detection and localization by subtracting the generated healthy image from the original image to reveal diseased regions, that is, **detection and localization by removal**. More importantly, our Fixed-Point GAN is trained using *only* image-level annotation. It is much easier to obtain image-level annotation than lesion-level annotation, because a large number of diseased and healthy images can be collected from PACS (picture archiving and communication systems), and labeled at the image level by analyzing their radiological reports through NLP (natural language processing). With the availability of large *well-organized* databases of medical images and their corresponding radiological reports in the future, we envision that Fixed-Point GAN will be able to detect and localize diseases more accurately—and eventually to segment diseases—using only image-level annotation.

All figures and images (including those in the main paper) better viewed on-line in color and zoomed in for details

Eyeglass Detection and Localization by Removal Using Only Image-Level Annotation of the CelebA Dataset

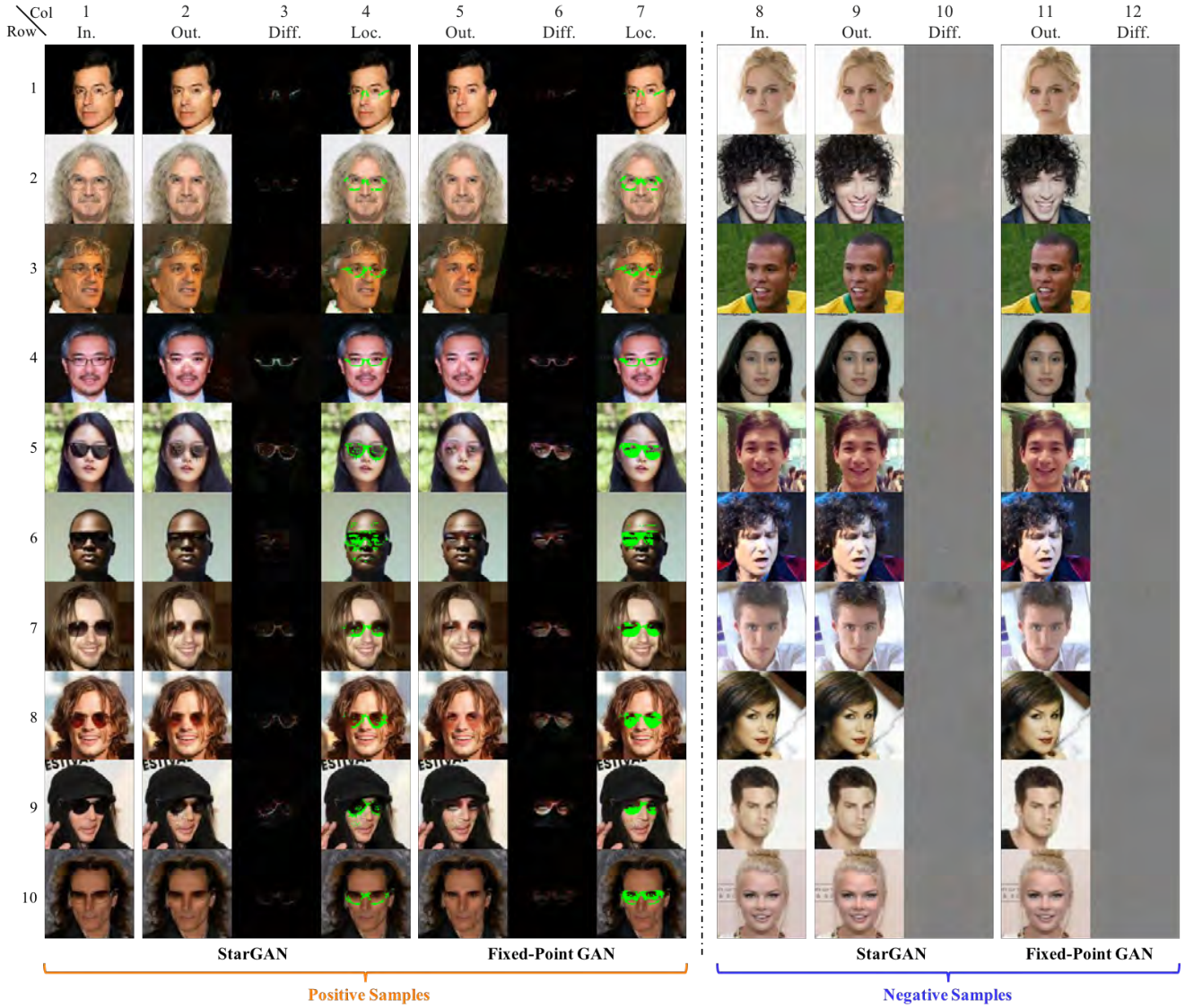


Fig. 7: [Continued from Fig. 2] Additional test results in glass detection and localization by removal. The difference map (Column 3 for StarGAN; Column 6 for Fixed-Point GAN) shows the absolute difference between the input (Column 1) and output (Column 2 for StarGAN; Column 5 for Fixed-Point GAN). Applying the k -means clustering algorithm on the difference map yields a localization map, which is then superimposed on the original image (Column 4 for StarGAN; Column 7 for Fixed-Point GAN), showing both StarGAN and Fixed-Point GAN attempt to remove eyeglasses. However, the former leaves noticeable white “inks” along eyeglass frames (Rows 1 and 4, Column 2), while our method preserves the face color better. Removing sunglasses (Rows 5–9) has proven to be challenging: both methods suffer from partial removal and artifacts. Nevertheless, Fixed-Point GAN tends to recover the face under the glasses and frames, but StarGAN changes only regions around the frames. More importantly, our method can “insert” eyes at proper positions, as revealed in the difference maps (Rows 5–9, Column 6), while StarGAN can hardly do so. To better visualize the subtle changes for negative samples (Column 8), instead of the absolute difference, we show the difference directly, where the gray color (*i.e.*, 0) means “no change”. In this way, it can be observed more easily that StarGAN does some unnecessary small changes on hair (Rows 7 and 9, Column 10) and eyes (Rows 7 and 10, Column 10), while Fixed-Point GAN generates smooth gray images (*i.e.*, close to 0 everywhere; Column 12). Please note that the CelebA Dataset currently does not have ground truth on the location and segmentation of glasses; therefore, a quantitative performance evaluation of the glass localization cannot be conducted. However, our quantitative performance evaluations of brain lesion localization and pulmonary embolism localization are included in Sec. 5.

Multi-Domain Image-to-Image Translation



Fig. 8: [Continued from Fig. 1] More test results in multi-domain image-to-image translation on CelebA dataset. Visually, Fixed-Point GAN outperforms StarGAN: Fixed-Point GAN (Columns 7-11) preserves better the background (Rows 1, 3, 4, 6, 8, and 9), face color (Rows 2-7), and facial features (Rows 7 and 9), whereas StarGAN (Columns 2-6) makes unnecessary changes. Furthermore, for same-domain translation, StarGAN introduces noticeable artifacts (outlined in red), while Fixed-Point GAN can leave all the details intact (outlined in green). It is worthy noting that the hair color of the facial image in the last input row (*i.e.*, Row 9, Column 1) belongs to Domain `gray hair`, which is not included in the training phase. As can be seen, Fixed-Point GAN successfully translates the input image to target domains by changing the unseen hair color to desired colors and maintaining the original hair color (gray) in hair-color-unrelated translations (Row 9, Columns 10-11). However, StarGAN produces unnatural images with artifacts (Row 9, Columns 2-4) and inconsistent white hair colors (Row 9, Columns 5-6). This example shows that Fixed-Point GAN is better than StarGAN in generalization.

Brain Lesion Detection and Localization by Removal Using Only Image-Level Annotation

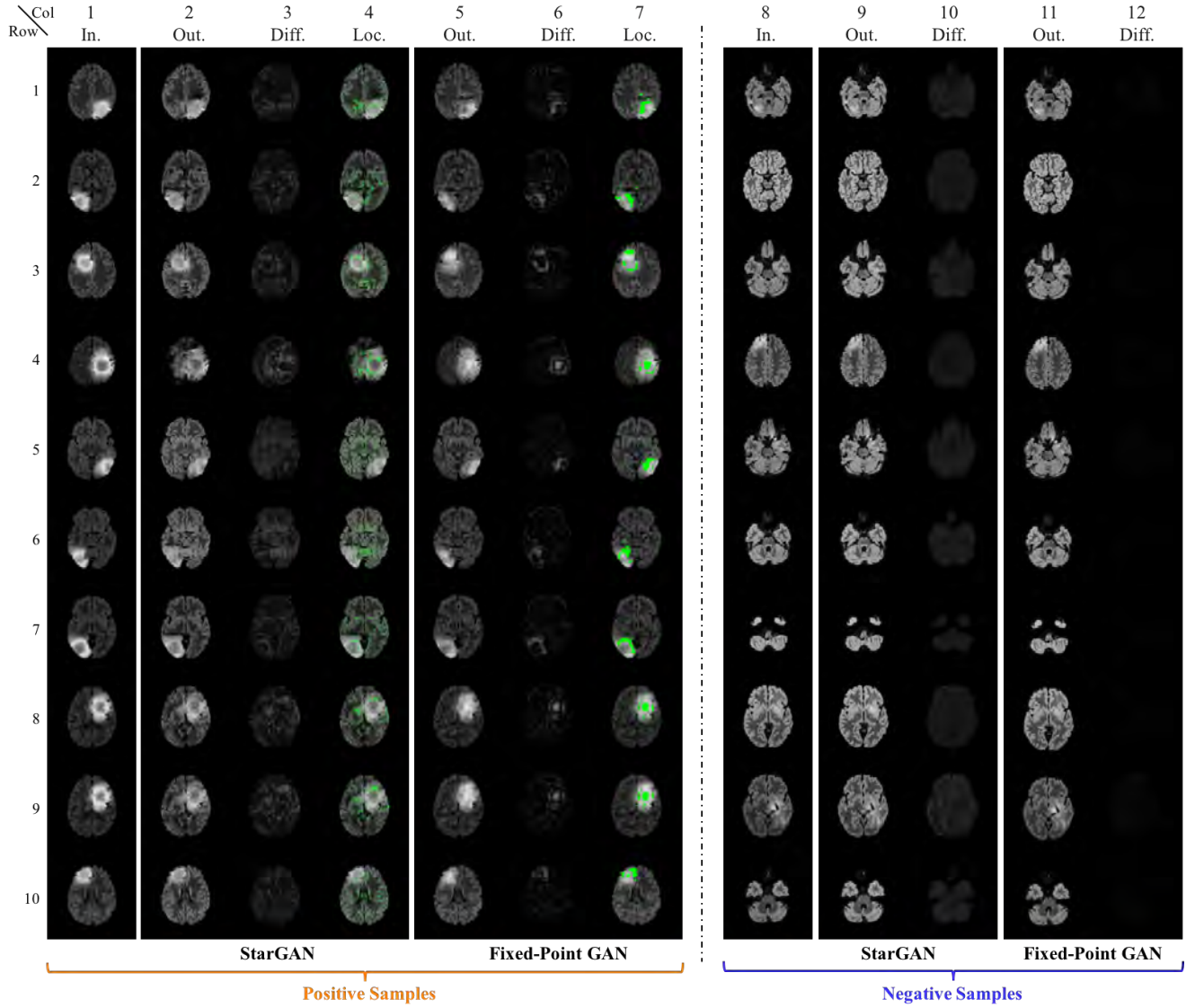


Fig. 9: [Continued from Fig. 2] Brain lesion detection and localization tested on additional positive samples (*i.e.* brain images with lesions; Column 1) and negative samples (*i.e.* brain images without lesions; Column 8). Fixed-Point GAN achieves much better detection performance, benefiting from the cleaner difference maps of negative samples (Column 12), while StarGAN highlights the brain regions in all cases and makes the difference maps of positive and negative samples indistinguishable (comparing Column 3 with Column 10). Although both methods fails to remove lesions completely, our method focuses on the lesion regions, and consequently, has a higher localization accuracy. By contrast, the localization map of StarGAN (Column 4) is very noisy and unsuitable for lesion localization. These comparisons demonstrate the superiority of Fixed-Point GAN in lesion detection and localization. For quantitative performance evaluations, please refer to Fig. 4a–c and Sec. 5.2.

Pulmonary Embolism Detection and Localization by Removal Using Only Image-Level Annotation

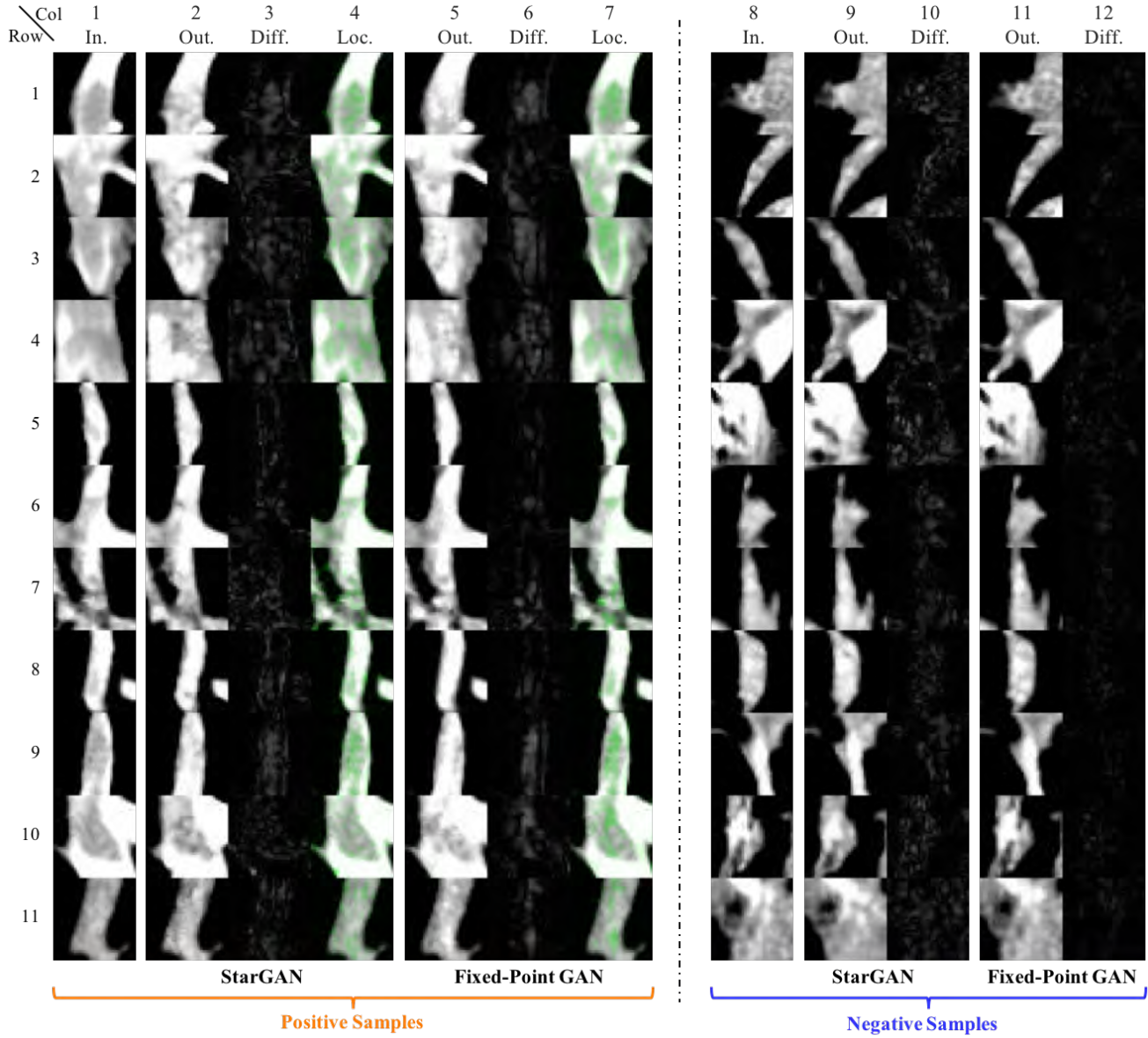


Fig. 10: [Continued from Fig. 2] Pulmonary Embolism (PE) detection and localization (longitudinal view) tested on additional positive samples (*i.e.* images with PEs; Column 1) and negative samples (*i.e.* images without PEs; Column 8). PE is a blood clot that creates blockage (appearing dark and centered in image) in pulmonary arteries (appearing white). The current candidate generator (*e.g.* [13]) produces lots of false positives (negative samples) during localization; therefore, our goal in this application is to reduce false positives through StarGAN and Fixed-Point GAN. Compared with StarGAN, the difference maps of negative samples from Fixed-Point GAN is clean and easy to be separated from the difference maps of positive samples, yielding better detection performance. For quantitative performance evaluations, please refer to Fig. 4d–f and Sec. 5.3.

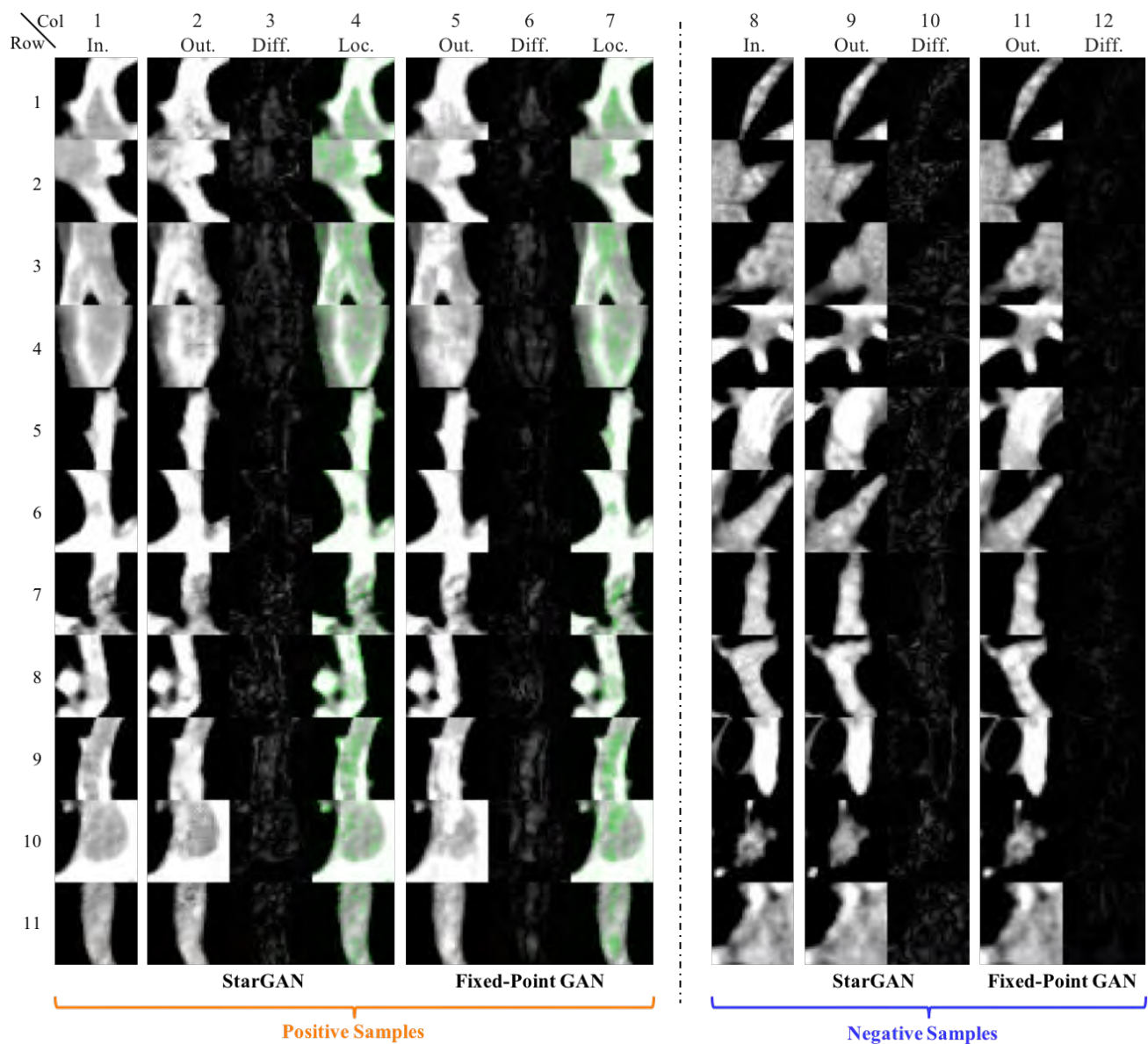


Fig. 11: Pulmonary Embolism (PE) detection and localization (longitudinal view). Notice the images are from same candidates as in Fig. 10 but the view direction is orthogonal to the angle used in Fig. 10.

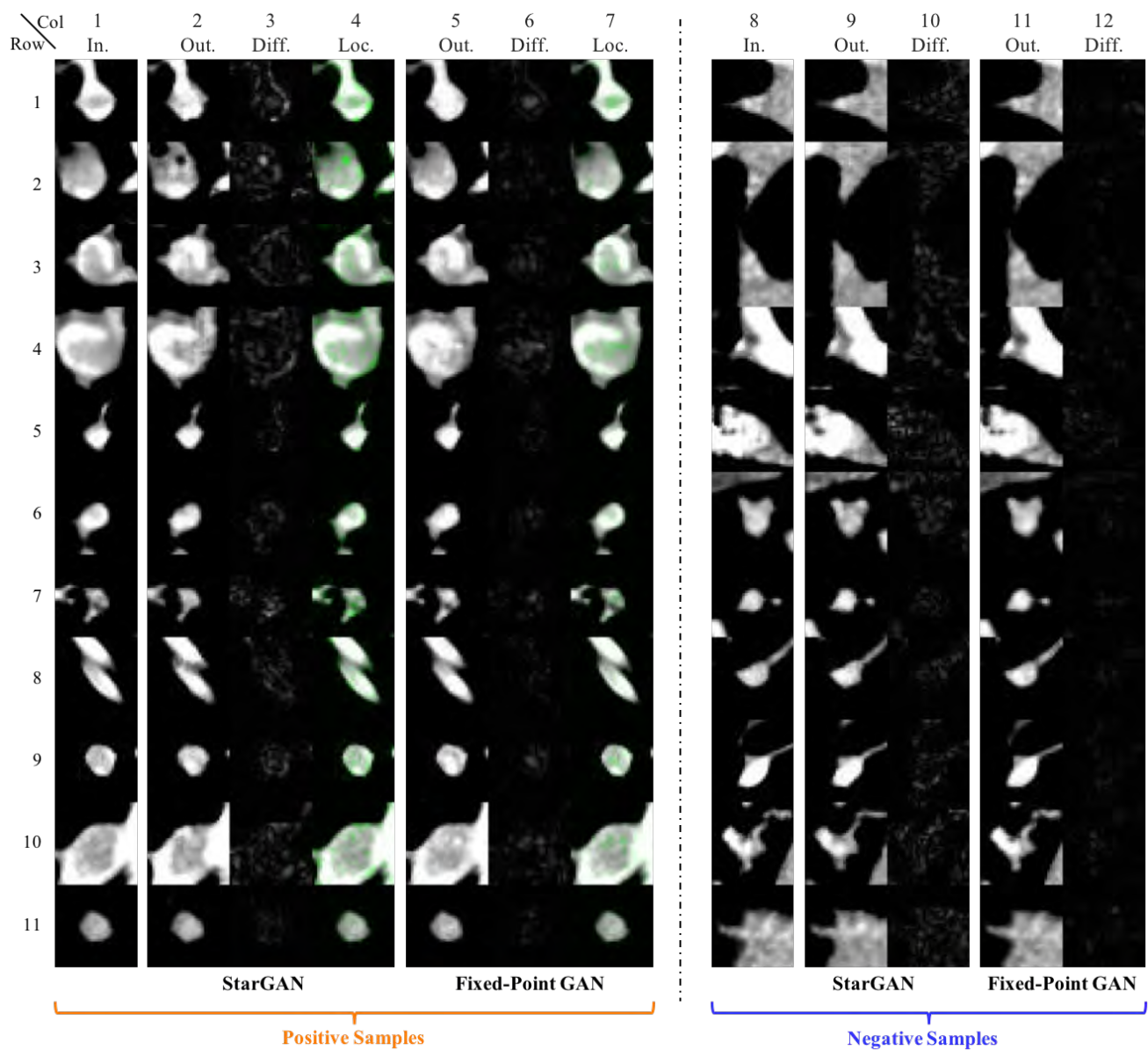


Fig. 12: Pulmonary Embolism (PE) detection and localization (cross-sectional view). Notice the images are from same candidates as in Fig. 10 but the view direction is cross-sectional.

Localization Using Class Activation Maps (CAMs)

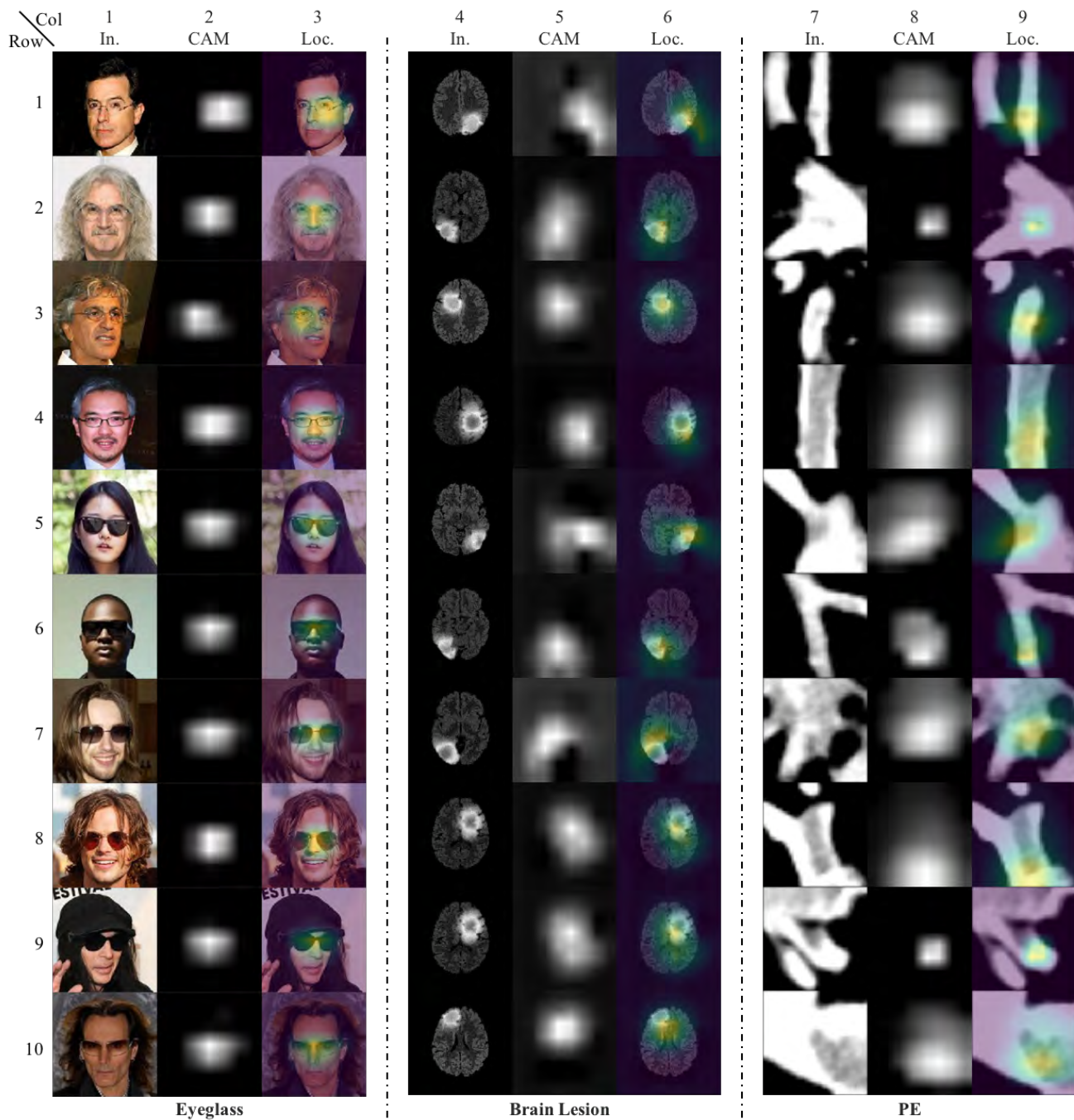


Fig. 13: [Continued from Fig. 2] Additional test results of localization using class activation maps (CAMs). CAMs for localizing glasses, brain lesion, and PE are obtained from ResNet-50 classifiers trained with corresponding datasets. Localization using CAMs is not as precise as Fixed-Point GAN as discussed in Sec. 6.

Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?

Nima Tajbakhsh, *Member, IEEE*, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang*, *Senior Member, IEEE*

Abstract—Training a deep convolutional neural network (CNN) from scratch is difficult because it requires a large amount of labeled training data and a great deal of expertise to ensure proper convergence. A promising alternative is to fine-tune a CNN that has been pre-trained using, for instance, a large set of labeled natural images. However, the substantial differences between natural and medical images may advise against such knowledge transfer. In this paper, we seek to answer the following central question in the context of medical image analysis: Can the use of pre-trained deep CNNs with sufficient fine-tuning eliminate the need for training a deep CNN from scratch? To address this question, we considered four distinct medical imaging applications in three specialties (radiology, cardiology, and gastroenterology) involving classification, detection, and segmentation from three different imaging modalities, and investigated how the performance of deep CNNs trained from scratch compared with the pre-trained CNNs fine-tuned in a layer-wise manner. Our experiments consistently demonstrated that 1) the use of a pre-trained CNN with adequate fine-tuning outperformed or, in the worst case, performed as well as a CNN trained from scratch; 2) fine-tuned CNNs were more robust to the size of training sets than CNNs trained from scratch; 3) neither shallow tuning nor deep tuning was the optimal choice for a particular application; and 4) our layer-wise fine-tuning scheme could offer a practical way to reach the best performance for the application at hand based on the amount of available data.

Index Terms—Carotid intima-media thickness, computer-aided detection, convolutional neural networks, deep learning, fine-tuning, medical image analysis, polyp detection, pulmonary embolism detection, video quality assessment.

I. INTRODUCTION

CONVOLUTIONAL neural networks (CNNs) have been used in the field of computer vision for decades [1]–[3]. However, their true value had not been discovered until the

ImageNet competition in 2012, a success that brought about a revolution through the efficient use of graphics processing units (GPUs), rectified linear units, new dropout regularization, and effective data augmentation [3]. Acknowledged as one of the top 10 breakthroughs of 2013 [4], CNNs have once again become a popular learning machine, now not only within the computer vision community but across various applications ranging from natural language processing to hyperspectral image processing and to medical image analysis. The main power of a CNN lies in its deep architecture [5]–[8], which allows for extracting a set of discriminating features at multiple levels of abstraction.

However, training a deep CNN from scratch (or full training) is not without complications [9]. First, CNNs require a large amount of labeled training data—a requirement that may be difficult to meet in the medical domain where expert annotation is expensive and the diseases (e.g., lesions) are scarce in the datasets. Second, training a deep CNN requires extensive computational and memory resources, without which the training process would be extremely time-consuming. Third, training a deep CNN is often complicated by overfitting and convergence issues, whose resolution frequently requires repetitive adjustments in the architecture or learning parameters of the network to ensure that all layers are learning with comparable speed. Therefore, deep learning from scratch can be tedious and time-consuming, demanding a great deal of diligence, patience, and expertise.

A promising alternative to training a CNN from scratch is to fine-tune a CNN that has been trained using a large labeled dataset from a different application. The pre-trained models have been applied successfully to various computer vision tasks as a feature generator or as a baseline for transfer learning [10]–[12]. Herein, we address the following central question in the context of medical image analysis: *Can the use of pre-trained deep CNNs with sufficient fine-tuning eliminate the need for training a deep CNN from scratch?* This is an important question because training deep CNNs from scratch may not be practical, given the limited labeled data in medical imaging. To answer this central question, we conducted an extensive set of experiments for 4 medical imaging applications: 1) polyp detection in colonoscopy videos, 2) image quality assessment in colonoscopy videos, 3) pulmonary embolism detection in computed tomography (CT) images, and 4) intima-media boundary segmentation in ultrasonographic images. We have chosen these applications to represent the most common clinically used imaging modality systems (i.e., CT, ultrasonography,

Manuscript received December 23, 2015; revised February 19, 2016; accepted February 21, 2016. Date of current version April 29, 2016. N. Tajbakhsh and J. Y. Shin have contributed equally. Asterisk indicates corresponding author.

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

N. Tajbakhsh and J. Y. Shin are with the Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA (e-mail: nima.tajbakhsh@asu.edu; sejong@asu.edu).

S. R. Gurudu is with Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, AZ 85259 USA (e-mail: gurudu.suryakanth@mayo.edu).

R. T. Hurst and C. B. Kendall are with Division of Cardiovascular Diseases, Mayo Clinic, Scottsdale, AZ 85259 USA (e-mail: hurst.r@mayo.edu; kendall.christopher@mayo.edu).

M. B. Gotway is with Department of Radiology, Mayo Clinic, Scottsdale, AZ 85259 USA (e-mail: gotway.michael@mayo.edu).

*J. Liang is with the Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA (e-mail: jianming.liang@asu.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMI.2016.2535302

and optical endoscopy) and the most common medical image analysis tasks (i.e., lesion detection, image segmentation, and image classification). For each application, we compared the performance of the pre-trained CNNs through fine-tuning with that of the CNNs trained from scratch entirely based on medical imaging data. We also compared the performance of the CNN-based systems with their corresponding handcrafted counterparts.

II. RELATED WORKS

Applications of CNNs in medical image analysis can be traced to the 1990s, when they were used for computer-aided detection of microcalcifications in digital mammography [13], [14] and computer-aided detection of lung nodules in CT datasets [15]. With revival of CNNs owing to the development of powerful GPU computing, the medical imaging literature has witnessed a new generation of computer-aided detection systems that show superior performance. Examples include automatic polyp detection in colonoscopy videos [16], [17], computer-aided detection of pulmonary embolism (PE) in CT datasets [18], automatic detection of mitotic cells in histopathology images [19], computer-aided detection of lymph nodes in CT images [20], and computer-aided anatomy detection in CT volumes [21]. Applications of CNNs in medical image analysis are not limited to only computer-aided detection systems, however. CNNs have recently been used for carotid intima-media thickness measurement in ultrasound images [22], pancreas segmentation in CT images [23], brain tumor segmentation in magnetic resonance imaging (MRI) scans [24], multimodality isointense infant brain image segmentation [25], neuronal membrane segmentation in electron microscopy images [26], and knee cartilage segmentation in MRI scans [27].

One important aspect of CNNs is the “transferability” of knowledge embedded in the pre-trained CNNs. Recent research conducted by Azizpour *et al.* [11] suggests that the success of knowledge transfer depends on the distance, or dissimilarity, between the database on which a CNN is trained and the database to which the knowledge is to be transferred. Although the distance between natural image and medical imaging databases is considerable, recent studies show the potential for knowledge transfer to the medical imaging domain.

The recent research on transfer learning in medical imaging can be categorized into two groups. The first group [28]–[30] consists of works wherein a pre-trained CNN is used as a feature generator. Specifically, the pre-trained CNN is applied to an input image and then the CNN outputs (features) are extracted from a certain layer of the network. The extracted features are then used to train a new pattern classifier. For instance, in [28], pre-trained CNNs were used as a feature generator for chest pathology identification. A similar study [29] by Ginneken *et al.* showed that although the use of pre-trained CNNs could not outperform a dedicated nodule detection system, the integration of CNN-based features with the handcrafted features enabled improved performance.

The second group [31]–[36] consists of works wherein a pre-trained CNN is adapted to the application at hand. For instance, in [33], the fully connected layers of a pre-trained CNN were replaced with a new logistic layer, and then the labeled data were used to train only the appended layer while keeping the rest of the network the same. This treatment yielded promising results for classification of unregistered multiview mammogram. Chen *et al.* [32] suggested the use of a fine-tuned pre-trained CNN for localizing standard planes in ultrasound images. In [35], the authors fine-tuned all layers of a pre-trained CNN for automatic classification of interstitial lung diseases. They also suggested an attenuation rescale scheme to convert 1-channel CT slices to RGB-like images needed for tuning the pre-trained model. Shin *et al.* [34] used fine-tuned pre-trained CNNs to automatically map medical images to document-level topics, document-level sub-topics, and sentence-level topics. In [36], fine-tuned pre-trained CNNs were used to automatically retrieve missing or noisy cardiac acquisition plane information from magnetic resonance imaging and predict the five most common cardiac views. Different from the previous approaches, Schlegl *et al.* [31] considered the fine-tuning of an unsupervised network. They explored unsupervised pre-training of CNNs to inject information from sites or image classes for which no annotations were available, and showed that such across site pre-training improved classification accuracy compared to random initialization of the model parameters.

III. CONTRIBUTIONS

In this paper, we systematically study knowledge transfer to medical imaging applications, making the following contributions:

- We demonstrated how fine-tuning a pre-trained CNN in a layer-wise manner leads to incremental performance improvement. This approach distinguishes our work from [28]–[30], which downloaded the features from the fully connected layers of a pre-trained CNN and then trained a separate pattern classifier. Our approach also differs from [31]–[33] wherein the entire pre-trained CNN underwent fine-tuning.
- We analyzed how the availability of training samples influences the choice between pre-trained CNNs and CNNs trained from scratch. To our knowledge, this issue has not yet been systematically addressed in the medical imaging literature.
- We compared the performance of pre-trained CNNs, not only against handcrafted approaches but also against CNNs trained from scratch using medical imaging data. This analysis is in contrast to [28], [29], who provided only limited performance comparisons between pre-trained CNNs and handcrafted approaches.
- We presented consistent results with conclusive outcomes for 4 distinct medical imaging applications involving classification, detection, and segmentation in 3 different medical imaging modalities, which add substantially to the state of the art where conclusions are based solely on 1 medical imaging application.

IV. CONVOLUTIONAL NEURAL NETWORKS (CNNs)

CNNs are so-named because of the convolutional layers in their architectures. Convolutional layers are responsible for detecting certain local features in all locations of their input images. To detect local structures, each node in a convolutional layer is connected to only a small subset of spatially connected neurons in the input image channels. To enable the search for the same local feature throughout the input channels, the connection weights are shared between the nodes in the convolutional layers. Each set of shared weights is called a *kernel*, or a *convolution kernel*. Thus, a convolutional layer with n kernels learns to detect n local features whose strength across the input images is visible in the resulting n feature maps. To reduce computational complexity and achieve a hierarchical set of image features, each sequence of convolution layers is followed by a *pooling layer*, a workflow reminiscent of simple and complex cells in the primary visual cortex [37]. The max pooling layer reduces the size of feature maps by selecting the maximum feature response in overlapping or non-overlapping local neighborhoods, discarding the exact location of such maximum responses. As a result, max pooling can further improve translation invariance. CNNs typically consist of several pairs of convolutional and pooling layers, followed by a number of consecutive fully connected layers, and finally a *softmax layer*, or *regression layer*, to generate the desired outputs. In more modern CNN architectures, computational efficiency is achieved by replacing the pooling layer with a convolution layer with a stride larger than 1.

Similar to multilayer perceptrons, CNNs are trained with the back-propagation algorithm by minimizing the following cost function with respect to the unknown weights W :

$$\mathcal{L} = -\frac{1}{|X|} \sum_i \ln(p(y^i|X^i)) \quad (1)$$

where $|X|$ denotes the number of training images, X^i denotes the i^{th} training image with the corresponding label y^i , and $p(y^i|X^i)$ denotes the probability by which X^i is correctly classified. Stochastic gradient descent is commonly used for minimizing this cost function, where the cost over the entire training set is approximated with the cost over mini-batches of data. If W_l^t denotes the weights in l^{th} convolutional layer at iteration t , and $\hat{\mathcal{L}}$ denotes the cost over a mini-batch of size N , then the updated weights in the next iteration are computed as follows:

$$\begin{aligned} \gamma^t &= \gamma^{\lfloor tN/|X| \rfloor} \\ V_l^{t+1} &= \mu V_l^t - \gamma^t \alpha_l \frac{\partial \hat{\mathcal{L}}}{\partial W_l} \\ W_l^{t+1} &= W_l^t + V_l^{t+1} \end{aligned} \quad (2)$$

where α_l is the learning rate of the l^{th} layer, μ is the momentum that indicates the contribution of the previous weight update in the current iteration, and γ is the scheduling rate that decreases learning rate α at the end of each epoch.

V. FINE-TUNING

The iterative weight update in (2) begins with a set of randomly initialized weights. Specifically, before the commencement of the training phase, weights in each convolutional layer of a CNN are initialized by values randomly sampled from a normal distribution with a zero mean and small standard deviation. However, considering the large number of weights in a CNN and the limited availability of labeled data, the iterative weight update, starting with a random weight initialization, may lead to an undesirable local minimum for the cost function. Alternatively, the weights of the convolutional layers can be initialized with the weights of a pre-trained CNN with the same architecture. The pre-trained net is generated with a massive set of labeled data from a different application. Training a CNN from a set of pre-trained weights is called *fine-tuning* and has been used successfully in several applications [10]–[12].

Fine-tuning begins with copying (transferring) the weights from a pre-trained network to the network we wish to train. The exception is the last fully connected layer whose number of nodes depends on the number of classes in the dataset. A common practice is to replace the last fully connected layer of the pre-trained CNN with a new fully connected layer that has as many neurons as the number of classes in the new target application. In our study, we deal with 2-class and 3-class classification tasks; therefore, the new fully connected layer has 2 or 3 neurons depending on the application under study. After the weights of the last fully connected layer are initialized, the new network can be fine-tuned in a layer-wise manner, starting with tuning only the last layer, then tuning all layers in a CNN.

Consider a CNN with L layers where the last 3 layers are fully connected layers. Also let α_l denote the learning rate of the l^{th} layer in the network. We can fine-tune only the last (new) layer of the network by setting $\alpha_l = 0$ for $l \neq L$. This level of fine-tuning corresponds to training a linear classifier with the features generated in layer $L - 1$. Likewise, the last 2 layers of the network can be fine-tuned by setting $\alpha_l = 0$ for $l \neq L, L - 1$. This level of fine-tuning corresponds to training an artificial neural network with 1 hidden layer, which can be viewed as training a nonlinear classifier using the features generated in layer $L - 2$. Similarly, fine-tuning layers $L, L - 1$, and $L - 2$ are essentially equivalent to training an artificial neural network with 2 hidden layers. Including the previous convolution layers in the update process further adapts the pre-trained CNN to the application at hand but may require more labeled training data to avoid overfitting.

In general, the early layers of a CNN learn low level image features, which are applicable to most vision tasks, but the late layers learn high-level features, which are specific to the application at hand. Therefore, fine-tuning the last few layers is usually sufficient for transfer learning. However, if the distance between the source and target applications is significant, one may need to fine-tune the early layers as well. Therefore, an effective fine-tuning technique is to start from the last layer and then incrementally include more layers in the update process until the desired performance is reached. We refer to tuning the last few convolutional layers as “shallow tuning” and we consider tuning all the convolutional layers as “deep tuning”. We would

like to note that the suggested fine-tuning scheme differs from [10], [12] wherein the network remains the same and serves as a feature generator, and also differs from [11] wherein the entire network undergoes fine-tuning at once.

VI. APPLICATIONS AND RESULTS

In our study, we considered 4 different medical imaging applications from 3 imaging modality systems. We study the performance of polyp detection and PE detection using a free-response operating characteristic (FROC) analysis, analyze the performance of frame classification by means of an ROC analysis, and evaluate the performance of boundary segmentation through a boxplot analysis. To perform statistical comparisons, we have computed the error bars corresponding to 95% confidence intervals for both ROC and FROC curves according to the method suggested in [38]. The error bars enable us to compare each pair of performance curves at multiple operating points from a statistical perspective. Specifically, if the error bars of a pair of curves do not overlap at a fixed false positive rate, then the two curves are statistically different ($p < .05$) at the given operating point. An appealing feature of this statistical analysis is that we can compare the performance curves at a clinically acceptable operating point rather than comparing the curves as a whole. While we have discussed the statistical comparisons throughout the paper, we have further summarized them in a number of tables in supplementary material, which can be found in the supplementary files/multimedia tab.

We used the Caffe library [39] for both training and fine-tuning CNNs. For consistency and ease of comparison, we used the AlexNet architecture for the 4 applications under study. Training and fine-tuning of each AlexNet took approximately 2–3 hours depending on the size of the training set. To ensure the proper convergence of each CNN, we monitored the area under the receiver operating characteristic curve. Specifically, for each experiment, we divided the training set into a smaller training set with 80% of the training data and a validation set with the remaining 20% of the training data and then computed area under the curve on the validation set. The training process was terminated when the highest accuracy on the validation set was observed. All training was performed using an NVIDIA GeForce GTX 980TI (6 GB on-board memory). The fully trained CNNs were initialized with random weights sampled from Gaussian distributions. We also experimented with other initialization techniques such as those suggested in [40] and [41], but we observed no significant performance gain after convergence, even though we noticed varying speed of convergence using these initialization techniques.

For both full training and fine-tuning scenarios, we used a stratified training set of image patches where the positive and negative classes were equally present. For this purpose, we randomly down-sampled the majority (negative) class, while keeping the minority class (positive) unchanged. For the fine-tuning scenario, we used the pre-trained AlexNet model provided in the Caffe library. The pre-trained AlexNet consists of approximately 5 million parameters in the convolution layers and about 55 million parameters in its fully connected layers, and is trained using 1.2 million images labeled with 1000 semantic classes. The model used in our study is the

TABLE I
THE ALEXNET ARCHITECTURE USED IN OUR EXPERIMENTS. OF NOTE, C IS THE NUMBER OF CLASSES, WHICH IS 3 FOR INTIMA-MEDIA INTERFACE SEGMENTATION AND IS 2 FOR COLONOSCOPY FRAME CLASSIFICATION, POLYP DETECTION, AND PULMONARY EMBOLISM DETECTION

layer	type	input	kernel	stride	pad	output
data	input	$3 \times 227 \times 227$	N/A	N/A	N/A	$3 \times 227 \times 227$
conv1	convolution	$3 \times 227 \times 227$	11×11	4	0	$96 \times 55 \times 55$
pool1	max pooling	$96 \times 55 \times 55$	3×3	2	0	$96 \times 27 \times 27$
conv2	convolution	$96 \times 27 \times 27$	5×5	1	2	$256 \times 27 \times 27$
pool2	max pooling	$256 \times 27 \times 27$	3×3	2	0	$256 \times 13 \times 13$
conv3	convolution	$256 \times 13 \times 13$	3×3	1	1	$384 \times 13 \times 13$
conv4	convolution	$384 \times 13 \times 13$	3×3	1	1	$384 \times 13 \times 13$
conv5	convolution	$384 \times 13 \times 13$	3×3	1	1	$256 \times 13 \times 13$
pool5	max pooling	$256 \times 13 \times 13$	3×3	2	0	$256 \times 6 \times 6$
fc6	fully connected	$256 \times 6 \times 6$	6×6	1	0	4096×1
fc7	fully connected	4096×1	1×1	1	0	4096×1
fc8	fully connected	4096×1	1×1	1	0	$C \times 1$

snapshot taken after 360,000 training iterations. As shown in Table I, AlexNet begins with 2 pairs of convolutional and pooling layers, mapping the 227×227 input images to 13×13 feature maps. This architecture then proceeds with a sequence of 3 convolutional layers that efficiently implement a convolutional layer with 9×9 kernels, yet with a larger degree of nonlinearity. The sequence of convolutional layers is then followed by a pooling layer and 3 fully connected layers. The first fully connected layer can be viewed as a convolution layer with 6×6 kernels and the other 2 fully connected layers as convolutional layers with 1×1 kernels.

Table II summarizes the learning parameters used for training and fine-tuning of AlexNet in our experiments. The listed parameters were tuned through an extensive set of trial and error experiments. According to our exploratory experiments, the learning rate and scheduling rate heavily influenced the convergence of CNNs. A learning rate of 0.001 however ensured proper convergence for all 4 applications. A smaller learning rate slowed down convergence and a larger learning rate often caused convergence failures. Our exploratory experiments also indicated that the value of γ depended on the speed of convergence. During a fast convergence, the learning rate can be safely decreased after a few epochs, allowing for the use of a small scheduling rate. However, during a slow convergence, a larger scheduling rate is required to maintain a relatively large learning rate. For all 4 applications, we found $\gamma = .95$ to be a reasonable choice.

A. Polyp Detection

Colonoscopy is the preferred technique for colon cancer screening and prevention. The goal of colonoscopy is to find and remove colonic polyps—precursors to colon cancer. Polyps, as shown in Fig. 1, can appear with substantial variations in color, shape, and size. The challenging appearance of polyps can often lead to misdetection, particularly during long and back-to-back colonoscopy procedures where fatigue negatively affects the performance of colonoscopists. Polyp miss-rates are estimated to be about 4% to 12% [43]–[46]; however, a more recent clinical study [47] is suggestive that this misdetection rate may be as high as 25%. Missed polyps can lead to the late diagnosis of colon cancer with an associated

TABLE II

LEARNING PARAMETERS USED FOR TRAINING AND FINE-TUNING OF ALEXNET IN OUR EXPERIMENTS. μ IS THE MOMENTUM, α IS THE LEARNING RATE OF THE WEIGHTS IN EACH CONVOLUTIONAL LAYER, AND γ DETERMINES HOW α DECREASES OVER EPOCHS. THE LEARNING RATE FOR THE BIAS TERM IS ALWAYS SET TWICE AS LARGE AS THE LEARNING RATE OF THE CORRESPONDING WEIGHTS. OF NOTE, “FINE-TUNED ALEXNET: LAYER1-LAYER2” INDICATES THAT ALL LAYERS BETWEEN AND INCLUDING THESE 2 LAYERS UNDERGO FINE-TUNING

CNNs	Parameters									
	μ	α_{conv1}	α_{conv2}	α_{conv3}	α_{conv4}	α_{conv5}	α_{fc6}	α_{fc7}	α_{fc8}	γ
Fine-tuned AlexNet:conv1-fc8	0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv2-fc8	0.9	0	0.001	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv3-fc8	0.9	0	0	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv4-fc8	0.9	0	0	0	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv5-fc8	0.9	0	0	0	0	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:fc6-fc8	0.9	0	0	0	0	0	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:fc7-fc8	0.9	0	0	0	0	0	0	0.001	0.01	0.95
Fine-tuned AlexNet:only fc8	0.9	0	0	0	0	0	0	0	0.01	0.95
AlexNet scratch	0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.95

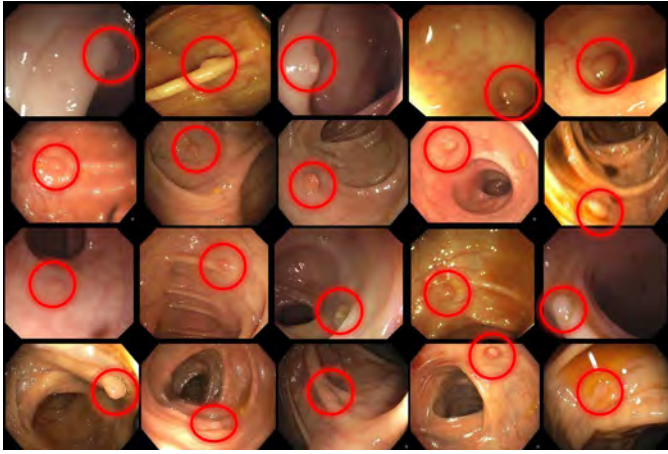


Fig. 1. Variations in shape and appearance of polyps in colonoscopy videos.

decreased survival rate of less than 10% for metastatic colon cancer [48]. Computer-aided polyp detection may enhance optical colonoscopy screening by reducing polyp misdetection.

Several computer-aided detection (CAD) systems have been suggested for automatic polyp detection in colonoscopy videos. The early systems [49]–[51] relied on polyp color and texture for detection. However, limited texture visibility on the surface of polyps and large color variations among polyps hindered the applicability of such systems. More recent systems [52]–[56] relied on temporal information and shape information to enhance polyp detection. Shape features proved more effective than color and texture in this regard; however, these features can be misleading without consideration of the context in which the polyp is found. In our previous works [57]–[59], culminated in [42], we attempted to overcome the limitation of approaches based solely on polyp shape. Specifically, we suggested a handcrafted approach for combining the shape and context information around the polyp boundaries and demonstrated the superiority of this approach over the other state-of-the-art methods.

For training and evaluation, we used our database of 40 short colonoscopy videos. Each colonoscopy frame in our database comes with a binary ground truth image. We randomly divided the colonoscopy videos into a training set containing 3,800 frames with polyps and 15,100 frames without polyps and into a test set containing 5,700 frames with polyps and 13,200 frames without polyps. We applied our handcrafted approach [42] to

the training and test frames to obtain a set of polyp candidates with the corresponding bounding boxes. At each candidate location, given the available bounding box, we extracted a set of image patches with data augmentation. Specifically, for each candidate, we extracted patches at 3 scales by enlarging the corresponding bounding box by a factor of $1.0 \times$, $1.2 \times$, and $1.5 \times$. At each scale, we extracted patches after we translated the candidate location by 10% of the resized bounding box in horizontal and vertical directions. We further rotated each resulting patch 8 times by horizontal and vertical mirroring and flipping. We then labeled a patch as positive if the underlying candidate fell inside the ground truth for polyps; otherwise, the candidate was labeled as negative. Because of the relatively large number of negative patches, we collected a stratified set of 100,000 training patches for training and fine-tuning the CNNs. During the test stage, all test patches extracted from a polyp candidate were fed to the trained CNN. We then averaged the probabilistic outputs of the test patches at the candidate level and performed an FROC analysis for performance evaluation.

Fig. 2(a) compares the FROC curve of our handcrafted approach with that of fine-tuned CNNs and a CNN trained from scratch. To avoid clutter in the figure, we have shown only a subset of representative FROC curves. Statistical comparisons between each pair of FROC curves at three operating points are also presented in Table S1 in the supplementary file. The handcrafted approach is significantly outperformed by all CNN-based scenarios ($p < .05$). This result is probably because our handcrafted approach used only geometric information to remove false-positive candidates. For fine-tuning, the lowest performance was obtained when only the last layer of AlexNet was updated with colonoscopy data. However, fine-tuning the last two layers (FT:fc7-fc8) achieved a significantly higher sensitivity ($p < .05$) at nearly all operating points compared to the pre-trained AlexNet with only 1 fine-tuned layer (FT:only fc8). We also observed incremental performance improvement when we included more convolutional layers in the fine-tuning process. Specifically, the pre-trained CNN with shallow fine-tuning (FT:fc7-fc8) was significantly outperformed by the pre-trained CNNs with a moderate level of fine-tuning (FT:conv5,4,3-fc8) at most of the operating points. Furthermore, the deeply-tuned CNNs (FT:conv1,2-fc8) achieved a significantly higher sensitivity than the pre-trained CNNs with a moderate level of fine-tuning particularly at low false positive rates. Also, as seen in Fig. 2(a), fine-tuning the

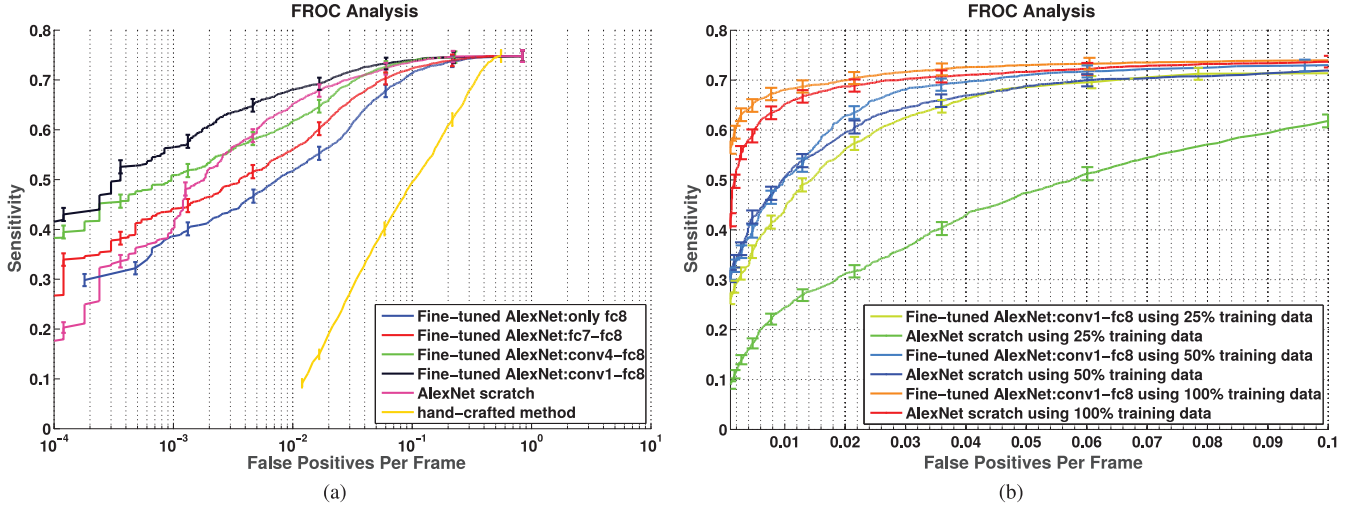


Fig. 2. FROC analysis for polyp detection. (a) Comparison between incremental fine-tuning, training from scratch, and a handcrafted approach [42]. (b) Effect of reduction in the training data on the performance of CNNs trained from scratch and deeply fine-tuned CNNs.

last few convolutional layers was sufficient to outperform an AlexNet model trained from scratch in a low false positive rate setting.

The performance gap between fully trained AlexNet model and their deeply fine-tuned counterparts becomes more evident when fewer training samples are used for training and tuning. To demonstrate this effect, we trained a CNN from scratch and fine-tuned the entire AlexNet using 50% and 25% of the entire training samples. We reduced training data at the video level to exclude a fraction of unique polyps from the training set. The results are shown in Fig. 2(b). With a 50% reduction in training data, a significant performance gap was observed between the CNN trained from scratch and the deeply fine-tuned CNN. With a 25% reduction in the training data, the fully trained CNN showed dramatic performance degradation, but the deeply fine-tuned CNN still exhibited relatively high performance. These findings strongly favor the use of the fine-tuning approach over full training of a CNN from scratch.

B. Pulmonary Embolism Detection

A PE is a blood clot that travels from a lower extremity source to the lung, where it causes blockage of the pulmonary arteries. The mortality rate of untreated PE may approach 30% [61], but it decreases to as low as 2% with early diagnosis and appropriate treatment [62]. CT pulmonary angiography (CTPA) is the primary means for PE diagnosis, wherein a radiologist carefully traces each branch of the pulmonary artery for any suspected PEs. CTPA interpretation is a time-consuming task whose accuracy depends on human factors, such as attention span and sensitivity to the visual characteristics of PEs. CAD can have a major role in improving PE diagnosis and decreasing the reading time of CTPA datasets.

We based our experiments on the PE candidates generated by our previous work [60] and the image representation that we suggested for PE in our recently published study [18]. Our candidate generation method is an improved version of the tobogganing algorithm [63] that aims to find an embolus as a dark region surrounded by a brighter background. Our image repre-

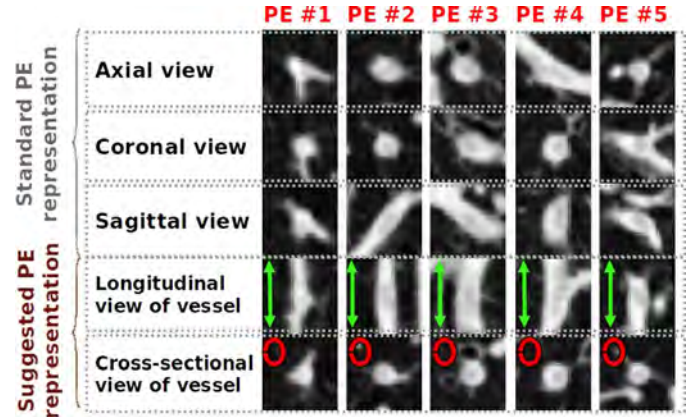


Fig. 3. 5 different PEs in the standard 3-channel representation and in our suggested 2-channel representation. PEs appear more consistently in our representation. We use our PE representation for the experiments presented herein because it achieves greater classification accuracy and enables improved convergence.

sentation consistently results in 2-channel image patches, which capture PEs in cross-sectional and longitudinal views of vessels (see Fig. 3). This unique representation dramatically decreases the variability in the appearance of PEs, enabling us to train more accurate CNNs. However, since the AlexNet architecture receives color images as its input, the 2-channel image patches must be converted to color patches. For this purpose, we simply repeated the second channel and produced 3-channel RGB-like image patches. The resulting patches were then used for training and fine-tuning an AlexNet. For performance comparison, we used a handcrafted approach [60], which is arguably one of the most, if not the most, accurate PE CAD system. The handcrafted approach utilizes the same candidate generation method [60], but uses vessel-based features along with Haralick [64] and wavelet-based features for PE characterization, and finally uses a multi-instance classifier for candidate classification.

For experiments, we used a database consisting of 121 CTPA datasets with a total of 326 PEs. We first applied the tobogganing algorithm to obtain a crude set of PE candidates. This

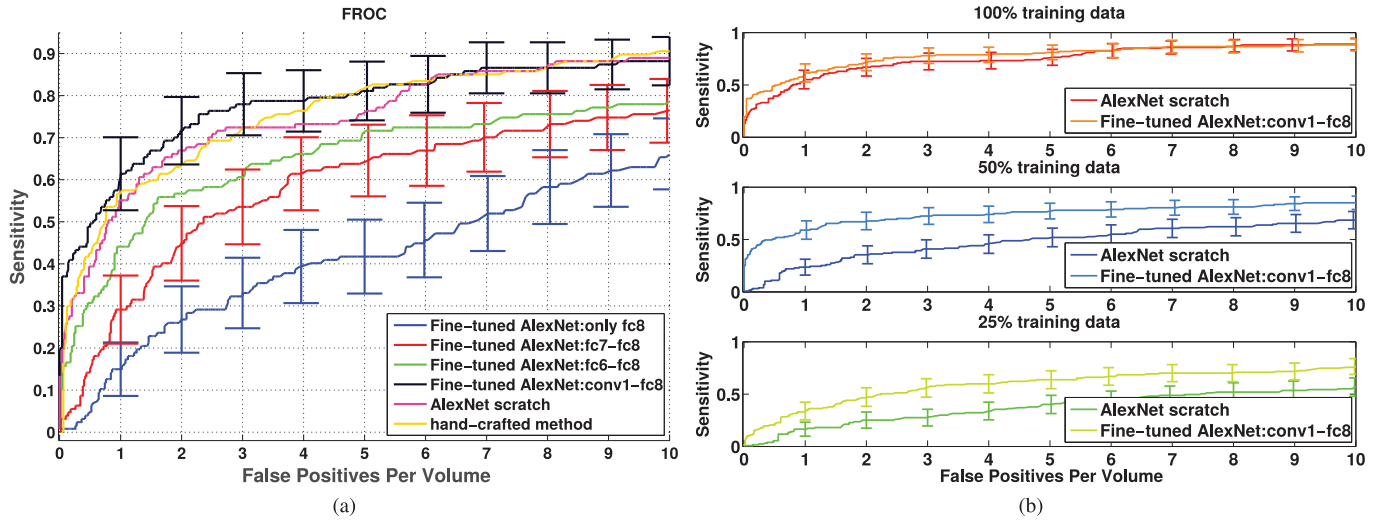


Fig. 4. FROC analysis for pulmonary embolism detection. (a) Comparison between incremental fine-tuning, training from scratch, and a handcrafted approach [60]. To avoid clutter in the figure, error bars are displayed for only a subset of plots. A more detailed analysis is presented in Table S2 in the supplementary file. (b) Effect of reduction in the training data on the performance of CNNs trained from scratch and deeply fine-tuned CNNs.

application resulted in 6,255 PE candidates, of which 5,568 were false positives and 687 were true positives. The number of true positives was far larger than the number of PEs because the tobogganing algorithm can cast several candidates for the same PE. We divided the collected candidates at the patient level into a training set with 434 true positives (199 unique PEs) and 3,406 false positives, and a test set with 253 true positives (127 unique PEs) and 2,162 false positives. For training the CNNs, we extracted patches of 3 different physical sizes, resulting in 10 mm-, 15 mm-, and 20 mm-wide patches. We also translated each candidate location along the direction of the affected vessel 3 times, up to 20% of the physical size of the patches. We further augmented the training dataset by rotating the longitudinal and cross-sectional vessel planes around the vessel axis, resulting in 5 additional variations for each scale and translation. We formed a stratified training set with 81,000 image patches for training and fine-tuning the CNNs. For testing, we performed the same data augmentation for each test candidate and then computed the overall PE probability by averaging the probabilistic scores generated for the data-augmented patches for each PE candidate.

For evaluation, we performed an FROC analysis by changing a threshold on the probabilistic scores generated for the test PE candidates. Fig. 4(a) shows the FROC curves for the handcrafted approach, a deep CNN trained from scratch, and a subset of representative pre-trained CNNs that are fine-tuned in a layer-wise manner. We have further summarized statistical comparisons between each pair of FROC curves in Table S2 in the supplementary file. As shown, the pre-trained CNN with two fine-tuned layers (FT:fc7-fc8) achieved a significantly higher sensitivity ($p < 0.05$) than that of the pre-trained CNN with only one fine-tuned layer (FT:only fc8). The improved sensitivity was observed at most of the operating points. However, inclusion of each new layer in the fine-tuning process resulted in only marginal performance improvement, even though the accumulation of such marginal improvements yielded a substantial margin between the deeply fine-tuned CNNs and those with 1, 2, or 3 fine-tuned layers. Specifically, the deeply

fine-tuned CNN (FT:conv1-fc8) yielded significantly higher sensitivity ($p < 0.05$) than that of the pre-trained CNN with 2 fine-tuned layers (FT:fc7-fc8) at the majority of the operating points shown in Fig. 4(a). At 3 false positives per volume, the deeply fine-tuned CNN also achieved significantly higher sensitivity ($p < 0.05$) than that of the pre-trained CNN with three fine-tuned layers (FT:fc7-fc8). From Fig. 4(a), it is also evident that the deeply fine-tuned CNN yielded a non-significant performance improvement over the handcrafted approach. This is probably because the handcrafted approach is an accurate system whose underlying features are specifically and incrementally designed to remove certain types of false detections. Yet, we find it interesting that an end-to-end learning machine can learn such a sophisticated set of features with minimal engineering effort. From Fig. 4(a), we also observed that the deeply fine-tuned CNN performs on a par with the CNN trained from scratch.

We further analyzed how the size of training samples influences the competitive performance between the CNN trained from scratch and the deeply fine-tuned CNN. For this purpose, we reduced the training samples at the PE-level to 50% and 25%. The results are shown in Fig. 4(b). With a 50% reduction in training data, a significant performance gap was observed between the CNN trained from scratch and the deeply tuned CNN in all the operating points. With a 25% reduction in the training data, we observed a decrease in the overall performance of both CNNs with a smaller yet significant gap between the two curves in most of the operating points. These findings not only favor the use of a deeply fine-tuned CNN but also underscore the importance of large training sets for effective training and fine-tuning of CNNs.

C. Colonoscopy Frame Classification

Image quality assessment can have a major role in objective quality assessment of colonoscopy procedures. Typically, a colonoscopy video contains a large number of non-informative

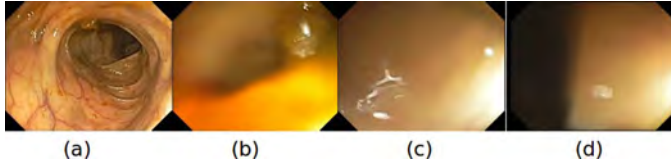


Fig. 5. (a) An informative colonoscopy frame. (b,c,d) Examples of non-informative colonoscopy images. The non-informative frames are usually captured during the rapid motion of the scope or during wall contact.

images with poor colon visualization that are not suitable for inspecting the colon or performing therapeutic actions. The larger the fraction of non-informative images in a video, the lower the quality of colon visualization, and thus the lower the quality of colonoscopy. Therefore, one way to measure the quality of a colonoscopy procedure is to monitor the quality of the captured images. Such quality assessment can be used during live procedures to limit low-quality examinations or in a post-processing setting for quality monitoring purposes.

Technically, image quality assessment at colonoscopy can be viewed as an image classification task whereby an input image is labeled as either *informative* or *non-informative*. Fig. 5 shows examples of non-informative and informative colonoscopy frames. In our previous work [65], we suggested a handcrafted approach based on local and global features that were pooled from the image reconstruction error. We showed that our handcrafted approach outperformed the other major methods [66], [67] for quality assessment in colonoscopy videos. In the current effort, we explored the use of deep CNNs as an alternative to a carefully engineered method. Specifically, we compared the performance of our handcrafted approach with that of a deep CNN trained from scratch and a pre-trained CNN that was fine-tuned using the labeled colonoscopy frames in a layer-wise manner.

For experiments, we used 6 complete colonoscopy videos. Considering the expenses associated with annotation of all video frames, we instead sampled each colonoscopy video by selecting 1 frame from every 5 seconds of each video and thereby removed many similar colonoscopy frames. The resulting set was further refined to create a balanced dataset of 4,000 colonoscopy images in which both informative and non-informative classes were represented equally. A trained expert then manually labeled the collected images as informative or non-informative. A gastroenterologist further reviewed the labeled images for corrections. We divided the labeled frames at the video-level into training and test sets, each containing approximately 2,000 colonoscopy frames. For data augmentation, we extracted 200 sub-images of size 227×227 pixels from random locations in each 500×350 colonoscopy frame, resulting in a stratified training set with approximately 40,000 sub-images. During the test stage, the probability of each frame being informative was computed as the average probabilities assigned to its randomly cropped sub-images.

We used an ROC analysis for performance comparisons between the CNN-based scenarios and handcrafted approach. The results are shown in Fig. 6(a). To avoid clutter in the figure, we have shown only a subset of representative ROC curves. We have, however, summarized the statistical comparisons be-

tween all ROC curves at 10%, 15%, and 20% false positive rates in Table S3 in the supplementary file. We observed that all CNN-based scenarios significantly outperformed the handcrafted approach in at least one of the above 3 operating points. We also observed that fine-tuning the pre-trained CNN halfway through the network (FT:conv4-fc8 and FT:conv5-fc8) not only significantly outperformed shallow-tuning but also was superior to a deeply fine-tuned CNN (FT:conv1-fc8) at 10% and 15% false positive rates. This was probably because the kernels learned in the early layers of the CNN were suitable for image quality assessment and thus their fine-tuning was unnecessary. Furthermore, while the CNN trained from scratch outperformed the pre-trained CNN with shallow fine-tuning (FT:only fc8), it was outperformed by the pre-trained CNN with a moderate level of fine-tuning (FT:conv5-fc8). Therefore, the fine-tuning scheme was superior to the full training scheme from scratch.

To examine how the performance of CNNs changes with respect to the size of the training data, we decreased the number of training samples by factors of 1/10, 1/20, and 1/100. Comparing these with other applications, we considered a further reduction in the size of the training dataset because a moderate decrease did not influence the performance of CNNs substantially. As shown in Fig. 6(b), both deeply fine-tuned CNNs and fully trained CNN showed insignificant performance degradation even when using 10% of the original training set. However, further reduction in the size of the training set substantially degraded the performance of fully trained CNNs and, to a largely less extent, the performance of deeply fine-tuned CNNs. The relatively high performance of the deeply fine-tuned CNNs, even with a limited training set, indicates the usefulness of the kernels learned from ImageNet for colonoscopy frame classification.

D. Intima-Media Boundary Segmentation

Carotid intima-media thickness (CIMT), a noninvasive ultrasonography method, has proven valuable for cardiovascular risk stratification. The CIMT is defined as the distance between the lumen-intima and media-adventitia interfaces at the far wall of the carotid artery (Fig. 7). The CIMT measurement is performed by manually tracing the lumen-intima and media-adventitia interfaces in a region of interest (ROI), followed by calculation of the average distance between the traced interfaces. However, manual tracing of the interfaces is time-consuming and tedious. Therefore, several methods [68]–[71] have been developed to allow automatic CIMT image interpretation. The suggested methods are more or less based on handcrafted techniques whose performance may vary according to image quality and the level of artifacts present within the images.

We formulated this interface segmentation task as a 3-class classification problem wherein the goal was to classify every pixel in the ROI into 3 categories: a pixel on the lumen-intima interface, a pixel on the media-adventitia interface, or a non-interface pixel. For this classification problem, we trained a 3-way CNN using the training patches collected from the lumen-intima interface and media-adventitia interface, as well as from other random locations far from the desired interfaces. Fig. 7 illustrates how these patches are extracted from an ultrasonography frame.

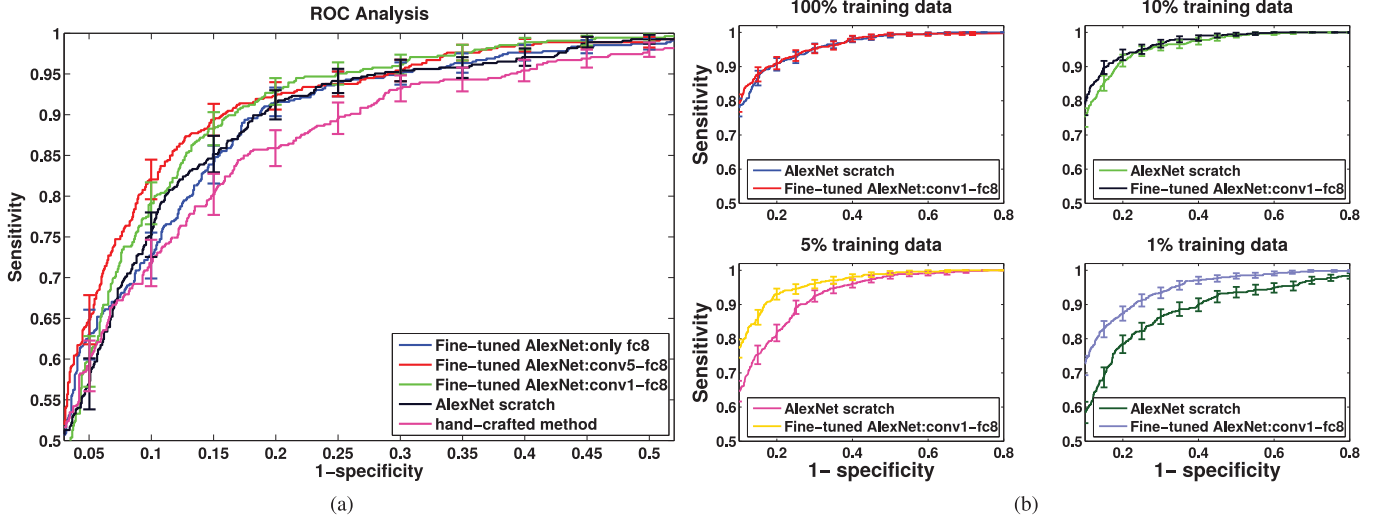


Fig. 6. ROC analysis for image quality assessment. (a) Comparison between incremental fine-tuning, training from scratch, and a handcrafted approach [65]. (b) Effect of reduction in the training data on the performance of convolutional neural networks (CNNs) trained from scratch vs deeply fine-tuned CNNs.

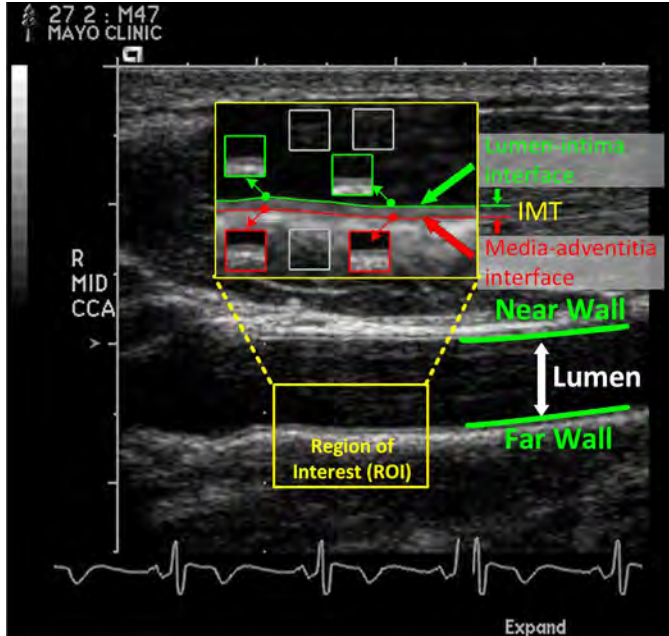


Fig. 7. Intima media thickness (IMT) is measured within a region of interest after the lumen-intima and media-adventitia interfaces are segmented. For automatic interface segmentation, we trained a 3-way convolutional neural network whose training patches were extracted from each of these interfaces (highlighted in red and green) and far from the interfaces (highlighted in gray).

Fig. 8 shows how a CNN-based system traces the interfaces for a given test ROI. The trained CNN is first applied to each pixel within the test ROI in a convolutional manner, generating 2 confidence maps of the same size as the ROI, with the first map showing the probability of a pixel residing on the lumen-intima interface and the second map showing the probability of a pixel residing on the media-adventitia interface. For visualization convenience, we merged these 2 confidence maps into 1 color-coded confidence map in which the green and red colors indicate the likelihood of being a lumen-intima interface and a media-adventitia interface, respectively. As

shown in Fig. 8(b), the probability band of each interface is too thick to accurately measure intima-media thickness. To resolve this issue, we obtained thinner interfaces by scanning the confidence map column by column to search for rows with the maximum response for each of the 2 interfaces, yielding a 1-pixel boundary with a step-like shape around each interface, as shown in Fig. 8(c). To smooth the boundaries, we used 2 active contour models (snakes) [72], one for the lumen-intima interface and one for the media-adventitia interface. The open snakes were initialized with the current step-like boundaries and then kept deforming until they took the actual shapes of the interfaces. Fig. 8(d) shows the converged snakes for the test ROI. We computed intima-media thickness as the average of the vertical distances between the 2 open snakes.

For the experiments, we used a database of 92 CIMT videos. The expert reviews each video to determine 3 ROIs for which the CIMT can be measured reliably. To create the ground truth, lumen-intima and media-adventitia interfaces were annotated as the consensus of 2 experts for each of the 276 ROIs. We divided the ROIs at the subject-level into a training set with 144 ROIs and a test set with 132 ROIs. For training and fine-tuning the CNNs, we extracted a stratified set of 200,000 training patches from the training ROIs. Because the AlexNet architecture used in our study required color patches as its input, each extracted gray-scale patch was converted to a color patch by repeating the gray channel thrice. Note that we did not perform data augmentation for the positive patches, for 2 reasons. First, 92×60 ROIs allow us to collect a large number of patches around the lumen-intima and media-adventitia interfaces, eliminating the need for any further data augmentation. Second, given the relatively small distance between the 2 interfaces, translation-based data augmentation would inject a large amount of label noise, which would negatively affect the convergence and the overall performance of the CNNs. In the test stage, we measured the error of interface segmentation as the average distance between the expert-annotated interfaces and those produced by the systems. For a more detailed analysis, we measured segmentation error for the lumen-intima and media-adventitia interfaces separately.



Fig. 8. The test stage of lumen-intima and media-adventitia interface segmentation. (a) A test region of interest. (b) The corresponding confidence map generated by the convolutional neural network. The green and red colors indicate the likelihood of a lumen-intima interface and media-adventitia interface, respectively. (c) The thick probability band around each interface is thinned by selecting the largest probability for each interface in each column. (d) The step-like boundaries are smoothed using 2 open snakes. (e) Interface segmentation from the ground truth.

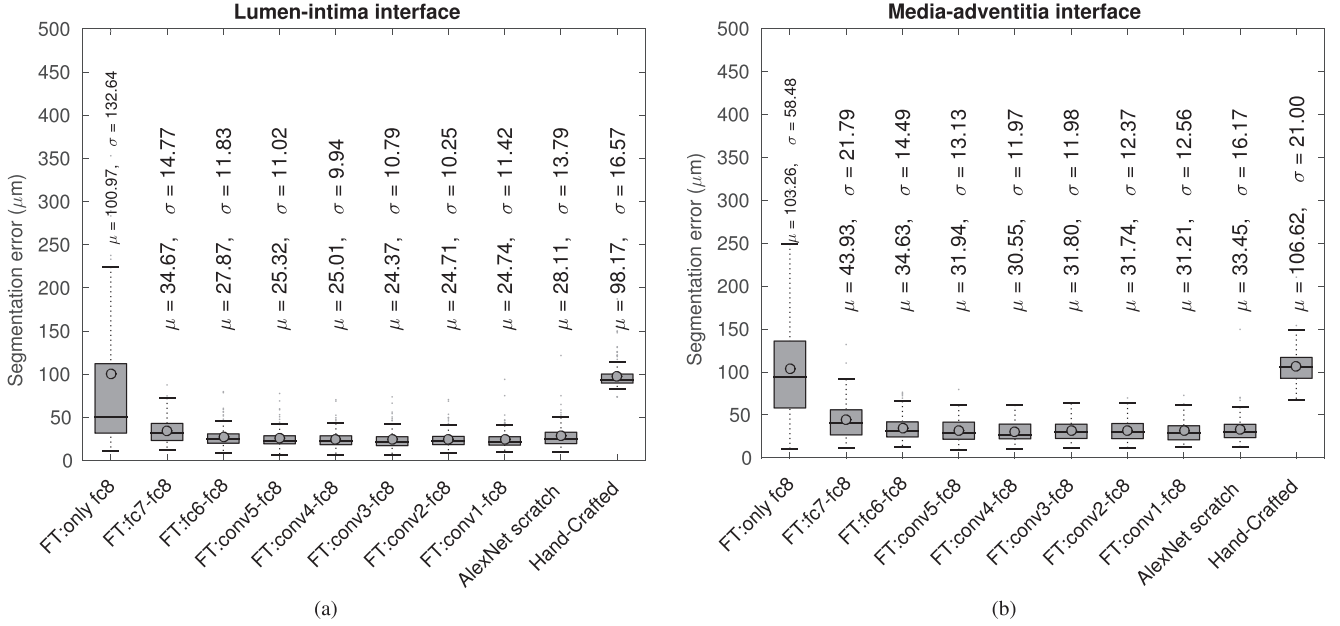


Fig. 9. Box plots of segmentation error for (a) the lumen-intima interface and (b) the media-adventitia interface.

Fig. 9 shows the box plots of segmentation error for each interface. The whiskers were plotted according to Tukey method. For easier quantitative comparisons, we have also shown the average and standard deviation of the localization error above each boxplot. The segmentation error for the media-adventitia interface was generally greater than the lumen-intima interface, which was expected because of the relatively more challenging image characteristics of the media-adventitia interface. For both interfaces, holding all the layers fixed except the last layer (FT: only fc8) resulted in the lowest performance, which was comparable to that of the handcrafted approach [73]. However, inclusion of layer fc7 in the fine-tuning process (FT:fc7-fc8) led to a significant decrease ($p < .0001$) in segmentation error for both interfaces. The reduced localization error was also significantly lower ($p < .0001$) than that of the handcrafted approach. We observed another significant drop ($p < .001$) in the localization error of both interfaces after fine-tuning layer fc6; however, this error was still significantly larger ($p < .001$) than that of the deeply fine-tuned AlexNet (FT:conv1-fc8). We observed a localization error comparable to that of the deeply fine-tuned AlexNet only after inclusion of layer conv5 in the fine-tuning process. With deeper fine-tuning, we obtained only marginal decrease in the localization error for both interfaces. Furthermore, the localization error obtained by the deeply fine-tuned CNN was significantly lower than that of the CNN trained from

scratch for media-adventitia interface ($p < .05$) and for Lumen-intima interface ($p < .0001$), indicating the superiority of the fine-tuning scheme over the training scheme from scratch. Of note, we observed no significant performance degradation for either deeply fine-tuned CNNs or fully trained CNNs, even after reducing the training patches to a single patient. This outcome resulted because each patient in our database provided approximately 12 ROIs, which enabled the extraction of a large number of distinct training patches that could be used for training and for fine-tuning the deep CNNs.

VII. DISCUSSION

In this study, to ensure generalizability of our findings, we considered 4 common medical imaging problems from 3 different imaging modality systems. Specifically, we chose PE detection as representative of computer-aided lesion detection in 3-dimensional volumetric images, polyp detection as representative of computer-aided lesion detection in 2-dimensional images, intima-media boundary segmentation as representative of machine learning-based medical image segmentation, and colonoscopy image quality assessment as representative of medical image classification. These applications differ because they require solving problems at different image scales. For instance, although intima-media boundary segmentation and PE detection may require the examination of a small sub-region

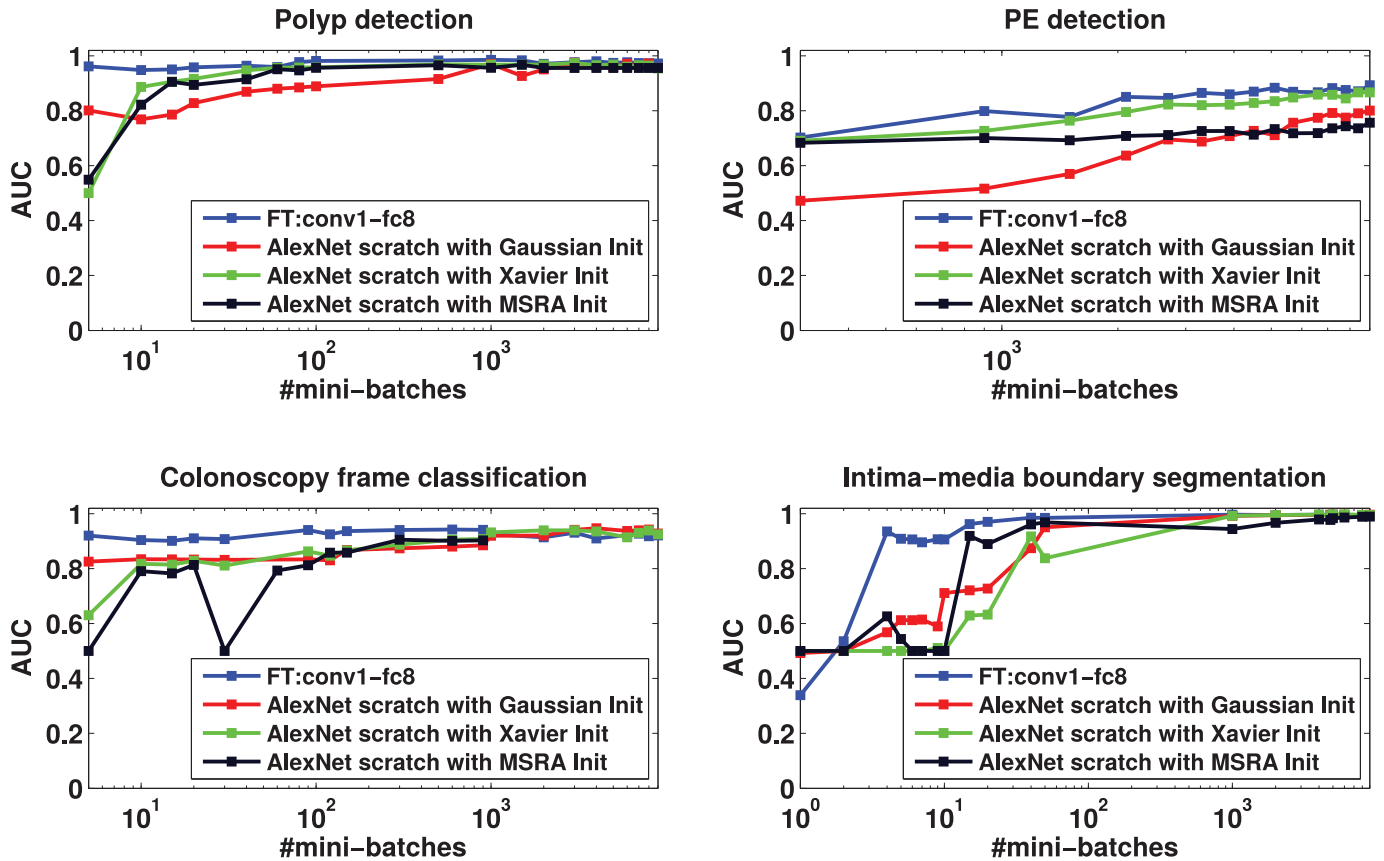


Fig. 10. Convergence speed for a deeply fine-tuned CNN and CNNs trained from scratch with three different initialization techniques.

within the images, polyp detection and frame classification demand far larger receptive fields. Therefore, we believe that the chosen applications encompass a variety of applications relevant to the field of medical imaging.

We thoroughly investigated the potential for fine-tuned CNNs in the context of medical image analysis as an alternative to training deep CNNs from scratch. We performed our analyses using both large training sets and reduced training sets. When using complete datasets, we observed that shallow tuning of the pre-trained CNNs most often led to a performance inferior to CNNs trained from scratch, whereas with deeper fine-tuning, we obtained performance comparable and even superior to CNNs trained from scratch. The performance gap between deeply fine-tuned CNNs and those trained from scratch widened when the size of training sets was reduced, which led us to conclude that fine-tuned CNNs should always be the preferred option regardless of the size of training sets available.

Another advantage of fine-tuned CNNs is the speed of convergence. To demonstrate this advantage, we compare the speed of convergence for a deeply fine-tuned CNN and a CNN trained from scratch in Fig. 10. For a thorough comparison, we used 3 different techniques to initialize the weights of the fully trained CNNs: 1) a method commonly known as Xavier, which was suggested in [40], 2) a revised version of Xavier called MSRA, which was suggested in [41], and a basic random initialization method based on Gaussian distributions. In this analysis, we computed the AUC on the validation data as a measure of convergence. Specifically, each snapshot of the model was ap-

plied to the patches of the validation set and then the classification performance was evaluated using an ROC analysis. Because we dealt with a 3-class classification problem for the task of intima-media boundary segmentation, we merged the 2 interface classes into a positive class and then computed the AUC for the resulting binary classification (interface vs. background). As shown, the fine-tuned CNN quickly reaches its maximum performance, but the CNNs trained from scratch require longer training in order to reach their highest performance. Furthermore, the use of different initialization techniques led to different trends of convergence, even though we observed no significant performance gain after complete convergence except for PE detection.

We observed that the depth of fine-tuning is fundamental to achieving accurate image classifiers. Although shallow tuning or updating the last few convolutional layers is sufficient for many applications in the field of computer vision to achieve state-of-the-art performance, we discovered that a deeper level of tuning is essential for medical imaging applications. For instance, we observed a marked performance gain using deeply fine-tuned CNNs, particularly for polyp detection and intima-media boundary segmentation, probably because of the substantial difference between these applications and the database with which the pre-trained CNN was constructed. However, we did not observe a similarly profound performance gain for colonoscopy frame classification, which we attribute to the relative similarity between ImageNet and the colonoscopy frames in our database. Specifically, both databases use high-resolution

images with similar low-level image information, which is why fine-tuning the late convolutional layers, which have application-specific features, is sufficient to achieve high-level performance for colonoscopy frame classification.

We based our experiments on the AlexNet architecture because a pre-trained AlexNet model was available in the Caffe library and that this architecture was deep enough that we could investigate the impact of the depth of fine-tuning on the performance of pre-trained CNNs. Alternatively, deeper architectures—such as VGGNet and GoogleNet—could have been used. Deeper architectures have recently shown relatively high performance for challenging computer vision tasks, but we do not anticipate a significant performance gain through the use of deeper architectures for medical imaging applications. We emphasize that the objective of this work was not to achieve the highest performance for a number of different medical imaging tasks but to examine the capabilities of fine-tuning in comparison with the training scheme from scratch. For these purposes, AlexNet is a reasonable architectural choice.

We would like to acknowledge that the performance curves reported for different models and applications may not be the best that we could achieve for each experiment. This sub-optimal performance is related to the choice of the hyper-parameters of CNNs that can influence the speed of convergence and final accuracy of a model. Although we attempted to find the working values of these parameters, finding the optimal values was not feasible given the large number of CNNs studied in our paper and that training each CNN was a time-consuming process even on the high-end GPUs. Nevertheless, this issue may not change our overall conclusions as the majority of the CNNs used in our comparisons are pre-trained models that may be less affected by the choice of hyper-parameters than the CNNs trained from scratch.

In this study, due to space constraints, we were not able to cover all medical imaging modalities. For instance, we did not study the performance of fine-tuning in MR images or histopathology images, for which full training of CNNs from scratch had shown promising performance. However, considering the successful knowledge transfer from natural images to CT, ultrasound, and endoscopy applications, we surmise that fine-tuning would succeed in other medical applications as well. Furthermore, our study was focused on fine-tuning of a pre-trained supervised model. However, a pre-trained unsupervised model such as those obtained by restricted Boltzmann machines (RBMs) or convolutional RBMs [74] could also be considered, even though the availability of ImageNet database with millions of labeled images from 1000 semantic classes may make the use of a pre-trained supervised model a natural choice for fine-tuning. Nevertheless, unsupervised models are still useful for 1D signal processing due to the absence of a large database of labeled 1D signals. For instance, fine-tuning of an unsupervised model was used in [75] for acoustic speech recognition and in [76] for detection of epilepsy in EEG recordings.

VIII. CONCLUSION

In this paper, we aimed to address the following central question in the context of medical image analysis: *Can the use of pre-trained deep CNNs, with sufficient fine-tuning,*

eliminate the need for training a deep CNN from scratch? Our extensive experiments, based on 4 distinct medical imaging applications from 3 different imaging modality systems, have demonstrated that deeply fine-tuned CNNs are useful for medical image analysis, performing as well as fully trained CNNs and even outperforming the latter when limited training data are available. Our results are important because they show that knowledge transfer from natural images to medical images is possible, even though the relatively large difference between source and target databases is suggestive that such application may not be possible. We also have observed that the required level of fine-tuning differed from one application to another. Specifically, for PE detection, we achieved performance saturation after fine-tuning the late fully connected layers; for colonoscopy frame classification, we achieved the highest performance through fine-tuning the late and middle layers; and for interface segmentation and polyp detection, we observed the highest performance by fine-tuning all layers in the pre-trained CNN. Our findings suggest that for a particular application, neither shallow tuning nor deep tuning may be the optimal choice. Through the layer-wise fine-tuning, one can learn the effective depth of tuning, as it depends on the application at hand and the amount of labeled data available for tuning. Layer-wise fine-tuning may offer a practical way to achieve the best performance for the application at hand based on the amount of available data. Our experiments further confirm the potential of CNNs for medical imaging applications because both deeply fine-tuned CNNs and fully trained CNNs outperformed the corresponding handcrafted alternatives.

REFERENCES

- [1] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.
- [2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [4] MIT Technol. Rev. "10 breakthrough technologies," [Online]. Available: <http://www.technologyreview.com/featuredstory/513696/deep-learning/>
- [5] C. Szegedy *et al.*, "Going deeper with convolutions" ArXiv, 2014 [Online]. Available: arXiv:1409.4842, to be published
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition" ArXiv, 2014 [Online]. Available: arXiv:1409.1556, to be published
- [7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Comput. Visi. ECCV*, 2014, pp. 818–833.
- [8] D. Eigen, J. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network" ArXiv, 2013 [Online]. Available: arXiv:1312.1847, to be published
- [9] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2009, pp. 153–160.
- [10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 512–519.
- [11] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition" ArXiv, 2014 [Online]. Available: arXiv:1406.5774, to be published
- [12] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, pp. 44–51.

- [13] W. Zhang *et al.*, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Med. Phys.*, vol. 21, no. 4, pp. 517–524, 1994.
- [14] H.-P. Chan, S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Med. Phys.*, vol. 22, no. 10, pp. 1555–1567, 1995.
- [15] S.-C. B. Lo *et al.*, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE Trans. Med. Imag.*, vol. 14, no. 4, pp. 711–718, Dec. 1995.
- [16] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *Information Processing in Medical Imaging*. New York: Springer, 2015, pp. 327–338.
- [17] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. IEEE 12th Int. Symp. o Biomed. Imag.*, 2015, pp. 79–83.
- [18] N. Tajbakhsh and J. Liang, "Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks," in *Proc. MICCAI*, 2015.
- [19] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. MICCAI*, 2013, pp. 411–418.
- [20] H. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. MICCAI*, P. Goll, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds., 2014, vol. 8673, LNCS, pp. 520–527.
- [21] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, "3d deep learning for efficient and robust landmark detection in volumetric data," in *Proc. MICCAI*, 2015, pp. 565–572.
- [22] J. Y. Shin, N. Tajbakhsh, R. T. Hurst, C. B. Kendall, and J. Liang, "Automating carotid intima-media thickness video interpretation with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, 2016.
- [23] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers, "Deep convolutional networks for pancreas segmentation in CT imaging," in *Proc. SPIE Med. Imag.*, 2015, pp. 94 131G–94 131G.
- [24] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks ArXiv, 2015 [Online]. Available: arXiv:1505.03540, to be published
- [25] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality iso-intense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.
- [26] D. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Red Hook, NY: Curran, 2012, vol. 25, pp. 2843–2851.
- [27] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a tri-planar convolutional neural network," in *Proc. MICCAI*, 2013, pp. 246–253.
- [28] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," in *Proc. SPIE Med. Imag.*, 2015, pp. 94 140V–94 140V.
- [29] B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, Apr. 2015, pp. 286–289.
- [30] J. Arevalo, F. Gonzalez, R. Ramos-Pollan, J. Oliveira, and M. G. Lopez, "Convolutional neural networks for mammography mass lesion classification," in *Proc. 37th Annu. Int. Conf. IEEE EMBC*, Aug. 2015, pp. 797–800.
- [31] T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in *Medical Computer Vision: Algorithms for Big Data*. New York: Springer, 2014, pp. 82–93.
- [32] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.
- [33] G. Carneiro, J. Nascimento, and A. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, vol. 9351, LNCS, pp. 652–660.
- [34] H.-C. Shin *et al.*, "Interleaved text/image deep mining on a very large-scale radiology database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1090–1099.
- [35] M. Gao *et al.*, "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks," in *1st Workshop Deep Learn. Med. Image Anal., Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Munich, Germany, 2015 [Online]. Available: www.research.rutgers.edu/~minggao/files/MingchenGao_MICCAIworkshop2015.pdf
- [36] J. Margea, A. Criminisi, R. C. Lozoya, D. C. Lee, and N. Ayache, "Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition," *Computer Methods Biomechan. Biomed. Eng., Imag. Visualizat.*, pp. 1–11, 2015.
- [37] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *J. Physiol.*, vol. 148, no. 3, pp. 574–591, 1959.
- [38] D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, "Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model," *Med. Phys.*, vol. 29, no. 12, pp. 2861–2870, 2002.
- [39] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding ArXiv, 2014 [Online]. Available: arXiv:1408.5093, to be published
- [40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 249–256.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification ArXiv, 2015 [Online]. Available: arXiv:1502.01852, to be published
- [42] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.
- [43] A. Pabby *et al.*, "Analysis of colorectal cancer occurrence during surveillance colonoscopy in the dietary polyp prevention trial," *Gastrointest Endosc.*, vol. 61, no. 3, pp. 385–391, 2005.
- [44] J. van Rijn *et al.*, "Polyp miss rate determined by tandem colonoscopy: A systematic review," *Am. Jo. Gastroenterol.*, vol. 101, no. 2, pp. 343–350, 2006.
- [45] D. H. Kim *et al.*, "CT colonography versus colonoscopy for the detection of advanced neoplasia," *N. Eng. J. Med.*, vol. 357, no. 14, pp. 1403–12, 2007.
- [46] D. Heresbach *et al.*, "Miss rate for colorectal neoplastic polyps: A prospective multicenter study of back-to-back video colonoscopies," *Endoscopy*, vol. 40, no. 4, pp. 284–290, 2008.
- [47] A. Leufkens, M. van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 05, pp. 470–475, 2012.
- [48] L. Rabeneck, H. El-Serag, J. Davila, and R. Sandler, "Outcomes of colorectal cancer in the united states: No change in survival (1986–1997)," *Am. J. Gastroenterol.*, vol. 98, no. 2, p. 471, 2003.
- [49] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.
- [50] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A comparative study of texture features for the discrimination of gastric polyps in endoscopic video," in *Proc. 18th IEEE Symp. Comput.-Based Med. Syst.*, 2005, pp. 575–580.
- [51] L. A. Alexandre, N. Nobre, and J. Casteleiro, "Color and position versus texture features for endoscopic polyp detection," in *Proc. Int. Conf. BioMedical Eng. Informat.*, 2008, vol. 2, pp. 38–42.
- [52] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *IEEE Int. Conf. Image Process.*, 2007, vol. 2, pp. II–465–II–468.
- [53] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.
- [54] J. Bernal, J. Sánchez, and F. Vilarino, "Impact of image preprocessing methods on polyp localization in colonoscopy frames," in *Proc. 35th Annu. Int. Conf. IEEE EMBC*, 2013, pp. 7350–7354.
- [55] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. de Groen, "Part-based multi-derivative edge cross-section profiles for polyp detection in colonoscopy," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1379–1389, Jul. 2014.
- [56] S. Y. Park, D. Sargent, I. Spofford, K. Vosburgh, and Y. A-Rahim, "A colon video analysis framework for polyp detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1408–1418, May 2012.
- [57] N. Tajbakhsh, S. Gurudu, and J. Liang, "A classification-enhanced vote accumulation scheme for detecting colonic polyps," in *Abdominal Imaging. Computation and Clinical Applications*. New York: Springer, 2013, vol. 8198, LNCS, pp. 53–62.

- [58] N. Tajbakhsh, C. Chi, S. R. Gurudu, and J. Liang, "Automatic polyp detection from learned boundaries," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, 2014, pp. 97–100.
- [59] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection using global geometric constraints and local intensity variation patterns," in *Proc. MICCAI*, 2014, pp. 179–187.
- [60] J. Liang and J. Bi, "Computer aided detection of pulmonary embolism with tobogganing and multiple instance classification in CT pulmonary angiography," in *Information Processing in Medical Imaging*. New York: Springer, 2007, pp. 630–641.
- [61] K. K. Calder, M. Herbert, and S. O. Henderson, "The mortality of untreated pulmonary embolism in emergency department patients," *Ann. Emerg. Med.*, vol. 45, no. 3, pp. 302–310, 2005.
- [62] G. Sadigh, A. M. Kelly, and P. Cronin, "Challenges, controversies, and hot topics in pulmonary embolism imaging," *Am. J. Roentgenol.*, vol. 196, no. 3, 2011.
- [63] J. Fairfield, "Toboggan contrast enhancement for contrast segmentation," in *Proc. 10th Int. Conf. Pattern Recognit.*, 1990, vol. 1, pp. 712–716.
- [64] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.
- [65] N. Tajbakhsh, "Automatic assessment of image informativeness in colonoscopy," in *Abdominal Imaging. Computational and Clinical Applications*. New York: Springer, 2014, pp. 151–158.
- [66] M. Arnold, A. Ghosh, G. Lacey, S. Patchett, and H. Mulcahy, "Indistinct frame detection in colonoscopy videos," in *Proc. 13th Int. Mach. Vis. Image Process. Conf.*, 2009, pp. 47–52.
- [67] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. C. de Groen, "Informative frame classification for endoscopy video," *Med. Image Anal.*, vol. 11, no. 2, pp. 110–127, 2007.
- [68] R.-M. Menchón-Lara and J.-L. Sancho-Gómez, "Fully automatic segmentation of ultrasound common carotid artery images based on machine learning," *Neurocomputing*, vol. 151, pp. 161–167, 2015.
- [69] R.-M. Menchón-Lara, M.-C. Bastida-Jumilla, A. González-López, and J. L. Sancho-Gómez, "Automatic evaluation of carotid intima-media thickness in ultrasounds using machine learning," in *Natural and Artificial Computation in Engineering and Medical Applications*. New York: Springer, 2013, pp. 241–249.
- [70] S. Petroudi, C. Loizou, M. Pantziaris, and C. Pattichis, "Segmentation of the common carotid intima-media complex in ultrasound images using active contours," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 11, pp. 3060–3069, Nov. 2012.
- [71] X. Xu, Y. Zhou, X. Cheng, E. Song, and G. Li, "Ultrasound intima-media segmentation using Hough transform and dual snake model," *Comput. Med. Imag. Graphics*, vol. 36, no. 3, pp. 248–258, 2012.
- [72] J. Liang, T. McInerney, and D. Terzopoulos, "United snakes," *Med. Image Anal.*, vol. 10, no. 2, pp. 215–233, 2006.
- [73] H. Sharma *et al.*, "ECG-based frame selection and curvature-based ROI detection for measuring carotid intima-media thickness," in *Proc. SPIE Med. Imag.*, 2014, pp. 904 016–904 016.
- [74] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.
- [75] O. Abdel-Hamid *et al.*, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.
- [76] D. Wulsin, J. Gupta, R. Mani, J. Blanco, and B. Litt, "Modeling electroencephalography waveforms with semi-supervised deep belief nets: Fast classification and anomaly measurement," *J. Neural Eng.*, vol. 8, no. 3, p. 036015, 2011.

SUPPLEMENTARY MATERIAL

TABLE S1: Statistical comparisons between the FROC curves shown in Fig. 2 for polyp detection (level of significance is $\alpha = 0.05$). The curves are compared at 0.01 and .001 false positives per frame, because they coincide with the elbows of the performance curves where they yield relatively higher sensitivity. A red cell indicates that a pair of curves are statistically different in neither of the chosen operating point whereas a green cell indicates at which operating points a statistically significant difference is observed.

	FT:only fc8	FT:fc7-fc8	FT:fc6-fc8	FT:conv5-fc8	FT:conv4-fc8	FT:conv3-fc8	FT:conv2-fc8	FT:conv1-fc8	AlexNet scratch
FT:only fc8									
FT:fc7-fc8	10 ⁻² , -3								
FT:fc6-fc8	10 ⁻² , -3								
FT:conv5-fc8	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3						
FT:conv4-fc8	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3						
FT:conv3-fc8	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻³					
FT:conv2-fc8	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3			
FT:conv1-fc8	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3			
AlexNet scratch	10 ⁻² , -3	10 ⁻²	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻³	10 ⁻³	10 ⁻² , -3	
Handcrafted [41]	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3	10 ⁻² , -3

TABLE S2: Statistical comparisons between the FROC curves shown in Fig. 4 for pulmonary embolism detection (level of significance is $\alpha=0.05$). Each cell presents a statistical comparison between a pair of FROC curves at 1, 2, 3, 4, and 5 false positives per volume. A red cell indicates that the two curves are not statistically different at any of the five operating points, but a green cell contains the operating points at which the two curves are statistically different.

	FT:only fc8	FT:fc7-fc8	FT:fc6-fc8	FT:conv5-fc8	FT:conv4-fc8	FT:conv3-fc8	FT:conv2-fc8	FT:conv1-fc8	AlexNet scratch
FT:only fc8									
FT:fc7-fc8	2,3,4,5								
FT:fc6-fc8	1,2,3,4,5								
FT:conv5-fc8	1,2,3,4,5	1,2							
FT:conv4-fc8	1,2,3,4,5	1,2,3							
FT:conv3-fc8	1,2,3,4,5	1,2,3,5	1						
FT:conv2-fc8	1,2,3,4,5	1,2,3,4,5							
FT:conv1-fc8	1,2,3,4,5	1,2,3,4,5	3						
AlexNet scratch	1,2,3,4,5	1,2,3							
Handcrafted [59]	1,2,3,4,5	1,2,3,5							

Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally*

Zongwei Zhou¹, Jae Shin¹, Lei Zhang¹, Suryakanth Gurudu², Michael Gotway², and Jianming Liang¹

¹Arizona State University

{zongweiz, sejong, lei.zhang.10, jianming.liang}@asu.edu

²Mayo Clinic

{gurudu.suryakanth, gotway.michael}@mayo.edu

Abstract

Intense interest in applying convolutional neural networks (CNNs) in biomedical image analysis is wide spread, but its success is impeded by the lack of large annotated datasets in biomedical imaging. Annotating biomedical images is not only tedious and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible. To dramatically reduce annotation cost, this paper presents a novel method called AIFT (active, incremental fine-tuning) to naturally integrate active learning and transfer learning into a single framework. AIFT starts directly with a pre-trained CNN to seek “worthy” samples from the unannotated for annotation, and the (fine-tuned) CNN is further fine-tuned continuously by incorporating newly annotated samples in each iteration to enhance the CNN’s performance incrementally. We have evaluated our method in three different biomedical imaging applications, demonstrating that the cost of annotation can be cut by at least half. This performance is attributed to the several advantages derived from the advanced active and incremental capability of our AIFT method.

1. Introduction

Convolutional neural networks (CNNs) [14] have brought about a revolution in computer vision thanks to large annotated datasets, such as ImageNet [6] and Places [27]. As evidenced by an IEEE TMI special issue [8] and two forthcoming books [28, 17], intense interest in applying CNNs in biomedical image analysis is wide spread,

but its success is impeded by the lack of such large annotated datasets in biomedical imaging. Annotating biomedical images is not only tedious and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible. Therefore, we seek to answer this critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging.* In doing so, we present a novel method called AIFT (active, incremental fine-tuning) to naturally integrate active learning and transfer learning into a single framework. Our AIFT method starts directly with a pre-trained CNN to seek “salient” samples from the unannotated for annotation, and the (fine-tuned) CNN is continuously fine-tuned by incrementally enlarging the training dataset with newly annotated samples. We have evaluated our method in three different applications including colonoscopy frame classification, polyp detection, and pulmonary embolism (PE) detection, demonstrating that the cost of annotation can be cut by at least half.

This outstanding performance is attributed to a simple yet powerful observation: To boost the performance of CNNs in biomedical imaging, multiple patches are usually generated automatically for each candidate through data augmentation; these patches generated from the same candidate share the same label, and are naturally expected to have similar predictions by the current CNN before they are expanded into the training dataset. As a result, their *entropy* and *diversity* provide a useful indicator to the “power” of a candidate in elevating the performance of the current CNN. However, automatic data augmentation inevitably generates “hard” samples for some candidates, injecting noisy labels; therefore, to significantly enhance the robustness of our method, we compute entropy and diversity by selecting only a portion of the patches of each candidate according to the predictions by the current CNN.

*This research has been supported partially by NIH under Award Number R01HL128785, by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Several researchers have demonstrated the utility of fine-tuning CNNs for biomedical image analysis, but they only performed one-time fine-tuning, that is, simply fine-tuning a pre-trained CNN once with available training samples involving no active selection processes (e.g., [4, 19, 5, 2, 21, 7, 18, 24]). To our knowledge, our proposed method is among the first to integrate active learning into fine-tuning CNNs in a continuous fashion to make CNNs more amicable for biomedical image analysis with an aim to cut annotation cost dramatically. Compared with conventional active learning, our AIFT method offers several advantages:

1. Starting with a completely empty labeled dataset, requiring no initial seed labeled samples (see Alg. 1);
2. Incrementally improving the learner through continuous fine-tuning rather than repeatedly re-training (see Sec. 3.1);
3. Naturally exploiting expected consistency among the patches associated for each candidate to select samples “worthy” of labeling (see Sec. 3.2);
4. Automatically handling noisy labels as only a portion (e.g., a quarter) of the patches in each candidate participate in the selection process (see Sec. 3.3);
5. Computing entropy and diversity locally on a small number of patches within each candidate, saving computation time considerably (see Sec. 3.3).

More importantly, our method has the potential to exert important impact on computer-aided diagnosis (CAD) in biomedical imaging, because the current regulations require that CAD systems be deployed in a “closed” environment, in which all CAD results be reviewed and errors if any be corrected by radiologists; as a result, all false positives are supposed to be dismissed and all false negatives supplied, an instant on-line feedback that may make CAD systems self-learning and improving possible after deployment given the continuous fine-tuning capability of our method.

2. Related work

2.1. Transfer learning for medical imaging

Gustavo *et al.* [2] replaced the fully connected layers of a pre-trained CNN with a new logistic layer and trained only the appended layer with the labeled data while keeping the rest of the network the same, yielding promising results for classification of unregistered multiview mammograms. In [5], a fine-tuned pre-trained CNN was applied for localizing standard planes in ultrasound images. Gao *et al.* [7] fine-tuned all layers of a pre-trained CNN for automatic classification of interstitial lung diseases. In [21], Shin *et al.* used fine-tuned pre-trained CNNs to automatically map medical images to document-level topics, document-level sub-topics, and sentence-level topics. In [18], fine-tuned pre-trained CNNs were used to automatically retrieve missing or noisy cardiac acquisition plane

information from magnetic resonance imaging and predict the five most common cardiac views. Schlegl *et al.* [19] explored unsupervised pre-training of CNNs to inject information from sites or image classes for which no annotations were available, and showed that such across site pre-training improved classification accuracy compared to random initialization of the model parameters. Tajbakhsh *et al.* [24] systematically investigated the capabilities of transfer learning in several medical imaging applications. However, they all performed *one-time fine-tuning*—simply fine-tuning a pre-trained CNN just once with available training samples, involving neither active selection processes nor continuous fine-tuning.

2.2. Integrating active learning with deep learning

The literature of general active learning and deep learning is rich and deep [8, 28, 17, 20, 9, 10, 26]. However, the research aiming to integrate active learning with deep learning is sparse: Wang and Shang [25] may be the first to incorporate active learning with deep learning, and based their approach on stacked restricted Boltzmann machines and stacked autoencoders. A similar idea was reported for hyperspectral image classification [15]. Stark *et al.* [22] applied active learning to improve the performance of CNNs for CAPTCHA recognition, while Al Rahhal *et al.* [1] exploited deep learning for active electrocardiogram classification. All these approaches are fundamentally different from our AIFT approach in that in each iteration they all *repeatedly re-trained the learner from scratch* while we continuously fine-tune the (fine-tuned) CNNs in an incremental manner, offering five advantages as listed in Sec. 1.

3. Proposed method

We present our AIFT method in the context of computer-aided diagnosis (CAD) in biomedical imaging. A CAD system typically has a candidate generator, which can quickly produce a set of candidates, among which, some are *true* positives and some are *false* positives. After candidate generation, the task is to train a classifier to eliminate as many false positives as possible while keeping as many true positives as possible. To train a classifier, each of the candidates must be labeled. We assume that each candidate takes one of $|Y|$ possible labels. To boost the performance of CNNs for CAD systems, multiple patches are usually generated automatically for each candidate through data augmentation; these patches generated from the same candidate inherit the candidate’s label. In other words, all labels are acquired at the candidate level. Mathematically, given a set of candidates, $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$, where n is the number of candidates, and each candidate $\mathcal{C}_i = \{x_i^1, x_i^2, \dots, x_i^m\}$ is associated with m patches, our AIFT algorithm iteratively selects a set of candidates for labeling (illustrated in Alg. 1).

Algorithm 1: Active incremental fine-tuning method.

Input: $\mathcal{U} = \{\mathcal{C}_i\}, i \in [1, n]$ $\{\mathcal{U}$ contains n candidates $\}$ $\mathcal{C}_i = \{x_i^j\}, j \in [1, m]$ $\{\mathcal{C}_i$ has m patches $\}$ \mathcal{M}_0 : pre-trained CNN b : batch size α : patch selection ratio**Output:** \mathcal{L} : labeled candidates \mathcal{M}_t : fine-tuned CNN model at Iteration t **Functions:** $p \leftarrow P(\mathcal{C}, \mathcal{M})$ $\{\text{outputs of } \mathcal{M} \text{ given } \forall x \in \mathcal{C}\}$ $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$ $\{\text{fine-tune } \mathcal{M}_{t-1} \text{ with } \mathcal{L}\}$ $a \leftarrow \text{mean}(p_i)$ $\{a = \frac{1}{m} \sum_{j=1}^m p_i^j\}$ **Initialize:** $\mathcal{L} \leftarrow \emptyset, t \leftarrow 1$

```
1 repeat
2   for each  $\mathcal{C}_i \in \mathcal{U}$  do
3      $p_i \leftarrow P(\mathcal{C}_i, \mathcal{M}_{t-1})$ 
4     if  $\text{mean}(p_i) > 0.5$  then
5        $S'_i \leftarrow$  top  $\alpha$  percent of the patches of  $\mathcal{C}_i$ 
6     else
7        $S'_i \leftarrow$  bottom  $\alpha$  percent of the patches of  $\mathcal{C}_i$ 
8     end
9     Build matrix  $R_i$  using Eq. 3 for  $S'_i$ 
10  end
11  Sort  $\mathcal{U}$  according to the numerical sum of  $R_i$ 
12  Query labels for top  $b$  candidates, yielding  $\mathcal{Q}$ 
13   $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{Q}; \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{Q}$ 
14   $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1}); t \leftarrow t + 1$ 
15 until classification performance is satisfactory;
```

3.1. Continuous fine-tuning

At the beginning, the labeled dataset \mathcal{L} is empty; we take a pre-trained CNN (e.g., AlexNet) and run it on \mathcal{U} to select b number of candidates for labeling. The newly labeled candidates will be incorporated into \mathcal{L} to continuously fine-tune the CNN incrementally until the performance is satisfactory. Several researchers have demonstrated that fine-tuning offers better performance and is more robust than training from scratch. From our experiments, we have found that continuously fine-tuning the CNN, which has been fine-tuned in the previous iteration, with enlarged datasets converges faster than repeatedly fine-tuning the original pre-trained CNN. We also found that continuously fine-tuning the CNN with only newly labeled data demands careful meta-parameter adjustments.

3.2. Active candidate selection

In active learning, the key is to develop a criterion for determining the “worthiness” of a candidate for annotation.

Our criterion is based on an observation: All patches generated from the same candidate share the same label; they are expected to have similar predictions by the current CNN. As a result, their *entropy* and *diversity* provide a useful indicator to the “power” of a candidate in elevating the performance of the current CNN. Intuitively, entropy captures the classification certainty—higher uncertainty values denote higher degrees of information; while diversity indicates the prediction consistency among the patches within a candidate—higher diversity values denote higher degrees of prediction inconsistency among the patches within a candidate. Therefore, candidates with higher entropy and higher diversity are expected to contribute more in elevating the current CNN’s performance. Formally, assuming the prediction of patch x_i^j by the current CNN is p_i^j , we define its entropy as:

$$e_i^j = - \sum_{k=1}^{|Y|} p_i^{j,k} \log p_i^{j,k} \quad (1)$$

and diversity between patches x_i^j and x_i^l of candidate \mathcal{C}_i as:

$$d_i(j, l) = \sum_{k=1}^{|Y|} (p_i^{j,k} - p_i^{l,k}) \log \frac{p_i^{j,k}}{p_i^{l,k}} \quad (2)$$

Entropy e_i^j denotes the information furnished by patch x_i^j of candidate \mathcal{C}_i in the unlabeled pool. Diversity $d_i(j, l)$, captured by the symmetric Kullback Leibler divergence [13], estimates the amount of information overlap between patches x_i^j and x_i^l of candidate \mathcal{C}_i . By definition, all the entries in e_i^j and $d_i(j, l)$ are non-negative. Further, $d_i(j, j) = 0$, $\forall j$, therefore, for notational simplicity, we combine e_i^j and $d_i(j, l)$ into a single matrix R_i for each candidate \mathcal{C}_i :

$$R_i(j, l) = \begin{cases} \lambda_1 e_i^j & \text{if } j = l, \\ \lambda_2 d_i(j, l) & \text{otherwise} \end{cases} \quad (3)$$

where λ_1 and λ_2 are trade-offs between entropy and diversity. We use two parameters for convenience, so as to easily turn on/off entropy or diversity during experiments.

3.3. Handling noisy labels via majority selection

Automatic data augmentation is essential to boost CNN’s performance, but it inevitably generates “hard” samples for some candidates as shown in Fig. 1 and Fig. 2 (c), injecting noisy labels; therefore, to significantly enhance the robustness of our method, we compute entropy and diversity by selecting only a portion of the patches of each candidate according to the predictions by the current CNN. Specially, for each candidate \mathcal{C}_i we first compute the average probabilistic prediction of all of its patches:

$$a_i = \frac{1}{m} \sum_{j=1}^m p_i^j \quad (4)$$

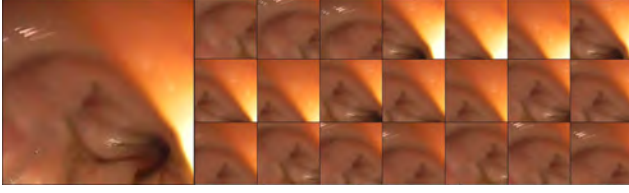


Figure 1: Experts label a frame based on the overall quality: if over 75% of a frame (i.e., candidate in this application) is clear, it is considered “informative”. For example, the whole frame (the leftmost image) is labeled as “informative”, but not all the patches associated with this frames are “informative”, although they inherit the “informative” label. This is the main motivation for the majority selection in our AIFT method.

where m is the number of patches within candidate \mathcal{C}_i , p_i^j is the prediction probability of patch x_i^j . If $a_i > 0.5$, we select the top α percent patches; otherwise, the bottom α percent patches. Based on the selected patches, we then use Eq. 3 to construct the score matrix R_i of size $\alpha m \times \alpha m$ for each candidate \mathcal{C}_i in \mathcal{U} . Our proposed majority selection method automatically excludes the patches with noisy labels because of their low confidences. We should note that the idea of combining entropy and diversity was inspired by [3], but there is a fundamental difference because they computed R across the whole unlabeled dataset with time complexity $\mathcal{O}(m^2)$, which is very computational expensive, while we compute $R_i(j, l)$ locally on the selected patches within each candidate, saving computation time considerably with time complexity $\mathcal{O}(\alpha^2 m^2)$, where $\alpha = 1/4$ in our experiments.

3.4. An illustration of prediction patterns

Given unlabeled candidates $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$ with $\mathcal{C}_i = \{x_i^1, x_i^2, \dots, x_i^m\}$, assuming the prediction of patch x_i^j by the current CNN is p_i^j , we call the histogram of p_i^j for $j \in [1, m]$ the prediction pattern of candidate \mathcal{C}_i . As shown in Column 1 of Tab. 1, there are seven typical prediction patterns:

- Pattern A: The patches’ predictions are mostly concentrated at 0.5, with a higher degree of uncertainty. Most active learning algorithms [20, 9] favor this type of candidate as it is good at reducing the uncertainty.
- Pattern B: It is flatter than Pattern A, as the patches’ predictions are spread widely from 0 to 1, yielding a higher degree of inconsistency. Since all the patches belonging to a candidate are generated via data argumentation, they (at least the majority of them) are expected to have similar predictions. This type of candidate has the potential to contribute significantly to enhancing the current CNN’s performance.
- Pattern C: The patches’ predictions are clustered at both ends, resulting in a higher degree of diversity. This type of candidate is most likely associated with noisy labels at the patch level as illustrated in Fig. 1, and it is the least favorable in active selection because it may cause confusion in fine-tuning the CNN.
- Patterns D and E: The patches’ predictions are clustered at one end (i.e., 0 or 1) with a higher degree of certainty. The annotation of these types of candidates at this stage should be postponed because the current CNN has most likely predicted them correctly; they would contribute very little to fine-tuning the current CNN. However, these candidates may evolve into different patterns worthy of annotation with more fine-tuning.
- Patterns F and G: They have higher degrees of certainty in some of the patches’ predictions and are associated with some outliers in the patches’ predictions. These types of candidates are valuable because they are capable of smoothly improving the CNN’s performance. Though they may not make significant contributions, they should not cause dramatic harm to the CNN’s performance.

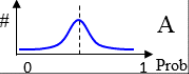
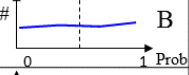
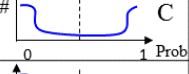
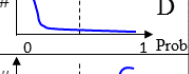
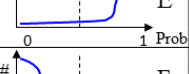
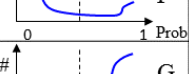
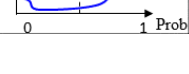
4. Applications

In this section, we apply our method to three different applications including colonoscopy frame classification, polyp detection, and pulmonary embolism (PE) detection. Our AIFT algorithm is implemented in the Caffe framework [11] based on the pre-trained AlexNet model [12]. In the following, we shall evaluate six variants of AIFT (active incremental fine-tuning) including Diversity^{1/4} (using diversity on 1/4 of the patches of each candidate), Diversity (using diversity on all the patches of each candidate), Entropy^{1/4}, Entropy, (Entropy+Diversity)^{1/4}, (Entropy+Diversity), and compare them with IFT Random (incremental fine-tuning with random candidate selection) and Learning from Scratch in terms of AUC (area under ROC curve).

4.1. Colonoscopy Frame Classification

Objective quality assessment of colonoscopy procedures is vital to ensure high-quality colonoscopy. A colonoscopy video typically contains a large number of non-informative images with poor colon visualization that are not ideal for inspecting the colon or performing therapeutic actions. The larger the fraction of non-informative images in a video, the lower the quality of colon visualization, thus the lower the quality of colonoscopy. Therefore, one way to measure the quality of a colonoscopy procedure is to monitor the quality of the captured images. Technically, image quality assessment at colonoscopy can be formulated as an image clas-

Table 1: Relationships among seven prediction patterns and six AIFT methods in active candidate selection. We assume that a candidate has 11 patches, and their probabilities predicted by the current CNN are listed in Column 2. AIFT Entropy $^\alpha$, Diversity $^\alpha$, and (Entropy+Diversity) $^\alpha$ operate on the top or bottom α percent of the candidate’s patches based on the majority prediction as described in Sec. 3.3. In this illustration, we choose α to be 1/4, meaning that the selection criterion (Eq. 3) is computed based on 3 patches within each candidate. The first choice of each method is highlighted in yellow and the second choice is in light yellow.

Prediction Pattern	Example	Entropy	Entropy ^{1/4}	Diversity	Diversity ^{1/4}	(Entropy+Diversity)	(Entropy+Diversity) ^{1/4}
 A	{0.4 0.4 0.4 0.5 0.5 0.5 0.5 0.5 0.5 0.6 0.6}	7.52	2.02	4.38	0.00	11.90	2.02
 B	{0.0 0.1 0.2 0.3 0.4 0.4 0.6 0.7 0.8 1.0 1.0}	4.57	0.83	1237.21	20.79	1241.77	21.62
 C	{0.0 0.0 0.0 0.1 0.1 0.9 0.9 1.0 1.0 1.0 1.0}	1.30	0.00	2816.66	0.00	2817.96	0.00
 D	{0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.1 0.1 0.1}	1.30	0.00	189.54	0.00	190.84	0.00
 E	{0.9 0.9 0.9 0.9 1.0 1.0 1.0 1.0 1.0 1.0 1.0}	1.30	0.00	189.54	0.00	190.84	0.00
 F	{0.0 0.0 0.1 0.1 0.1 0.1 0.2 0.2 0.3 0.9 1.0}	3.24	0.33	1076.87	13.54	1080.11	13.86
 G	{0.0 0.1 0.7 0.8 0.8 0.9 0.9 0.9 0.9 1.0 1.0}	3.24	0.33	1076.87	13.54	1080.11	13.86

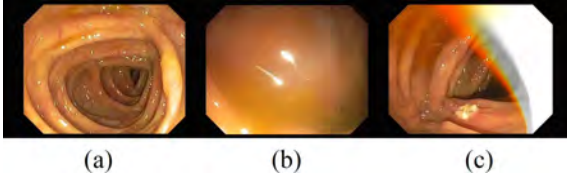


Figure 2: Three colonoscopy frames, (a) informative, (b) non-informative, and (c) ambiguous but labeled “informative” because it is mostly clear. The ambiguous frames contain both clear and blur parts, and generate noisy labels at the patch level via automatic data argumentation. Our AIFT method aims to automatically handle the label noise.

sification task whereby an input image is labeled as either informative or non-informative.

For the experiments, 4,000 colonoscopy frames are selected from 6 complete colonoscopy videos. A trained expert then manually labeled the collected images as informative or non-informative. A gastroenterologist further reviewed the labeled images for corrections. The labeled frames at the video level are separated into training and test sets, each containing approximately 2,000 colonoscopy frames. For data augmentation, we extracted 21 patches

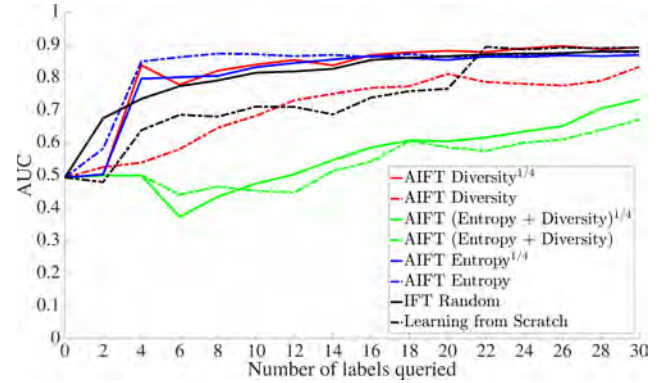


Figure 3: Comparing 8 methods in colonoscopy frame classification (see text for a detailed analysis).

from each frame.

In all three applications, our AIFT begins with an empty training dataset and directly uses AlexNet pre-trained on ImageNet. Fig. 3 shows that at the first step (with 2 labels queried), IFT Random yields the best performance. There are two possible reasons: (1) random selection gives the samples with the positive/negative ratio compatible with the test dataset; (2) the pre-trained AlexNet gives poor predic-

tions on our dataset, as it was trained by natural images instead of biomedical images. Its output probabilities are mostly confused or even incorrect, yielding poor selection scores. However, AIFT Diversity^{1/4}, Entropy, Entropy^{1/4} quickly surpass IFT Random after the first fine-tuning, as they select important samples for fine-tuning, making the training process more efficient than just randomly selecting from the remaining training dataset. AIFT Entropy and Diversity^{1/4} with only 4 label queries can achieve the performance of IFT Random with 18 label queries, and that of Learning from Scratch with 22 randomly selected frames. Thereby, more than 75% labeling cost could be saved from IFT Random and 80% from Learning from Scratch.

AIFT Diversity works even poorer than IFT Random because of noisy labels generated through data augmentation. AIFT Diversity strongly favors frames whose prediction pattern resembles Pattern C (see Tab. 1). Naturally, it will most likely select an ambiguous frame such as Fig. 1 and Fig. 2 (c), because predictions of its patches are highly diverse. All patches generated from the same frame inherit the same label as the frame; therefore, at the patch level, the labels are very noisy for the ambiguous frames. AIFT Entropy, Entropy^{1/4}, and Diversity^{1/4} can automatically exclude the noisy label, naturally yielding outstanding performance. Given the outstanding performance of AIFT Entropy, Entropy^{1/4}, and Diversity^{1/4}, one may consider combining entropy and diversity, but unfortunately, combinations do not always give better performance, because finding a nice balance between entropy and diversity is tricky as shown in our example analysis in Tab. 1 and supplementary material.

4.2. Polyp Detection

Colonoscopy is the preferred technique for colon cancer screening and prevention. The goal of colonoscopy is to find and remove colonic polyps—precursors to colon cancer—as shown in Fig. 4. For polyp detection, our database contains 38 short colonoscopy videos from 38 different patients, and they are separated into the training dataset (21 videos; 11 with polyps and 10 without polyps) and the testing dataset (17 videos; 8 videos with polyps and 9 videos without polyps). There are no overlaps between the training dataset and testing dataset at the patient level. Each colonoscopy frame in the data set comes with a binary ground truth image. 16300 candidates and 11950 candidates were generated from the training dataset and testing dataset, respectively.

At each polyp candidate location with the given bounding box, we perform a data augmentation by a factor $f \in \{1.0, 1.2, 1.5\}$. At each scale, we extract patches after the candidate is translated by 10 percent of the resized bounding box in vertical and horizontal directions. We further rotate each resulting patch 8 times by mirroring and flipping. The



Figure 4: Polyps in colonoscopy videos with different shape and appearance.

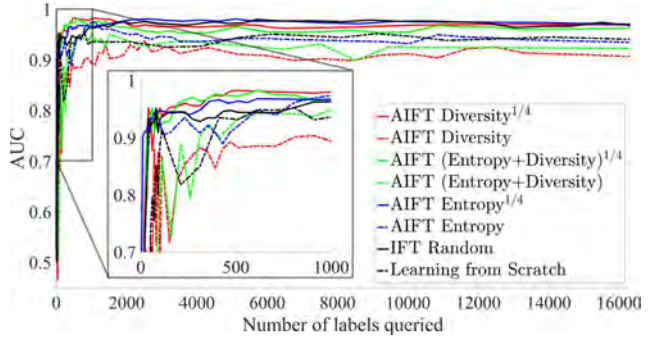


Figure 5: Comparing 8 methods in polyp detection (see text for a detailed analysis).

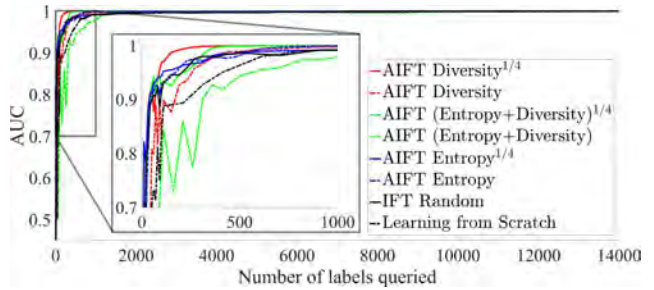


Figure 6: Monitor the performance of the proposed method on the remaining training dataset. Using 5% of the whole training dataset (800/16300), the CNN can predict almost perfectly on the remaining 95% dataset.

patches generated by data augmentation belong to the same candidate.

Fig. 5 shows that AIFT (Entropy+Diversity)^{1/4} and Diversity^{1/4} reach the peak performance with 610 label queries, while IFT Random needs 5711 queries, indicating that AIFT can cut nearly 90% of the annotation cost required by IFT Random. The fast convergence of AIFT (Entropy+Diversity)^{1/4} and Diversity^{1/4} is attributed to the majority selection method, which can efficiently select the informative and representative candidates while excluding those with noisy labels. When the queried number is about 5000, the AIFT Entropy^{1/4} reaches its peak performance. The reason is that the entropy can only measure the informativeness so the queried sample is very likely to be similar to each other. It needs more queries to select most

of the informative candidates. AIFT Diversity and (Entropy+Diversity) cannot perform as well as the counterparts with the majority selection due to noisy labels. Learning from Scratch never achieves the performance of fine-tuning even if all training samples are used, which is in agreement with [24].

To gain further insights, we also monitor the performance of the 8 methods on the remaining training dataset. Each time after we have fine-tuned the previous CNN, we test it on the remaining training dataset. We have observed that only 800 candidates are needed to reach the maximum performance. As is shown in Fig. 6, the candidates selected by our method, which are only 5% (800/16300) of all the candidates, can represent the remaining dataset, because in colonoscopy videos consecutive frames are usually similar to each other.

4.3. Pulmonary Embolism Detection

Our experiments are based on the PE candidates generated by the method proposed in [16] and the image representation introduced in [23] as shown in Fig. 7. We adopt the 2-channel representation because it consistently captures PEs in cross-sectional and longitudinal views of vessels, achieving greater classification accuracy and accelerating CNN training process. In order to feed the RGB-like patches into CNN, the 2-channel patches are converted to 3-channel RGB-like patches by duplicating the second channel. For experiments, we use a database consisting of 121 CTPA datasets with a total number of 326 PEs. The tobogganing algorithm [16] is applied to obtain a crude set of PE candidates. 6255 PE candidates are generated, of which 5568 are false positives and 687 are true positives. To train CNN, we extract patches of 3 different physical sizes, i.e., 10 mm-, 15 mm-, and 20 mm-wide. Then, we translate each candidate location along the direction of the affected vessel 3 times, up to 20% of the physical size of each patch. Then, data augmentation for training dataset is performed by rotating the longitudinal and cross-sectional vessel planes around the vessel axis, resulting in 5 additional variations for each scale and translation.

Finally, a stratified training dataset with 434 true positive PE candidates and 3406 false positive PE candidates would be generated for training and incrementally fine-tuning the CNN and a testing dataset with 253 true positive PE candidates and 2162 false positive PE candidates. The overall PE probability is calculated by averaging the probabilistic prediction generated for the patches within PE candidate after data augmentation.

Fig. 8 compares the 8 methods on the testing dataset. The performance of each method becomes saturated after 2000 labels queried. AIFT (Entropy+Diversity)^{1/4} and Diversity^{1/4} converge the fastest among the 8 methods and yields the best overall performance, attributed to

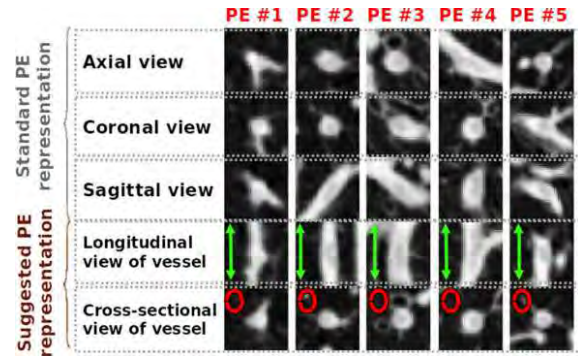


Figure 7: Five different PEs in the standard 3-channel representation, as well as in the 2-channel representation [23], which was adopted in this work because it achieves greater classification accuracy and accelerates CNN training convergence. The figure is used with permission.

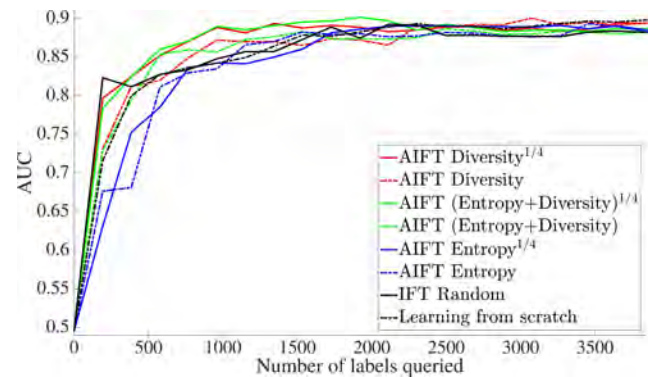


Figure 8: Comparing 8 methods in pulmonary embolism detection (see text for a detailed analysis).

majority selection method proposed in this work. AIFT (Entropy+Diversity)^{1/4} and Diversity^{1/4} with only 1000 labels required can achieve the performance of random selecting 2200 labels fine-tune from AlexNet (IFT Random). Note that even AIFT Diversity reach its peak performance when about 3100 samples queried because PE data set injected little noisy labels. Since entropy favors the uncertain ambiguous samples, both AIFT Entropy^{1/4} and Entropy perform bad at the beginning. IFT Random outperforms at the first few steps as analysed in Sec. 4.1, but increase slowly overall. Based on this analysis, the cost of annotation can be cut at least half by the our method.

4.4. Observations on selected patterns

We meticulously monitored the active selection process and examined the selected candidates, as an example, we include the top 10 candidates selected by the six AIFT methods at Iteration 3 in colonoscopy frame classification in the

supplementary material (see Fig. 10). From this process, we have observed the following:

- Patterns A and B are dominant in the earlier stages of AIFT as the CNN has not been fine-tuned properly to the target domain.
- Patterns C, D and E are dominant in the later stages of AIFT as the CNN has been largely fine-tuned on the target dataset.
- The majority selection—AIFT Entropy^{1/4}, Diversity^{1/4}, or (Entropy+Diversity)^{1/4}—is effective in excluding Patterns C, D, and E, while AIFT Entropy (without the majority selection) can handle Patterns C, D, and E reasonably well.
- Patterns B, F, and G generally make good contributions to elevating the current CNN's performance.
- AIFT Entropy and Entropy^{1/4} favor Pattern A because of its higher degree of uncertainty as shown in Fig. 10.
- AIFT Diversity^{1/4} prefers Pattern B while AIFT Diversity prefers Pattern C (Fig. 10). This is why AIFT Diversity may cause sudden disturbances in the CNN's performance and why AIFT Diversity^{1/4} should be preferred in general.
- Combining entropy and diversity would be highly desirable, but striking a balance between them is not trivial, because it demands application-specific λ_1 and λ_2 (see Eq. 3) and requires further research.

5. Conclusion, discussion and future work

We have developed an active, incremental fine-tuning method, integrating active learning with transfer learning, offering several advantages: It starts with a completely empty labeled dataset, and incrementally improves the CNN's performance through continuous fine-tuning by actively selecting the most informative and representative samples. It also can automatically handle noisy labels via majority selection and it computes entropy and diversity locally on a small number of patches within each candidate, saving computation time considerably. We have evaluated our method in three different biomedical imaging applications, demonstrating that the cost of annotation can be cut by at least half. This performance is attributed to the advanced active and incremental capability of our AIFT method.

We based our experiments on the AlexNet architecture because a pre-trained AlexNet model is available in the Caffe library and its architecture strikes a nice balance in depth: it is deep enough that we can investigate the impact of AIFT on the performance of pre-trained CNNs, and it is also shallow enough that we can conduct experiments quickly. Alternatively, deeper architectures such as VGG, GoogleNet, and Residual network could have been used and have shown relatively high performance for challenging computer vision tasks. However, the purpose of this work is

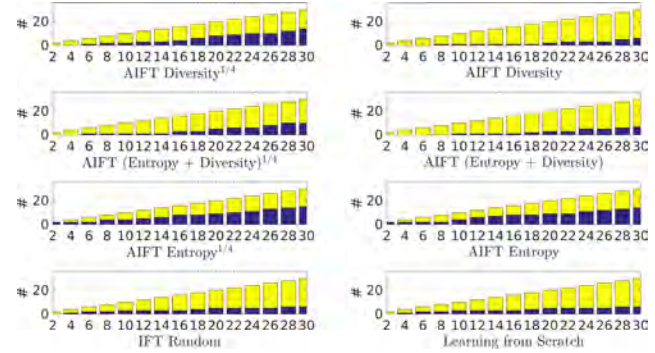


Figure 9: Positive/negative ratio in the samples selected by six methods. Yellow bar represents the negatives and blue bar represents the positives.

not to achieve the highest performance for different biomedical image tasks but to answer the critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging.* The architecture and learning parameters are reported in the supplementary material.

In the real world, datasets are usually unbalanced. In order to achieve good classification performance, both classes of samples should be used in training. Fig. 9 shows the positive/negative label ratio of the samples selected by the six methods in each iteration in colonoscopy quality application. For random selection, the ratio is nearly the same as whole training dataset, a reason that IFT Random has stable performance at the cold-start. AIFT Diversity^{1/4}, Entropy^{1/4} and Entropy seem capable of keeping the dataset balanced automatically, a new observation that deserves more investigation in the future.

We choose to select, classify and label samples at the candidate level. Labeling at the patient level would certainly reduce the cost of annotation more but introduce more severe label noise; labeling at the patch level would cope with the label noise but impose a much heavier burden on experts for annotation. We believe that labeling at the candidate level offers a sensible balance in our three applications.

Finally, in this paper, we use only entropy and diversity as the criteria. In theory, a large number of active selection methods may be designed, but we have found that there are only seven fundamental patterns as summarized in the Sec. 3.4. As a result, we could conveniently focus on comparing the seven patterns rather than the many methods. Multiple methods may be used to select a particular pattern: for example, entropy, Gaussian distance, and standard deviation would seek Pattern A, while diversity, variance, and divergence look for Pattern C. We would not expect significant performance differences among the methods within each group, resulting in six major selection methods for deep comparisons based on real-world clinical applications.

References

- [1] M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016. [2](#)
- [2] G. Carneiro, J. Nascimento, and A. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 652–660. Springer International Publishing, 2015. [2](#)
- [3] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1945–1958, 2015. [4](#)
- [4] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–514. Springer, 2015. [2](#)
- [5] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *Biomedical and Health Informatics, IEEE Journal of*, 19(5):1627–1636, Sept 2015. [2](#)
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. [1](#)
- [7] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. In *the 1st Workshop on Deep Learning in Medical Image Analysis, International Conference on Medical Image Computing and Computer Assisted Intervention, at MICCAI-DLMIA'15*, 2015. [2](#)
- [8] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016. [1](#), [2](#)
- [9] I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov. *JMLR Workshop and Conference Proceedings (Volume 16): Active Learning Challenge*. Microtome Publishing, 2011. [2](#), [4](#)
- [10] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008. [2](#)
- [11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. [4](#)
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [4](#)
- [13] M. Kukar. Transductive reliability estimation for medical diagnosis. *Artificial Intelligence in Medicine*, 29(1):81–106, 2003. [3](#)
- [14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. [1](#)
- [15] J. Li. Active learning for hyperspectral image classification with a stacked autoencoders based neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1062–1065, Sept 2016. [2](#)
- [16] J. Liang and J. Bi. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in ct pulmonary angiography. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 630–641. Springer, 2007. [7](#)
- [17] L. Lu, Y. Zheng, G. Carneiro, and L. Yang. *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*. Springer, 2016. [1](#), [2](#)
- [18] J. Margeta, A. Criminisi, R. Cabrera Lozoya, D. C. Lee, and N. Ayache. Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–11, 2015. [2](#)
- [19] T. Schlegl, J. Ofner, and G. Langs. Unsupervised pre-training across image domains improves lung tissue classification. In *Medical Computer Vision: Algorithms for Big Data*, pages 82–93. Springer, 2014. [2](#)
- [20] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11. [2](#), [4](#)
- [21] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2015. [2](#)
- [22] F. Stark, C. Hazirbas, R. Triebel, and D. Cremers. Captcha recognition with active deep learning. In *Workshop New Challenges in Neural Computation 2015*, page 94. Citeseer, 2015. [2](#)
- [23] N. Tajbakhsh, M. B. Gotway, and J. Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 62–69. Springer, 2015. [7](#)
- [24] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. [2](#), [7](#)
- [25] D. Wang and Y. Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, July 2014. [2](#)
- [26] H. Wang, Z. Zhou, Y. Li, Z. Chen, P. Lu, W. Wang, W. Liu, and L. Yu. Comparison of machine learning methods for

classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 f-fdg pet/ct images. *EJNMMI research*, 7(1):11, 2017. 2

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 1

[28] K. Zhou, H. Greenspan, and D. Shen. *Deep Learning for Medical Image Analysis*. Academic Press, 2016. 1, 2

Supplementary material

The AlexNet architecture and learning parameters used in our experiments

As discussed in Sec. 5, the purpose of this work is not to achieve the highest performance for different biomedical image tasks but to answer the critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging*. For this purpose, we base our experiments on AlexNet, whose architecture is shown in Table 2, as it is deep enough that we can investigate the impact of AIFT on the performance of pre-trained CNNs, and also small enough that we can conduct experiments quickly. Learning parameters used for the training and fine-tuning of AlexNet in our experiments are summarized in Table 3.

Table 2: The AlexNet architecture used in our experiments. Of note, C is 2 as all our three applications are binary classifications by nature.

layer	type	input	kernel	stride	pad	output
data	input	3x227x227	N/A	N/A	N/A	3x227x227
conv1	convolution	3x227x227	11x11	4	0	96x55x55
pool1	max pooling	96x55x55	3x3	2	0	96x27x27
conv2	convolution	96x27x27	5x5	1	2	256x27x27
pool2	max pooling	256x27x27	3x3	2	0	256x13x13
conv3	convolution	256x13x13	3x3	1	1	384x13x13
conv4	convolution	384x13x13	3x3	1	1	384x13x13
conv5	convolution	384x13x13	3x3	1	1	256x13x13
pool5	max pooling	256x13x13	3x3	2	0	256x6x6
fc6	fully connected	256x6x6	6x6	1	0	4096x1
fc7	fully connected	4096x1	1x1	1	0	4096x1
fc8	fully connected	4096x1	1x1	1	0	Cx1

Table 3: Learning parameters used for the training and fine-tuning of AlexNet in our experiments. μ is the momentum, α_{fc8} is the learning rate of the weights in the last layer, α is the learning rate of the weights in the rest layers, and γ determines how α decreases over epochs. The learning rate for the bias term is always set twice as large as the learning rate of the corresponding weights. “Epochs” indicates the number of epochs used in each AIFT iteration. AIFT₁ indicates the first iteration of AIFT while AIFT₊ indicates all the following iterations of AIFT.

Application	Method	μ	α	α_{fc8}	γ	epochs
Colonoscopy Frame Classification	AIFT ₁	0.9	0.0001	0.001	0.95	20
	AIFT ₊	0.9	0.0001	0.0001	0.95	15
	Learning from Scratch	0.9	0.0001	0.001	0.95	20
Polyp Detection	AIFT ₁	0.9	0.001	0.01	0.95	5
	AIFT ₊	0.9	0.0001	0.001	0.10	3
	Learning from Scratch	0.9	0.001	0.01	0.95	10
Pulmonary Embolism Detection	AIFT ₁	0.9	0.001	0.01	0.95	10
	AIFT ₊	0.9	0.001	0.01	0.10	5
	Learning from Scratch	0.9	0.001	0.01	0.95	20

¹ Polyp Detection AIFT Diversity₊: 0.9 | 0.001 | 0.01 | 0.50 | 3

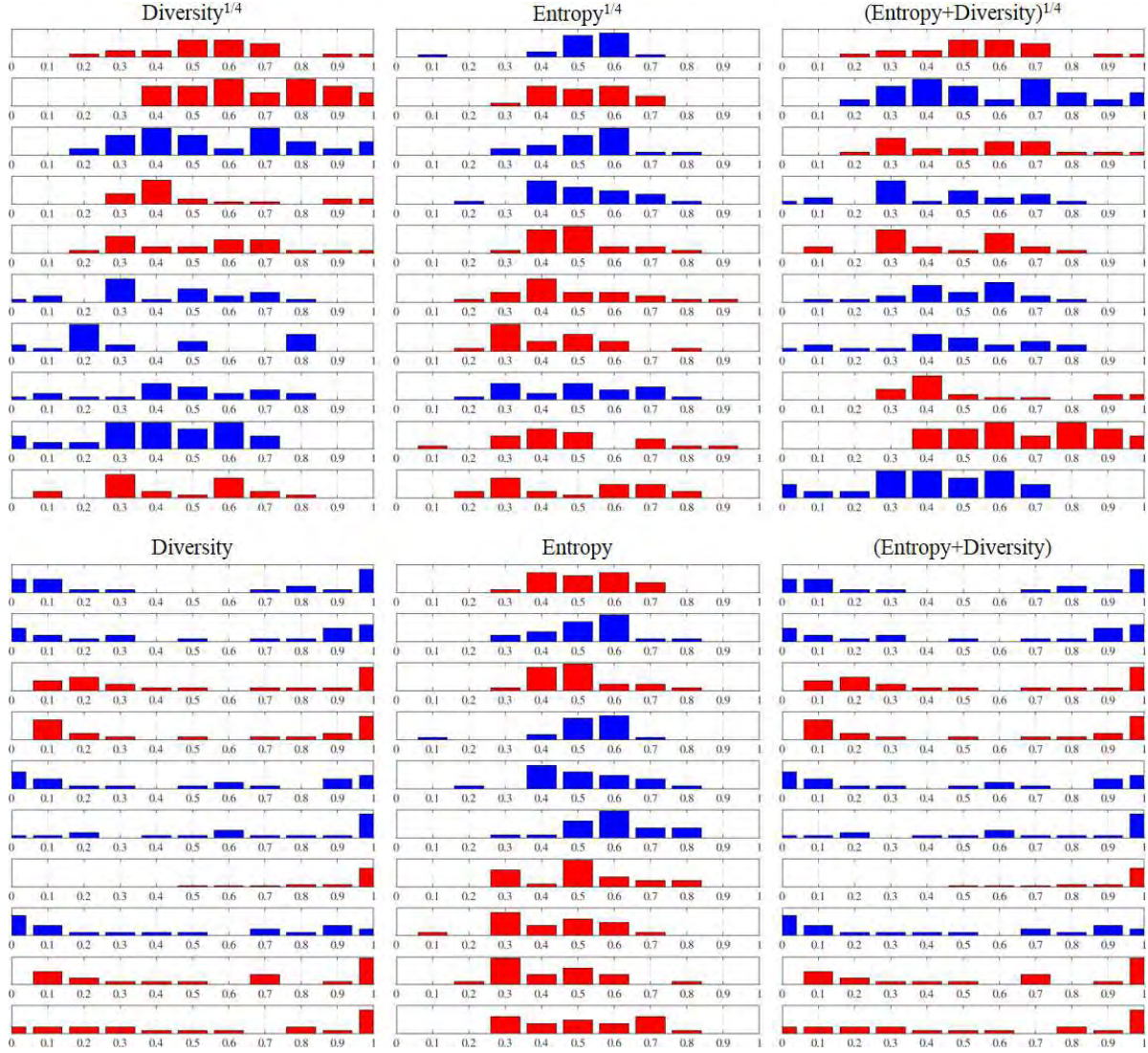



Figure 10: Top 10 candidates selected by the six AIFT methods at Iteration 3 in colonoscopy frame classification. Positive candidates are in red and negative candidates are in blue. Both AIFT Entropy and AIFT Entropy^{1/4} favor Pattern A because of its higher degrees of uncertainty. AIFT Diversity^{1/4} prefers Pattern B while AIFT Diversity suggests Pattern C. With $\lambda_1 = \lambda_2 = 1$ (Eq. 3), diversity is dominant in AIFT (Entropy+Diversity) and (Entropy+Diversity)^{1/4}.



Integrating Active Learning and Transfer Learning for Carotid Intima-Media Thickness Video Interpretation

Zongwei Zhou¹ · Jae Shin¹ · Ruibin Feng¹ · R. Todd Hurst² · Christopher B. Kendall² · Jianming Liang¹ 

© Society for Imaging Informatics in Medicine 2018

Abstract

Cardiovascular disease (CVD) is the number one killer in the USA, yet it is largely preventable (World Health Organization 2011). To prevent CVD, carotid intima-media thickness (CIMT) imaging, a noninvasive ultrasonography method, has proven to be clinically valuable in identifying at-risk persons before adverse events. Researchers are developing systems to automate CIMT video interpretation based on deep learning, but such efforts are impeded by the lack of large annotated CIMT video datasets. CIMT video annotation is not only tedious, laborious, and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible. To dramatically reduce the cost of CIMT video annotation, this paper makes three main contributions. Our first contribution is a new concept, called Annotation Unit (AU), which simplifies the entire CIMT video annotation process down to six simple mouse clicks. Our second contribution is a new algorithm, called AFT (active fine-tuning), which naturally integrates active learning and transfer learning (fine-tuning) into a single framework. AFT starts directly with a pre-trained convolutional neural network (CNN), focuses on selecting the most informative and representative AUs from the unannotated pool for annotation, and then fine-tunes the CNN by incorporating newly annotated AUs in each iteration to enhance the CNN's performance gradually. Our third contribution is a systematic evaluation, which shows that, in comparison with the state-of-the-art method (Tajbakhsh et al., IEEE Trans Med Imaging 35(5):1299–1312, 2016), our method can cut the annotation cost by >81% relative to their training from scratch and >50% relative to their random selection. This performance is attributed to the several advantages derived from the advanced active, continuous learning capability of our AFT method.

Keywords Active learning · Transfer learning · Cardiovascular disease

Introduction

Cardiovascular disease (CVD) is the leading cause of death in the USA: every 40 s, one American dies of CVD; nearly one-half of these deaths occur suddenly and one-third of them occur in patients younger than 65 years, but CVD is preventable [1]. To prevent CVD, the key is to identify at-risk persons, so that scientifically proven and efficacious preventive care can be prescribed appropriately. Carotid intima-media thickness (CIMT) imaging, a noninvasive ultrasonography method, has proven to be clinically valuable for predicting individual CVD risk [8, 22, 31]. It quantifies subclinical atherosclerosis, adds predictive value to traditional risk factors (e.g., the Framingham Risk Score), and has several advantages over computed tomography (CT) coronary artery calcium score: safer (no radiation exposure), more sensitive in a young population, and more accessible to the primary care setting. However, the CIMT imaging protocol (see the “[CIMT Imaging Protocol](#)” section) requires to acquire four videos for each subject, and interpretation of each CIMT video involves three

✉ Jianming Liang
jianming.liang@asu.edu

Zongwei Zhou
zongweiz@asu.edu

Jae Shin
sejong@asu.edu

Ruibin Feng
rfeng12@asu.edu

R. Todd Hurst
hurst.r@mayo.edu

Christopher B. Kendall
kendall.christopher@mayo.edu

¹ Arizona State University, 13212 E Shea Blvd, Scottsdale, AZ 85259, USA

² Mayo Clinic, 13400 E Shea Blvd, Scottsdale, AZ 85259, USA

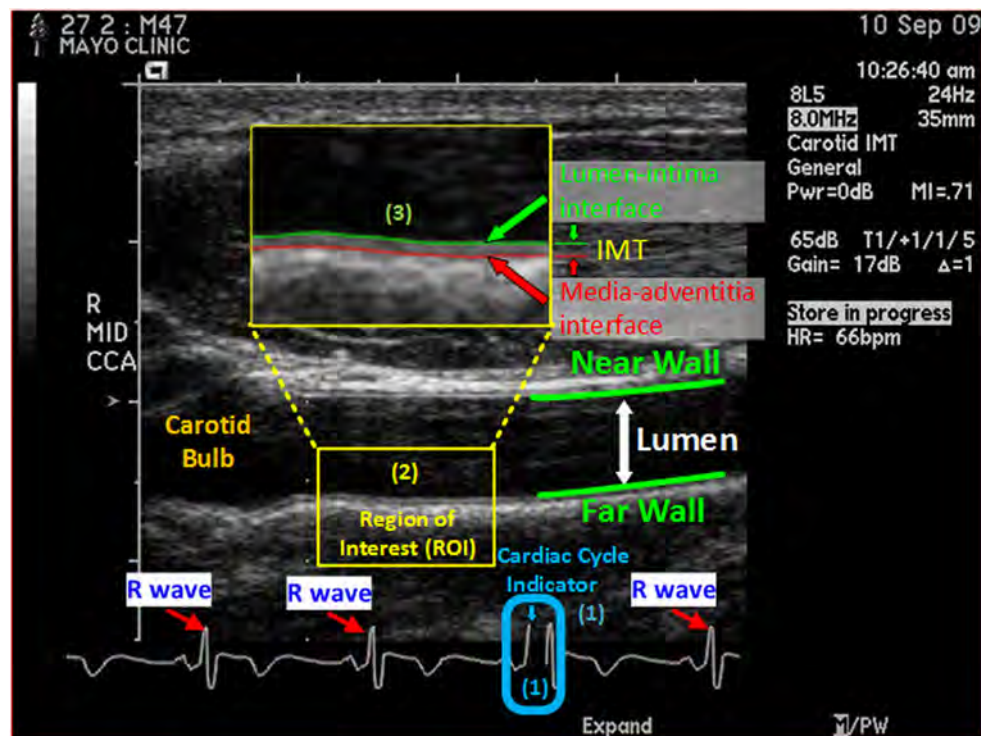


Fig. 1 End-diastolic ultrasound frame (EUF), showing a longitudinal view of a common carotid artery in an ultrasound B-scan image. EUFs are selected based on the cardiac cycle indicator, a black line, which indicates to where in the cardiac cycle the current frame corresponds. CINT is the distance between the lumen-intima interface (in green) and the media-adventitia interface (in red) at an EUF, and it is determined in a region of interest (ROI) approximately 1 cm distal from the carotid bulb at the EUF. In a CINT exam, the sonographer examines the common carotid arteries on both sides of the neck from the two angles, yielding 4 CINT ultrasound videos for each subject. Interpreting *each* CINT video involves three manual steps: (1) select 3 EUFs in

each video based on the cardiac cycle indicator; (2) localize an ROI in each selected EUF according to the carotid bulb; (3) trace the lumen-intima and the media-adventitia interfaces within the localized ROI and compute the minimum, maximum, and average of the distance between the traced lumen-intima and the media-adventitia interfaces. The final CINT report of a subject is a statistical summary of all CINT measurements on the 12 (4×3) EUFs from the 4 CINT videos acquired for the subject. This figure is used with permission [28] under IEEE license number 4407260599014

manual steps (illustrated Fig. 1), which are not only tedious and laborious but also subjective to large interoperator variability if guidelines are not properly followed, hindering the widespread utilization of CINT in clinical practice. Therefore, it is highly desirable to have a system that can automate the CINT video interpretation.

The tedious and laborious manual operations also mean significant work in expert annotation when developing such systems based on machine learning. This paper is not to develop such a system but rather to present a new idea: how to minimize the cost of expert annotation for building such systems that can automate CINT video interpretation based on deep learning. In this research, we make the following three contributions:

Our first contribution is a new concept, called Annotation Unit (AU), which naturally groups the objects to be annotated into sets, and all the objects in each set can be conveniently labeled once with as few operations as possible. This concept significantly simplifies the entire

CINT video annotation process down to six mouse clicks as detailed in the “[Annotation Units](#)” section and illustrated in Fig. 2. Our second contribution is a new algorithm, called AFT (active fine-tuning), which naturally integrates active learning and transfer learning into a single framework (see Algorithm 1) to focus on selecting the most informative and representative AUs for annotation, thereby dramatically reducing the cost of annotation in CINT. AFT starts directly with a pre-trained CNN to seek “worthy” samples from the unannotated pool for annotation, and then fine-tunes the CNN by incorporating newly annotated samples in each iteration to enhance the CNN’s performance gradually. Compared with conventional active learning, AFT offers four advantages: (1) it starts with a completely empty labeled dataset, requiring no initial seed-labeled training samples; (2) it incrementally improves the learner through fine-tuning rather than repeatedly re-training; (3) it can automatically handle multiple classes; and (4) it is applicable to many biomedical image analysis tasks [37],

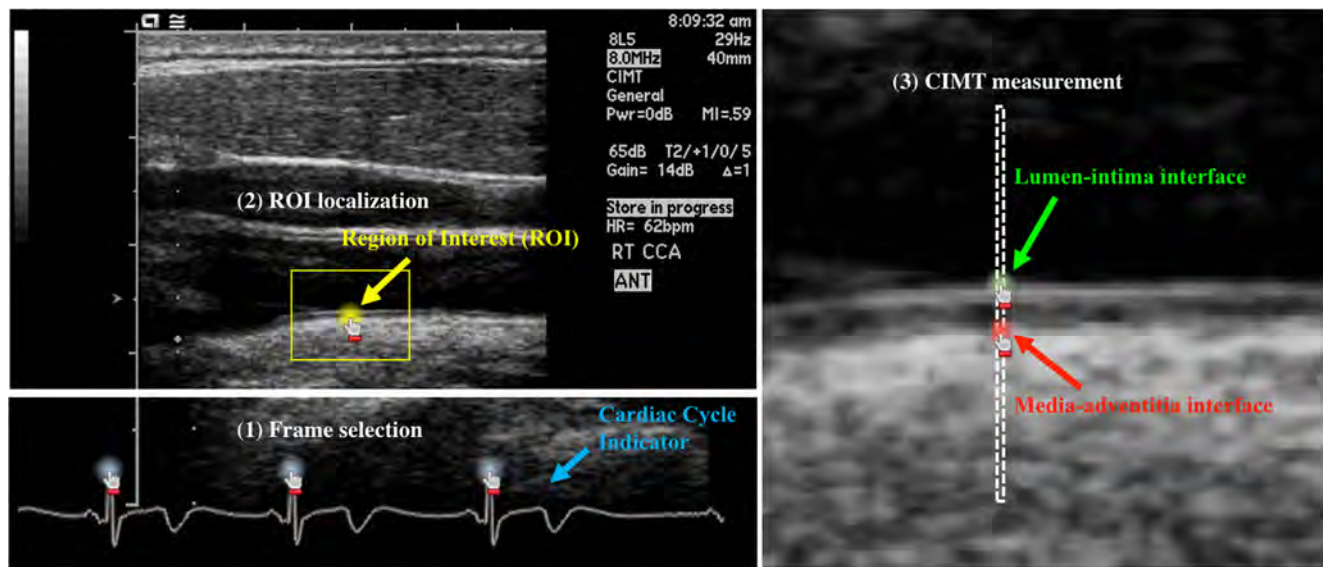


Fig. 2 We simplify the annotation process for each CIMA video down to six mouse clicks. As illustrated, the annotation of EUF selection is made at the video level with three mouse clicks on the three R waves of an ECG signal, while for ROI localization, the annotation is made at the frame level with one mouse click at the center of the ROI. Manually tracing the lumen-intima interface or the media-adventitia interface

is tedious and laborious. To reduce workload, we eliminate the tracing by two mouse clicks on the lumen-intima and media-adventitia interfaces between two vertical dashed lines, only when requested by our proposed AFT algorithm (see the “Annotation Units” section for details)

including detection, classification, and segmentation. Our third contribution is a systematic evaluation of our proposed method, which demonstrates that, with AFT, the cost of annotation for CIMA can be cut by at least half in comparison with FT (fine-tuning with random selection) and by >81% relative to their training from scratch as detailed in the “Experiments” section. This result is significant for enhancing the system performance for automating CIMA video interpretation [28, 34]. Given the current performance of our system [28], it is very difficult to improve its performance by randomly annotating new CIMA videos. We must focus on annotating the most informative and representative videos; otherwise, we will have to annotate many new videos but gain very little in boosting its performance.

CIMA Imaging Protocol

The CIMA exams utilized in our research were performed with B-Mode ultrasound using an 8–14-MHz linear array transducer utilizing fundamental frequency only (Acuson SequoiaTM, Mountain View, CA, USA). The carotid screening protocol begins with scanning bilateral carotid arteries in a transverse manner from the proximal aspect to the proximal internal and external carotid arteries. The probe is then turned to obtain the longitudinal view of the distal common carotid artery (Fig. 1). The sonographer

optimizes the 2D images of the lumen-intima and media-adventitia interfaces at the level of the common carotid artery by adjusting overall gain, time gain, compensation, and focus position. Once the parameters are optimized, the sonographer captures two CIMA videos focused on the common carotid artery from two optimal angles of incidence, and ensures that each CIMA video covers at least three cardiac cycles. The same procedure is repeated for the other side of the neck, resulting in a total of four CIMA videos for each subject.

CIMA stands for carotid intima-media thickness, but in the literature, it may refer to the imaging method, the ultrasonography examination, or the examination results. For clarify, we define some terms used in this paper. By CIMA imaging, we refer to the noninvasive ultrasonography examination procedure described above, yielding four CIMA videos for each subject. The CIMA video interpretation is a process to analyze all four CIMA videos acquired for a subject and produce a CIMA report, which includes a statistical summary of all CIMA measurements performed on the three end-diastolic ultrasound frames (EUFs) selected from each of the four CIMA videos acquired for the subject. EUFs are selected based on the cardiac cycle indicator as shown in Fig. 1, and there are 12 ($= 4 \times 3$) EUFs for each subject. A CIMA measurement on an EUF includes the minimum, maximum, and average of the distance between the lumen-intima and the media-adventitia interfaces (see Fig. 1),

thereby requiring the tracing of lumen-intima and the media-adventitia interfaces. The interpretation of each CIMT video involves three manual steps: (1) select three EUFs in each video based on the cardiac cycle indicator; (2) localize an ROI in each selected EUF according to the carotid bulb; and (3) trace the lumen-intima and the media-adventitia interfaces within the localized ROI and compute the minimum, maximum, and average of the distance between the traced lumen-intima and the media-adventitia interfaces. Certainly, we may adopt the full CIMT interpretation process to annotate CIMT videos as required by machine learning algorithms. However, to dramatically reduce the annotation efforts, we will introduce a separate CIMT video annotation process in conjunction with our proposed AFT algorithm.

Related Work

Carotid Intima-Media Thickness Video Interpretation

As discussed in the “CIMT Imaging Protocol” section, to measure CIMT, the lumen-intima and the media-adventitia interfaces must be traced first. Naturally, the earlier approaches are focused on analyzing the intensity profile and distribution, computing the gradient, or combining various edge properties through dynamic programming [19]. Recent approaches [5, 20] are mostly based on active contours (a.k.a snakes) or their variations [15]. Most recently, researchers are focusing on developing algorithms based on machine learning for CIMT video interpretation. For example, Menchón-Lara et al. employed a committee of standard multi-layer perceptron in [24] and a single standard multi-layer perceptron with an auto-encoder in [25] for CIMT video interpretation. Shin et al. [28] presented a unified framework based on convolutional neural networks (CNNs) for automating the entire CIMT video interpretation process, and Tajbakhsh et al. [34] further demonstrated that the measurement errors are within the interobserver variation. However, none of the aforementioned publications has mentioned the cost of expert annotation in their system development. To our knowledge, we are among the first to minimize the cost of annotation by integrating active learning with the fine-tuning of CNNs for building systems that automate the CIMT video interpretation.

Transfer Learning for Medical Imaging

Gustavo et al. [3] replaced the fully connected layers of a pre-trained CNN with a new logistic layer and

trained only the appended layer with the labeled data while keeping the rest of the network the same, yielding promising results for classification of unregistered multi-view mammogram. In [4], a fine-tuned pre-trained CNN was applied for localizing standard planes in ultrasound images. Gao et al. [7] fine-tuned all layers of a pre-trained CNN for automatic classification of interstitial lung diseases. In [27], Shin et al. used fine-tuned pre-trained CNNs to automatically map medical images to document-level topics, document-level sub-topics, and sentence-level topics. In [23], fine-tuned pre-trained CNNs were used to automatically retrieve missing or noisy cardiac acquisition plane information from magnetic resonance imaging and predict the five most common cardiac views. Schlegl et al. [26] explored unsupervised pre-training of CNNs to inject information from sites or image classes for which no annotations were available, and showed that such across-site pre-training improved classification accuracy compared to random initialization of the model parameters. Several researchers [9, 12, 33] have demonstrated that fine-tuning offers better performance and is more robust than training from scratch, especially in biomedical imaging tasks that labels are not easily accessible. However, none of these works involves active selection processes as our AFT method does, and they all performed *one-time fine-tuning*, that is, simply fine-tuned a pre-trained CNN just once with available training samples.

Integrating Active Learning with Deep Learning

Research in this area is sparse: Wang and Shang [35] may be the first to incorporate active learning with deep learning, and based their approach on stacked restricted Boltzmann machines and stacked auto-encoders. A similar idea was reported for hyperspectral image classification [17]. Stark et al. [30] applied active learning to improve the performance of CNNs for CAPTCHA recognition, while Al Rahhal et al. [2] exploited deep learning for active electrocardiogram classification. All these approaches are fundamentally different from our AFT approach in that in each iteration, they all *repeatedly re-trained the learner from scratch* while we fine-tune pre-trained CNNs, dramatically cutting the cost of annotation further by combining active learning with fine-tuning. Yang et al. [36] adopted active learning into fully convolutional network (FCN) [21] by extracting representative samples into training dataset but training a segmentation network requiring accurate object contour while our ROI localization is only a coarse-labeled location (a single click around the center of the ROI as shown in yellow in Fig. 2). Most recently, Zhou et al. [37] integrated active learning and deep learning based on continuous fine-

tuning but their method is limited to binary classification and requires that all patches within each AU share the same label. Therefore, their method is not applicable to this CIMT application, which requires three-way classifiers.

The Proposed Method

The aim of this research is not to develop methods for automating the interpretation process, rather to investigate how to minimize the cost of expert annotation required for creating such systems that can automate CIMT video interpretation based on CNNs.

Annotation Units

We could follow the same process as illustrated in Fig. 1 [28] to create the ground truth as required to train CNNs. However, these three steps, and in particular the CIMT measurement, are not only tedious and laborious but also subjective to large inter-operator variability if guidelines are not properly followed. To accelerate the annotation process, we introduce a new concept, Annotation Unit (AU), which is defined as a set of objects that the annotator can associate with multiple labels at a time with as few operations as possible during the annotation process. The benefits of AU have two folds. First, the objects to be annotated are grouped into sets, and each set can be easily labeled with as few operations as possible. Taking CIMT measurement annotation as an example, instead of tracing the entire lumen-intima and media-adventitia interfaces within an ROI, we define an one-pixel-wide column in the ROI as an AU, so that all pixels within the column can be labeled once with two mouse clicks: one on the lumen-intima interface and one on the media-adventitia interface. Second, with the aforementioned properties, all the objects in an AU can be correctly associated with their labels once after the required operations. Using the CIMT measurement example again, after the two clicks, the first clicked pixel is associated with class 1 (lumen-intima), the second clicked pixel is with class 2 (media-adventitia), and all the rest

pixels are with class 0 (background). It should be noticed that when all AUs are labeled, the interpretation quality is identical or at least similar to the standard process in [28], but our goal is to annotate as few AUs as possible during the annotation process; therefore, the annotation process may not result in a complete interpretation for a subject. In other words, the annotation process is designed for annotation (as little as possible) only, and it is not intended for clinical use, which requires a complete interpretation for each subject.

With the definition of AU, the CIMT video annotation process can be simplified down to just six mouse clicks as illustrated in Fig. 2. The annotation for EUF selection is made at the video level. With three mouse clicks on the R waves of the ECG signal, three end-diastolic ultrasound frames (EUFs) are determined and annotated as class 1, while all the rest frames are automatically labeled as class 0 (non-EUF). For ROI localization in an EUF, the annotation is made at the frame level with one mouse click on the EUF, giving the center of the ROI. Given the anatomical constraint that ROI should be approximately 1 cm distal from carotid bulb, the latter's location can be automatically estimated. For data argumentation and classification robustness, all pixels within 15 mm from the selected center are considered as class 1 (ROI), and those within 15 mm from the estimated bulb location are as class 2, while all the rest pixels belong to class 0 automatically. For CIMT measurement, it would be too tedious and laborious for the annotator to manually trace the lumen-intima and media-adventitia interfaces. To reduce workload, two vertical dashed lines are drawn to indicate an AU (see Fig. 2) and the annotator makes two mouse clicks on the two interfaces between the two dashed lines. The top pixel and bottom pixel are regarded as the lumen-intima interface (class 1) and media-adventitia interface (class 2), respectively, while all the rest pixels between the two lines are considered as background (class 0). The optimal distance between the two dashed lines can be determined based on experiments, and we set it at one pixel (0.99 mm) currently. We summarize the objects of AU, annotation labels, and required operations per AU in each step of CIMT video annotation process in Table 1.

Table 1 The AU, annotated labels, and required operations per AU in each step of CIMT video annotation process

	EUF selection	ROI localization	CIMT measurement
AU	ECG signal	EUF frame	One-pixel-wide column in ROI
Labels	EUF	ROI	Lumen-intima interface
	Non-EUF	Carotid bulb	Lumen-intima interface
		Background	Background
Operations	3 clicks	1 click	2 clicks

Algorithm 1 Active fine-tuning

Input:
 $\mathcal{U} = \{\mathcal{C}_i\}, i \in [1, n]$ $\{\mathcal{U}$ contains n AUs}
 $\mathcal{C}_i = \{x_i^j\}, j \in [1, m]$ $\{\mathcal{C}_i$ has m objects}
 \mathcal{M} : a pre-trained CNN
 b : batch size

Output:
 \mathcal{L} : the labeled AUs
 \mathcal{M}_t : the fine-tuned CNN model at Iteration t

```

1  $\mathcal{L} \leftarrow \emptyset$ 
2 repeat
3   for each  $\mathcal{C}_i \in \mathcal{U}$  do
4      $p_i \leftarrow P(\mathcal{C}_i, \mathcal{M}_{t-1})$  {outputs of  $\mathcal{M}_{t-1}$  given  $\forall x_i \in \mathcal{C}_i$ }
5      $\mathcal{E}_i \leftarrow E(\mathcal{C}_i)$  {compute entropy  $\mathcal{E}_i$  for  $\mathcal{C}_i$  using Eq. 1}
6   end
7    $\mathcal{U}' \leftarrow S(\mathcal{U}, \mathcal{E})$  {sort  $\mathcal{C}_i \in \mathcal{U}$  according to the value of  $\mathcal{E}_i \in \mathcal{E}$ }
8    $\mathcal{Q} \leftarrow Q(\mathcal{U}', b)$  {associate labels for the top  $b$  AUs in the sorted  $\mathcal{U}'$ }
9    $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{Q}; \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{Q}; t \leftarrow t+1$ 
10   $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M})$  {fine-tune  $\mathcal{M}$  with  $\mathcal{L}$ }
11 until classification performance is satisfactory;

```

Active Fine-Tuning

Mathematically, given a set of AUs, $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$, where n is the number of AUs, and each $\mathcal{C}_i = \{x_i^1, x_i^2, \dots, x_i^m\}$ is associated with m objects, our AFT algorithm iteratively selects a subset of AUs for annotation as illustrated in Algorithm 1. From annotation, each object (in each selected AU) will be associated with one of Y number of possible classes. At the beginning, the labeled dataset \mathcal{L} is empty; we take a pre-trained CNN from ImageNet [6] (e.g., AlexNet) as initialization of the network and run it on \mathcal{U} to select b number of AUs for labeling. The newly labeled AUs will be incorporated into \mathcal{L} to fine-tune the CNN until the performance is satisfactory. From our experiments, we have found that continuously fine-tuning the CNN, which has been fine-tuned in the previous iteration, with enlarged datasets converges faster than repeatedly fine-tuning the original pre-trained CNN, but the latter offers better generalization. We have also found that continuously fine-tuning the CNN with only newly labeled data demands careful meta-parameter adjustments. Therefore, in this paper, our AFT fine-tunes the original pre-trained CNN with the labeled dataset enlarged with the newly labeled data in each iteration to achieve better performance. To determine the “worthiness” of an AU, we use entropy, as intuitively, entropy captures the

classification certainty—higher uncertainty values denote higher degrees of information. Assuming the prediction of object x_i^j in \mathcal{C}_i by the current CNN is p_i^j , we define the entropy of \mathcal{C}_i as the average information furnished by all objects x_i^j in \mathcal{C}_i from the unlabeled pool:

$$\mathcal{E}_i = -\frac{1}{m} \sum_{j=1}^m \sum_{k=1}^Y p_i^{j,k} \log p_i^{j,k}. \quad (1)$$

Experiments

In our experiments, we use a fully interpreted (annotated) database and simulate the active learning process (Algorithm 1) by retrieving labels for the samples selected based on selection criterion as present in Eq. 1. In this way, our approach can be validated without “physically” involving the experts in the loop.

Dataset

Due to space, we focus on the two most important tasks: ROI localization and CIMT measurement. Our AFT algorithm is implemented in Caffe [14] based on the pre-trained AlexNet [16]. In the following, we shall compare our method AFT (active fine-tuning) with the state-of-the-art method [33]: FT (fine-tuning with random selection) and LS (learning from scratch) in each task. We utilize 23 patients from UFL MCAEL CIMT research database [13]. Each patient has four videos (two on each side) [31], resulting in a total of 92 CIMT videos with 8,021 frames. Each video covers at least three cardiac cycles and thus a minimum of three EUFs. We randomly divide the CIMT videos at patient level into training, validation, and test datasets (no overlaps). The training dataset contains 44 CIMT videos of 11 patients with a total of 4,070 frames, the validation dataset contains 4 videos of 1 patient with 386 frames, and the test dataset contains 44 CIMT videos of 11 patients with 3,565 frames. From the perspective of active learning, the training dataset is the “unlabeled pool” for active selection; when an AU is selected, the label of each object will be provided. The fine-tuned CNN from each iteration is always evaluated with the test dataset, so that we can monitor the performance enhancement across AUs. Please note that we do not need many patients as we have many CIMT frames for each patient and we can generate a large number of patches for training deep models in each experiment. For example, in our ROI localization experiments, one AU practically provides 1715 labeled patches (297 as *background*, 709 as *bulb*, and 709 as *ROI*). Random translation and flipping data augmentation were applied when training the models.

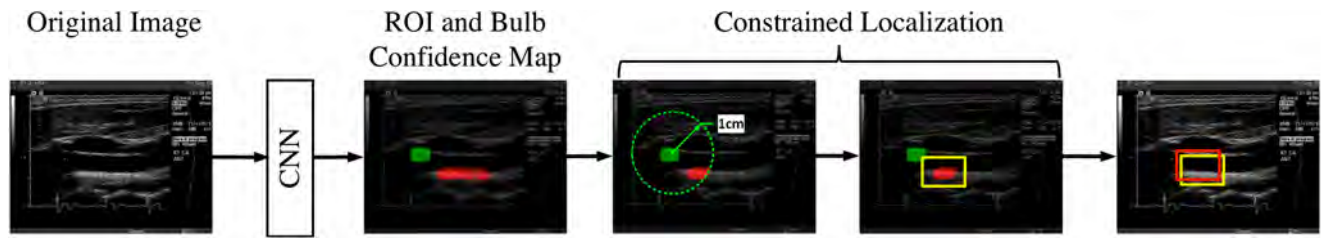


Fig. 3 ROI localization process (see text for details). The detected ROI, ground truth, and carotid bulb are in yellow, red, and green, respectively. The ROI is constrained by green circle with a 1-cm radius

ROI Localization

Accurate localization of the ROI is challenging because, as illustrated by Shin et al. [28] in their figure 1, no notable differences can be observed in image appearance among the ROIs on the far wall of the carotid artery. To overcome this challenge, we use the location of the carotid bulb as a contextual constraint. We choose this constraint for two reasons: (1) the carotid bulb appears as a distinct dark area in the ultrasonographic frame and thus can be uniquely identified; and (2) according to the consensus statement of the American Society of Echocardiography for Cardiovascular Risk Assessment [31], the ROI should be placed approximately 1 cm from the carotid bulb on the far wall of the common carotid artery. The former motivates the use of the carotid bulb location as a constraint from a technical point of view, and the latter justifies this constraint from a clinical standpoint. We incorporate this constraint by simultaneously localizing both ROI and carotid bulb and then refine the estimated location of the ROI given the location of the carotid bulb. As illustrated in Fig. 3, we first determine the location of the carotid bulb as the centroid of the largest connected component within the confidence map for the carotid bulb and then localize the centroid of constrained ROI area using the following formula:

$$l_{\text{roi}} = \frac{\sum_{p \in C^*} M(p) \cdot p \cdot I(p)}{\sum_{p \in C^*} M(p) \cdot I(p)} \quad (2)$$

where M denotes the confidence map of being the ROI, C^* is the largest connected component in M that is nearest to the carotid bulb, and $I(p)$ is an indicator function for pixel $p = [p_x, p_y]$ that is defined as

$$I(p) = \begin{cases} 1, & \text{if } \|p - l_{\text{cb}}\| < 1 \text{ cm} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where l_{cb} is the centroid of the carotid bulb. Basically, the indicator function excludes the pixels located farther than 1 cm from the carotid bulb location. This choice of the distance threshold is motivated by the fact that the ROI is located within 1 cm to the right of the carotid bulb.

CIMT Measurement

To automatically measure intima-media thickness, the lumen-intima and media-adventitia interfaces of the carotid artery must be detected within the ROI. Although the lumen-intima interface is relatively easy to detect, the detection of the media-adventitia interface is challenging, because of the faint image gradients around its boundary. We formulate this interface segmentation problem as a three-class classification task with the goal to classify each pixel within the ROI into three categories: (1) a pixel on the lumen-intima interface, (2) a pixel on the media-adventitia interface, and (3) a background pixel. During testing, the trained CNN is applied to a given test ROI in a convolutional manner, generating two confidence maps with the same size as the ROI. The first confidence map shows the probability of each pixel being on the lumen-intima interface; the second confidence map shows the probability of each pixel being on the media-adventitia interface. A relatively thick high-probability band is apparent along each interface, which hinders the accurate measurement of intima-media thickness. To thin the detected interfaces, we scan the confidence map column by column, searching for the rows with the maximum response for each of the two interfaces. By doing so, we obtain a 1-pixel-thick boundary with a step-like shape around each interface. To further refine the boundaries, we use two active contour models (a.k.a., snakes) [18], one for the lumen-intima interface and one for the media-adventitia interface. The open snakes are initialized with the current step-like boundaries and then deform solely based on the probability maps generated by the CNN rather than the original image content.

Results and Discussions

To evaluate AFT performance on ROI localization, in each iteration, we compute two criteria across all test patients: (1) the average ROI localization error (the Euclidean distance between the detected ROI and expert-annotated ROI) and (2) the predicted confidence of each expert-annotated ROI. Figure 4a shows the average ROI localization error over

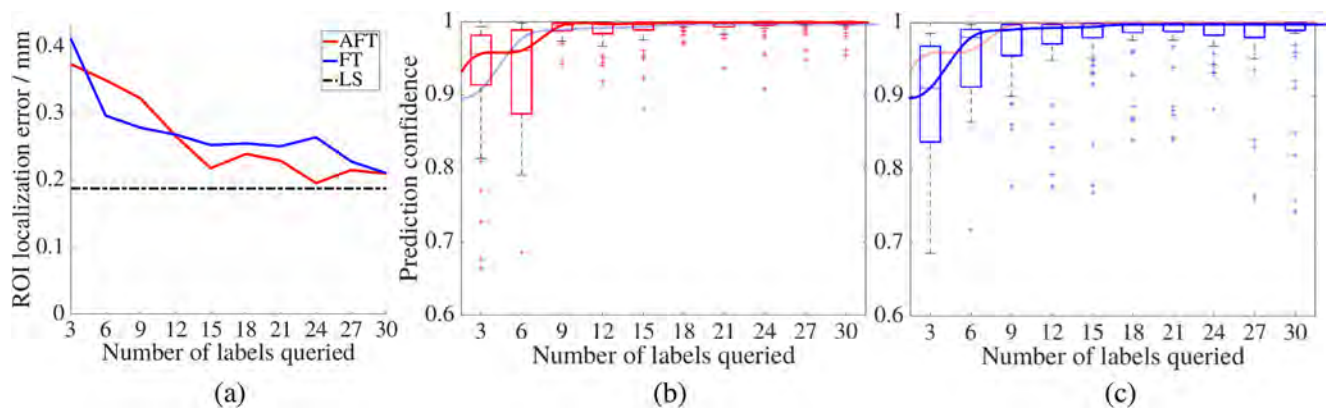


Fig. 4 **a** The average ROI localization errors of AFT, FT, and LS on the test patients over 30 AUs. The ROI confidence predicted by AFT **b** and FT **c**, respectively, on the test patients over 30 AUs. The trendlines

denote active selection (in red) and random selection (in blue), and they are duplicated with different transparencies for their easy performance comparison in **b** and **c**

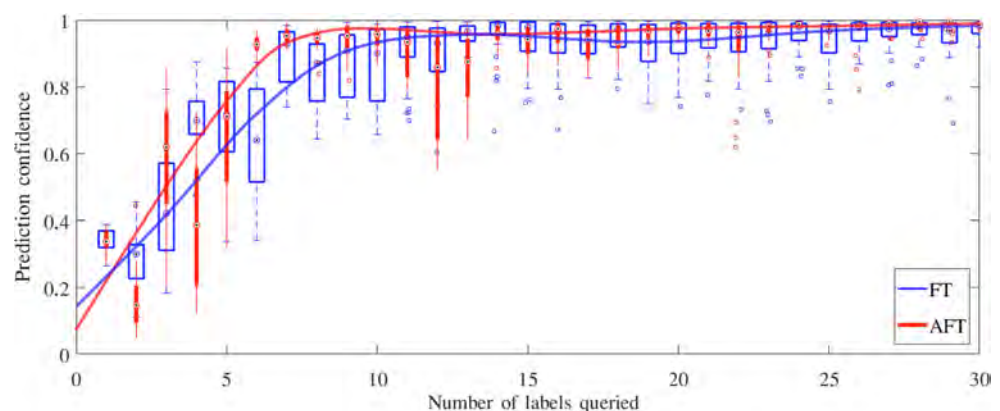
30 AUs (automatically generating 51,450 labeled patches) demonstrating that our AFT dramatically reduces the labeling cost in comparison with FT and LS. Black dashed line represents the ROI localization error of LS, where the CNN was trained with the entire training dataset (132 AUs) without fine-tuning. We should note that at the earlier stage (less than 12 AUs), FT learns faster and yields better performance than AFT, a well-known phenomenon in active learning [10]. However, AFT quickly surpasses FT after a few times of fine-tuning. With only 24 AUs, AFT can nearly achieve the performance of LS with 132 AUs; with 15 AUs, AFT achieves that of FT with 30 annotations. Thereby, the cost of annotation can be cut by at least half in comparison with FT and by more than 81% in comparison with LS. To increase the robustness, the predicted confidence of each expert-annotated ROI is computed as the average of the predicted scores of all pixels within 15 pixels from the ROI center. Figure 4b, c is the box plots of the ROI confidence across all the test patients. Clearly, the more AUs used from the training dataset, the higher the ROI confidence with the test dataset. In terms of mean and standard deviation of ROI confidence, with just 9 AUs, AFT offers the same confidence

as FT at 30 AUs. Moreover, ROI confidence from AFT can quickly converge to 1.0, while even using 30 AUs, FT still has many outliers.

We evaluate our AFT on CIMT measurement in the same way as in ROI localization. However, due to the post-processing with snakes, AFT and FT give the similar localization errors at the lumen-intima and media-adventitia interfaces; therefore, we focus on the CIMT measurement confidence. Figure 5 is the box plot of the CIMT measurement confidence on the test dataset. AFT significantly outperforms FT, especially when a limited number of training samples are used. For example, actively selecting only 7 AUs can approximate the performance by randomly selecting 14 AUs. In addition, with only 7 AUs selected by our AFT algorithm, we can nearly achieve the accuracy offered by the entire dataset (12,144 AUs).

In our experiments, we adopted the AlexNet architecture because a pre-trained AlexNet model is available in the Caffe library and its architecture strikes a nice balance in depth: it is deep enough that we can investigate the impact of AFT on the performance of pre-trained CNNs, and it is also shallow enough that we can conduct experiments quickly.

Fig. 5 The CIMT measurement confidence predicted by AFT and FT, respectively, on the test patients over 30 AUs



Alternatively, deeper architectures, such as VGG [29], GoogleNet [32], and ResNet [11], could have been used and have shown relatively high performance for challenging computer vision tasks. However, the purpose of this work is not to achieve the highest performance for the CIMT video interpretation but to answer a critical question: *How much the cost of annotation can be reduced when applying CNNs for CIMT video interpretation.* For this purpose, AlexNet is a reasonable architectural choice. Nevertheless, we plan to investigate the performance of AFT on different deep architectures. Also, our algorithm aims to select the most informative and representative AUs for annotation with our proposed six-click strategy. As a result, the process will not generate full interpretations for all patients, that is, the six-click strategy is only applicable in the context of our proposed algorithm for reducing annotation efforts (as little as possible), and it is not designed and should not be used for clinical practice, where a complete interpretation is required for each patient.

Conclusions

We have developed an active fine-tuning method for CIMT video interpretation. It integrates active learning and transfer learning, offering two advantages: It starts with a completely empty labeled dataset, and incrementally improves the CNN's performance via fine-tuning by actively selecting the most informative and representative samples. To accelerate the CIMT video annotation process, we introduced a new concept, Annotation Unit, which simplifies the CIMT video annotation process down to six mouse clicks. We have demonstrated that the cost of CIMT video annotation can be cut by at least half. This performance is attributed to the advanced active fine-tuning capability of our AFT method. In the future, we plan to explore possible algorithms in assisting sonographers to acquire high-quality CIMT videos more quickly and integrate our AFT algorithm into the process for collecting the most informative and representative CIMT videos to enhance our system performance.

Acknowledgements This research has been supported partially by NIH under Award Number R01HL128785 and partially by ASU and Mayo Clinic through the Discovery Translation Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. World Health Organization. Global atlas on cardiovascular disease prevention and control. Available at www.who.int/cardiovascular_diseases/publications (September 19, 2011)
2. Al Rahhal Ms, Bazi Y, AlHichri H, Alajlan N, Melgani F, Yager R: Deep learning approach for active classification of electrocardiogram signals. *Inf Sci* 345:340–354, 2016
3. Carneiro G, Nascimento J, Bradley A: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: Navab N, Hornegger J, Wells WM, Frangi AF Eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science*, vol. 9351, pp. 652–660. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24574-4_78
4. Chen H, Ni D, Qin J, Li S, Yang X, Wang T, Heng PA: Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE J Biomed Health Inform* 19(5):1627–1636, 2015
5. Delsanto S, Molinari F, Giustetto P, Liboni W, Badalamenti S, Suri JS: Characterization of a completely user-independent algorithm for carotid artery segmentation in 2-d ultrasound images. *IEEE Trans Instrum Meas* 56(4):1265–1274, 2007
6. Deng J, Dong W, Socher R, Li LJ, Li K: Fei-fei, L.: ImageNet: A Large-Scale Hierarchical Image Database CVPR09, 2009
7. Gao M, Bagci U, Lu L, Wu A, Buty M, Shin HC, Roth H, Papadakis GZ, Depeursinge A, Summers RM, et al: Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. In: *The 1st workshop on deep learning in medical image analysis, international conference on medical image computing and computer assisted intervention, at MICCAI-DLMIA'15*, 2015
8. Gepner AD, Young R, Delaney JA, Tattersall MC, Blaha MJ, Post WS, Gottesman RF, Kronmal R, Budoff MJ, Burke GL, et al: A comparison of coronary artery calcium presence, carotid plaque presence, and carotid intima-media thickness for cardiovascular disease prediction in the multi-ethnic study of atherosclerosis (mesa). *Circulation Cardiovascular Imaging* 8(1):e002262, 2015
9. Greenspan H, van Ginneken B, Summers RM: Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* 35(5):1153–1159, 2016
10. Guyon I, Cawley G, Dror G, Lemaire V, Statnikov A (2011) *JMLR Workshop and conference proceedings (volume 16): Active learning challenge microtome publishing*
11. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, 2016
12. Hoo-Chang S, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 35(5):1285, 2016
13. Hurst RT, Burke RF, Wissner E, Roberts A, Kendall CB, Lester SJ, Somers V, Goldman ME, Wu Q, Khandheria B: Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players. *Am J Cardiol* 105(8):1107–1111, 2010
14. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*
15. Kass M, Witkin A, Terzopoulos D: Snakes: Active contour models. *Int J Comput Vis* 1(4):321–331, 1988
16. Krizhevsky A, Sutskever I, Hinton GE: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp 1097–1105, 2012
17. Li J: Active learning for hyperspectral image classification with a stacked autoencoders based neural network. In: *2016 IEEE International conference on image processing (ICIP)*, pp 1062–1065, 2016. <https://doi.org/10.1109/ICIP.2016.7532520>

18. Liang J, McInerney T, Terzopoulos D: United snakes. *Med Image Anal* 10(2):215–233, 2006
19. Liang Q, Wendelhag I, Wikstrand J, Gustavsson T: A multiscale dynamic programming procedure for boundary detection in ultrasonic artery images. *IEEE Trans Med Imaging* 19(2):127–142, 2000
20. Loizou CP, Pattichis CS, Pantziaris M, Nicolaides A: An integrated system for the segmentation of atherosclerotic carotid plaque. *IEEE Trans Inf Technol Biomed* 11(6):661–667, 2007
21. Long J, Shelhamer E, Darrell T: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 3431–3440, 2015
22. Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M: Prediction of clinical cardiovascular events with carotid intima-media thickness. *Circulation* 115(4):459–467, 2007
23. Margeta J, Criminisi A, Cabrera Lozoya R, Lee DC, Ayache N: Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition. *Comput Methods Biomech Biomed Eng: Imaging Visual* 5(5):339–349, 2017
24. Menchón-Lara RM, Bastida-Jumilla MC, González-López A, Sancho-Gómez JL: Automatic evaluation of carotid intima-media thickness in ultrasounds using machine learning. In: *Natural and artificial computation in engineering and medical applications*, pp 241–249. Springer, 2013
25. Menchón-Lara RM, Sancho-Gómez JL: Fully automatic segmentation of ultrasound common carotid artery images based on machine learning. *Neurocomputing* 151:161–167, 2015
26. Schlegl T, Ofner J, Langs G: Unsupervised pre-training across image domains improves lung tissue classification. In: *Medical computer vision: Algorithms for big data*, pp 82–93. Springer, 2014
27. Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM: Interleaved text/image deep mining on a very large-scale radiology database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 1090–1099, 2015
28. Shin J, Tajbakhsh N, Todd Hurst R, Kendall CB, Liang J: Automating carotid intima-media thickness video interpretation with convolutional neural networks. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, 2016
29. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. In: *ICLR*, 2015
30. Stark F, Hazırbas C, Triebel R., Cremers D.: Captcha recognition with active deep learning. In: *Workshop new challenges in neural computation 2015*, p 94. Citeseer, 2015
31. Stein JH, Korcarz CE, Hurst RT, Lonn E, Kendall CB, Mohler ER, Najjar SS, Rembold CM, Post WS: Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: a consensus statement from the american society of echocardiography carotid intima-media thickness task force endorsed by the society for vascular medicine. *J Am Soc Echocardiogr* 21(2):93–111, 2008
32. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: Going deeper with convolutions. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, 2015
33. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans Med Imaging* 35(5):1299–1312, 2016
34. Tajbakhsh N, Shin JY, Hurst RT, Kendall CB, Liang J: Automatic interpretation of carotid intima-media thickness videos using convolutional neural networks. In: *Deep learning for medical image analysis*, pp 105–131. Elsevier, 2017
35. Wang D, Shang Y: A new active labeling method for deep learning. In: *2014 International joint conference on neural networks (IJCNN)*, pp 112–119, 2014. <https://doi.org/10.1109/IJCNN.2014.6889457>
36. Yang L, Zhang Y, Chen J, Zhang S, Chen DZ: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, pp 399–407. Springer, 2017
37. Zhou Z, Shin J, Zhang L, Gurudu S, Gotway M, Liang J: Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: *IEEE Conference on computer vision and pattern recognition, hawaii*, pp 7340–7349, 2017

UNet++: A Nested U-Net Architecture for Medical Image Segmentation

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee,
Nima Tajbakhsh, and Jianming Liang

Arizona State University
{zongweiz,mrahmans,ntajbakh,jianming.liang}@asu.edu

Abstract. In this paper, we present UNet++, a new, more powerful architecture for medical image segmentation. Our architecture is essentially a deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks. We argue that the optimizer would deal with an easier learning task when the feature maps from the decoder and encoder networks are semantically similar. We have evaluated UNet++ in comparison with U-Net and wide U-Net architectures across multiple medical image segmentation tasks: nodule segmentation in the low-dose CT scans of chest, nuclei segmentation in the microscopy images, liver segmentation in abdominal CT scans, and polyp segmentation in colonoscopy videos. Our experiments demonstrate that UNet++ with deep supervision achieves an average IoU gain of 3.9 and 3.4 points over U-Net and wide U-Net, respectively.

1 Introduction

The state-of-the-art models for image segmentation are variants of the encoder-decoder architecture like U-Net [9] and fully convolutional network (FCN) [8]. These encoder-decoder networks used for segmentation share a key similarity: skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network. The skip connections have proved effective in recovering fine-grained details of the target objects; generating segmentation masks with fine details even on complex background. Skip connections is also fundamental to the success of instance-level segmentation models such as Mask-RCNN, which enables the segmentation of occluded objects. Arguably, image segmentation in natural images has reached a satisfactory level of performance, but do these models meet the strict segmentation requirements of medical images?

Segmenting lesions or abnormalities in medical images demands a higher level of accuracy than what is desired in natural images. While a precise segmentation mask may not be critical in natural images, even marginal segmentation errors in medical images can lead to poor user experience in clinical settings. For instance,

the subtle spiculation patterns around a nodule may indicate nodule malignancy; and therefore, their exclusion from the segmentation masks would lower the credibility of the model from the clinical perspective. Furthermore, inaccurate segmentation may also lead to a major change in the subsequent computer-generated diagnosis. For example, an erroneous measurement of nodule growth in longitudinal studies can result in the assignment of an incorrect Lung-RADS category to a screening patient. It is therefore desired to devise more effective image segmentation architectures that can effectively recover the fine details of the target objects in medical images.

To address the need for more accurate segmentation in medical images, we present UNet++, a new segmentation architecture based on nested and dense skip connections. The underlying hypothesis behind our architecture is that the model can more effectively capture fine-grained details of the foreground objects when high-resolution feature maps from the encoder network are gradually enriched prior to fusion with the corresponding semantically rich feature maps from the decoder network. We argue that the network would deal with an easier learning task when the feature maps from the decoder and encoder networks are semantically similar. This is in contrast to the plain skip connections commonly used in U-Net, which directly fast-forward high-resolution feature maps from the encoder to the decoder network, resulting in the fusion of semantically dissimilar feature maps. According to our experiments, the suggested architecture is effective, yielding significant performance gain over U-Net and wide U-Net.

2 Related Work

Long *et al.* [8] first introduced fully convolutional networks (FCN), while U-Net was introduced by Ronneberger *et al.* [9]. They both share a key idea: skip connections. In FCN, up-sampled feature maps are summed with feature maps skipped from the encoder, while U-Net concatenates them and add convolutions and non-linearities between each up-sampling step. The skip connections have shown to help recover the full spatial resolution at the network output, making fully convolutional methods suitable for semantic segmentation. Inspired by DenseNet architecture [5], Li *et al.* [7] proposed H-denseunet for liver and liver tumor segmentation. In the same spirit, Drozdza *et al.* [2] systematically investigated the importance of skip connections, and introduced short skip connections within the encoder. Despite the minor differences between the above architectures, they all tend to fuse semantically dissimilar feature maps from the encoder and decoder sub-networks, which, according to our experiments, can degrade segmentation performance.

The other two recent related works are GridNet [3] and Mask-RCNN [4]. GridNet is an encoder-decoder architecture wherein the feature maps are wired in a grid fashion, generalizing several classical segmentation architectures. GridNet, however, lacks up-sampling layers between skip connections; and thus, it does not represent UNet++. Mask-RCNN is perhaps the most important meta framework for object detection, classification and segmentation. We would like to note that

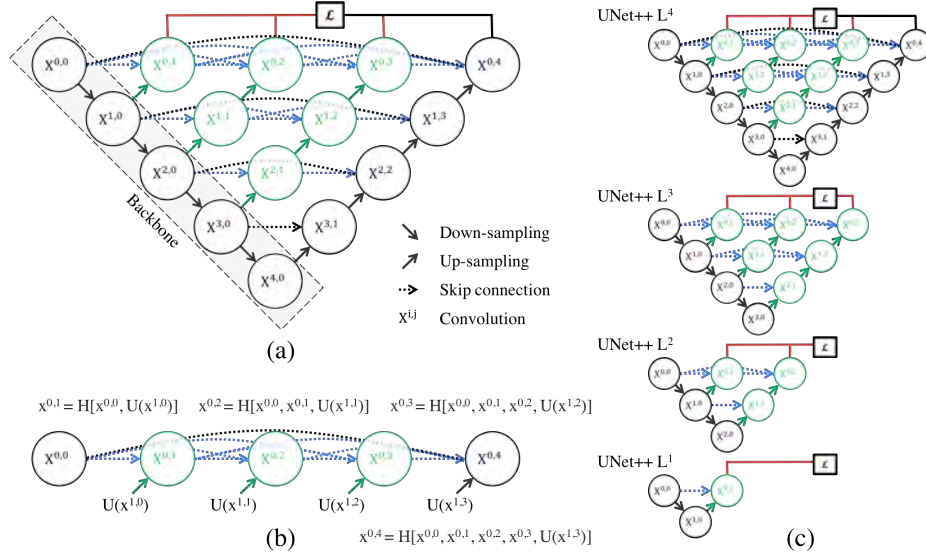


Fig. 1: (a) UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks. The main idea behind UNet++ is to bridge the semantic gap between the feature maps of the encoder and decoder prior to fusion. For example, the semantic gap between $(X^{0,0}, X^{1,3})$ is bridged using a dense convolution block with three convolution layers. In the graphical abstract, black indicates the original U-Net, green and blue show dense convolution blocks on the skip pathways, and red indicates deep supervision. Red, green, and blue components distinguish UNet++ from U-Net. (b) Detailed analysis of the first skip pathway of UNet++. (c) UNet++ can be pruned at inference time, if trained with deep supervision.

UNet++ can be readily deployed as the backbone architecture in Mask-RCNN by simply replacing the plain skip connections with the suggested nested dense skip pathways. Due to limited space, we were not able to include results of Mask RCNN with UNet++ as the backbone architecture; however, the interested readers can refer to the supplementary material for further details.

3 Proposed Network Architecture: UNet++

Fig. 1a shows a high-level overview of the suggested architecture. As seen, UNet++ starts with an encoder sub-network or backbone followed by a decoder sub-network. What distinguishes UNet++ from U-Net (the black components in Fig. 1a) is the re-designed skip pathways (shown in green and blue) that connect the two sub-networks and the use of deep supervision (shown red).

3.1 Re-designed skip pathways

Re-designed skip pathways transform the connectivity of the encoder and decoder sub-networks. In U-Net, the feature maps of the encoder are directly received in the decoder; however, in UNet++, they undergo a dense convolution block whose number of convolution layers depends on the pyramid level. For example, the skip pathway between nodes $X^{0,0}$ and $X^{1,3}$ consists of a dense convolution block with three convolution layers where each convolution layer is preceded by a concatenation layer that fuses the output from the previous convolution layer of the same dense block with the corresponding up-sampled output of the lower dense block. Essentially, the dense convolution block brings the semantic level of the encoder feature maps closer to that of the feature maps awaiting in the decoder. The hypothesis is that the optimizer would face an easier optimization problem when the received encoder feature maps and the corresponding decoder feature maps are semantically similar.

Formally, we formulate the skip pathway as follows: let $x^{i,j}$ denote the output of node $X^{i,j}$ where i indexes the down-sampling layer along the encoder and j indexes the convolution layer of the dense block along the skip pathway. The stack of feature maps represented by $x^{i,j}$ is computed as

$$x^{i,j} = \begin{cases} \mathcal{H}(x^{i-1,j}), & j = 0 \\ \mathcal{H}\left(\left[x^{i,k}\right]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1})\right), & j > 0 \end{cases} \quad (1)$$

where function $\mathcal{H}(\cdot)$ is a convolution operation followed by an activation function, $\mathcal{U}(\cdot)$ denotes an up-sampling layer, and $[\]$ denotes the concatenation layer. Basically, nodes at level $j = 0$ receive only one input from the previous layer of the encoder; nodes at level $j = 1$ receive two inputs, both from the encoder sub-network but at two consecutive levels; and nodes at level $j > 1$ receive $j + 1$ inputs, of which j inputs are the outputs of the previous j nodes in the same skip pathway and the last input is the up-sampled output from the lower skip pathway. The reason that all prior feature maps accumulate and arrive at the current node is because we make use of a dense convolution block along each skip pathway. Fig. 1b further clarifies Eq. 1 by showing how the feature maps travel through the top skip pathway of UNet++.

3.2 Deep supervision

We propose to use deep supervision [6] in UNet++, enabling the model to operate in two modes: 1) accurate mode wherein the outputs from all segmentation branches are averaged; 2) fast mode wherein the final segmentation map is selected from only one of the segmentation branches, the choice of which determines the extent of model pruning and speed gain. Fig. 1c shows how the choice of segmentation branch in fast mode results in architectures of varying complexity.

Owing to the nested skip pathways, UNet++ generates full resolution feature maps at multiple semantic levels, $\{x^{0,j}, j \in \{1, 2, 3, 4\}\}$, which are amenable to

Table 1: The image segmentation datasets used in our experiments.

Dataset	Images	Input Size	Modality	Provider
cell nuclei	670	96×96	microscopy	Data Science Bowl 2018
colon polyp	7,379	224×224	RGB video	ASU-Mayo [10,11]
liver	331	512×512	CT	MICCAI 2018 LiTS Challenge
lung nodule	1,012	64×64×64	CT	LIDC-IDRI [1]

Table 2: Number of convolutional kernels in U-Net and wide U-Net.

encoder / decoder	$X^{0,0}/X^{0,4}$	$X^{1,0}/X^{1,3}$	$X^{2,0}/X^{2,2}$	$X^{3,0}/X^{3,1}$	$X^{4,0}/X^{4,0}$
U-Net	32	64	128	256	512
wide U-Net	35	70	140	280	560

deep supervision. We have added a combination of binary cross-entropy and dice coefficient as the loss function to each of the above four semantic levels, which is described as:

$$\mathcal{L}(Y, \hat{Y}) = -\frac{1}{N} \sum_{b=1}^N \left(\frac{1}{2} \cdot Y_b \cdot \log \hat{Y}_b + \frac{2 \cdot Y_b \cdot \hat{Y}_b}{Y_b + \hat{Y}_b} \right) \quad (2)$$

where \hat{Y}_b and Y_b denote the flatten predicted probabilities and the flatten ground truths of b^{th} image respectively, and N indicates the batch size.

In summary, as depicted in Fig. 1a, UNet++ differs from the original U-Net in three ways: 1) having convolution layers on skip pathways (shown in green), which bridges the semantic gap between encoder and decoder feature maps; 2) having dense skip connections on skip pathways (shown in blue), which improves gradient flow; and 3) having deep supervision (shown in red), which as will be shown in Section 4 enables model pruning and improves or in the worst case achieves comparable performance to using only one loss layer.

4 Experiments

Datasets: As shown in Table 1, we use four medical imaging datasets for model evaluation, covering lesions/organs from different medical imaging modalities. For further details about datasets and the corresponding data pre-processing, we refer the readers to the supplementary material.

Baseline models: For comparison, we used the original U-Net and a customized wide U-Net architecture. We chose U-Net because it is a common performance baseline for image segmentation. We also designed a wide U-Net with similar number of parameters as our suggested architecture. This was to ensure that the performance gain yielded by our architecture is not simply due to increased number of parameters. Table 2 details the U-Net and wide U-Net architecture.

Implementation details: We monitored the Dice coefficient and Intersection over Union (IoU), and used *early-stop* mechanism on the validation set. We also

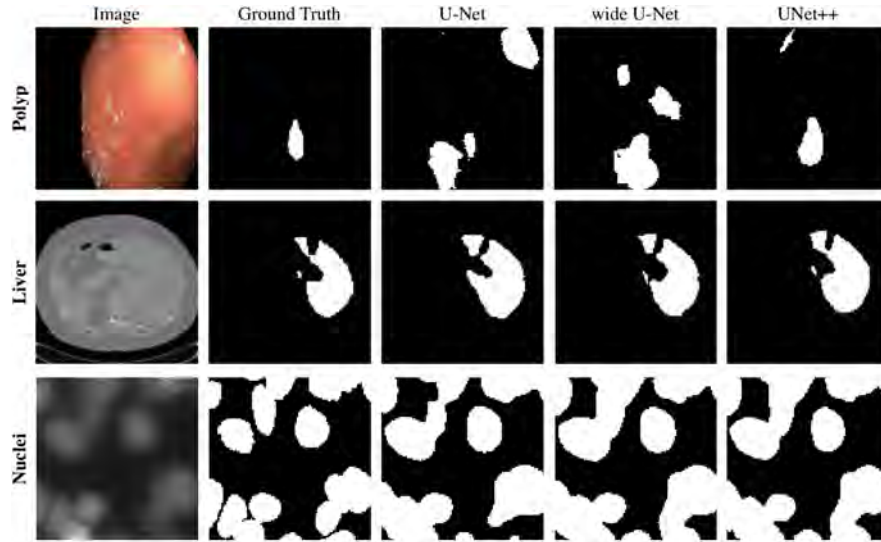


Fig. 2: Qualitative comparison between U-Net, wide U-Net, and UNet++, showing segmentation results for polyp, liver, and cell nuclei datasets (2D-only for a distinct visualization).

used Adam optimizer with a learning rate of $3e-4$. Architecture details for U-Net and wide U-Net are shown in Table 2. UNet++ is constructed from the original U-Net architecture. All convolutional layers along a skip pathway ($X^{i,j}$) use k kernels of size 3×3 (or $3 \times 3 \times 3$ for 3D lung nodule segmentation) where $k = 32 \times 2^i$. To enable deep supervision, a 1×1 convolutional layer followed by a sigmoid activation function was appended to each of the target nodes: $\{x^{0,j} \mid j \in \{1, 2, 3, 4\}\}$. As a result, UNet++ generates four segmentation maps given an input image, which will be further averaged to generate the final segmentation map. More details can be found at github.com/Nested-UNet.

Results: Table 3 compares U-Net, wide U-Net, and UNet++ in terms of the number parameters and segmentation accuracy for the tasks of lung nodule segmentation, colon polyp segmentation, liver segmentation, and cell nuclei segmentation. As seen, wide U-Net consistently outperforms U-Net except for liver segmentation where the two architectures perform comparably. This improvement is attributed to the larger number of parameters in wide U-Net. UNet++ without deep supervision achieves a significant performance gain over both U-Net and wide U-Net, yielding average improvement of 2.8 and 3.3 points in IoU. UNet++ with deep supervision exhibits average improvement of 0.6 points over UNet++ without deep supervision. Specifically, the use of deep supervision leads to marked improvement for liver and lung nodule segmentation, but such improvement vanishes for cell nuclei and colon polyp segmentation. This is because polyps and liver appear at varying scales in video frames and CT

Table 3: Segmentation results (IoU: %) for U-Net, wide U-Net and our suggested architecture UNet++ with and without deep supervision (DS).

Architecture	Params	Dataset			
		cell nuclei	colon polyp	liver	lung nodule
U-Net [9]	7.76M	90.77	30.08	76.62	71.47
Wide U-Net	9.13M	90.92	30.14	76.58	73.38
UNet++ w/o DS	9.04M	92.63	33.45	79.70	76.44
UNet++ w/ DS	9.04M	92.52	32.12	82.90	77.21

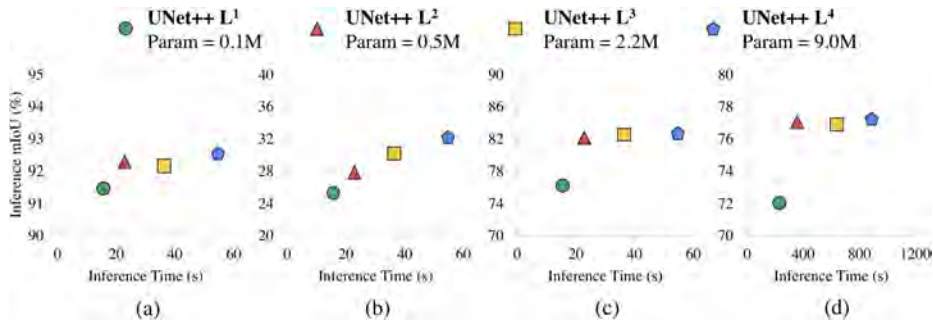


Fig. 3: Complexity, speed, and accuracy of UNet++ after pruning on (a) cell nuclei, (b) colon polyp, (c) liver, and (d) lung nodule segmentation tasks respectively. The inference time is the time taken to process **10k** test images using one NVIDIA TITAN X (Pascal) with 12 GB memory.

slices; and thus, a multi-scale approach using all segmentation branches (deep supervision) is essential for accurate segmentation. Fig. 2 shows a qualitative comparison between the results of U-Net, wide U-Net, and UNet++.

Model pruning: Fig. 3 shows segmentation performance of UNet++ after applying different levels of pruning. We use UNet++ L^i to denote UNet++ pruned at level i (see Fig. 1c for further details). As seen, UNet++ L^3 achieves on average 32.2% reduction in inference time while degrading IoU by only 0.6 points. More aggressive pruning further reduces the inference time but at the cost of significant accuracy degradation.

5 Conclusion

To address the need for more accurate medical image segmentation, we proposed UNet++. The suggested architecture takes advantage of re-designed skip pathways and deep supervision. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks, resulting in a possibly simpler optimization problem for the optimizer

to solve. Deep supervision also enables more accurate segmentation particularly for lesions that appear at multiple scales such as polyps in colonoscopy videos. We evaluated UNet++ using four medical imaging datasets covering lung nodule segmentation, colon polyp segmentation, cell nuclei segmentation, and liver segmentation. Our experiments demonstrated that UNet++ with deep supervision achieved an average IoU gain of 3.9 and 3.4 points over U-Net and wide U-Net, respectively.

Acknowledgments This research has been supported partially by NIH under Award Number R01HL128785, by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.
2. M. Drozdal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.
3. D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf. Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*, 2017.
4. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
5. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.
6. C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.
7. X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng. H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from ct volumes. *arXiv preprint arXiv:1709.07330*, 2017.
8. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
9. O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
10. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
11. Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7340–7351, 2017.

Computer-Aided Pulmonary Embolism Detection Using a Novel Vessel-Aligned Multi-planar Image Representation and Convolutional Neural Networks

Nima Tajbakhsh¹, Michael B. Gotway², and Jianming Liang¹

¹ Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ, USA
 {Nima.Tajbakhsh, Jianming.Liang}@asu.edu

² Department of Radiology, Mayo Clinic, Scottsdale, AZ, USA
 Gotway.Michael@mayo.edu

Abstract. Computer-aided detection (CAD) can play a major role in diagnosing pulmonary embolism (PE) at CT pulmonary angiography (CTPA). However, despite their demonstrated utility, to achieve a clinically acceptable sensitivity, existing PE CAD systems generate a high number of false positives, imposing extra burdens on radiologists to adjudicate these superfluous CAD findings. In this study, we investigate the feasibility of convolutional neural networks (CNNs) as an effective mechanism for eliminating false positives. A critical issue in successfully utilizing CNNs for detecting an object in 3D images is to develop a “right” image representation for the object. Toward this end, we have developed a vessel-aligned multi-planar image representation of emboli. Our image representation offers three advantages: (1) efficiency and compactness—concisely summarizing the 3D contextual information around an embolus in only 2 image channels, (2) consistency—automatically aligning the embolus in the 2-channel images according to the orientation of the affected vessel, and (3) expandability—naturally supporting data augmentation for training CNNs. We have evaluated our CAD approach using 121 CTPA datasets with a total of 326 emboli, achieving a sensitivity of 83% at 2 false positives per volume. This performance is superior to the best performing CAD system in the literature, which achieves a sensitivity of 71% at the same level of false positives. We have further evaluated our system using the entire 20 CTPA test datasets from the PE challenge. Our system outperforms the winning system from the challenge at 0mm localization error but is outperformed by it at 2mm and 5mm localization errors. In our view, the performance at 0mm localization error is more important than those at 2mm and 5mm localization errors.

Keywords: Computer-aided detection, pulmonary embolism, convolutional neural networks, vessel-aligned image representation.

1 Introduction

Pulmonary embolism (PE) is a thrombus, occasionally colloquially referred to as a blood clot, that travels from the legs, or rarely other parts of the body, to the lungs where it obstructs central, lobar, segmental, or subsegmental pulmonary arteries depending on the size of the embolus. The untreated mortality rate of PE may approach

30%. However, with early diagnosis and treatment, the mortality rate decreases to as low as 2% to 11%. CT pulmonary angiography (CTPA) is the primary means for the evaluation of suspected PE. At CTPA, an embolus appears as a dark region surrounded by the brighter, contrast-enhanced vessel lumen. CTPA dataset interpretation demands a radiologist to carefully trace each branch of the pulmonary artery for any suspected PEs. Therefore, PE diagnosis often requires extensive reading time, and the accuracy of CTPA interpretation depends on the radiologists' experience, attention span, eye fatigue, and their sensitivity to visual characteristics of PEs.

Computer-aided detection (CAD) can play a major role in detecting and diagnosing PEs. Recent clinical studies have shown that CAD systems can help radiologists increase their sensitivity for PE detection [3]. However, despite their demonstrated utility, existing CAD systems still require a relatively high false positive rate in order to achieve a clinically acceptable PE sensitivity. The false positives generated by CAD systems prolong the reading time of CTPA studies, because each CAD finding must be examined by a radiologist and adjudicated. It is therefore highly desirable to develop a CAD system that can achieve higher sensitivity while maintaining a clinically acceptable false positive range (between 1 to 5 false positives per CTPA study).

This paper investigates the feasibility of convolutional neural networks (CNNs) as an effective tool for eliminating false positive detections. We have found that the effective utilization of CNNs for detecting PEs and removing false detections in 3D CTPA datasets is contingent on an effective image representation of PEs. As such, a key finding from our work is a vessel-aligned multi-planar image representation of emboli that offers three advantages: (1) our proposed image representation is *efficient* and *compact* because it concisely summarizes the 3D contextual information around an embolus in only 2 image channels; (2) our proposed image representation is *consistent* because it automatically aligns the embolus in the 2-channel images according to the orientation of the affected vessel; and (3) our proposed image representation is *expandable* because it naturally supports data augmentation for training a CNN. We have evaluated our CAD system using 121 CTPA datasets containing a total of 326 emboli, achieving a sensitivity of 83% at 2 false positives per volume. This performance is superior to the best performing CAD system in the literature, which achieves a sensitivity of 71% at the same level of false positives. We have further evaluated our system with the entire 20 CTPA test datasets from the PE challenge [1]. Our system outperforms MeVis', the best reported system, at 0mm localization error but is outperformed by MeVis' at 2mm and 5mm localization errors. In our view, the performance at 0mm localization error is more important than those at 2mm and 5mm localization errors.

2 Related Work

CAD systems for PE typically consist of four stages: 1) extracting a volume of interest (VOI) from the original dataset by performing lung segmentation [5,11,8] or vessel segmentation [7,11,2]; 2) generating a set of PE candidates within the VOI using algorithms such as tobogganing [5]; 3) extracting hand-crafted features from each PE candidate (e.g., [6]), and 4) computing a confidence score for each of the candidates using a rule based classifier [7], neural networks and a nearest neighbor classifier [11,8],

or a multi-instance classifier [5]. However, current CAD systems either produce many false positives to achieve a high detection sensitivity [7], or yield acceptable false positive rates but with only limited sensitivity levels [8,2,11] (see Table 1 for a detailed performance comparison). We hypothesize that inadequate modeling of PEs based on hand-crafted features results in suboptimal CAD performance, and therefore investigate the use of a new image representation for PEs, coupled with CNNs, to improve state-of-the-art performance.

3 Proposed Method

Given a CTPA dataset, our method first segments lungs and then generates a set of PE candidates within the lung area using the tobogganing algorithm [5]. Our method then uses our vessel-aligned multi-planar image representation to produce a 2-channel image representation for each PE candidate. The resulting 2-channel patches are then fed to a CNN to classify the underlying candidates into PE or non-PE categories. Please refer to [5] for the tobogganing algorithm and to [4] for the CNN. In the following, we shall focus on our suggested vessel-aligned multi-planar image representation.

3.1 Vessel-Aligned Multi-planar Image Representation

The success of CNNs for object detection in 3D volumetric datasets such as CT images heavily relies on the representation of the object of interest [9,10]. We have experimentally found that a suitable 3D image representation for CNNs must meet three requirements: (1) compactness and efficiency, (2) consistency across instances, and (3) expandability for data augmentation. With these requirements in mind, we propose an image representation, called vessel-aligned multi-planar image representation, for PE, which has these three critical properties. In the following, we first describe our unique image representation and then explain how it meets the above requirements.

To obtain our image representation, we first estimate the orientation of the vessel that contains the candidate. For this purpose, a $15 \times 15 \times 15$ mm neighborhood is extracted around the PE candidate. In the resulting subvolume, the PE appears as a filling defect, because PEs are relatively darker than the contrast-enhanced vessel. To minimize the influence of the filling defect on vessel orientation estimation, the vessel-like intensity value of 100 HU (Hounsfield units) is assigned to the PE voxels within the subvolume. Note that the tobogganing algorithm [4] has already labeled the PE voxels associated with each candidate. Next, a principle component analysis is performed in the connected component (≥ 100 HU) that contains the PE. If v_1, v_2, v_3 denote the eigen vectors of the analyzed component ($\lambda_1 \geq \lambda_2 \geq \lambda_3$), then interpolating the volume along $\{v_1, v_2\}$ or $\{v_1, v_3\}$ results in the longitudinal view of the PE (the first channel of our image representation) and interpolating the volume along $\{v_2, v_3\}$ results in the cross-sectional view of the PE (the second channel of our image representation).

Our image representation is compact because it concisely summarizes the 3D contextual information around PEs in only 2 image channels. While it is theoretically possible to train a CNN using subvolumes with an arbitrary number of slices, the performance of such networks have been reported to be inferior to the CNNs that have been trained

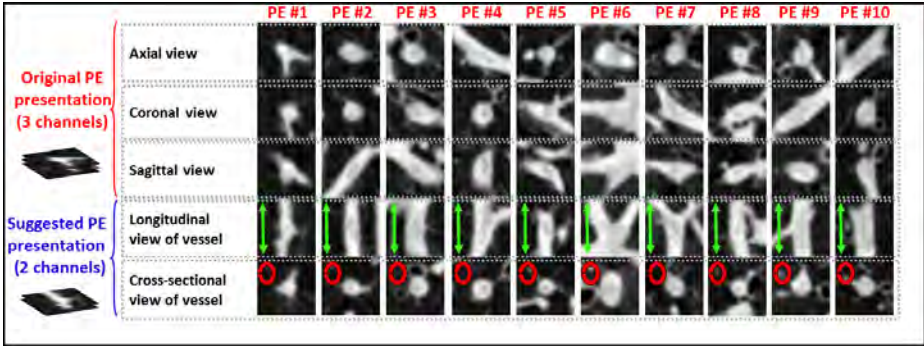


Fig. 1. The suggested 2-channel image representation characterizes emboli more consistently than the original axial, sagittal, and coronal views. As seen, in nearly all cases, the suggested scheme consistently captures PEs within the containing vessel as elongated and circular structures in the first and second channels, respectively. The three standard views do not provide this property given the varying orientation of the containing vessels. A consistent image appearance is the key to training an accurate image classifier.

using samples with a fewer number of slices [9]. In fact, the information embedded in the additional image slices has been shown to degrade classification performance [9]. This phenomenon is attributed to the curse of dimensionality, where a large number of image channels corresponds to learning a far larger number of network parameters, which in turn leads to over-fitting to the training samples and thus poor generalization performance. It is therefore desirable to efficiently represent the 3D context around the object of interest using a low dimensional image representation.

Our image representation consistently describes PEs and the containing vessels. In general, emboli can affect pulmonary arteries in any orientation. As a result, images extracted from the axial, sagittal, coronal planes exhibit a significant variation in the appearance of emboli. This in turn complicates the classification task and hinders effective utilization of CNNs. With the benefit of vessel alignment, our image representation allows for a consistent image representation whereby emboli consistently appear as elongated structures in the longitudinal vessel view and as circular structures in the cross-sectional vessel view. Fig. 1 illustrates variations in PE appearances using the suggested vessel-aligned image representation and a standard image representation based on sagittal, coronal and axial views.

Our image representation amenable supports data augmentation, which is essential for effective training and testing of CNNs. In 2D applications, data augmentation is performed by applying arbitrary in-plane rotations and then collecting samples at multiple scales and translations. A 3D representation must also support the above operations to enable data augmentation. While it is straightforward to extend translation and scale to a 3D space, the rotation operation can be problematic. Our image representation is based on longitudinal and cross-sectionals planes; however, rotating such planes along a random axis will result in the arbitrary appearance of the same PE in the resulting 2-channel images (Fig. 2(a)). The major challenge is how to perform 3D rotation such that the PE representation remains consistent. Our image representation accommodates this need by rotating the planes around the vessel axis v_1 . By doing so, we obtain two

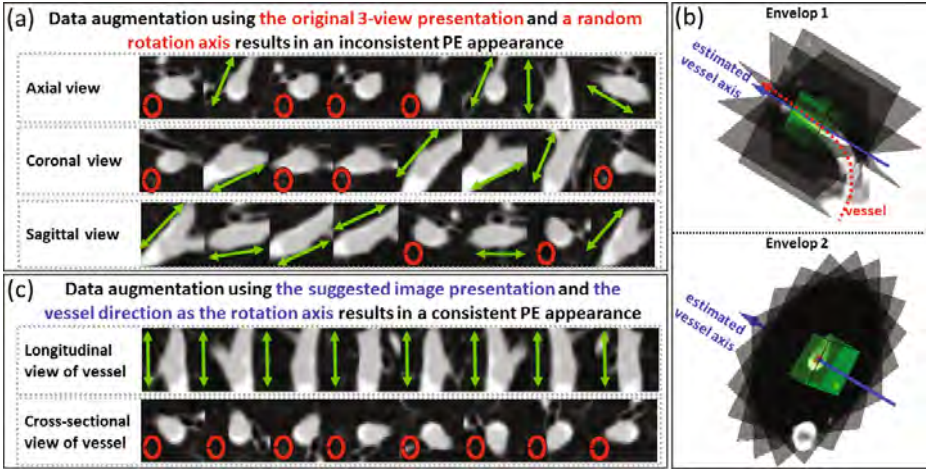


Fig. 2. (a) Data augmentation using random rotation axes, as suggested in [10], results in inconsistent PE appearance. (b) The suggested image representation uses two envelopes of planes to achieve consistency for data augmentation. (c) Consistent PE appearance after data augmentation using the suggested envelopes of planes. The green double arrows and red ellipses represent the shapes of PEs and the containing vessels.

envelopes of image planes (see Fig. 2(b)) where the first envelope contains the planes that all intersect at the vessel axis and the second envelope contains the image planes whose normals are the vessel axis. By selecting any pairs of planes from the two envelopes, one can generate a new PE instance while retaining the consistency. Fig. 2(c) illustrates consistency in appearance of PEs after data augmentation using the suggested envelopes of planes.

3.2 Convolutional Neural Networks (CNNs)

CNNs are deep learning machines that can potentially eliminate the need for designing hand-crafted features—they learn the features and train the classifier simultaneously. CNNs are so-named for their convolutional layers that learn discriminative patterns of the training samples at multiple scales. In this work, we employ the GPU-based open-source implementation of CNNs [4] and use the layout shown in Fig. 3. We have experimented with more sophisticated network architectures but observed no significant performance gain.

4 Experiments

We have evaluated our CAD system using 2 databases: (1) our private database consisting of 121 CTPA datasets with a total of 326 emboli, and (2) the test datasets from the PE challenge [1] consisting of 20 CTPA datasets with a total of 133 emboli.

Evaluations Using Our Database. The candidate generation module of our CAD system produces a total of 8585 PE candidates in the 121 CTPA datasets, of which 7722

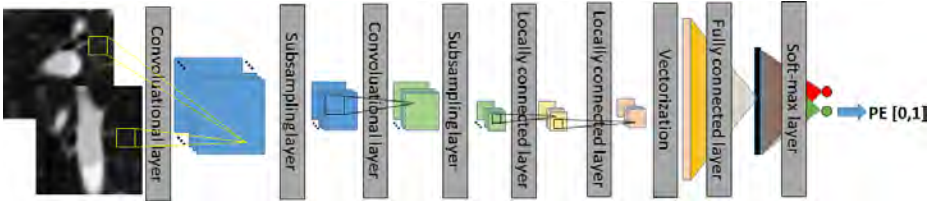


Fig. 3. The layout of the CNN used in our experiments.

are false positives and 863 are true positives. It is possible for a CAD system to produce multiple detections for a single large PE and that explains why the number of our true detections is greater than the number of PEs in the database. According to the available ground truth, the candidate generation module achieves a sensitivity of 93% for PE detection while producing, on average, 65.8 false positives per patient.

Our goal is to use CNNs to minimize the number of false positives while maintaining a high sensitivity for PE detection. To train CNNs, we randomly split the collected detections at the patient level into 3 groups, enabling a 3-fold cross validation of our CAD system. We then used the false positive detections as negative candidates and the true detections as positive candidates. Given the limited number of candidates, we formed the training set by performing data augmentation. For this purpose, we collected $N = N_r \times N_t \times N_s$ samples from each candidate location based on our vessel-aligned multi-planar PE representation, where N_r is the number of rotations, N_t is the number of translations, and N_s is the number of image scaling. To produce rotated patches, we rotated the longitudinal and cross-sectional vessel planes around the vessel axis $N_r = 5$ times. For scaling, we extracted patches at $N_s = 3$ different scales, resulting in 10mm, 15mm, and 20mm wide patches. In each scale, we have performed image interpolation so that the resulting patches are all 32x32 pixels. For translation, we shifted the candidate location along the vessel direction $N_t = 3$ times, up to 20% of the physical width of the patches. With data augmentation, we can increase the size of the training set by a factor of $N = 45$, which is sufficiently large to train CNNs. Given a test CTPA dataset, we first obtain a set of candidates, and then apply the trained CNN on N 2-channel image patches extracted from each candidate location. The confidence values for the underlying candidate is then computed as the average of the resulting N confidence values. Once all the test candidates are processed, we obtain an FROC curve by changing a threshold on the corresponding confidence values.

Fig. 4 shows the FROC curve of the suggested system. For comparison, we have computed the FROC curve of [5] using the prediction results provided by the corresponding author. We have chosen [5] for performance comparison because their suggested system has achieved the best performance reported in the literature on a reasonably large CTPA database (see Table 1). For further comparison, we have replaced our suggested image presentation with a 2.5D image representation as suggested in [10]. For fair comparisons, we have kept all the other stages the same. As seen in Fig. 4, our system outperforms [5], which is a CAD system based on a carefully designed set of hand-crafted features [6] and a multi-instance classifier. In addition, we observed that the CNN trained using a 2.5D image representation results in a performance which is not only inferior to our suggested image representation but also to the hand-crafted approach, demonstrating the

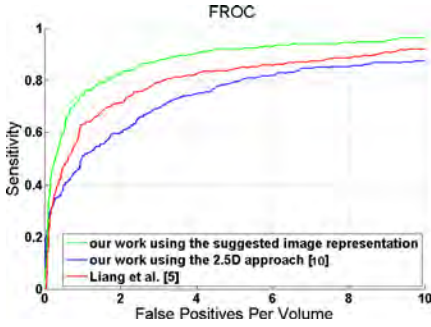


Fig. 4. Our CAD system using the suggested image representation outperforms the best hand-crafted approach [5] and also a CNN powered by a 2.5D approach [10].

Method	Sensitivity	FPs/vol	#datasets	#PEs
Liang et al. [5]	70.0%	2.0	132	716
Bouma et al. [2]	58%	4.0	19	116
Park et al. [8]	63.2	18.4	20	44
Ozkan et al. [7]	61%	8.2	33	450
Wang et al. [11]	62%	17.1	12	24
This work	83.4%	2.0	121	326
This work (2.5D)	60.4%	2.0	121	326
Liang et al. [5]	71.7%	2.0	121	326

Table 1. (top) Performance of the existing PE CAD systems obtained through different datasets. (bottom) Performance comparison based on our database of 121 CTPA datasets. Operating points are taken from Fig. 4.

significant contribution of our effective image representation in achieving the improved performance. Table 1 contrasts the performance of our proposed CAD system with that of the other CAD systems suggested in the literature.

Evaluations Using PE Challenge Database. We have further trained a CNN, powered by our unique image representation, using all 121 CTPA datasets from our database and then evaluated our CAD system using the test database from the PE challenge [1]. Since the ground-truth was not available on the website, our detection results were evaluated by the organizers. At 0mm localization error, our CAD system achieves a sensitivity of 34.6% at 2 FPs/vol, which outperforms the winning team (a commercial CAD system designed MeVis Medical Solutions) with a sensitivity of 28.4% at the same false positive rate. Our CAD system is, however, outperformed by MeVis’ at 2mm and 5mm localization errors. For more detailed comparisons, please refer to [1]. Despite the demonstrated superiority at 0mm localization error, our CAD system exhibits a notable performance degradation compared to the results obtained using our database. We attribute this to faulty lung segmentation, which results in many PE candidates in the colon and diaphragm. Since such false positives had not been observed in our training sets, the trained CNN did not perform optimally in removing such false positives.

5 Conclusions and Discussions

In this work, we investigated the possibility of a unique PE representation, coupled with CNNs, to produce a more accurate PE CAD system. Our system contrasts with existing systems, wherein a traditional hand-crafted feature design is used for characterizing PEs. We evaluated our system in comparison with the most robust hand-crafted approach [5] and a learning-based approach using CNNs powered by a 2.5D PE representation, demonstrating a marked performance improvement. Our method was also tested using the test database from the PE challenge where it outperformed the academic systems at the three localization errors and also outperformed a commercial CAD system at 0mm localization error. Moving forward, we intend to improve the accuracy

of our CAD system using additional training cases to address the issue of faulty lung segmentation resulting from non-pulmonary candidates.

Acknowledgment. This project is supported by a seed grant awarded by Arizona State University and Mayo Clinic. We thank German Gonzalez, the organizer of the CAD PE challenge, for evaluating our results with the test datasets from the challenge.

References

1. <http://www.cad-pe.org>
2. Bouma, H., Sonnemans, J.J., Vilanova, A., Gerritsen, F.A.: Automatic detection of pulmonary embolism in cta images. *IEEE Transactions on Medical Imaging* 28(8), 1223–1230 (2009)
3. Das, M., Mühlenbruch, G., Helm, A., Bakai, A., Salganicoff, M., Stanzel, S., Liang, J., Wolf, M., Günther, R.W., Wildberger, J.E.: Computer-aided detection of pulmonary embolism: influence on radiologists detection performance with respect to vessel segments. *European Radiology* 18(7), 1350–1355 (2008)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
5. Liang, J., Bi, J.: Computer aided detection of pulmonary embolism with tobogganing and mutple instance classification in CT pulmonary angiography. In: Karssemeijer, N., Lelieveldt, B. (eds.) *IPMI 2007. LNCS*, vol. 4584, pp. 630–641. Springer, Heidelberg (2007)
6. Liang, J., Bi, J.: Local characteristic features for computer aided detection of pulmonary embolism in ct angiography. In: *Proceedings of the First MICCAI Workshop on Pulmonary Image Analysis*, pp. 263–272 (2008)
7. Özkan, H., Osman, O., Şahin, S., Boz, A.F.: A novel method for pulmonary embolism detection in cta images. *Computer Methods and Programs in Biomedicine* 113(3), 757–766 (2014)
8. Park, S.C., Chapman, B.E., Zheng, B.: A multistage approach to improve performance of computer-aided detection of pulmonary embolisms depicted on CT images: Preliminary investigation. *IEEE Transactions on Biomedical Engineering* 58(6), 1519–1527 (2011)
9. Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M.: Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) *MICCAI 2013, Part II. LNCS*, vol. 8150, pp. 246–253. Springer, Heidelberg (2013)
10. Roth, H.R., et al.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) *MICCAI 2014, Part I. LNCS*, vol. 8673, pp. 520–527. Springer, Heidelberg (2014)
11. Wang, X., Song, X., Chapman, B.E., Zheng, B.: Improving performance of computer-aided detection of pulmonary embolisms by incorporating a new pulmonary vascular-tree segmentation algorithm. In: *SPIE Medical Imaging*, pp. 83152U–83152U. International Society for Optics and Photonics (2012)

Automating Carotid Intima-Media Thickness Video Interpretation with Convolutional Neural Networks*

Jae Y. Shin*, Nima Tajbakhsh*, R. Todd Hurst, Christopher B. Kendall, and Jianming Liang†

Abstract

Cardiovascular disease (CVD) is the leading cause of mortality yet largely preventable, but the key to prevention is to identify at-risk individuals before adverse events. For predicting individual CVD risk, carotid intima-media thickness (CIMT), a noninvasive ultrasound method, has proven to be valuable, offering several advantages over CT coronary artery calcium score. However, each CIMT examination includes several ultrasound videos, and interpreting each of these CIMT videos involves three operations: (1) select three end-diastolic ultrasound frames (EUF) in the video, (2) localize a region of interest (ROI) in each selected frame, and (3) trace the lumen-intima interface and the media-adventitia interface in each ROI to measure CIMT. These operations are tedious, laborious, and time consuming, a serious limitation that hinders the widespread utilization of CIMT in clinical practice. To overcome this limitation, this paper presents a new system to automate CIMT video interpretation. Our extensive experiments demonstrate that the suggested system significantly outperforms the state-of-the-art methods. The superior performance is attributable to our unified framework based on convolutional neural networks (CNNs) coupled with our informative image representation and effective post-processing of the CNN outputs, which are uniquely designed for each of the above three operations.

***Shorter version:** J. Y. Shin, N. Tajbakhsh, R. T. Hurst, C. B. Kendall, and J. Liang. Automating carotid intima-media thickness video interpretation with convolutional neural networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR16), Pages 2526-2535; **Extended version:** N. Tajbakhsh, J. Y. Shin, R. T. Hurst, C. B. Kendall, and J. Liang. Automatic interpretation of carotid intima-media thickness videos using convolutional neural networks. Deep Learning for Medical Image Analysis. edited by Kevin Zhou, Hayit Greenspan and Dinggang Shen, Academic Press, 2017.

†J. Y. Shin, N. Tajbakhsh and J. Liang are with the Department of Biomedical Informatics, Arizona State University, 13212 East Shea Boulevard, Scottsdale, AZ 85259, USA (e-mail: {Sejong.Nima.Tajbakhsh, Jianming.Liang}@asu.edu). Nima Tajbakhsh and Jae Y. Shin have contributed equally.

‡R. T. Hurst and C. Kendall are with the Division of Cardiovascular Diseases of Mayo Clinic, 13400 E. Shea Blvd., Scottsdale, AZ 85259, USA (e-mail: {Hurst.R, Kendall.Christopher}@mayo.edu).

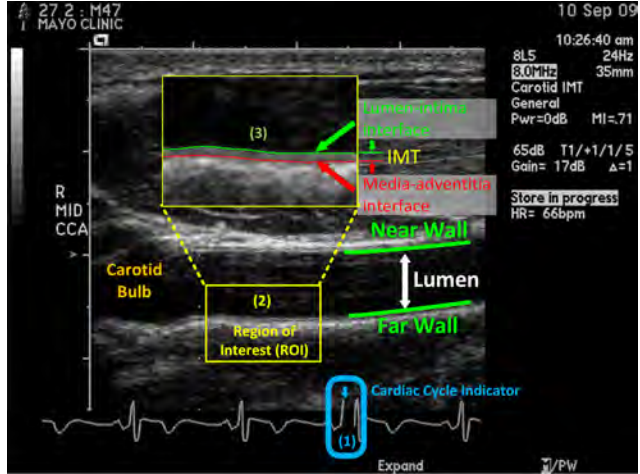


Figure 1: Longitudinal view of the carotid artery in an ultrasound B-scan image. CIMT is defined as the distance between the lumen-intima interface and the media-adventitia interface, measured approximately 1 cm distal from the carotid bulb on the far wall of the common carotid artery at the end of the diastole; therefore, interpreting a CIMT video involves three operations: (1) select three end-diastolic ultrasound frames (EUFs) in each video (the cardiac cycle indicator, a black line, shows to where in the cardiac cycle the current frame corresponds); (2) localize a region of interest (ROI) approximately 1 cm distal from the carotid bulb in the selected EUF; (3) measure the CIMT within the localized ROI. This paper aims to automate these three operations simultaneously through a unified framework based on convolutional neural networks.

1. Introduction

Given the clinical significance of carotid intima-media thickness (CIMT) as an early and reliable indicator of cardiovascular risk, several methods have been developed for CIMT image interpretation. The CIMT is defined as the distance between the lumen-intima and media-adventitia interfaces at the far wall of the carotid artery (Figure 1). Therefore, to measure CIMT, the lumen-intima and the media-adventitia interfaces must be identified. As a re-

sult, the earlier approaches are focused on analyzing the intensity profile and distribution, computing the gradient [24, 27, 5], or combining various edge properties through dynamic programming [13, 3, 25]. Recent approaches [16, 4, 23, 28, 7, 2] are mostly based on active contours (aka, snakes) or their variations [9]. Some of these approaches require user interaction, while other approaches aim for complete automation through integrating with various image processing algorithms, such as Hough transform [21] and dynamic programming [25]. Most recently, Menchn-Lara et al. employed a committee of standard multilayer perceptrons in [18] and a single standard multilayer perceptron with an auto-encoder in [19] for CIMT image interpretation, but both methods did not outperform the snake-based methods from the same research group [2, 1]. For a more complete survey of methods for automatic CIMT measurements, please refer to the review studies conducted by Molinari et al. [22] and Loizou et al. [15].

However, nearly all the aforementioned methods are focused on only the third operation: CIMT measurement, ignoring the two preceding operations, i.e., frame selection and ROI localization. To our knowledge, the only system that simultaneously automates the three operations is the work [26], an extension of [29], which automatically selects the EUF frame, localizes the ROI in each selected EUF frame, and provides the CIMT measurement in the selected ROI. However, as with other works, this method is based on hand-crafted algorithms, which often lack the desired robustness for routine clinical use, a weakness that we aim to overcome in this paper.

A key contribution of this paper is a new system that accelerates CIMT video interpretation by automating all the three operations in a novel unified framework based on convolutional neural networks (CNNs). We will show that with proper pre-processing and post-processing, our proposed CNN-based approach can significantly outperform the existing methods in all aspects of CIMT image interpretation including frame selection, ROI localization, and CIMT measurements, making the following specific contributions:

- A unified framework based on CNNs that automates the entire CIMT interpretation process. This is in contrast to the prior works where only the very last step of the CIMT interpretation process was automated. The performance of the suggested system significantly outperforms the hand-crafted approach [26], which, to our knowledge, is the only system in the literature that aimed to automate all the above three tasks.
- A novel frame selection method based on the ECG signals at the bottom of ultrasound frames. The suggested method utilizes effective pre-processing of patches and post processing of CNN outputs, enabling a significant increase in the performance of a baseline CNN.
- A new method that localizes the ROI for CIMT in-

terpretation. The suggested method combines the discriminative power of a CNN with a contextual constrain to accurately localize the ROIs in the selected frames. We demonstrate that the suggested contextually-constrained CNN outperforms the performance of a baseline CNN.

- A framework that combines CNNs with active contour models for accurate boundary segmentation. Specifically, given a localized ROI, the CNN initializes two open snakes, which further deform to acquire the shapes of intima-media boundaries. We show that the segmentation accuracy of the suggested method is far higher than the state-of-the-art methods.
- Extensive evaluation of each stage of the suggested CIMT interpretation system. Specifically, we perform leave-one-patient-out cross-validation¹ using only the training CIMT videos to tune the parameters of the suggested system, and then thoroughly evaluate the performance of our system using a large number of independent test CIMT videos.

2. CIMT Protocol

The CIMT exams utilized in this paper were performed with B-Mode ultrasound using an 8-14MHz linear array transducer utilizing fundamental frequency only (Acuson Sequoia™, Mountain View, CA, USA) [6]. The carotid screening protocol begins with scanning bilateral carotid arteries in a transverse manner from the proximal aspect to the proximal internal and external carotid arteries. The probe is then turned to obtain the longitudinal view of the distal common carotid artery. The sonographer optimizes the 2D images of the lumen-intima and media-adventitia interfaces at the level of the common carotid artery by adjusting overall gain, time gain, compensation and focus position. Once the parameters are optimized, the sonographer captures two CIMT videos focused on the common carotid artery from two optimal angles of incidence. The same procedure is repeated for the other side of neck, resulting in a total of 4 CIMT videos for each subject.

3. Method

Our goal is to automate the three operations in CIMT video interpretation, i.e, given a CIMT video, our method will automatically identify three EUFs (Section 3.1), localize an ROI in each EUF (Section 3.2), and segment the lumen-intima and media-adventitia interfaces within each ROI (Section 3.3).

3.1. Frame Selection

We select the EUFs based on the ECG signal embedded at the bottom part of a CIMT video. The cardiac cycle indi-

¹We leave all the videos from one patient out for validation.

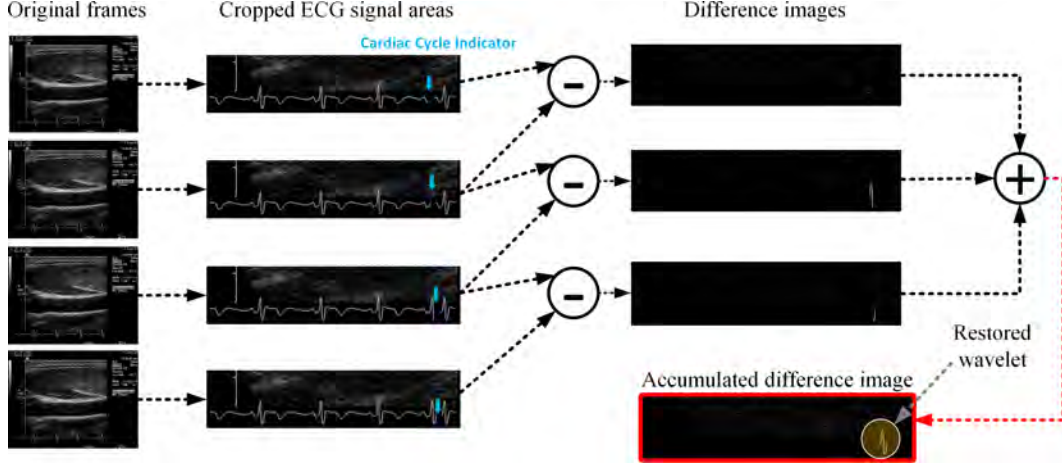


Figure 2: An accumulated difference image is generated by adding up three neighboring difference images.

cator is represented by a moving black line in each frame. Since the ECG signal is overlaid on the ultrasound image, there is quite bit of noise around the indicator. The challenge is to reconstruct the original ECG signal from noisy frames and to detect the R peaks from the ECG signal, as the R-peaks correspond to the EUFs. To do so, we introduce accumulated difference images that carry sufficient information for CNN to learn and distinguish R-peaks from non-R-peaks.

Training Phase: Let I^t denote an image subregion selected from the lower part of an ultrasound frame so that it contains the ECG signal. We first construct a set of difference images d^t by subtracting every consecutive pairs of images, $d^t = |I^t - I^{t+1}|$, and then form accumulated difference images by adding up every three neighboring difference images, $D^t = \sum_{i=0}^2 d^{t-i}$. Accumulated difference image D^t can capture the cardiac cycle indicator at frame t . Figure 2 illustrates how an accumulated difference image is generated.

Next, we determine the location of the restored wavelet in each accumulated difference image. For this purpose, we find the weighted centroid $c = [c_x, c_y]$ of each accumulated difference image D^t as follows:

$$c = \frac{1}{Z_t} \sum_{p \in D^t} D^t(p_x, p_y) \times p$$

where $p = [p_x, p_y]$ is a pixel in the accumulated difference image and $Z_t = \sum_{p \in D^t} D^t(p_x, p_y)$ is a normalization factor that ensures the weighted centroid stays within the image boundary. Once centroids are identified, we extract patches of size 32×32 around the centroid locations. Specifically, we extract patches with up to 2 pixel translations from each centroid. However, we do not scale the patches in data augmentation, because doing so would inject label noise in the training set. For instance, a small restored wavelet may take the appearance of an R-peak after expanding or an R-peak

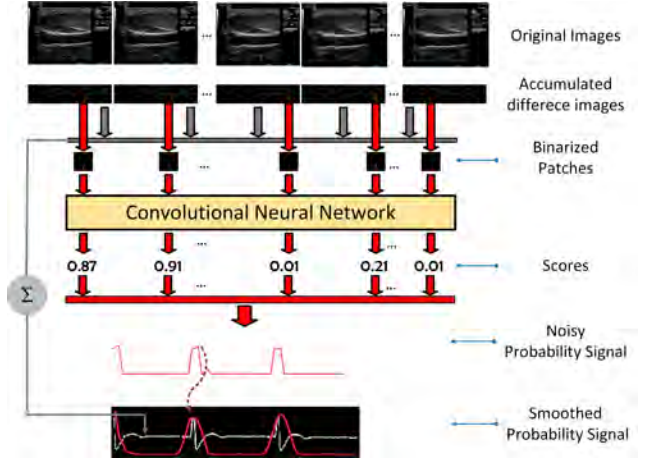


Figure 3: The test stage of our automatic frame selection scheme.

wavelet may look like a non-R-peak wavelet after shrinking. Nor do we perform rotation-based patch augmentation, because we do not expect the restored wavelets to appear with rotation in the test image patches. Once collected, patches are binarized using Otsu’s method. In Section 4, we discuss the choice of binarization method through an extensive set of experiments. Each binary patch is then labeled as positive if it corresponds to an EUF (i.e., an R-peak); otherwise negative. Basically, given a patch, we first determine the accumulated difference image from which the patch is extracted. We then trace back to the underlying difference images and check whether they are related to the EUF or not. Once the patches are labeled, we form a stratified set with 96,000 patches to train a 2-way CNN for frame selection.

Testing Phase: Figure 3 shows our frame selection system given a test video. We first compute an accumulated differ-

ence image for each frame in the video. We then extract image patches from the weighted centroids of the accumulated difference images. The probability of each frame being the EUF is measured as the average probabilities assigned by the CNN to the corresponding patches. By concatenating the resulting probabilities for all the frames in the video, we obtain a probability signal whose local maxima indicate the locations of the EUFs. However, the generated probability signals often exhibit abrupt changes, which can cause too many local maxima along the signal. We therefore first smooth the probability signal using a Gaussian function, and then find the EUFs by locating the local maxima of the smoothed signals. In Figure 3, for illustration purposes, we have also shown the reconstructed ECG signal, which is computed as the average of the accumulated difference images, $\frac{1}{N} \sum_{t=1}^N D^t$ with N being the number of frames in the video. As seen, the probability of being the EUF reaches its maximum around the R peaks of the QRS complexes (as desired) and then smoothly decays as it distances from the R peaks. By mapping the locations of the local maxima to the frame numbers, we can identify the EUFs in the test video.

3.2. ROI Localization

Accurate localization of the ROI is challenging, because, as seen in Figure 1, there are no significant differences that can be observed in image appearance among the ROIs on the far wall of the carotid artery. To overcome this challenge, we utilize the location of the carotid bulb as a contextual constraint. We choose this constraint for two reasons: 1) the carotid bulb appears as a distinct dark area in the ultrasound frame and thus can be uniquely identified; 2) according to the consensus statement of American society of Electrophysiology for cardiovascular risk assessment, the ROI should be placed approximately 1 cm from the carotid bulb on the far wall of the common carotid artery. While the former motivates the use of the carotid bulb location as a constraint from a technical point of view, the latter justifies this constraint from a clinical standpoint.

Training Phase: We incorporate this constraint in the suggested system by training a 3-way CNN that simultaneously localizes both ROI and carotid bulb, and then refines the estimated location of the ROI given the location of the carotid bulb. Figure 9 in the supplementary material illustrates how the image patches are extracted from a training frame. We perform data augmentation by extracting the training patches within a circle around the locations of the carotid bulbs and the ROIs. The negative patches are extracted from a grid of points sufficiently far from the locations of the carotid bulbs and the ROIs. Note that the above translation-based data augmentation is sufficient for this application, because our database provides a relatively large number of training EUFs, from which a large set of training patches can be collected. Once the patches are collected, we form

a stratified training set with approximately 410,000 patches to train a 3-way CNN for constrained ROI localization.

Testing Phase: Referring to Figure 4, during the test stage, the trained CNN is applied to all the pixels in the EUF, generating two confidence maps with the same size as the EUF. The first confidence map shows the probability of a pixel being the carotid bulb and the second confidence map shows the probability of a pixel being the ROI. One way to localize the ROI is to find the center of the largest connected component within the ROI confidence map without considering the detected location of the carotid bulb. However, this naive approach may fail to accurately localize the ROI. For instance, a long-tale connected component along the far wall of the carotid artery may cause substantial ROI localization error. To compound the problem, the largest connected component of the ROI confidence map may appear far from the actual location of the ROI, resulting in a complete detection failure. To overcome these limitations, we constraint the ROI location l_{roi} by the location of the carotid bulb l_{cb} . For this purpose, we first determine the location of the carotid bulb as the centroid of the largest connected component within the first confidence map, and then localize the ROI using the following formula

$$l_{roi} = \frac{\sum_{p \in C^*} M(p) \cdot p \cdot I(p)}{\sum_{p \in C^*} M(p) \cdot I(p)} \quad (1)$$

where l_{roi} denotes the ROI location, l_{cb} denotes the center of the carotid bulb, M denotes the confidence map of being the ROI, C^* is the largest connected component in M that is the nearest to the carotid bulb, and $I(p)$ is an indicator function for pixel $p = [p_x, p_y]$ that is defined as

$$I(p) = \begin{cases} 1, & \text{if } \|p - l_{cb}\| < 1 \text{ cm} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$(3)$$

The indicator function $I(p)$ is binary function that simply includes pixel when the value is 1 as in Eq. 2, otherwise excludes pixel when the value is 0 as in Eq. 3.

3.3. Intima-Media Thickness Measurement

Measuring intima-media thickness require a continuous and one-pixel precise boundary for lumen-intima and media-adventitia. Lumen-intima is relatively easier to detect because of strong gradient change at the border, however, detecting media-adventitia interface is quite challenging due to its subtle image gradients and noise around its border. We approach this problem as a 3-way classification task: 1) lumen-intima interface, 2) media-adventitia interface, and 3) background.

Training Phase: To train 3-way CNN, we collected sparse background patches and then pixel-by-pixel image patches around lumen-intima interface and media-adventitia interface with additional patches ± 3 pixels from the ground

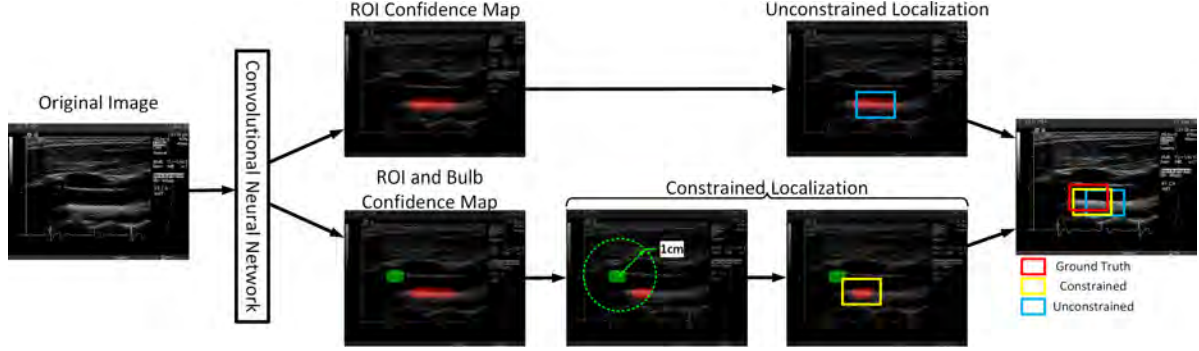


Figure 4: The test stage of our ROI localization method. In the unconstrained scenario, we only use the ROI confidence map, which results in relatively large localization error. In the constrained mode, given the estimated location of the carotid bulb, we localize the ROI more accurately.

truth. Using ± 3 pixels for additional patches around intima-media boundary was necessary to balance number of patches with background patches and produced better results. Figure 10 in the supplementary material illustrates how the training patches are collected from an ROI.

Testing Phase: Figure 5 illustrates the testing process. The 3-way trained CNN is applied in a sliding-window fashion for a given test ROI and generates two confidence maps (Figure 5(b)) with the same size as the ROI. Since confidence map is thicker than a pixel, we choose the maximum response column-by-column and generate a new binary image as shown in Figure 5(c). Finally, we use two active contour models (a.k.a, snakes) [12] for segmenting lumen-intima and media-adventitia interfaces. Figure 5(d) shows two final converged snakes and we take measurements as the average vertical distance between the two snakes.

4. Experiments

We use a database of 92 CIMT videos captured from 23 subjects with 2 CIMT videos from the left and 2 CIMT videos from the right carotid artery of each subject. The ground truth for each video contains the EUF number, the locations of ROI, and the segmentation of lumen-intima and media-adventitia interfaces. For consistency, we use the same training set and the same test set (no overlap with training) for all three tasks. Our training set contains 48 CIMT videos of 12 subjects with a total of 4,456 frames and our test set contains 44 CIMT videos of 11 subjects with a total of 3,565 frames. For each task, we perform leave-one-patient-out cross-validation based on the *training subjects* to tune the parameters, and then evaluate the performance of the tuned system using the test subjects.

Architecture: As shown in Table 1, we employ a CNN architecture with 2 convolutional layers, 2 subsampling layers, and 2 fully connected layers (see Section 5 for our justifications). We also append a softmax layer to the last

fully connected layer so as to generate probabilistic confidence values for each class. Our CNN architecture has input patches of size 32×32 , and we resize the collected patches to 32×32 prior to the training process. For the CNNs used in our experiments, we employ a learning rate of $\alpha = 0.001$, a momentum of $\mu = 0.9$, and a constant scheduling rate of $\gamma = 0.95$.

Pre- and post-processing for frame selection: We have experimentally found out that binarized image patches improve the quality of convergence and accuracy of frame selection. Furthermore, we have observed that the standard deviation of the Gaussian function used for smoothing the probability signals, can also substantially influence frame selection accuracy. Therefore, we have conducted leave-one-patient-out cross-validation based on the training subjects to find the best binarization method and the optimal standard deviation of the Gaussian function. For binarization, we have considered a fixed set of thresholds and adaptive thresholding using Otsu’s method. For smoothing, we have considered a Gaussian function with different standard deviation (σ_g) as well as the scenario where no smoothing is applied. For each configuration of parameters, we have done a free-response ROC (FROC) analysis. We consider a selected frame a true positive, if it is found within one frame from the expert-annotated EUF; otherwise, a false positive.

Our leave-one-patient-out cross-validation study, summarized in Figure 11 in the supplementary material, indicates that the use of a Gaussian function with $\sigma_g = 1.5$ for smoothing the probability signals and adaptive thresholding using Otsu’s method achieve the highest performance. Figure 6 shows the FROC curve of our system for the test subjects using the above parameters. For comparison, we have also shown the operating point of the hand-crafted approach [26], which is significantly outperformed by the suggested system.

Constrained ROI Localization: We conduct a leave-one-



Figure 5: The test stage of lumen-intima and media-adventitia interface detection. (a) a test ROI. (b) The trained CNN generates a confidence map where the green and red colors indicate the likelihood of lumen-intima interface and media-adventitia interface, respectively. (c) The thick probability band around each interface is thinned by selecting the largest probability for each interface in each column. (d) The step-like boundaries are refined through two open snakes.

Table 1: The CNN architecture used in our experiments. Note that C is the number of classes, which is 2 for frame selection and 3 for both ROI localization and intima-media thickness measurements.

layer	type	input	kernel	stride	pad	output
0	input	32x32	N/A	N/A	N/A	32x32
1	convolution	32x32	5x5	1	0	64x28x28
2	max pooling	64x28x28	3x3	2	0	64x14x14
1	convolution	64x14x14	5x5	1	0	64x10x10
2	max pooling	64x10x10	3x3	2	0	64x5x5
2	fully connected	64x5x5	5x5	1	0	250x1
2	fully connected	250x1	1x1	1	0	Cx1

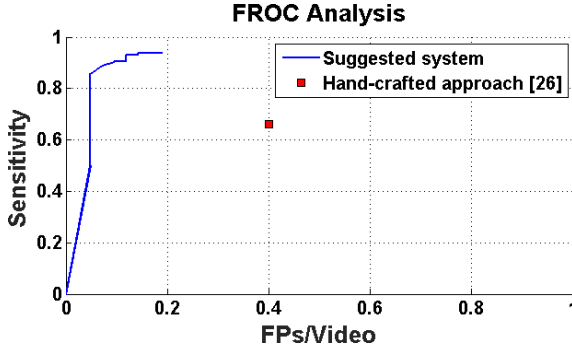


Figure 6: FROC curve of our frame selection system for the test subjects using the tuned parameters. For comparison, we have also shown the operating point of the prior hand-crafted approach [26], which is significantly outperformed by the suggested system.

patient-out cross-validation study based on the training subjects to find the optimal size of the training patches. Our cross-validation analysis, summarized in Figure 12 in the supplementary material, indicates that the use of 1.8×1.8 cm patches achieves the most stable performance, yielding low ROI localization error with only a few outliers. Fig-

ure 7 shows the ROI localization error of our system for the test subjects using the optimal size of training patches. To demonstrate the effectiveness of our constrained ROI localization method, we have also included the performance of the unconstrained counterpart. In the constrained mode, we use Eq. 1 for ROI localization whereas in the unconstrained mode we localize the ROI as the center of the largest connected component in the corresponding confidence map without considering the location of the carotid bulb. Our method achieves an average localization error of 0.19 mm and 0.35 mm in the constrained and unconstrained modes, respectively. The decrease in localization error is statistically significant ($p < 0.01$). Also as seen in Figure 7, our method in the unconstrained mode has resulted in 3 complete localization failures (outliers), which have been corrected in the constrained mode. Furthermore, compared with the hand-crafted approach [26], our system in the constrained mode shows a decrease of 0.1 mm in ROI localization error, which is statistically significant ($p < .00001$).

Intima-Media Thickness Measurement: We determined the optimal image patch size by leave-one-patient-out cross-validation using various image patch sizes and found that $360 \times 360 \mu\text{m}$ achieved slightly lower localization error and fewer outliers (see Figs. 12-13 in the supplementary ma-

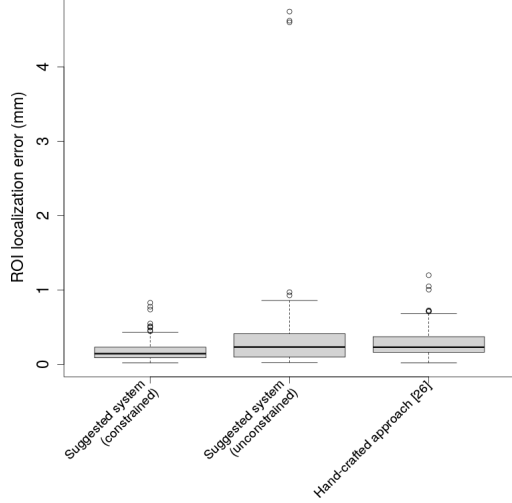


Figure 7: ROI localization error for the test subjects. Our method in the constrained mode outperforms both the unconstrained counterpart and the prior hand-crafted approach [26].

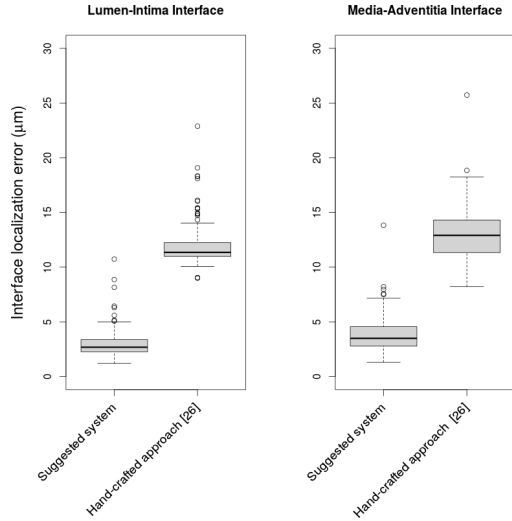


Figure 8: Localization error of the lumen-intima and media-adventitia interfaces for the suggested system and the prior hand-crafted approach [26]. The results are obtained for the test subjects.

material). Figure 8 shows the interface localization error of our system on the test subjects, where we break down the overall localization error for lumen-intima and that of the media-adventitia interface as well as the hand-crafted approach [26] for each interface. We further analyzed agreement between our system and the expert with the Bland-Altman plot (see Figure 14 in the supplementary material).

5. Discussions

In Section 4, we investigated how the choice of patch binarization and degree of Gaussian smoothing affect the accuracy of frame selection. Here, we would like to discuss our findings and provide insights about our choices. We choose to binarize the patches, because it reduces appearance variability and suppress the low-magnitude noise content in the patches. Without patch binarization, one can expect a large amount of variability in the appearance of wavelets can deteriorate the performance of the subsequent CNN (see Figure 11 in the supplementary material). The choice of binarization threshold is another important factor. The use of a high threshold results in the partial appearance of the wavelets in the resulting binary patches, reducing the discriminatory appearance features of the patches. A low threshold, on the other hand, can intensify noise content in the images, which decreases the quality of training samples and consequently a drop in classification performance. According to our analyses, it is difficult to find a fixed threshold that can both suppress the noise content and keep the shapes of the restored wavelets intact in all the collected patches. Otsu’s method seems to overcome this limitation by adaptively selecting a binarization threshold according to the intensity distribution of each individual patch. For patches with intensity values between 0 and 1, the adaptive thresholds have a mean of 0.15 and standard deviation of 0.05. The wide range of adaptive thresholds explains why a constant threshold may not perform as desirably.

Gaussian smoothing of the probability signals is also essential for accurate frame selection. This is because the probability signals prior to smoothing exhibit high frequency fluctuations, which may complicate the localization of the local maxima in the signals. The first cause of such high frequency changes is patch misplacement in the accumulated difference images. Recall that we extract the patches around the weighted centroids of the accumulated difference images. However, a large amount of noise content in the difference images may cause the weighted centroid to deviate from the center of the restored wavelet. In this case, the extracted patch may partially or completely miss the restored wavelet. This can manifest itself as a sudden change in the CNN output and as a result in the corresponding probability signal. The second cause of high frequency changes is the inherited high variance of CNNs. Use of ensemble of CNNs and data augmentation can alleviate this problem at a significant computation cost. Alternatively, we choose to mitigate these undesirable fluctuations using Gaussian smoothing for computational efficiency.

As described in Section 3.2, we constrain our ROI localization method by the location of the carotid bulb. This is because the bulb area appears as a relatively distinct dark area in the ultrasound frame. The distinct appearance of the carotid bulb is also confirmed by our experiments, where

we obtain the average bulb localization error of 0.26 mm for the test subjects with only one failure case, which is more favorable than the average unconstrained ROI localization error of 0.38 mm with 3 failure cases. Therefore, the localization of the bulb area can be done more reliably than the localization of the ROI, which motivates the use of the bulb location as a guide for more accurate ROI localization. We integrate this constraint into our localization system through a post-processing mechanism (see Eq. 1). Alternatively, we could train a regression CNN where each pixel in the image directly votes for the location of the ROI. However, this approach may be hindered by lack of stable anatomical structures in noisy ultrasound images. We will explore a regression CNN for ROI localization as future work.

In Section 4, we showed a high level of agreement between our system and the expert for the assessment of intima-media thickness. The suggested system achieves a mean absolute error of $2.8\mu\text{m}$ with a standard deviation of $2.1\mu\text{m}$ for intima-media thickness measurements. However, this level of measurement error cannot hurt the interpretation of the vascular age, because there exists a minimum difference of $400\mu\text{m}$ between the average intima-media thickness of healthy and high-risk population ($600\mu\text{m}$ for healthy and $\geq 1000\mu\text{m}$ for high-risk population) [8]. To further put the performance of our system into perspective, in Table 2, we have compared the accuracy of intima-media thickness measurements produced by our system with those of the other automatic methods recently suggested in the literature. As seen, our method yields a lower level of mean absolute error and smaller standard deviation.

We used a LeNet-like CNN architecture in our study, but it does not limit the suggested framework to this architecture. In fact, we have experimented with deeper CNN architectures such as AlexNet [10] in both training and fine-tuning modes; however, we did not observe any significant performance gain. This was probably because the higher level semantic features detected by the deeper networks are not very relevant to the tasks in our CIMT applications. Meanwhile, the concomitant computational cost of deep architectures may hinder the applicability of our system, because it lowers the speed—a key usability factor of our system. We also do not envision that a shallower architecture can offer the performance required for clinical practice. This is because a network shallower than the LeNet has only one convolutional layer and thus limited to learning primitive edge like features. Detecting the carotid bulb and the ROI, and segmenting intima-media boundaries are relatively challenging tasks, requiring more than primitive edge-like features. Similarly, for frame selection, classifying the restored wavelets into R-peak and non-R-peak categories is similar to digit recognition, for which LeNet is a common choice of architecture. Therefore, LeNet-like CNN architecture seems to represent an optimal balance be-

Table 2: CIMT error for our system and the other state-of-the-art methods.

Author	Year	Thickness error (μm)
Current work	—	2.8 ± 2.1
Bastida-Jumillacite [1]	2015	13.8 ± 31.9
Ilea [7]	2013	80 ± 40
Loizou [14]	2013	30 ± 30
Molinari [20]	2011	43 ± 93

tween efficiency and accuracy for CIMT video analysis.

We should note that throughout this paper, all performance evaluations were performed without involving any user interactions. However, our goal is not to exclude the user (sonographer) from the loop rather to relieve him from the three tedious, laborious, and time consuming operations by automating them while still offering the user a highly, user-friendly interface to bring his indispensable expertise onto CIMT interpretation through refining the automatic results easily at the end of each of the automated operations. For instance, our system is expected to automatically locate a EUF within one frame, which is clinically acceptable, but in case the automatic selected EUF is not the exact one as desired, the user can simply press an arrow key to move one frame forward or backward. From our experience, the automatically localized ROI is acceptable even if there is a small distance from the ground truth location, but the user still can easily drag the ROI and move it around as desired. Finally, in refining the automatically identified lumen-intima and media-adventitia interfaces, the original snake formulation comes with spring forces for user interaction [9], but given the small distance between the lumen-intima and media-adventitia interfaces, we have found that “movable” hard constraints as proposed in [12] are far more effective than the spring forces in measuring CIMT.

6. Conclusion

In this paper, we presented a unified framework to fully automate and accelerate CIMT video interpretation. Specifically, we suggested a computer-aided CIMT measurement system with three components: (1) automatic frame selection in CIMT videos, (2) automatic ROI localization within the selected frames, (3) automatic intima-media boundary segmentation within the localized ROIs. We based each of the above components on a CNN with a LeNet-like architecture and then boosted the performance of the employed CNNs with effective pre- and post-processing techniques. For frame selection, we demonstrated that how patch binarization as a pre-processing step and smoothing the probability signals as a post-processing step improve the results generated by the CNN. For ROI localization, we experimentally proved that the location of the carotid bulb, as a constraint in a post-processing setting, significantly improves

ROI localization accuracy. For intima-media boundary segmentation, we employed open snakes as a post processing step to further improve the segmentation accuracy. We compared the results produced by the suggested system with those of the major prior works, demonstrating more accurate frame selection, ROI localization, and CIMT measurements. This superior performance is attributed to the effective use of CNNs coupled with pre- and post- processing steps, uniquely designed for the three CIMT tasks.

References

- [1] M. Bastida-Jumilla, R. Menchón-Lara, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz, and J. Sancho-Gómez. Frequency-domain active contours solution to evaluate intima-media thickness of the common carotid artery. *Biomedical Signal Processing and Control*, 16:68–79, 2015.
- [2] M. C. Bastida-Jumilla, R. M. Menchón-Lara, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz, and J. L. Sancho-Gómez. Segmentation of the common carotid artery walls based on a frequency implementation of active contours. *Journal of digital imaging*, 26(1):129–139, 2013.
- [3] D.-C. Cheng and X. Jiang. Detections of arterial wall in sonographic artery images using dual dynamic programming. *Information Technology in Biomedicine, IEEE Transactions on*, 12(6):792–799, 2008.
- [4] S. Delsanto, F. Molinari, P. Giustetto, W. Liboni, S. Badalamenti, and J. S. Suri. Characterization of a completely user-independent algorithm for carotid artery segmentation in 2-d ultrasound images. *Instrumentation and Measurement, IEEE Transactions on*, 56(4):1265–1274, 2007.
- [5] F. Faita, V. Gemignani, E. Bianchini, C. Giannarelli, L. Ghiadoni, and M. Demi. Real-time measurement system for evaluation of the carotid intima-media thickness with a robust edge operator. *Journal of Ultrasound in Medicine*, 27(9):1353–1361, 2008.
- [6] R. T. Hurst, R. F. Burke, E. Wissner, A. Roberts, C. B. Kendall, S. J. Lester, V. Somers, M. E. Goldman, Q. Wu, and B. Khandheria. Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players. *The American journal of cardiology*, 105(8):1107–1111, 2010.
- [7] D. E. Ilea, C. Duffy, L. Kavanagh, A. Stanton, and P. F. Whelan. Fully automated segmentation and tracking of the intima media thickness in ultrasound video sequences of the common carotid artery. *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on*, 60(1), 2013.
- [8] D. Jacoby, I. Mohler, EmileR., and D. Rader. Noninvasive atherosclerosis imaging for predicting cardiovascular events and assessing therapeutic interventions. *Current Atherosclerosis Reports*, 6(1):20–26, 2004.
- [9] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [11] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. Citeseer, 1990.
- [12] J. Liang, T. McInerney, and D. Terzopoulos. United snakes. *Medical image analysis*, 10(2):215–233, 2006.
- [13] Q. Liang, I. Wendelhag, J. Wikstrand, and T. Gustavsson. A multiscale dynamic programming procedure for boundary detection in ultrasonic artery images. *Medical Imaging, IEEE Transactions on*, 19(2):127–142, 2000.
- [14] C. Loizou, T. Kasparis, C. Spyrou, and M. Pantziaris. Integrated system for the complete segmentation of the common carotid artery bifurcation in ultrasound images. In H. Papadopoulos, A. Andreou, L. Iliadis, and I. Maglogiannis, editors, *Artificial Intelligence Applications and Innovations*, volume 412 of *IFIP Advances in Information and Communication Technology*, pages 292–301. Springer Berlin Heidelberg, 2013.
- [15] C. P. Loizou. A review of ultrasound common carotid artery image and video segmentation techniques. *Medical & biological engineering & computing*, 52(12):1073–1093, 2014.
- [16] C. P. Loizou, C. S. Pattichis, M. Pantziaris, and A. Nicolaides. An integrated system for the segmentation of atherosclerotic carotid plaque. *Information Technology in Biomedicine, IEEE Transactions on*, 11(6):661–667, 2007.
- [17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2014.
- [18] R.-M. Menchón-Lara, M.-C. Bastida-Jumilla, A. González-López, and J. L. Sancho-Gómez. Automatic evaluation of carotid intima-media thickness in ultrasounds using machine learning. In *Natural and Artificial Computation in Engineering and Medical Applications*, pages 241–249. Springer, 2013.
- [19] R.-M. Menchón-Lara and J.-L. Sancho-Gómez. Fully automatic segmentation of ultrasound common carotid artery images based on machine learning. *Neurocomputing*, 151:161–167, 2015.
- [20] F. Molinari, K. Meiburger, G. Zeng, A. Nicolaides, and J. Suri. Caudles-ef: Carotid automated ultrasound double line extraction system using edge flow. *Journal of Digital Imaging*, 24(6):1059–1077, 2011.
- [21] F. Molinari, K. M. Meiburger, L. Saba, G. Zeng, U. R. Acharya, M. Ledda, A. Nicolaides, and J. S. Suri. Fully automated dual-snake formulation for carotid intima-media thickness measurement a new approach. *Journal of Ultrasound in Medicine*, 31(7):1123–1136, 2012.
- [22] F. Molinari, G. Zeng, and J. S. Suri. A state of the art review on intima-media thickness (imt) measurement and wall segmentation techniques for carotid ultrasound. *Computer methods and programs in biomedicine*, 100(3):201–221, 2010.
- [23] S. Petroudi, C. Loizou, M. Pantziaris, and C. Pattichis. Segmentation of the common carotid intima-media complex in ultrasound images using active contours. *Biomedical Engineering, IEEE Transactions on*, 59(11):3060–3069, 2012.

- [24] P. Pignoli and T. Longo. Evaluation of atherosclerosis with b-mode ultrasound imaging. *The Journal of nuclear medicine and allied sciences*, 32(3):166–173, 1987.
- [25] A. C. Rossi, P. J. Brands, and A. P. Hoeks. Automatic localization of intimal and adventitial carotid artery layers with noninvasive ultrasound: a novel algorithm providing scan quality control. *Ultrasound in medicine & biology*, 36(3):467–479, 2010.
- [26] H. Sharma, R. G. Golla, Y. Zhang, C. B. Kendall, R. T. Hurst, N. Tajbakhsh, and J. Liang. Ecg-based frame selection and curvature-based roi detection for measuring carotid intima-media thickness. In *SPIE Medical Imaging*, pages 904016–904016. International Society for Optics and Photonics, 2014.
- [27] P.-J. Touboul, P. Prati, P.-Y. Scarabin, V. Adrai, E. Thibout, and P. Ducimetière. Use of monitoring software to improve the measurement of carotid wall thickness by b-mode imaging. *Journal of hypertension*, 10:S37–S42, 1992.
- [28] X. Xu, Y. Zhou, X. Cheng, E. Song, and G. Li. Ultrasound intima-media segmentation using hough transform and dual snake model. *Computerized Medical Imaging and Graphics*, 36(3):248–258, 2012.
- [29] X. Zhu, C. B. Kendall, R. T. Hurst, and J. Liang. A user friendly system for ultrasound carotid intima-media thickness image interpretation. In *SPIE Medical Imaging*, pages 79681G–79681G. International Society for Optics and Photonics, 2011.

Supplementary material

Convolutional Neural Networks

As with multi-layer pceptrons, convolutional neural networks are trained using the back-propagation algorithm. If D denotes a set of training images, W denotes a matrix containing the weights of the convolutional layers, and $f_W(D^{(i)})$ denotes the loss for the i^{th} training image, the loss over the entire training set is then computed as

$$\mathcal{L}(W) = \frac{1}{|D|} \sum_i^{|D|} f_W(X^{(i)}) \quad (4)$$

Gradient descent is commonly used for minimizing the above loss function with respect to the unknown weights W . However, the modern massively parallelized implementations of CNNs are limited by the amount of memory on GPUs; therefore, one cannot evaluate the loss function based on the entire training set D at once. Instead, the loss function is approximated with the loss over the mini-batches of training images of size $N \ll |D|$. A common choice of the mini-batch size is 128, which is a reasonable trade-off between loss noise suppression and memory management. Given the size of mini-batches, one can approximate the loss function as $\mathcal{L}(W) \approx \frac{1}{N} \sum_{i=1}^N f_W(X^{(i)})$, and iteratively update the weights of the network with the following equations:

$$\begin{aligned} \gamma_t &= \gamma \lfloor \frac{tN}{|D|} \rfloor \\ V_{t+1} &= \mu V_t - \gamma_t \alpha \Delta L(W_t) \\ W_{t+1} &= W_t + V_{t+1} \end{aligned} \quad (5)$$

where α is the learning rate, μ is the momentum that indicates the contribution of the previous weight update in the current iteration, and γ is the scheduling rate that decreases learning rate α at the end of each epoch.

Figures

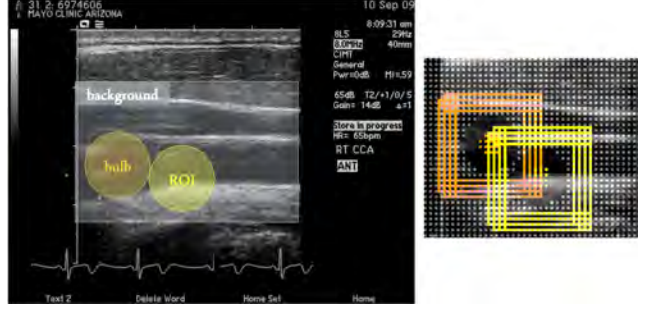


Figure 9: For constrained ROI localization, we use a 3-way CNN whose training image patches are extracted from a grid of points on the background and around the ROI and the carotid bulb locations.

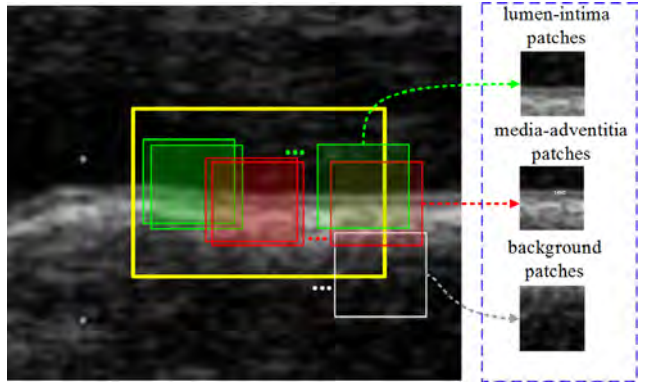


Figure 10: For lumen-intima and media-adventitia interface segmentation, we use a 3-way CNN whose training image patches are extracted from the background and around the lumen-intima and media-adventitia interfaces.

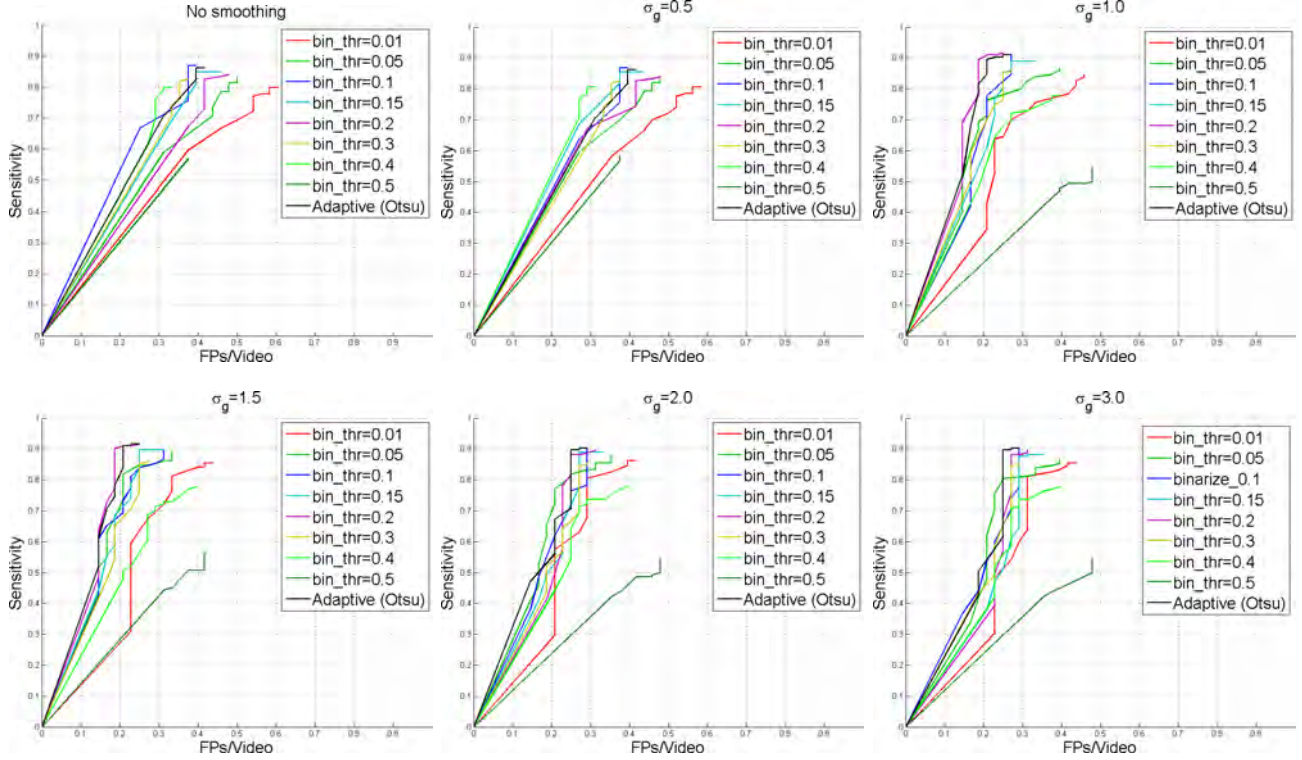


Figure 11: FROC curves of our system for automatic frame selection. Each plot shows FROC curves for different binarization thresholds and different levels of Gaussian smoothing. The results are obtained using leave-one-patient-out cross-validation based on the training subjects. As seen, no smoothing or a small degree of Gaussian smoothing leads to relatively low frame selection accuracy. This is because a trivial level of smoothing may not properly handle the fluctuations in the probability signals, causing a large number of false positives around an EUF. On the other hand, a large degree of smoothing may decrease the sensitivity of frame selection as the locations of the local maxima may be found more than one frame away from the expert-annotated EUFs. We therefore use a Gaussian function with $\sigma_g = 1.5$ for smoothing the probability signals. Our results also indicate that the adaptive thresholding method and a fixed threshold of 0.2 achieve the highest frame selection accuracy. However, we choose to use adaptive thresholding because it decreases the parameters of our system by one and that it performs more consistently at different levels of Gaussian smoothing.

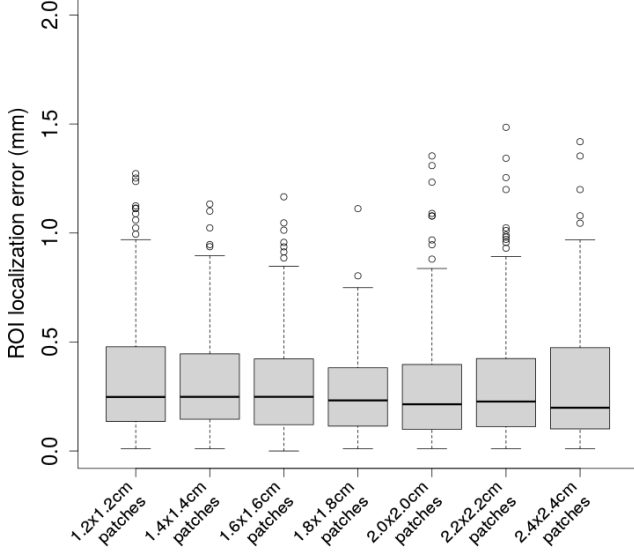


Figure 12: ROI localization error of our system for different sizes of patches. The results are obtained using leave-one-patient-out cross-validation based on the training subjects. In our analyses, we measure the localization error as the Euclidean distance between the estimated ROI location and the one provided by the expert. As can be seen, the use of 1.8×1.8 cm patches achieves the most stable performance, yielding low ROI localization error with only a few outliers.

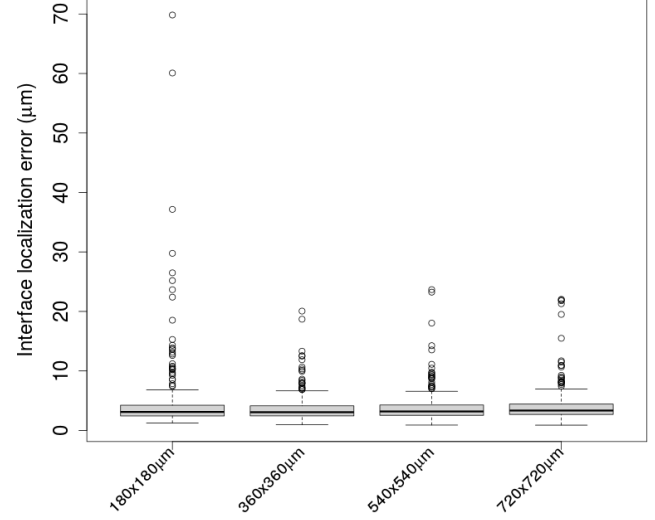


Figure 13: Combined interface localization error for different sizes of patches. The results are produced through a leave-one-patient-out cross-validation study based on the training subjects. Each box plots show the combined localization error of lumen-intima and media-adventitia interfaces for a different size of patches. In our analyses, we determine the localization error as the average of absolute vertical distances between our detected boundaries and the expert-annotated boundaries for the interfaces. As can be seen, while our system shows a high degree of robustness against different sizes of input patches, the use of patches of size $360 \times 360 \mu\text{m}$ achieves slightly lower localization error and fewer outliers. Furthermore, this choice of patches yields higher computational efficiency compared to the larger counterpart patches.

Agreement Analysis

We further analyze agreement between our system and the expert for the assessment of intima-media thickness. To this end, we use the Bland-Altman plot, which is a well-established technique to measure agreement between different observers. We have shown the Bland-Altman plot for the test subjects in Figure 14, where each circle represents a pair of thickness measurements, one from our method and one from the expert. As seen, the majority of circles fall within 2 standard deviations from the mean error, which suggests a large agreement between the automatically computed thickness measurements and those of the expert. Furthermore, Pearson product-moment correlation coefficient for the average and difference measurements is -0.097, indicating that the agreement between our method and the expert does not depend on intima-media thickness.

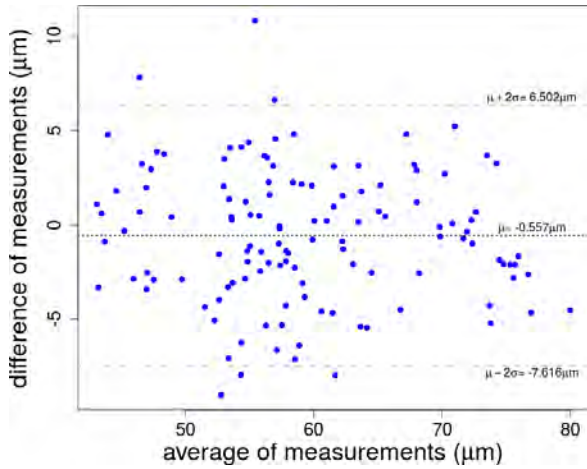


Figure 14: The Bland-Altman plot shows high agreement between our system and the expert for the assessment of intima-media thickness. Each circle in this plot represents a pair of thickness measurements from our method and the expert for a test ROI. In this plot, we have a total of 126 circles corresponding to 44 test videos.

Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information

Nima Tajbakhsh*, *Member, IEEE*, Suryakanth R. Gurudu, and Jianming Liang, *Senior Member, IEEE*

Abstract—This paper presents the culmination of our research in designing a system for computer-aided detection (CAD) of polyps in colonoscopy videos. Our system is based on a hybrid context-shape approach, which utilizes context information to remove non-polyp structures and shape information to reliably localize polyps. Specifically, given a colonoscopy image, we first obtain a crude edge map. Second, we remove non-polyp edges from the edge map using our unique feature extraction and edge classification scheme. Third, we localize polyp candidates with probabilistic confidence scores in the refined edge maps using our novel voting scheme. The suggested CAD system has been tested using two public polyp databases, CVC-ColonDB, containing 300 colonoscopy images with a total of 300 polyp instances from 15 unique polyps, and ASU-Mayo database, which is our collection of colonoscopy videos containing 19,400 frames and a total of 5,200 polyp instances from 10 unique polyps. We have evaluated our system using free-response receiver operating characteristic (FROC) analysis. At 0.1 false positives per frame, our system achieves a sensitivity of 88.0% for CVC-ColonDB and a sensitivity of 48% for the ASU-Mayo database. In addition, we have evaluated our system using a new detection latency analysis where latency is defined as the time from the first appearance of a polyp in the colonoscopy video to the time of its first detection by our system. At 0.05 false positives per frame, our system yields a polyp detection latency of 0.3 seconds.

Index Terms—Optical colonoscopy, polyp detection, boundary classification, edge voting, detection latency.

I. INTRODUCTION

C OLORECTAL cancer is the second-highest cause of cancer-related deaths in the United States with approximately 50,000 deaths in 2015 [1]. However, colon cancer is preventable using effective screening tests. Colonoscopy is the preferred technique for colon cancer screening and prevention. The goal of colonoscopy is to remove colonic polyps before they develop into colon cancer. Colonoscopy has been a successful preventative procedure and has contributed to a 30% decline in the incidence of colorectal cancer [2]. However, colonoscopy is an operator dependent procedure. Human

factors, such as lack of sensitivity to visual characteristics of polyps, fatigue, and insufficient attentiveness during colon examination, can lead to the miss-detection of polyps. Polyp miss-rates are estimated around 4–12% [3]–[6]; however, a more recent clinical study [7] projects this as 25%. Missed polyps can lead to the late diagnosis of colon cancer with the survival rate of less than 10% [8]. Computer-aided polyp detection may help colonoscopists reduce their polyp miss-rates.

We proposed our initial system in [9] where Haar features and a mixture of random forest classifiers were used to obtain a refined edge map and then a new voting scheme was applied to localize polyp candidates within the refined edge maps. We improved our system in [10] by replacing Haar features with a new patch descriptor and replacing a mixture of random forest classifiers with a 2-stage edge classification scheme. We further improved our system in [11] by introducing the notion of narrow bands and isocontours to assign a probabilistic score to each polyp candidate. Our current work presents an improved presentation and rigorous evaluation of our system that was suggested in [11]. Specifically, we have included new pseudocodes and illustrations to improve the presentation of the system, employed a significantly larger database of polyps to strengthen our evaluations, included new sensitivity analyses for parameters of our system, performed more detailed performance comparisons, and introduced a new performance curve that overcomes the limitation of the free-response receiver operating characteristic (FROC) curves.

This paper represents the culmination of our research in this area, it is self-contained, and summarizes our key contributions from this research as follows:

- *An efficient yet powerful patch descriptor:* We present a new feature extraction method in Section III.B that is designed to operate at high speed and low computational complexity. Our descriptor is both rotation invariant and robust against linear illumination changes.
- *An effective edge classification scheme:* We suggest a 2-stage edge classification framework in Section III.C that is able to enhance low level image features prior to classification. Our scheme fuses the information extracted from a pair of patches to not only detect the desired edges but to determine, on which side of the detected edges, the desired structures reside.
- *A robust voting scheme:* We propose a voting scheme in Section III.D that is designed to robustly detect polyps as objects with curvy boundaries in the fragmented edge maps. Unlike the existing alternatives, our scheme is not limited to detecting objects with a specific parametric model.

Manuscript received May 07, 2015; revised July 26, 2015; accepted August 30, 2015. Date of publication October 08, 2015; date of current version February 01, 2016. This work was supported by the seed grant awarded by Arizona State University and Mayo Clinic. Asterisk indicates corresponding author.

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

*N. Tajbakhsh is with the Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA (e-mail: nima.tajbakhsh@asu.edu).

S. R. Gurudu is with the Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, AZ 85259 USA (e-mail: gurudu.suryakanth@mayo.edu).

J. Liang is with the Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259 USA (e-mail: jianming.liang@asu.edu).

Digital Object Identifier 10.1109/TMI.2015.2487997

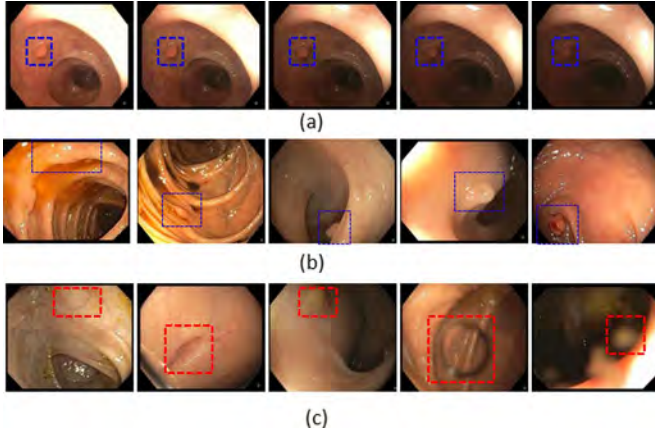


Fig. 1. (a) Significant color variations between different instances of the same polyp due to varying lighting conditions. (b) Significant color variations between different polyps. (c) Colonoscopy videos contain structures or artifacts with similar geometric characteristics to polyps. These structures can mislead a polyp detector that relies on context-free shape features.

- *A flexible probability assignment technique:* We present a probability assignment technique in Section III.E that produces a probabilistic output for each polyp candidate. Our method does not require any predefined parametric model of shapes or any information about the size of polyps.

II. RELATED WORKS

The early work of Karkanis *et al.* [12] was published in 2003 where color wavelet features coupled with a sliding window was used for detecting polyps in colonoscopy images. This article inspired more research [13]–[16] on polyp detection using texture and color descriptors. However, these methods are limited by (1) partial texture visibility caused by the relatively large distance between the polyps and the single-focus camera, and (2) large color variations among polyps. Fig. 1(a) shows how lighting conditions can cause color variations among different instances of the same polyp. As seen, the same polyp appears in different shades, ranging from dark to saturated colors. Fig. 1(b) shows color variations among different polyps. The reliability of color and texture based methods was also questioned in [17].

The other category of techniques for polyp detection employed shape and appearance features. Hwang *et al.* [18] suggested elliptical shape features for detecting the shots of polyps in colonoscopy videos. However, they did not consider image context when extracting the shape clues, leaving the door open for false positive detections around other elliptical structures in the complex endoluminal scenes. Fig. 1(c) shows examples of non-polyp structures that can mislead a polyp detector relying on context-free shape features. Our previous works [9]–[11] aimed to address this drawback by eliminating such misleading structures from the images using contextual features. Bernal *et al.* [17] employed valley information and a region growing approach to find polyps in colonoscopy images. Bernal *et al.* further improved their method in [19] by reducing the number of false positives around vascular structures and specular reflections, and presented further evaluations in [20].

The more recent systems have considered spatio-temporal features, boundary features, and imbalanced learning.

Park *et al.* [21] suggested the use of spatio-temporal features for polyp detection. Their method would require information from the past and future frames for polyp localization at the current frame, generating delayed feedback on the locations of polyps. Tajbakhsh *et al.* [22], [23] employed convolutional neural networks for learning discriminative spatial and temporal features. However, unlike [21], their method was not reliant on the future frames, avoiding the delayed feedback on the locations of polyps. Wang *et al.* [24] used edge cross-sectional profiles for detecting protruding polyps. Their method was designed to capture shape, texture, protrusion, and smoothness of the polyp surface. Finally, Bae and Yoon [25] proposed a polyp detection system based on imbalanced learning and discriminative feature learning.

Polyp detection and classification have also been considered in CT colonography [26]–[32], wireless capsule endoscopy [33]–[38], and narrow band imaging [39]–[41]. However, the challenges posed by these imaging modalities differ from that of colonoscopy. To design a polyp detection system for colonoscopy, one needs to consider the effects of varying lighting conditions, specular reflections, spontaneous colon deformation, and diverse view angles of the camera. However, such challenges do not or partially apply to the other imaging modalities. For instance, while texture is not reliable for polyp detection in colonoscopy, the pit patterns of polyps are heavily used in narrow-band imaging; or while shape and curvature clues have been successfully used for polyp detection in CT colonography, they are misleading in the complex colonoscopy images if not combined with the context clues.

III. PROPOSED METHOD

We propose a hybrid context-shape approach for polyp detection, because a pure shape-based approach may mislead a polyp detector towards other polyp-like structures such as fecal content and reflection spots, and a pure context-based approach may not capture the discriminative geometric information of polyps.

We use distinct image appearance around polyp boundaries as context clues. We have illustrated the distinct boundary appearance around polyps in Fig. 2(a) by comparing average appearance of polyp boundaries with that of vessels, lumen areas, and specular reflections. To obtain average image appearance, we collect oriented patches along the boundaries of vessels, specular spots, lumen areas, and polyps (see Fig. 2(b)) and then average the resulting patches for each structure of interest. We use context information when designing our patch descriptor and edge classification schemes.

We use curvature of polyp boundaries as shape clues. As seen in Fig. 2(c), although polyps appear in different shapes, they most often have a curvy segment in their boundaries. We have highlighted these curvy segments with the blue rectangles. We utilize shape clues through our voting scheme, which is designed to localize polyps as objects with curvy heads.

As shown in Fig. 3, our polyp detection system consists of four stages: (1) constructing an edge map for an input image, (2) refining the edge map by classifying every edge pixel into polyp and non-polyp categories using context information, (3) localizing polyp candidates from the refined edge maps using shape information, and (4) placing a band around each polyp

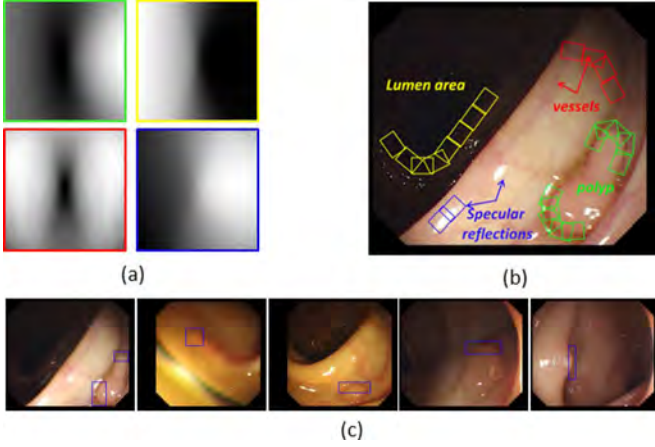


Fig. 2. (a) From left to right: average appearance of polyps, lumen areas, vessels, and specular reflection across thousands of image patches. As seen, the image appearance around polyp boundaries is distinct. (b) To obtain average image appearance, we collect oriented patches along the boundaries of vessels, specular spots, lumen areas, and polyps and then average the resulting patches for each structure of interest. (c) Although polyps appear in different shapes, they most often have a curvy segment in their boundaries. We have highlighted these curvy segments with the blue rectangles.

candidate to measure the probability of being a polyp. In the following, we describe each stage of the suggested method for automatic polyp detection.

A. Constructing Edge Maps

We apply Canny's method on the three color channels of the input images to extract as many edges as possible, followed by estimating gradient orientations for all the edge pixels in the map. We later use the gradient orientations to extract oriented patches around the edge pixels. Canny's algorithm computes gradient directions based on the local image gradients in horizontal and vertical directions; however, such estimations are often not accurate, leading to a non-smooth gradient direction map. Alternatively, we estimate gradient orientations by performing ball tensor voting [42].

B. Feature Extraction

Our patch descriptor begins with extracting an oriented patch around an edge pixel such that the edge segment appears vertically in the middle of the patch. This representation allows us to characterize intensity variation patterns across the edges independent of their orientations. We then divide each patch to sub-patches of size $n \times m$ with 50% overlap along horizontal and vertical directions. Each sub-patch is then averaged vertically, resulting in a 1D intensity signal S , embedding intensity variations along the horizontal axis. To summarize the patterns of intensity variations, we then apply a 1D discrete cosine transform (DCT) to the extracted signal and select the first few salient DCT coefficients:

$$C_k = \frac{2}{n} w(k) \sum_{i=0}^{n-1} S[i] \cos\left(\frac{2i+1}{2n} \pi k\right) \quad (1)$$

where

$$w(k) = 1/\sqrt{2}, \quad k = 0 \text{ and } w(k) = 1, \quad 1 \leq k \leq n-1.$$

However, such a compact representation of the signal based on DCT coefficients lacks robustness against changes in lighting conditions: the DC coefficient, $S[0]$, is readily affected by a constant change in the intensity of the patch, and intensity scaling directly affects all the DCT coefficients. To overcome these drawbacks, we remove the DC component and normalize the rest of the DCT components using their L^2 -norm. For computational efficiency, we compute the L^2 -norm of only the selected coefficients rather than all the resulting DCT coefficients. Finally, the salient coefficients selected from each sub-patch are concatenated to form a feature vector for the extracted patch.

The suggested patch descriptor has 4 advantages. First, it is fast because compressing each sub-patch into a 1D signal eliminates the need for the expensive 2D DCT. In addition, only a few 1D DCT coefficients are computed from each intensity signal, which further accelerates the feature extraction process. Second, due to the normalization treatment applied to the DCT coefficients, our descriptor achieves invariance to linear illumination changes and partial tolerance against nonlinear illumination variations over the entire patch, particularly if the nonlinear change can be decomposed to a set of linear illumination changes on the local sub-patches. Third, our descriptor provides a rotation invariant presentation of the intensity variation patterns thanks to the consistent appearance of the edge segments in the middle of oriented patches. Fourth, our descriptor handles small positional changes, which is essential to cope with patch variations due to edge mislocalization. In practice, spurious edges around polyp boundaries and Gaussian smoothing prior to Canny edge detector can cause inaccurate edge localization where an edge pixel is found a few pixels away from the actual location of the boundary. It is important for a patch descriptor to provide a consistent image presentation in the presence of such positional changes. We decrease positional variability by selecting and averaging overlapping sub-patches in both horizontal and vertical directions.

C. Edge Classification

The purpose of edge classification is two-fold: (1) discarding as many non-polyp edges as possible and (2) determining on which side of the retained edges the polyps are present. As shown in Fig. 4, our classification scheme analyzes a pair of oriented patches around each edge pixel and then depending on the appearance of the image pair classifies the underlying edge into either the polyp or non-polyp category and in the case of a polyp edge, identifies on which side of the boundary the polyp is present. In the following, we first explain how the image pairs are collected and then describe the suggested classification scheme.

After ball tensor voting, each edge pixel will be assigned a structure tensor whose dominant eigenvector \vec{e} indicates the gradient orientation. However, since the gradient orientation has no particular directions, as shown in Fig. 5, one can assume two normal directions for a given edge pixel, $\{\vec{n}_i^1 = \vec{e}_i, \vec{n}_i^2 = -\vec{e}_i\}$. The image is then interpolated along the normal and edge directions at each edge location, resulting in a pair of oriented patches $\{p_i^1, p_i^2\}$ given the two possible normal directions. Our classification scheme operates on each pair of patches and then combines their information not only to classify the underlying

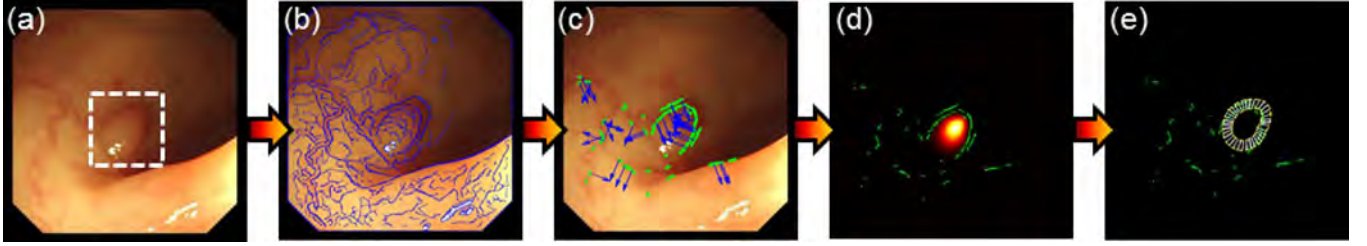


Fig. 3. Our polyp detection system given a test image. (a) The input image with a polyp inside. (b) An initial edge map is obtained (Section III.A). (c) The edge map undergoes a classification scheme where the goal is to filter out non-polyp boundary edges (Sections III.B and III.C). In this stage, a voting direction is also inferred by the classifier for each of the retained edges (blue arrows), which points to the polyp location. (d) Shape and curvature information of the retained edges modulated by the inferred voting directions is employed in the suggested voting scheme (Section III.D) for polyp localization. As shown in the heat map, the votes are accumulated in the region surrounded by the high curvature boundary. The pixel with maximum vote accumulation is considered as the polyp candidate. (e) A band (a set of line segments in its discrete form) is automatically determined around the candidate to measure the probability of being a polyp (Section III.E). The fraction of the line segments that hit the retained edges and meet some requirements determine the polyp likelihood for the generated candidate. For illustration purposes, only a subset of the line segments is displayed. Our method is summarized as an algorithm in Fig. S1 in the supplementary material.

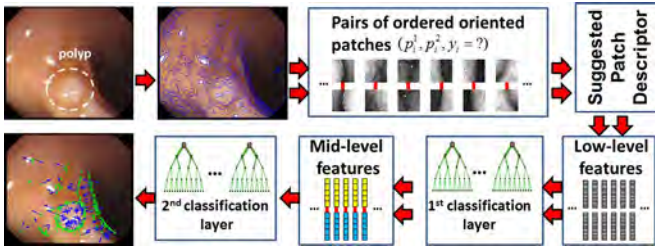


Fig. 4. The test stage of the suggested classification scheme. Pairs of oriented patches extracted around each edge pixel are fed to the suggested feature extraction and classification schemes. In the end result, the green pixels indicate the edges that have passed the classification stage and the blue arrows point to the possible location of a polyp.

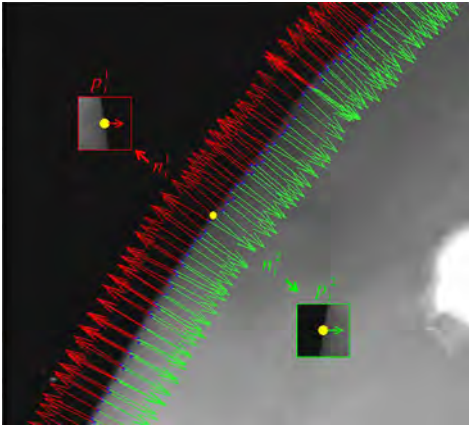


Fig. 5. The green and red arrows show edge normal directions for a subset of edge pixels on a boundary. As seen, the angle between each pair of normal vectors is 180° , $\angle \vec{n}_i^2 = \angle \vec{n}_i^1 + \pi$. At each edge pixel, we extract two oriented patches according to the two normal directions $(\vec{n}_i^1, \vec{n}_i^2)$. Each pair of patches (p_i^1, p_i^2) are horizontally mirrored to each other.

edge into polyp and non-polyp categories, but also to determine which normal direction among \vec{n}_i^1 and \vec{n}_i^2 points towards the polyp location. We refer to the determined normal direction as “voting direction” in the rest of this article.

Our classification scheme has 2 stages. In the first stage, we learn the image appearance around the boundaries of the structures of interest in contrast to random structures in colonoscopy images. The structures of interest are chosen through a prior misclassification analysis and consist of polyps, vessels, lumen

areas, and specular reflections. We train a 5-class classifier based on the features generated by our patch descriptor to distinguish between such boundaries in complex endoluminal scenes. In other words, the classifier learned in the first stage measures the similarities between the input patches and the predefined structures by learning a non-linear metric in the low level feature space. The first layer can be also viewed as a feature enhancer that takes low-level image features from the proposed patch descriptor and produces mid-level similarity features.

The key to train the above classifier is to have a consistent presentation for each of the structures of interest. A consistent presentation is defined as a set of dedicated patches that all have the structure of interest on one side (e.g., right) and the background on the other side (e.g., left). For this purpose, one must choose the normal directions prior to patch extraction such that they all point towards or away from the structures of interest. Choosing normal directions in an arbitrary manner results in the arbitrary appearance of the structure of interest on the left and right side of the collected patches. Fig. 5 shows how the choice of the normal direction places the gray region in the left and right side of the resulting image patches.

To achieve patch consistency, we use the ground truth that we have created for the structures of interest. The ground truth definition depends on the structure of interest. For polyps, the ground truth is a binary image where the polyp region is white and background is black. When extracting polyp patches, we choose the normal direction such that it points towards the polyp region. For lumen areas and specular spots, the ground truth is still a binary image but the white region corresponds to the dark lumen and the bright specular reflection, respectively. For vessels, the binary ground truth shows only the locations of vessels—we do not assume any preferred normal directions, because the image appearance around the vessels is most often symmetric. For random structures, we collect image patches at random edge locations in colonoscopy images with arbitrary choice of normal directions.

The goal of the second classifier is to group the underlying edges into polyp and non-polyp categories and determine the desired normal directions. For this purpose, we train the second classifier based on the pairs of patches in the mid-level feature space, which is generated by the first classifier. We use pairs of patches because for a new image no information about the polyp

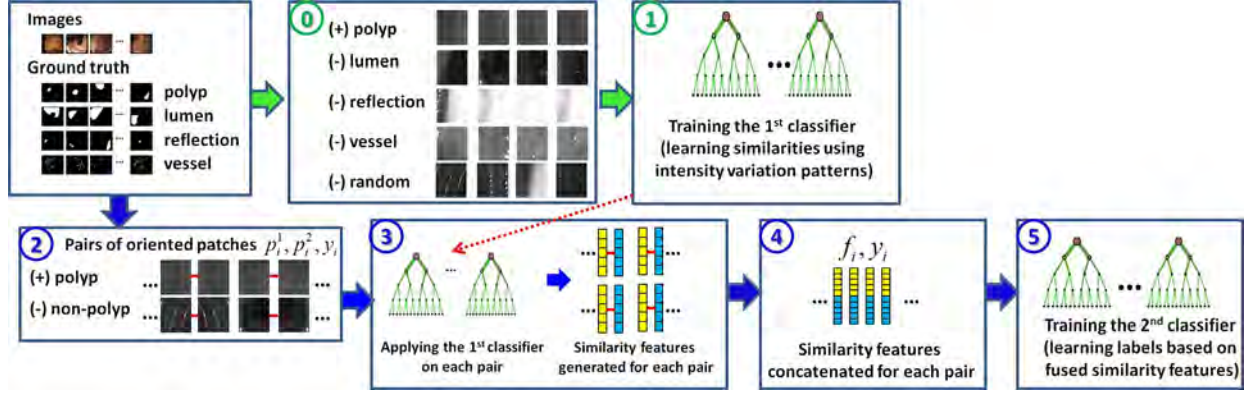


Fig. 6. The training stage of the suggested classification scheme.

location nor about the desired normal directions is available. The classifier learns the desired normal directions by combining the information from each image pair. The training process of the suggested classification scheme is illustrated in Fig. 6, is summarized in Fig. S2 in the supplementary material, and is further explained as follows:

1) *Layer 1:*

- **Step 0:** We collect a stratified set of N_1 oriented patches around the boundaries of polyps, vessels, lumen areas, specular reflections, and random structures in the training images. Mathematically,

$$S^1 = \{(p_i^*, y_i) | y_i \in \{p, v, l, s, r\}, i = 1, 2, \dots, N_1\}.$$

Note that asterisk indicates that the patches are extracted according to the desired normal directions, which are available given the ground truth for the training images.

- **Step 1:** Once patches are extracted, we train a five-class classifier in the low level feature space created by the proposed patch descriptor. The trained classifier generates 5 probabilistic values for each input low level feature vector, which reflect to what degree the underlying patch resembles the appearance of the five predefined structures.

2) *Layer 2:*

- **Step 2:** We select N_2 edge pixels from boundaries of polyps and other structures in the training images. From i th edge pixel, we extract an ordered pair of patches $\{p_i^1, p_i^2\}$ where $\angle \vec{n}_i^1 < \angle \vec{n}_i^2$. This convention is to keep patches in a consistent order. Each pair of patches is then assigned a label, $S^2 = \{(p_i^1, p_i^2, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, \dots, N_2\}$, which is determined as follows:

$$y_i = \begin{cases} 0 & \text{a non-polyp edge} \\ 1 & \text{a polyp edge and } \vec{n}_i^1 = \text{voting direction} \\ 2 & \text{a polyp edge and } \vec{n}_i^2 = \text{voting direction} \end{cases} \quad (2)$$

- **Step 3:** We extract low level features from each pair of patches using the suggested patch descriptor and then apply the classifier trained in Step 2, resulting in two arrays of mid-level features.
- **Step 4:** We generate a mid-level feature vector for the underlying edge by concatenating the two arrays of mid-level features, $\{(f_i, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, \dots, N_2\}$.

- **Step 5:** We train a 3-class classifier based on concatenated mid-level features. The classifier determines if an edge belongs to a polyp boundary and if so determines the normal direction.

During the test stage, the underlying edge of each image pair is declared as a polyp edge if

$$p(y_i = c) > p(y_i \neq c), c \in \{1, 2\},$$

which is met if and only if

$$p(y_i = c) > 0.5, c \in \{1, 2\}.$$

Therefore, the underlying edge of each pair of patches is classified according to the following rule:

$$\begin{cases} \text{"polyp"} \text{ and } \vec{n}_i^* \leftarrow \vec{n}_i^1 & \text{if } p(y_i = 1) > 0.5 \\ \text{"polyp"} \text{ and } \vec{n}_i^* \leftarrow \vec{n}_i^2 & \text{if } p(y_i = 2) > 0.5 \\ \text{"non-polyp"} & \text{otherwise,} \end{cases} \quad (3)$$

where \vec{n}_i^* is the desired normal direction or voting direction. The other alternative to (3) is to assign the edge pixel to the class with maximum probability, but this cannot handle uncertain situations where the probability associated with one class is only slightly larger than each of the other two individual classes. Once all the edges within an edge map are classified, non-polyp edges are removed from the edge map and the remaining edges along with their corresponding voting directions are sent to the voting scheme for polyp localization.

D. Voting Scheme

Our voting scheme is designed to localize polyps by identifying their curvy heads. This is achieved by generating a heat map where higher temperature is assigned to the regions that are surrounded by curved boundaries.

The voters that participate in our voting scheme are the edges that have passed the classification stage. Therefore, the i th voter comes with a voting direction \vec{n}_i^* and a classification confidence c_i . The vote cast by the voter v_i at a receiver pixel $r = [x, y]$ is computed as

$$M_i(x, y) = \begin{cases} c_i \exp\left(\frac{-\|\vec{v}_i \vec{r}\|^2}{\sigma_F}\right) \cos(\angle \vec{n}_i^* \vec{v}_i \vec{r}), & \text{if } \angle \vec{n}_i^* \vec{v}_i \vec{r} < \pi/2 \\ 0, & \text{if } \angle \vec{n}_i^* \vec{v}_i \vec{r} \geq \pi/2 \end{cases} \quad (4)$$

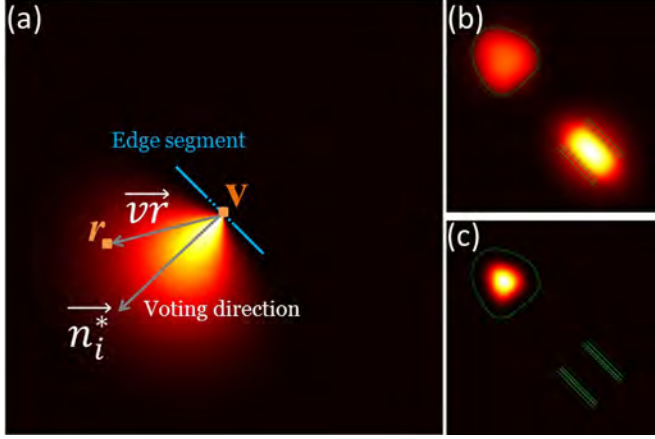


Fig. 7. (a) A voting function for an edge pixel. (b) The resultant voting map for a synthetic scene when the generated votes are accumulated in one voting map. Undesirably, large vote accumulation is observed around the parallel lines. (c) The resultant voting map for the same scene when the generated votes are accumulated in $K = 4$ voting maps and then multiplied (5). As seen, votes are desirably accumulated only in the object with the curvy boundary.

where the only unknown parameter is σ_F , which controls the size of the voting field. We have shown in Fig. 7(a) the voting function for a voter on an edge segment. As seen, the voter casts votes according to its voting direction \vec{n}_i^* , and thus the voting field appears only one side of the edge segment. Furthermore, the magnitude of votes smoothly decreases along the radial and angular directions due to the exponential and cosinusoidal terms in the voting function.

Our voting scheme begins by dividing the voters into K categories based on their voting directions, $V^k = \{v_i | (k\pi/K) < \text{mod}(\angle \vec{n}_i^*, \pi) < ((k+1)\pi/K)\}$, $k = 0 \dots K-1$. We then perform the voting process for each category of the voters and collect the resulting votes from each category in a separate voting map. Next, we multiply the K voting maps and select the pixel with maximum vote accumulation (MVA) as a polyp candidate. Mathematically,

$$\text{MVA} = \arg \max_{x,y} \prod_{k=0}^{K-1} \sum_{v \in V^k} M_v(x,y). \quad (5)$$

It is essential for our voting scheme to prevent vote accumulation in the regions that are surrounded by low curvature boundaries, because such regions in general cannot represent the curvy heads of polyps. This was achieved in our voting scheme by grouping edges prior to vote casting and multiplying the resultant voting maps. The rationale is that pixels on low curvature boundaries contribute to only a small fraction of the K to-be-multiplied voting maps. To clarify this, we generate a synthetic image and compare the resulting voting maps with and without the map multiplication strategy. The synthetic image consists of edge pixels that are arranged on a polyp-like structure and on a set of parallel lines. Fig. 7(b) shows the vote accumulation map when the votes cast by the voters are all accumulated in one voting map, $\sum_v M_v(x,y)$. As seen, the votes are accumulated in two regions: inside the curvy structure which

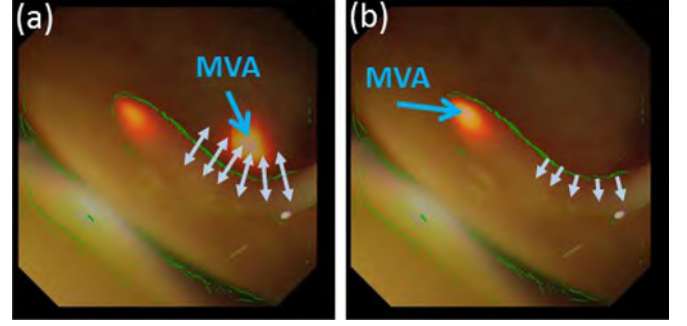


Fig. 8. Incorporating voting directions improves the accuracy of polyp localization. The edges retained after classification are shown in green. (a) A colonoscopy image and its corresponding voting map when the voters cast votes along both possible normal directions. As seen, the polyp candidate (MVA) is placed outside the polyp region. (b) The same colonoscopy image and its corresponding voting map when the voters cast votes only along the inferred voting directions. The polyp has been localized successfully.

is desirable, and between the parallel lines which is undesirable. Fig. 7(c) shows the voting map for the same image when edge grouping and map multiplication are employed (see (5)). As seen, the accumulator assigns low values to the region between the parallel lines, and high values to the region inside the polyp-like structure.

Another important characteristic of our voting scheme is the utilization of voting directions that, as shown in Fig. 7(a), limits a voter to cast votes only along its assigned voting direction. Ignoring voting directions (\vec{n}_i^*) and allowing the voters to vote along both possible normal directions (\vec{n}_i^1, \vec{n}_i^2) result in vote accumulation on both sides of the boundaries, which often leads to polyp mislocalization. This is illustrated in Fig. 8(a) where no selectivity in voting directions causes polyp mislocalization, but including such a selectivity allows for correct polyp localization (Fig. 8(b)). Our voting scheme is summarized in Step 3 of Fig. S1 in the supplementary material.

E. Probability Assignment

The magnitude of vote, accumulated at a polyp candidate, is not suitable for inferring a probabilistic score. This is because the magnitude of vote changes proportional to the scale of polyps. The larger the polyp, the larger the number of voters and thus the larger the magnitude of accumulated votes. Alternatively, we estimate a probabilistic score for a polyp candidate by determining the contributing voters within a narrow band around each polyp candidate. If the narrow band around a polyp candidate contains contributing voters in a larger number of directions, the candidate will have a higher likelihood of being a polyp.

We parametrize the narrow band B as a set of radial lines ℓ_θ :

$$B : \theta \rightarrow \ell_\theta$$

$$\ell_\theta : \text{MVA} + t[\cos(\theta), \sin(\theta)]^T, t \in \left[t_\theta - \frac{\delta}{2}, t_\theta + \frac{\delta}{2} \right] \quad (6)$$

where the unknown parameters of the band are the bandwidth, δ , and a set of distances, t_θ , between the candidate location and the corresponding point on the band skeleton Γ . We have shown an example of the narrow band in Fig. 9. Once the

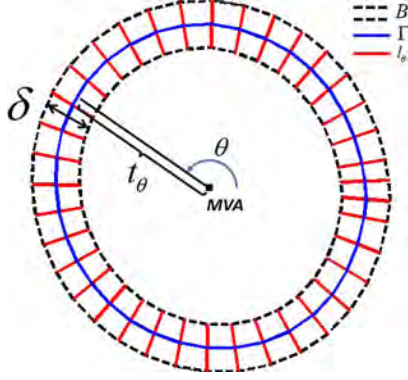


Fig. 9. A narrow band is determined around each polyp candidate to generate a probabilistic score. The blue contour is the band skeleton and the red lines are a subset of the radial line segments that represent the band in its discrete form.

band is formed, the probability assigned to a polyp candidate is computed as

$$\frac{1}{|S_\theta|} \sum_{\theta \in S_\theta} (I_\theta \vee I_{\theta+180}), \quad (7)$$

where S_θ denotes the set of angles along which the voters are searched and $|S_\theta|$ is the cardinality of S_θ . We consider the discrete set $S_\theta = \{\theta | 0 \leq \theta < 179\}$ for probability computation. In (7), I_θ is an indicator variable that takes 1 if the line segment ℓ_θ hits at least a voter v whose voting direction n_v^* points toward the candidate location:

$$I_\theta = \begin{cases} 1 & \text{if } (\exists v \text{ on } \ell_\theta) \wedge (\vec{n}_v^* \cdot [\tan(\theta), 1]^T < 0) \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

We estimate the unknown parameters of the bands from the isocontour of the voting maps. The isocontour Φ_c of the voting map V is defined as $\Phi_c = \{(x, y) | V(x, y) = c \times M\}$ where M denotes the maximum of the voting map and c is a constant between 0 and 1. Fig. 10(a) shows a synthetic shape, its corresponding voting map, and the isocontours for $c = \{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$. As seen, the isocontours become increasingly similar to the synthetic shape as the constant c decreases. This suggests that the isocontours that are farther away from MVA are more suitable for predicting the shape of the synthetic object and thus the parameters of the narrow band. However, in practice, such isocontours may be affected by other nearby voters in the scene. On the other hand, the isocontours that are located very close to MVA do not follow the shape of the object and thus are not suitable for our purpose. We therefore obtain a set of isocontours and then take their median shape as the representative isocontour $\bar{\Phi}$ of the voting map (Fig. 10(b)). We choose this set of isocontours so that their corresponding level c uniformly covers the range (0,1). We have experimented with different sets of isocontours and found out that as long as they are uniformly selected, the resulting representative isocontour serves the desired purpose. We use the isocontours shown in Fig. 10(a) for the rest of the experiments.

Let d_i^{iso} denote the distance between the i th point on the representative isocontour $\bar{\Phi}$ and the candidate location. We use d_i^{iso} to predict d_i^{obj} , the distance between the corresponding point on the object boundary and the candidate location (see Fig. 10(b)).

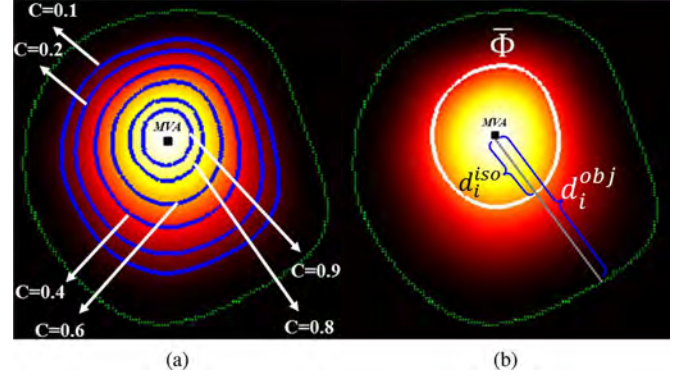


Fig. 10. (a) Isocontours of a synthetic shape for $c = \{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$. The isocontours, shown in blue, become increasingly similar to the synthetic shape as the constant c decreases. (b) The white contour shows the representative isocontour $\bar{\Phi}$, which is computed as the median shape of the blue isocontours. We use the representative isocontour to localize a band around the boundary of the synthetic shape.

For this purpose, we employ a second order polynomial regression model

$$d_i^{\text{obj}} = b_0 + b_1 (d_i^{\text{iso}}) + b_2 (d_i^{\text{iso}})^2, \quad (9)$$

where b_0 , b_1 , and b_2 are the regression coefficients that we estimate using a least square approach. The regression model in (9) yields the distance to the object boundary with a prediction interval. We choose the narrow band so that its skeleton approximates the object boundary. As a result, d_i^{obj} s are the unknowns t_θ s and the prediction interval is the unknown bandwidth δ . Therefore, given the candidate location and the representative isocontour, one can estimate the unknowns (t_θ s and δ) and then compute the probability using (7). We have shown in Fig. 11 a pseudocode that details how the band is formed and how a probabilistic score is generated given a voting map.

IV. EXPERIMENTS

We use a publicly available polyp database *CVC-ColonDB* [43] and our collection of short colonoscopy videos to evaluate our polyp detection system. We first employ *CVC-ColonDB* [43] for both tuning and evaluating the suggested system and then further evaluate our polyp detection system using our collected videos.

A. Evaluation Using *CVC-ColonDB* [43]

CVC-ColonDB is a public polyp database that contains 300 colonoscopy images each with a polyp inside. These images are selected from 15 short colonoscopy videos such that the images show maximum variation in scale and view angles of the polyps. Each image in this database comes with a binary ground truth image where the polyp and background regions are shown in white and black, respectively.

1) *Edge Detection*: To determine the degree of Gaussian smoothing σ_g prior to performing the Canny edge detector, we performed a set of experiments and investigated how changes in Gaussian smoothing can affect the percentage of polyp edges that can be detected by the Canny in each of the 300 images. For this purpose, we compare the resulting edge maps against the ground truth for polyps. To do so, for each boundary pixel in

Input:

- A voting map V
- A pre-constructed regression model \mathcal{RM}

Output:

- Probability of being a polyp p

Probability computation

Determine the candidate location (MVA)

$$\text{MVA} \leftarrow \arg \max_{x,y} V(x, y)$$

Obtain isocontours for $c = \{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$

Compute the representative isocontour

$$\bar{\Phi} = \text{median}(\Phi_c)$$

for θ from 0 up to 359

$Pt \leftarrow$ find the point on $\bar{\Phi}$ at angle θ wrt MVA

$$d_{\theta}^{iso} \leftarrow \|MVA - Pt\|$$

Apply the regression model $[t_{\theta}, \delta] = \mathcal{RM}(d_{\theta}^{iso})$

Form line ℓ_{θ} given t_{θ}, δ using Eq. 6

Determine I_{θ} according to Eq. 8

end for

$$\text{Compute probability } p = \frac{1}{180} \sum_{\theta=0}^{179} (I_{\theta} \vee I_{\theta+180})$$

Fig. 11. This pseudocode shows how a probabilistic score is computed for a polyp candidate.

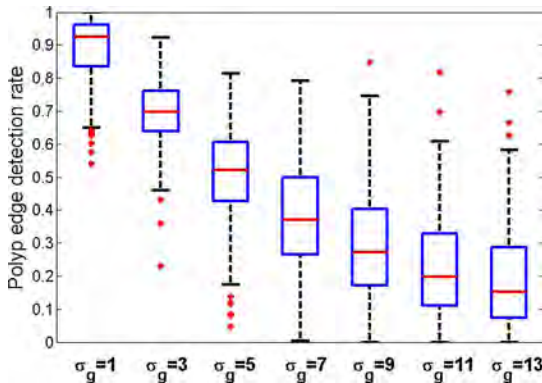


Fig. 12. Effect of Gaussian smoothing on the sensitivity of the Canny edge detector. Each box plot shows the percentage of polyp edges detected by Canny in the 300 colonoscopy images of *CVC-ColonDB*.

the ground truth, we find the closest edge pixel in the edge map and if the distance is less than 10 pixels, we mark that polyp boundary pixel as detected. Note that we always have some degrees of edge mislocalization due to Gaussian smoothing before applying the Canny. Once all the polyp boundary pixels are labeled, we can measure polyp edge detection rate.

We have shown in Fig. 12 the polyp edge detection results for different amounts of Gaussian smoothing. The whiskers are plotted according to the Tukey's method [44]. The red crosses below the box plots correspond to the polyps that have faint edge segments in their boundaries. As seen, the Canny edge detector can effectively capture a high percentage of polyp edges particularly for the small values of σ_g . However, in practice, small

values of σ_g result in very cluttered edge maps, complicating edge direction estimation with ball tensor voting. Therefore, we use $\sigma_g = 3$ for the rest of our experiments.

2) *Feature Extraction and Edge Classification*: We employed 5-fold cross validation to train and test our 2-stage classification system. We collected $N_1 = 100,000$ oriented image patches of size 64×64 with approximately 20,000 samples for each of the five chosen structures to train a random forest classifier for the first stage, and $N_2 = 100,000$ pairs of oriented image patches of size 64×64 to train another random forest classifier for the second stage. We have selected patches of size 64×64 , because they are more suitable for capturing context around the boundaries. In the supplementary material, we have analyzed overall performance for different sizes of patches. The choice of a random forest classifier is motivated by its recent success in a variety of computer vision and medical image analysis applications where it outperformed other widely-used classifiers such as AdaBoost and support vector machines [45]. The two main ingredients of a random forest classifier are bagging of a large number of fully grown decision trees and random feature selection at each node while training the trees, which together achieve low generalization error and high quality probabilistic outputs. In our experiments, we kept adding decision trees to the random forest classifiers until the decreasing trend of out-of-bag error converged. According to our experiments, 100 fully grown decision trees achieved a stable out-of-bag error for both random forest classifiers.

Fig. 13(a) shows the receiver operating characteristic (ROC) curves of the first classification layer for the suggested patch descriptor and the other widely used descriptors such as HoG¹ [46], LBP² [47], and Daisy³ [48]. The first stage classifier is trained for a 5-class classification problem; however, to avoid clutter in this figure, we have shown only the ROC curves corresponding to “polyp vs. rest” classification scenario. As seen, our descriptor surpasses HOG and LBP with a large margin and outperforms Daisy with a smaller yet statistically significant margin⁴ ($p < .0001$). In addition to superior classification performance, our descriptor runs approximately 30 times faster than its closest competitor (Daisy), which makes it further amenable to the suggested classification scheme.

For fair comparisons between the patch descriptors, we used the same training set for parameter tuning and the same test set for performance evaluation. For our descriptor, we used 8×16 sub-patches in each image patch and extracted 3 DCT coefficients from each sub-patch, which yielded a good balance between the size of feature vectors and the discrimination power. This configuration resulted in a feature vector with 315 elements for each image patch. For LBP, we divided each patch into cells of size 8×8 and for each cell we computed the normalized histogram of rotation invariant uniform patterns using a 3×3 neighborhood around the pixels. The resulting 10-bin histograms from all the cells were then concatenated to form the final feature vector. For HoG, we used cells of size 8×8 pixels

¹lear.inrialpes.fr/pubs/2005/DT05/

²www.cse.oulu.fi/CMV/Downloads/LBPMatlab

³cvlab.epfl.ch/software/daisy

⁴DeLong's method using MedCalc for Windows, version 13.3 (MedCalc Software, Ostend, Belgium)

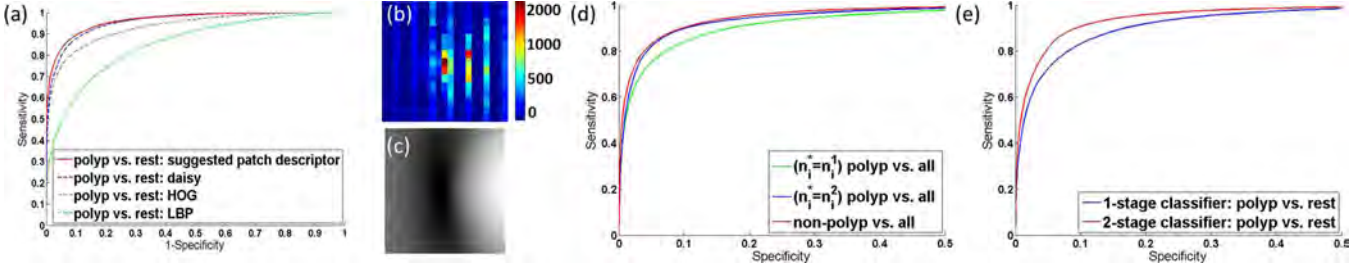


Fig. 13. (a) ROC curves of the first-stage edge classifier for “polyp vs. rest” classification scenario. Our patch descriptor surpasses HOG and LBP with a large margin and outperforms Daisy with a smaller yet statistically significant margin. Our descriptor runs approximately 30 times faster than Daisy. (b) Feature importance computed by the first-stage random forest classifier. Comparison with (c), which is the average image appearance around polyp boundaries, suggests that important features are more densely located inside the polyp region and across the polyp boundary. The characteristic stripes in the importance map, on average, show a decreasing trend, suggesting that lower frequency DCT coefficients are more important. (d) ROCs of the second-stage edge classifier for the three possible “one vs. rest” classification scenarios. (e) ROCs of the 1-stage and 2-stage edge classification scheme in the “polyp vs. rest” classification scenario.

and blocks of size 2×2 cells or 16×16 pixels. We computed a gradient histogram with 9 orientation bins for each block and then concatenated the resulting histograms. For Daisy, we defined three concentric rings around the center of the patch and then selected 8 equally spaced points on each ring. Next, we concatenated 8-bin gradient histograms computed at each of the selected points.

To further analyze our patch descriptor, we visualize the variable importance computed by the random forest classifier for each of the extracted features. Random forest calculates the importance of feature f_i in each tree and then takes their average to compute the overall importance of feature f_i . To measure importance of feature f_i in each tree, random forest permutes the values of this feature in the out-of-bag samples randomly and then measures the subsequent increase in the misclassified samples. A feature is considered more important if the corresponding permutations further increase the. We collect the variable importance of all the 315 features and then reshape them into a matrix form such that each feature gets mapped to the part from which it has been extracted.

We have shown the importance matrix in Fig. 13(b). For easier comparison with the average appearance around polyp boundaries (see Fig. 13(c)), we scale up the importance matrix to the same size as the input image patches. As seen, while the important features are found all over the importance map, they are more densely located inside the polyp region and across the polyp boundary. The relatively less number of important features on the background side (left side) of the patches can be explained by the large variability in the backgrounds of polyps. The importance matrix also contains characteristic stripes that show the importance of the 3 DCT coefficients extracted from each subpatch. Averaging these stripes across the importance matrix reveals a decreasing trend in the importance of the DCT coefficients, where the 1st and 3rd coefficients are, respectively, the most and least important features. This can be explained by the susceptibility of higher frequency DCT coefficients to noise and other degradation factors that may appear in the images patches.

Fig. 13(d) shows the ROC curves for the edge classifier of the second classification layer in 3 possible “one vs. rest” classification scenarios. Recall that the polyp class and the desired voting directions are embedded in 3 labels. To avoid clutter in this figure, we have shown only the ROCs corresponding to the suggested patch descriptor, which outperformed the other

TABLE I
POLYP DETECTION RATES AT 0.05 FALSE POSITIVES PER FRAME
FOR DIFFERENT CONFIGURATIONS OF THE VOTING SCHEME

		Size of the voting field			
		$\sigma_F=70$	$\sigma_F=80$	$\sigma_F=90$	$\sigma_F=100$
# Edge groups	$K=3$	80.7%	81%	80%	75%
	$K=4$	83%	84.6%	81.7%	80%
	$K=5$	79%	81.6%	81.7%	78.7%
	$K=6$	83%	82.6%	82%	81%

competing feature extraction methods in the first classification stage. As seen, our edge classification scheme tends to underperform for polyp edges with $\vec{n}^* = \vec{n}^1$. This is because these edges most often lie on the low gradient boundaries of polyps, where they attach to the colon wall. As a result, the corresponding patches show indistinct appearance features, causing edge misclassification.

We have further compared edge classification performance with and without the second classification layer. We perform this comparison in “polyp vs. rest” classification scenario using the stratified set of 100,000 pairs of oriented patches. The comparison is shown Fig. 13(e). As seen, employing the second classification layer leads to a significantly higher area under ROC curve ($p < 0.0001$), demonstrating the effectiveness of the suggested two-stage classification scheme. In Section IV.B, we will further demonstrate that employing the second classification layer significantly improves polyp detection performance in colonoscopy videos.

3) *Voting Scheme and Probability Assignment*: We trained a number of regression models for different values of σ_F and K , and investigated how the choice of these parameters affected polyp detection performance. Because the goal of the voting scheme was to detect the curvy heads of polyps, we designed a shape generator model that could produce objects resembling the curvy heads of polyps. We used these synthetic shapes to collect pairs of $(d_i^{\text{obj}}, d_i^{\text{iso}})$ and then constructed a regression model for each combination of σ_F and K . We have explained our shape generator model and the protocol for training the regression models in the supplementary material. Table I compares the polyp detection rates of these models at 0.05 false positives per frame. As seen in Table I, we achieved relatively stable results for a wide range of σ_F and K , but obtained the best result using $\sigma_F = 80$ and $K = 4$.



Fig. 14. Successful polyp localization. The edges retained after classification are shown in green. The blue line segments indicate the radial lines that have reached the contributing voters within the estimated bandwidth.

Fig. 14 shows examples of successful polyp localization using our voting scheme. For better visualization, we superimpose the voting heat maps on the original images and show only a number of the constituent line segments of the discrete bands. We use color coding with blue indicating the line segments that hit at least a voter with the desired voting direction (see (8)), and red otherwise. As seen, our polyp detection system is able to localize polyps of different shapes, scales, and colors.

To demonstrate the effectiveness of the suggested voting scheme, we compare the polyp localization accuracy of our system using our voting scheme and the phase-coded Hough transform [49]. For fair comparisons, we applied both algorithms on the same refined edge maps and chose the maximum response in the corresponding map as the location of a polyp candidate. Also, the radius range of the Hough transform was tuned to detect the smallest and biggest polyps in the database. According to our experiments, the phase-coded Hough transform placed the polyp candidates inside 246 out of the 300 polyps with 54 false positives, which was outperformed by our voting scheme with 262 true positives and 38 false positives.

We have also compared our voting scheme with the phase-coded Hough transform using an FROC analysis. For this purpose, we changed a threshold on the scores assigned to the polyp candidates by both algorithms and then computed the sensitivity and the number of false positives at each threshold. For the Hough transform, the scores were selected from the Hough voting maps at the candidate locations. For our method, the scores were generated based on the narrow bands. As shown in Fig. 15, our voting scheme significantly outperforms the Hough transform in all the operating points. For a more detailed comparison, we have also included the operating points of other polyp detection systems⁵ suggested in the literature. As seen, our CAD system outperforms the other methods with a large margin.

B. Evaluation Using Our Collected Videos

Our database of colonoscopy videos⁶ is, to our knowledge, the largest annotated polyp database. We have selected 10 positive shots and 10 negative shots from our database. A positive

⁵Note that the operating point shown for [17] was not available in their published manuscript—it was kindly provided by the corresponding author.

⁶available at <http://tinyurl.com/polyp2015>

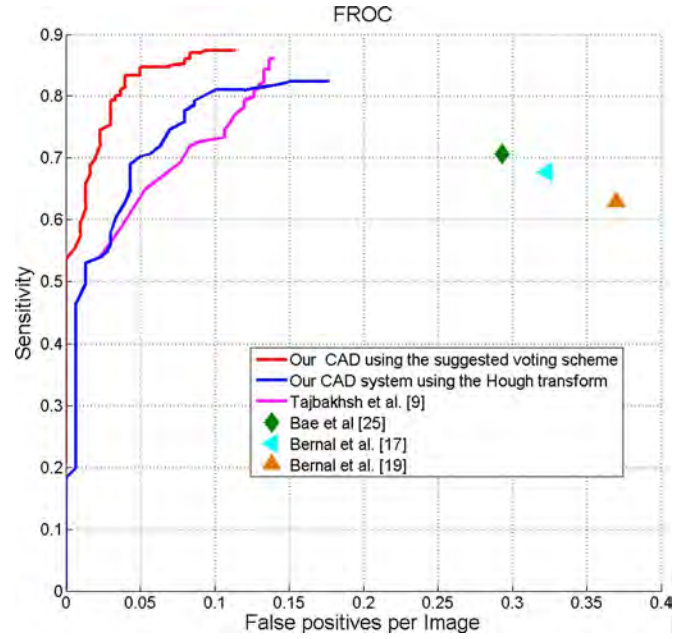


Fig. 15. FROC curves of our system and the other competing methods for *CVC-ColonDB*. Our CAD system using the suggested voting scheme excels in all the operating points.

shot is a segment of a colonoscopy video that displays a unique polyp at multiple scales and from different viewing angles. A negative shot is a segment of a colonoscopy video that does not contain a polyp. The selected shots consists of 5200 frames with polyps and 14,200 frames with no polyps inside. These images show a large degree of variability in the colonoscopic view, including varying levels of colon preparation, different colonoscopic events, narrow band imaging, and different amounts of motion and interlacing artifacts. As with *CVC-ColonDB*, each frame in our database comes with a binary ground truth image where the polyp region is shown in white. If an image contains no polyp, the corresponding ground truth will be a completely black image. To create the ground truth, our expert colonoscopists first determined the locations and extents of a few polyp instances in each video. Then, following the expert annotations, a number of trained volunteers created the ground truth for the remaining frames in each video. The resulting truth was then reviewed by our experts.

For video-based evaluation, we trained our system on the entire *CVC-ColonDB* using the previously tuned parameters $\sigma_g = 3$ and $K = 4$. The shaded FROC curve shown in Fig. 16 displays the variation in our system's performance when σ_F changes between 70 and 100. As seen, our system using the suggested 2-stage edge classifier yields relatively stable performance over a wide range of voting field sizes, generating on average 0.11 false positives per frame (FPPF) at 50% sensitivity. Similar to *CVC-ColonDB*, the best FROC curve is obtained using $\sigma_F = 80$, demonstrating the robustness of the suggested system across different databases.

In Fig. 16, we further compare the performance of our polyp detection system using 1) the suggested 2-stage edge classifier, and 2) an alternative 1-stage edge classification scheme. The latter is realized by simply discarding the second classifier in our classification scheme. Basically, after a pair of patches passes

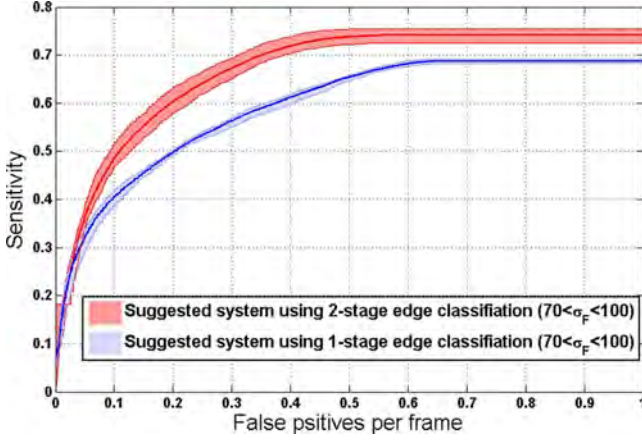


Fig. 16. FROC curves of our polyp detection system based on 20 short colonoscopy videos. The suggested 2-stage edge classification scheme significantly improves the overall polyp detection performance compared to the scenario where our system employs a 1-stage edge classifier.

the first classifier, we obtain two sets of probabilities (mid-level features). To determine a polyp edge, one can compare the polyp probabilities between the two patches and check whether the larger probability is above the classification threshold of 0.5. The desired normal direction is also determined as the normal direction associated with the patch with the larger probability. As seen in Fig. 16, the suggested 2-stage edge classifier significantly outperforms the 1-stage edge classifier in nearly all operating points. Another observation is that the maximum sensitivity achieved by the 2-stage edge classifier is higher than that of the 1-stage counterpart, indicating that the refined edge maps produced by our suggested classification scheme can better retain polyp boundaries and are thus more suitable for polyp localization.

One limitation of the FROC analysis may be that it does not account for the factor of time, simply measuring sensitivity to polyps in the entire videos. However, polyps that are missed in colonoscopy videos most often appear in the colonoscopic view briefly. Therefore, a polyp CAD system with high sensitivity yet with large detection latency may provide limited clinical value. Hence, we propose a variant of the FROC analysis, called the detection latency analysis, which replaces the sensitivity on the vertical axis with the median detection latency of the positive shots. Let t_1 denote the arrival frame of a polyp in a video and t_2 denote the frame where our CAD system detects the polyp. We measure the detection latency of a positive shot as $\Delta T = (t_2 - t_1)/(fps)$ where fps is the frame rate of the video. Fig. 17 shows the detection latency of the suggested system using the previously tuned parameters $\sigma_g = 3$ and $K = 4$. The shaded FROC curves shown in Fig. 16 display the variation in our system's performance when σ_F changes between 70 and 100. As seen, our system achieves low detection latency in a wide range of operating points, detecting polyps in less than 1 second upon their appearance in the videos. In addition, given a fixed number of false positives, our system using the suggested 2-stage classifiers achieves shorter polyp detection latency compared to the 1-stage classification scenario.

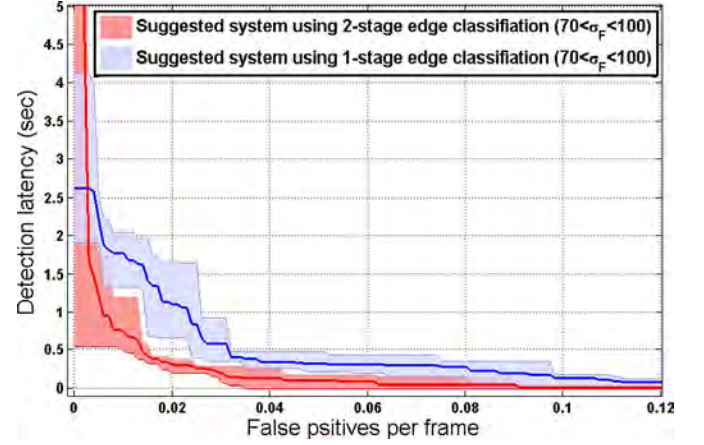


Fig. 17. Detection latency of our polyp detection system based on 20 short colonoscopy videos. The suggested 2-stage classification scheme yields significantly shorter polyp detection latency particularly in low false positive rates.

V. DISCUSSION

In the previous sections, we presented a CAD system for detecting colonic polyps in colonoscopy videos, and evaluated it on 2 polyp databases: (1) a public polyp database, *CVC-ColonDB*, containing 300 colonoscopy images with a total of 300 polyp instances, and (2) our collection of colonoscopy videos containing 19,400 frames and a total of 5,200 polyp instances. We characterized the performance of our CAD system using FROC and detection latency analyses. At 0.1 false positives per frame, we obtained a sensitivity of 88.0% for *CVC-ColonDB* and a sensitivity of 48% for our collection of colonoscopy videos. We also observed a detection latency of 0.3 seconds at 0.05 false positives per frame. The performance variation between these two databases could be caused by the insufficient number of images in *CVC-ColonDB*, and more importantly, absence of images that have no polyps inside. This limitation is overcome in our database by including 14,200 frames with no polyps inside.

Table II summarizes the operating points for our CAD system and the other methods that have recently been suggested in the literature. As seen, the majority of the recent systems have used different databases for evaluation. We can fairly compare our system against only the systems suggested in [17], [19], [25], [9], because they also use *CVC-ColonDB* for performance evaluation. The tabulated operating points for these systems are selected from Fig. 15. *CVC-ColonDB* allows us to draw fair performance comparisons; however, this database is not sufficiently large for a solid evaluation of polyp detection systems. To enable more rigorous evaluation and facilitate future performance comparisons, we have released our annotated polyp database to the public.

The work of Bernal *et al.* [17] is probably the most similar to our approach. Fig. 15 shows that our system outperforms this method as well as the other competing systems. Here, we would like to intuitively discuss the reasons behind our superior performance. In [17], valley information is used for polyp localization, which corresponds to the features that we extract from the middle part of the oriented patches. However, as seen in Fig. 13(b), features from other parts of the patches can also

TABLE II

RECENT POLYP DETECTION METHODS DESIGNED FOR OPTICAL COLONOSCOPY. AS SEEN, THE EXISTING METHODS HAVE BEEN EVALUATED USING DIFFERENT DATASETS AND THEIR RESULTS HAVE BEEN REPORTED BASED ON DIFFERENT PERFORMANCE METRICS (FPPF: FALSE POSITIVES PER FRAME, FPR: FALSE POSITIVE RATE, FPPS: FALSE POSITIVES PER SHOT). OUR WORK CAN BE FAIRLY COMPARED AGAINST [17], [19], [9], [25], BECAUSE ALL THE FIVE SYSTEMS HAVE BEEN EVALUATED USING THE SAME PUBLIC DATABASE

Author	Year	Feature	Dataset # images	Result
			with polyps / without polyps	
Current work	—	Shape and context information	<i>CVC-ColonDB</i> (300 / 0) 5200 / 14200	88% sensitivity @ 0.1 FPPF 48% sensitivity @ 0.1 FPPF
Bae et al. [25]	2015	Discriminative feature learning	<i>CVC-ColonDB</i> (300 / 0)	70.6 % sensitivity @ 0.3 FPPF
Bae et al. [25]	2015	Discriminative feature learning	1123/140	24% sensitivity @ 0.1 FPPF
Wang et al. [24]	2013	Edge cross section profile	1025 / 488	34% sensitivity @ 0.1 FPPF
Bernal et al. [19]	2013	Modified valley information	<i>CVC-ColonDB</i> (300 / 0)	67.6 % sensitivity @ 0.3 FPPF
Tajbakhsh et al.[9]	2013	Shape in context	<i>CVC-ColonDB</i> (300 / 0)	73.3% sensitivity @ 0.1 FPPF
Bernal et al. [17]	2012	Valley information	<i>CVC-ColonDB</i> (300 / 0)	67.6% sensitivity @ 0.32 FPPF
Park et al. [21]	2012	Temporal and appearance features	35 complete videos	56% recall @ 10% FPR
Cheng et al. [16]	2011	Texture features	37 / 0	86.2% recall @ 1.26 FPPF
Ameling et al. [15]	2009	Texture and color features	1736 / 0	56% recall @ 10% FPR
Hwang et al. [18]	2007	Temporal and elliptical shape features	1 video	96% recall @ 1 FPPS

provide discriminative power for polyp boundary classification. Their work further assumes polyps as circular structures whose radii vary in a pre-specified range, but our approach does not make such assumptions and automatically estimates the shapes and scales of polyps using the narrow bands.

In the design of our patch descriptor, we prefer the DCT over other transforms such as Discrete Sine Transform (DST) and Discrete Fourier Transform (DFT), because the DCT is more suitable for compressing patch information. Specifically, the DFT assumes that the intensity signal S is a part of a periodic signal; therefore, if the intensity values at both ends of the intensity signal are not equal ($S[0] \neq S[n-1]$), S will appear as a part of non-continuous periodic function to the DFT, yielding large high-frequency Fourier coefficients, and preventing the information (energy) of the signal to be compressed in a few Fourier coefficients. Large high frequency components will also appear in the case of the DST, if the intensity signals have non-zero values at their both ends ($S[0] \neq 0, S[n-1] \neq 0$). In contrast, the DCT relaxes these constraints, requiring only smooth behavior at both ends of the intensity signals. This property is more amenable to our application, because the intensity signals, which are obtained by averaging the corresponding sub-patches, are usually smooth without abrupt changes at both ends.

In Section IV.A2, we experimentally found out that the first three DCT coefficients (excluding the DC component) are the most suitable for feature extraction. To intuitively explain this choice of coefficients, Fig. 18 shows the basis functions corresponding to the first 4 DCT coefficients: the first DCT basis function corresponds to the DC component, the second one measures whether the intensity signal S is monotonically decreasing (increasing) or not, the third one measures the similarity of the intensity signal against a valley (ridge), and finally the fourth one checks for the existence of both a valley and a ridge in the signal. The higher order basis functions are more suitable for modeling high frequency changes, which we rarely observe in our smooth intensity signals. Therefore, the number of desired coefficients can be intuitively determined without resorting to more complicated feature selection algorithms.

Our edge classification scheme consists of two stages where the first stage serves as a feature enhancer and the second stage performs the main classification task. However, one may dis-

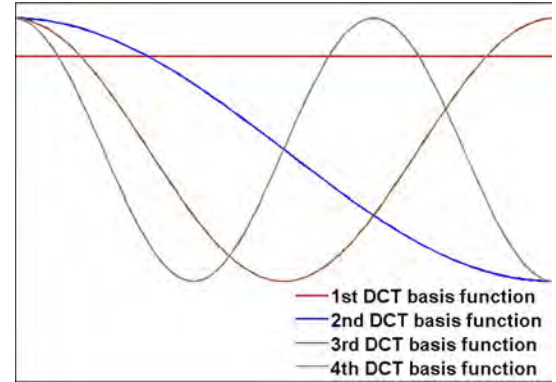


Fig. 18. Basis functions corresponding to low frequency DCT coefficients.

card the second classifier, and perform edge classification and determine voting directions merely based on the outputs of the first classifier. In Section IV.B, we demonstrated that this deteriorates the performance of the edge classification scheme. Here, we would like to discuss the necessity of the second classification layer from a different perspective. Consider the following two sets of probabilities generated by the first classifier for a pair of image patches extracted from the i th edge pixel, $\{0.6, 0.1, 0.0, 1.0, 0.2\}$ and $\{0.0, 0.3, 0.1, 0.6, 0.0\}$, where the probabilities, respectively, measure the similarity of each patch to boundary appearance of polyps, vessels, lumen area, specular reflections, and other random structures. The resulting probabilities suggest that patch p_i^1 resembles the appearance of a polyp boundary and that p_i^2 resembles the boundary appearance of specular reflections. This produces uncertainty as to the decision regarding the underlying edge pixel. The choice is to either rely on the first patch and declare a polyp edge with edge normal being n_i^1 or consider information from the counterpart patch and declare a non-polyp edge. One way to resolve this issue is to define a number of decision rules that may only achieve a sub-optimal solution. Alternatively, we choose to train a second classifier in the mid-level feature space, exploring all such decision rules in a systematic manner.

In Section III.C, we explained how the suggested classification scheme inferred the voting directions for a given edge map. However, one may argue that the voting direction for

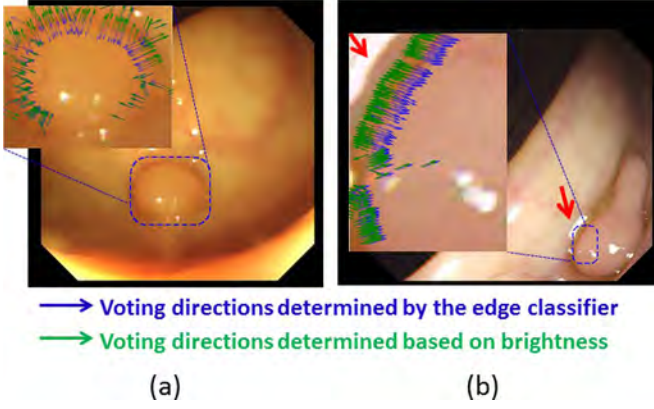


Fig. 19. The suggested edge classifier determines voting directions more reliably than a brightness-based approach. (a) The voting directions determined by the edge classifier correctly point inward (toward the polyp region). (b) Specular reflections, highlighted by the red arrow, result in erroneous voting directions when using brightness-based approach. This drawback is overcome by the suggested edge classifier.

an edge pixel could be simply chosen as the normal direction that would result in a patch with higher intensity values on its right side. This argument could stem from Fig. 13(b), which suggests that polyps, on average, appear brighter than their surrounding regions. While this simplistic approach may be effective in some cases, there are many polyp instances for which this approach fails to yield the correct voting directions. Fig. 19 shows two such examples. Fig. 19(a) display a polyp that appears darker than the colon surface. As seen, while the voting directions determined by the classifier correctly point inward (towards the polyp), the green arrows determined by the brightness-based approach point both inward and outward. Furthermore, existence of specular reflection around the boundaries of polyps can also complicate the estimation of voting directions based on average brightness. This is shown in Fig. 19(b) where the brightness-based method influenced by the specular spot (shown by the red arrow) results in wrong voting directions whereas the learning-based method still yields correct voting directions. Therefore, average brightness may not be reliable for determining the voting directions.

The size of voting fields, controlled by σ_F , determines the size of polyps that can be detected by our CAD system. In Sections IV.B and IV.A3, we extensively studied the choice of σ_F and found out that $\sigma_F = 80$ works the best for both databases. This is probably because the majority of polyp instances in both *CVC-ColonDB* and our collection of videos appear in small or moderate sizes in the colonoscopic view. This choice of σ_F can potentially lead to the misdetection of the polyps that appear large in the videos, but we do not consider that as a drawback for our polyp detection system. This is because polyps that appear large in the videos probably have already been detected and are under examination by colonoscopists; therefore, there are no clinical needs for computer-aided polyp detection. Furthermore, the physical size of a polyp differs from what “appears” in the videos: the former is fixed but the latter varies depending on the distance between the polyp and the camera. Therefore, our CAD system can also detect physically large polyps when they have not been reached by the camera and thus appear in small or moderate size in the colonoscopic view.

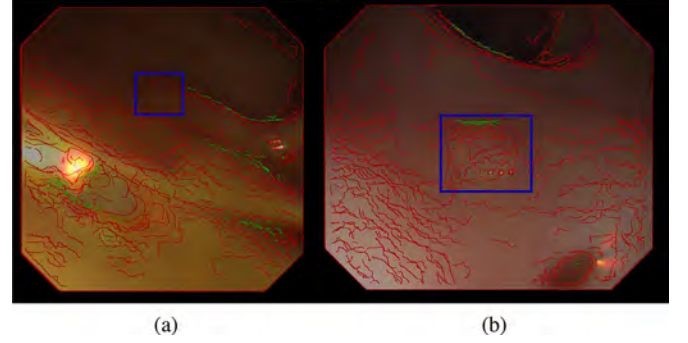


Fig. 20. Unsuccessful polyp localization. (a) Polyp mislocalization due to the failure of the Canny edge detector in capturing the curvy head of the polyp. This is caused by the weak boundaries in the region highlighted with the blue rectangle. (b) Polyp mislocalization due to the failure of the edge classification stage. Majority of the polyp boundary pixels are classified as non-polyp, because intensity variation patterns around polyp boundaries are corrupted by the interlacing artifact.

Our method can detect multiple polyps in each frame if we consider all the local maxima in (5). However, since a patient may have only a small number polyps (if any) in his/her entire colon, there is a slender chance that they all cluster in a small area of the colon. Therefore, we chose to find the global maximum of the voting map in (5), which allows for detection of at most one polyp in each frame. However, this does not mean that our method cannot detect multiple polyps in a patient. For a more intuitive discussion, consider the scenario where two polyps, for instance, Polyp A and Polyp B, have appeared in the same frame. If these two polyps are relatively far apart, then one of the polyps, Polyp A (or B), has certainly appeared first in the colonoscopic view alone and thus will exist earlier from the colonoscopic view during the scope withdrawal, leaving the other Polyp, Polyp B (or A) alone in the colonoscopic view. Also, if the two polyps appear in the same frame yet very close, there may be no clinical needs to detect both of them, because the feedback on one polyp will automatically bring the other one into colonoscopist's attention.

The suggested CAD system may fail to detect the polyps that have faint gradients around their boundaries. A faint polyp boundary is likely to be missed by the edge detector, causing the polyp to appear as a non-curvy structure in the resulting edge map, resulting in a polyp localization failure. Fig. 20(a) shows an example of a missed polyp where the curvy head of the polyp is not captured by the Canny edge detector. Unsuccessful edge classification can also cause localization failures. Colonoscopy frames may contain motion blurriness and interlacing artifacts that can corrupt the desired intensity variation patterns around the polyp boundaries. As a result, polyp boundary pixels will be classified as non-polyp edges, causing the polyps to appear as a non-curvy structure to the voting scheme. Fig. 20(b) shows an example of a missed polyp where the boundary is corrupted with the interlacing artifact. While it would be interesting to perform a comprehensive robustness analysis against different levels of such artifacts, it might be more effective to remove or mitigate these artifacts prior to polyp detection. This underlines the importance of image quality assessment in colonoscopy, a topic that we have considered in our previous work [50].

On a desktop computer with a 2.4 GHz quad core Intel, our CAD system processes each colonoscopy image at 2.6 seconds on average, which is significantly faster than [24] with run-time of 7.1 seconds and [17] with run-time of 19 seconds. In addition, the MEX implementation of our patch descriptor processes 36,000 image patches in a second. Considering that the edge map of a colonoscopy image after applying Gaussian smoothing contains on average 20,000 edge pixels, our descriptor can process each image in approximately 0.5 seconds. By converting our Matlab-MEX implementation to C/C++ and employing parallel computing optimization, we expect a significant increase in the speed of the suggested system.

Our CAD system with some modifications could also be used for detecting polyps in capsule endoscopy images. In contrast to optical colonoscopy, capsule endoscopy is not a live process so the CAD system could be employed in an off-line fashion to scan the images for polyps or other types of lesions in the gastrointestinal tracts. We will consider the application of our CAD system to capsule endoscopy in our future work. Another use case of the suggested CAD system would be to annotate stored colonoscopy videos in a post-exam setting. Such a video annotation mechanism in conjunction with other objective quality documentation systems [50], [51] can be used for more effective colonoscopy reimbursement.

ACKNOWLEDGMENT

The authors would like to thank Saiswathi Javangula, Ireen Khan, Kamran Bodushev, Sarah Fallah-Adl, and Tracy Phan for their diligent effort in creating the ground truth. They are also grateful to Chang-ching Chi for performing extensive comparative performance analyses and Anirudh Som for proofreading our manuscript. They would also like to thank Jorge Bernal, the corresponding author of [17], for providing us with the detection results.

REFERENCES

- [1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA A Cancer J. Clin.*, vol. 65, no. 1, pp. 5–29, 2015.
- [2] D. Lieberman, "Quality and colonoscopy: A new imperative," *Gastrointest. Endosc.*, vol. 61, no. 3, pp. 392–394, Mar. 2005.
- [3] A. Pabby *et al.*, "Analysis of colorectal cancer occurrence during surveillance colonoscopy in the dietary polyp prevention trial," *Gastrointest. Endosc.*, vol. 61, no. 3, pp. 385–391, 2005.
- [4] J. van Rijn *et al.*, "Polyp miss rate determined by tandem colonoscopy: A systematic review," *Am. J. Gastroenterol.*, vol. 101, no. 2, pp. 343–350, 2006.
- [5] D. H. Kim *et al.*, "CT colonography versus colonoscopy for the detection of advanced neoplasia," *N. Eng. J. Med.*, vol. 357, no. 14, pp. 1403–1412, 2007.
- [6] D. Heresbach *et al.*, "Miss rate for colorectal neoplastic polyps: A prospective multicenter study of back-to-back video colonoscopies," *Endoscopy*, vol. 40, no. 4, pp. 284–290, 2008.
- [7] A. Leufkens, M. van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 05, pp. 470–475, 2012.
- [8] L. Rabeneck, H. El-Serag, J. Davila, and R. Sandler, "Outcomes of colorectal cancer in the united states: No change in survival (1986–1997)," *Am. J. Gastroenterol.*, vol. 98, no. 2, p. 471, 2003.
- [9] N. Tajbakhsh, S. Gurudu, and J. Liang, "A classification-enhanced vote accumulation scheme for detecting colonic polyps," in *Abdominal Imaging. Computation and Clinical Applications*, 2013, vol. 8198, Lecture Notes Comput. Sci., pp. 53–62.
- [10] N. Tajbakhsh, C. Chi, S. R. Gurudu, and J. Liang, "Automatic polyp detection from learned boundaries," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, 2014, pp. 97–100.
- [11] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection using global geometric constraints and local intensity variation patterns," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2014*. New York: Springer, 2014, pp. 179–187.
- [12] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.
- [13] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A comparative study of texture features for the discrimination of gastric polyps in endoscopic video," in *Proc. 18th IEEE Symp. Comput.-Based Med. Syst.*, 2005, pp. 575–580.
- [14] L. A. Alexandre, N. Nobre, and J. Casteleiro, "Color and position versus texture features for endoscopic polyp detection," in *Proc. IEEE Int. Conf. BioMed. Eng. Informat.*, 2008, vol. 2, pp. 38–42.
- [15] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung fr die Medizin 2009*, ser. Informatik aktuell, H.-P. Meinzer, T. Deserno, H. Handels, and T. Tolxdorff, Eds. Berlin: Springer, 2009, pp. 346–350.
- [16] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, and X. Jiang, "Automatic detection of colorectal polyps in static images," *Biomed. Eng.: Appl. Basis Commun.*, vol. 23, no. 5, pp. 357–367, 2011.
- [17] J. Bernal, J. Snchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.
- [18] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 2, pp. II-465–II-468.
- [19] J. Bernal, J. Sánchez, and F. Vilarino, "Impact of image preprocessing methods on polyp localization in colonoscopy frames," in *Proc. 35th Annu. Int. Conf. IEEE EMBC*, 2013, pp. 7350–7354.
- [20] J. Bernal *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.
- [21] S. Y. Park, D. Sargent, I. Spofford, K. Vosburgh, and Y. A-Rahim, "A colon video analysis framework for polyp detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1408–1418, May 2012.
- [22] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, 2015, pp. 79–83.
- [23] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *Information Processing in Medical Imaging*. New York: Springer, 2015, pp. 327–338.
- [24] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. de Groen, "Part-based multi-derivative edge cross-section profiles for polyp detection in colonoscopy," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1379–1389, Jul. 2014.
- [25] S. Bae and K. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. : 2379–2393, Nov. 2015.
- [26] L. Zhao *et al.*, "Lines of curvature for polyp detection in virtual colonoscopy," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 12, no. 5, pp. 885–892, 2006.
- [27] J.-G. Lee, J. H. Kim, S. H. Kim, H. S. Park, and B. I. Choi, "A straightforward approach to computer-aided polyp detection using a polyp-specific volumetric feature in {CT} colonography," *Comput. Biol. Med.*, vol. 41, no. 9, pp. 790–801, 2011.
- [28] J. Ong and A.-K. Seghouane, "From point to local neighborhood: Polyp detection in CT colonography using geodesic ring neighborhoods," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1000–1010, Apr. 2011.
- [29] H. Zhu, Y. Fan, and Z. Liang, "Improved curvature estimation for shape analysis in computer-aided detection of colonic polyps," in *Virtual Colonoscopy and Abdominal Imaging. Computational Challenges and Clinical Opportunities*. Berlin: Springer, 2011, vol. 6668, pp. 9–14.
- [30] K. Suzuki, J. Zhang, and J. Xu, "Massive-training artificial neural network coupled with laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in ct colonography," *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1907–1917, Nov. 2010.
- [31] V. Van Ravesteijn *et al.*, "Computer-aided detection of polyps in CT colonography using logistic regression," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 120–131, Jan. 2010.
- [32] C. van Wijk, V. Van Ravesteijn, F. Vos, and L. van Vliet, "Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow," *IEEE Trans. Med. Imag.*, vol. 29, no. 3, pp. 688–698, Mar. 2010.

- [33] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, pp. 1–11, 2013.
- [34] I. N. Figueiredo, S. Kumar, and P. N. Figueiredo, "An intelligent system for polyp detection in wireless capsule endoscopy images," *VIPIIMAGE 2013*, p. 229, 2013.
- [35] S. Hwang and M. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 678–681.
- [36] A. Mamonov, I. Figueiredo, P. Figueiredo, and Y.-H. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Trans. Med. Imag.*, vol. 33, no. 7, pp. 1488–1502, Jul. 2014.
- [37] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy," in *Proc. 7th Int. Conf. Biomed. Eng. Informat.*, 2014, pp. 237–241.
- [38] B. Li and M. Q.-H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10952–10958, 2012.
- [39] R. Kwitt, "Learning pit pattern concepts for gastroenterological training," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2011*, ser. Lecture Notes in Computer Science, G. Fichtinger, A. Martel, and T. Peters, Eds. Berlin: Springer, 2011, vol. 6893, pp. 280–287.
- [40] S. Gross, S. Palm, J. J. W. Tischendorf, A. Behrens, C. Trautwein, and T. Aach, "Automated classification of colon polyps in endoscopic image data pp. 83 150W–83 150W-8," 2012.
- [41] T. Tamaki *et al.*, "Computer-aided colorectal tumor classification in {NBI} endoscopy using local features," *Med. Image Anal.*, vol. 17, no. 1, pp. 78–100, 2013.
- [42] P. Mordohai and G. Medioni, *Tensor Voting: A Perceptual Organization Approach to Computer Vision and Machine Learning*, ser. Synthesis Lectures on Image, Video, And Multimedia Processing. San Rafael, CA: Morgan Claypool, 2007.
- [43] CVC-Databascolon: A Database for Assessment of Polyp Detection [Online]. Available: <http://mv.cvc.uab.es/projects/colon-qa/cvc-colondb> 2011
- [44] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *Am. Statist.*, vol. 32, no. 1, pp. 12–16, 1978.
- [45] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. New York: Springer, 2013.
- [46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.
- [47] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.
- [48] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
- [49] T. Atherton and D. Kerbyson, "Using phase to represent radius in the coherent circle Hough transform," *IEE Colloquium Hough Transforms*, pp. 5–1, 1993.
- [50] N. Tajbakhsh, C. Chi, Q. Wu, S. R. Gurudu, and J. Liang, "Automatic assessment of image informativeness in colonoscopy," in *Abdominal Imaging. Computation and Clinical Applications*, 2014, Lecture Notes Comput. Sci..
- [51] J. Oh *et al.*, "Measuring objective quality of colonoscopy," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 9, pp. 2190–2196, Sep. 2009.

SUPPLEMENTARY MATERIAL

Input:

- A colonoscopy image I
- Trained random forest classifiers $RF_i|_{i=1}^2$

Output:

- polyp probability

Detection process

{Step 1: Collect edges and normals}

$$E = \{(e_i, n_i) \mid \angle n_i \in [0, \pi), i = 1, 2, \dots, N\}$$

{Step 2: Refine the edge map via classification}

for $i = 1, 2, \dots, N$ //for each edge

{Step 2.1: Extract a pair of patches}

//assuming two normals

$$\{n_i^1 \leftarrow n_i, n_i^2 \leftarrow -n_i\}$$

p_i^1 oriented patch with n_i^1 being the normal

p_i^2 oriented patch with n_i^2 being the normal

{Step 2.2: Extract features}

$$d_i^1 \leftarrow \mathcal{F}(p_i^1), d_i^2 \leftarrow \mathcal{F}(p_i^2)$$

{Step 2.3: Classify edges}

{A. Generate mid-level features}

$$f_i^1 \leftarrow RF_1(d_i^1)$$

$$f_i^2 \leftarrow RF_1(d_i^2)$$

$$f_i \leftarrow c(f_i^1, f_i^2) \text{ //concatenation}$$

{B. Fuse patch information}

$$S \leftarrow RF_2(f_i) \text{ //1x3 array}$$

if $S[1] > 0.5$

$$y_i \leftarrow 1, n_i^* \leftarrow n_i^1 \text{ //edge accepted}$$

else if $S[2] > 0.5$

$$y_i \leftarrow 2, n_i^* \leftarrow n_i^2 \text{ //edge accepted}$$

else

$$y_i \leftarrow 0 \text{ //edge rejected}$$

end if

end for

{Step 3: Localize polyps through voting}

{Step 3.1: Group edges}

$$V^k = \{e_i \mid y_i \notin 0 \wedge \frac{k\pi}{4} < \text{mod}(\angle n_i^*, \pi) < \frac{(k+1)\pi}{4}\}$$

{Step 3.2: Generate the voting map}

for $k = 0, 1, 2, 3$

$$\mathcal{M}^k = \sum_{v_i \in V^k} M_{v_i}(x, y)$$

end for

$$MVA \leftarrow \underset{x, y}{\operatorname{argmax}} \prod_{k=0}^3 \mathcal{M}^k \text{ //candidate location}$$

{Step 3.3: Compute the polyp probability}

$$p(\text{polyp} \mid MVA) \leftarrow \text{pr} \text{ // see Fig. 11}$$

Input:

- A set of training images $\mathcal{I} = \{I^1, I^2, \dots, I^m\}$
- Ground truth images $\mathcal{G} = \{G^1, G^2, \dots, G^m\}$
 - Truth make up $G(x, y) \in \{1, 2, 3, 4, 5\}$

1: polyp, 2: vessel, 3: lumen, 4: specular reflection, 5: random

Output:

- Trained random forest classifiers RF_1, RF_2

Learning process

{Layer 1: Train the 1st classifier}

Step0: collect labeled edges

$E = \{\}$ //set of edges

$L = \{\}$ //set of labels

$N = \{\}$ //set of desired normals

for $i=1\dots m$ //for each image

$$I_{bin} = \text{edge}(I^i)$$

$$E = E \cup \{e \mid I_{bin}(e_x, e_y) = 1\}$$

$$L = L \cup \{l = G^i(e_x, e_y) \mid I_{bin}(e_x, e_y) = 1\}$$

$$N = N \cup \{\vec{n}_{x,y}^* \mid I_{bin}(e_x, e_y) = 1\}$$

// n_i^* adjusted to point towards the ROI

end for

//Extract N_1 oriented patches using n_i^*

$$P = \{(p_i, l_i) \mid l_i \in \{1, 2, 3, 4, 5\}, i = 1 \dots N_1\}$$

Step1: Extract low-level features and train the 1st random forest classifier

$$d_i \leftarrow \mathcal{F}(p_i) \text{ //low-level feature vector}$$

$$\{(d_i, l_i) \mid l_i \in \{1, 2, 3, 4, 5\}, i = 1 \dots N_1\} \implies RF_1$$

{Layer 2: Train the 2nd classification layer}

Step2: Collect N_2 pairs of patches

$$\{(e_i, l_i, n_i) \mid l_i \in \{0, 1, 2\} \wedge \angle n_i \in [0, \pi), i = 1 \dots N_2\}$$

0: a non-polyp edge,

1: a polyp edge where n_i points toward the polyp,

2: a polyp edge where $-n_i$ points toward the polyp

for $i = 1 \dots N_2$ //for each edge

//Assume two normals

$$\{n_i^1 \leftarrow n_i, n_i^2 \leftarrow -n_i\}$$

p_i^1 oriented patch with n_i^1 being the normal

p_i^2 oriented patch with n_i^2 being the normal

Step3: Extract features

$$d_i^1 \leftarrow \mathcal{F}(p_i^1), d_i^2 \leftarrow \mathcal{F}(p_i^2)$$

//Apply the first classifier RF_1

$$f_i^1 \leftarrow RF_1(d_i^1), f_i^2 \leftarrow RF_1(d_i^2)$$

Step4: Concatenate features

$$f_i \leftarrow c(f_i^1, f_i^2)$$

end for

Step5: Train the 2nd random forest classifier

$$\{(f_i, l_i) \mid l_i \in \{0, 1, 2\}, i = 1 \dots, N_2\} \implies RF_2$$

Fig. S1: This pseudocode explains how the suggested polyp detection system operates when a new test image is provided.

Fig. S2: This pseudocode explains how the suggested edge classification pipeline is trained.

SHAPE GENERATOR

Our stochastic shape model is to generate objects that resemble the curvy heads of polyps. In our stochastic shape model, a shape is parameterized as a curve Γ with the position vector v :

$$\begin{aligned}\Gamma : \Omega &\rightarrow \mathbb{R}^2 \\ \Theta \rightarrow v(\Theta) &= [x(\Theta), y(\Theta)]^T\end{aligned}$$

where $\Theta = \{\theta, \mu_r, \sigma_r, \mu_a, \sigma_a\}$. In the above equation, $x(\Theta)$ and $y(\Theta)$ are determined as follows:

$$\begin{aligned}x(\Theta) &= x_C + r \times a \times \cos(\theta) \\ y(\Theta) &= y_C + r \times \sin(\theta)\end{aligned}\quad (10)$$

where $[x_C, y_C]^T$ is the shape center, θ is the angle with respect to the center, and radius r and aspect ratio a are drawn from $N(\mu_r, \sigma_r)$ and $N(\mu_a, \sigma_a)$, respectively. Since this model does not pose any constraint on the first and second derivatives of the contours, the resultant shapes are not smooth. To overcome this, we concatenate the $x(\Theta)$ s of the points on a contour to produce the signal X , and the $y(\Theta)$ s to produce the signal Y . We then apply 1D FFT on the generated signals, remove the high frequency components, and reconstruct the signals using the remaining low frequency coefficients, \hat{X} and \hat{Y} . To compensate for the unwanted shrinking caused by the smoothness process, we scale the smoothed shapes up to the original size by the following linear transformation:

$$\begin{aligned}x(\Theta) &= x_C + (\hat{x}(\Theta) - x_C) \frac{\sum_i X_i}{\sum_i \hat{X}_i} \\ y(\Theta) &= y_C + (\hat{y}(\Theta) - y_C) \frac{\sum_i Y_i}{\sum_i \hat{Y}_i}\end{aligned}$$

CONSTRUCTING REGRESSION MODELS

To construct a regression model for a fixed K and σ_F , we take the following steps:

- 1) We generate 3000 objects at three different scales corresponding to small, medium, and large polyps. To do so, we use our shape generator model and choose $\mu_r \in \{20, 40, 60\}$, and set $\sigma_r = 0.2\mu_r$, $\mu_a = 1$, and $\sigma_a = 0.1$.
- 2) We perform the voting scheme for each generated object based on the selected K and σ_F . For each object, the set of voters consists of all the edge pixels that form the object contour. To initiate the voting process, each voter must be assigned a voting direction. We first obtain the edge direction for an edge pixel using ball tensor voting and then determine its voting direction such that it points towards inside the corresponding object.
- 3) We find the representative isocontour of each voting map and then collect pairs of (d_i^{obj}, d_i^{iso}) from the boundaries of the objects and the representative isocontours.
- 4) We construct a regression model based on the collected pairs of (d_i^{obj}, d_i^{iso}) .

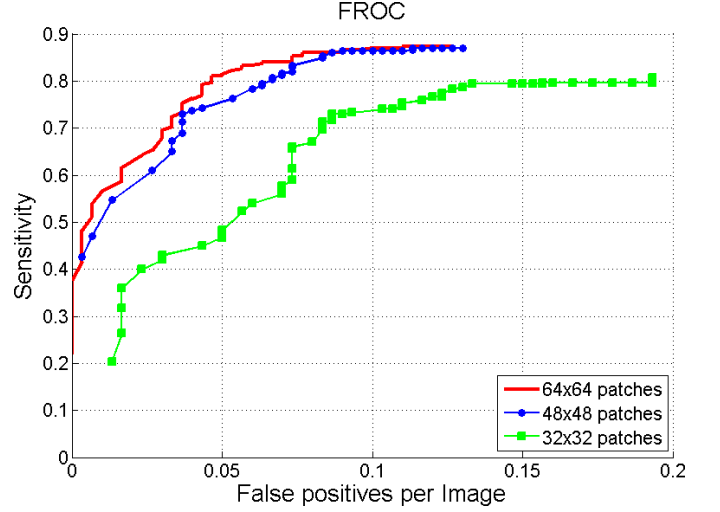


Fig. S3: FROC curves of our system for different sizes of patches. *CVC-ColonDB* is used for evaluation.

EFFECT OF PATCH SIZE

We have chosen patches of size 64x64 (32 pixels on each side of the boundary), because they were suitable for capturing the context in 512x512 colonoscopy frames. To investigate how the size of patches influences the performance, we have conducted new experiments based on 32x32 and 48x48 image patches. The results are shown in Fig. S3. As seen, 32x32 patches are not large enough to capture sufficient discriminatory features from the context, but this limitation is overcome by increasing the size of image patches.

A Comprehensive Computer-Aided Polyp Detection System for Colonoscopy Videos

Nima Tajbakhsh^{1(✉)}, Suryakanth R. Gurudu², and Jianming Liang¹

¹ Department of Biomedical Informatics, Arizona State University,
Scottsdale, AZ, USA

{Nima.Tajbakhsh, Jianming.Liang}@asu.edu

² Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, AZ, USA
Gurudu.Suryakanth@mayo.edu

Abstract. Computer-aided detection (CAD) can help colonoscopists reduce their polyp miss-rate, but existing CAD systems are handicapped by using either shape, texture, or temporal information for detecting polyps, achieving limited sensitivity and specificity. To overcome this limitation, our key contribution of this paper is to fuse all possible polyp features by exploiting the strengths of each feature while minimizing its weaknesses. Our new CAD system has two stages, where the first stage builds on the robustness of shape features to reliably generate a set of candidates with a high sensitivity, while the second stage utilizes the high discriminative power of the computationally expensive features to effectively reduce false positives. Specifically, we employ a unique edge classifier and an original voting scheme to capture geometric features of polyps in context and then harness the power of convolutional neural networks in a novel score fusion approach to extract and combine shape, color, texture, and temporal information of the candidates. Our experimental results based on FROC curves and a new analysis of polyp detection latency demonstrate a superiority over the state-of-the-art where our system yields a lower polyp detection latency and achieves a significantly higher sensitivity while generating dramatically fewer false positives. This performance improvement is attributed to our reliable candidate generation and effective false positive reduction methods.

1 Introduction

Colon cancer most often develop from colonic polyps. However, polyp grow slowly and it typically take years for polyps to develop into cancer, making colon cancer amenable to prevention. Colonoscopy is the preferred procedure for preventing colon cancer. The goal of colonoscopy is to find and remove polyps before turning into cancer. Despite its demonstrated utility, colonoscopy is not a perfect procedure. A recent clinical study [5] reports that a quarter of polyps are missed during colonoscopy. Computer-aided polyp detection can help colonoscopists reduce their polyp miss-rate, in particular, during long and back-to-back procedures where fatigue and inattentiveness may result in miss detection of polyps.

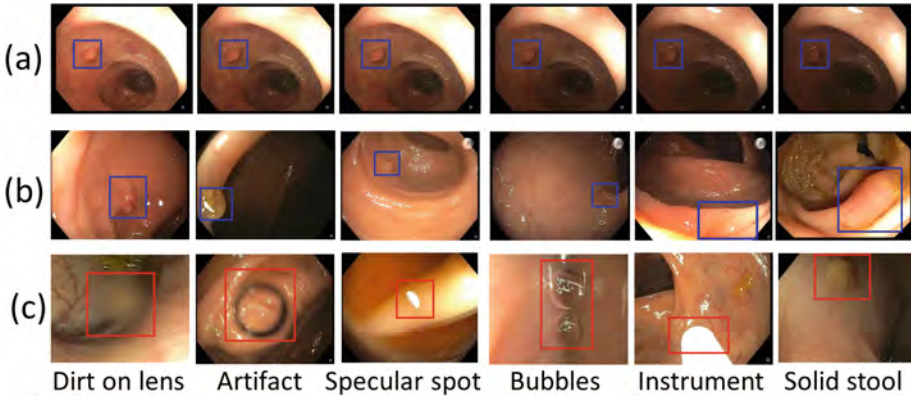


Fig. 1. Significant variation in visual characteristics of polyps. (a) Color and appearance variation of the same polyp due to varying lighting conditions. (b) Texture and shape variation among polyps. Note how the distance between the polyps and colonoscopy camera determines the availability of polyp texture. (c) Other polyp-like structures in the colonoscopic view (Color figure online).

However, designing a high-performance system for computer-aided polyp detection is challenging: (1) Polyps appear differently in color, and even the same polyp, as shown in Fig. 1(a), may look differently due to varying lighting conditions. (2) Polyps have large inter- and intra-morphological variations. As shown in Fig. 1(b), the shapes of polyps vary considerably from one to another. The intra-shape variation of polyps is caused by various factors, including the viewing angle of the camera and the spontaneous spasms of the colon. (3) Visibility of the texture on the surface of polyps is also varying due to biological factors and distance between the polyps and the colonoscopy camera. This can be seen in Fig. 1(b) where texture visibility decrease as the polyps distance from the capturing camera. The significant variations among polyps suggest that there is no single feature that performs the best for detecting all the polyps.

As a result, to achieve a reliable polyp detection system, it is critical to fuse all possible features of polyps, including shapes, color, and texture. Each of these features has strengths and weaknesses. Among these features, geometric shapes are most robust because polyps, irrespective of their morphology and varying levels of protrusion, have at least one curvilinear head at their boundaries. However, this property is not highly specific to polyps. This is shown in Fig. 1(c) where non-polyp structures exhibit similar geometric characteristics to polyps. Texture features have the weakness of limited availability; however, when visible, they can distinguish polyps from some non-polyp structures such as specular spots, dirt, and fecal matter. In addition, temporal information is available in colonoscopy and may be utilized to distinguish polyps from bubbles or other artifacts that only briefly appear in colonoscopy videos.

Our key contribution of this paper is an idea to exploit the strengths of each feature and minimize its weaknesses. To realize this idea, we have developed

a new system for polyp detection with two stages. The first stage builds on the robustness of shape features of polyps to reliably generate a set of candidate detections with a high sensitivity, while the second stage utilizes the high discriminative power of the computationally expensive features to effectively reduce false positive detections. More specifically, we employ a unique edge classifier coupled with a voting scheme to capture geometric features of polyps in context and then harness the power of convolutional deep networks in a novel score fusion approach to capture shape, color, texture, and temporal information of the candidates. Our experimental results based on the largest annotated polyp database demonstrate that our system achieves high sensitivity to polyps and generates significantly less number of false positives compared to state-of-the-art. This performance improvement is attributed to our reliable candidate generation and effective false positive reduction methods.

2 Related Works

Automatic polyp detection in colonoscopy videos has been the subject of research for over a decade. Early methods, e.g., [1, 3] for detecting colonic polyps utilized hand-crafted texture and color descriptors such as LBP and wavelet transform. However, given large color variation among polyps and limited texture availability on the surface of polyps (See Fig. 1), such methods could offer only a partial solution. To avoid such limitations, more recent techniques have considered temporal information [6] and shape features [2, 7, 9–11], reporting superior performance over the early polyp detection systems. Despite significant advancements, state-of-the-art polyp detection methods fail to achieve a clinically acceptable performance. For instance, to achieve the polyp sensitivity of 50%, the system suggested by Wang et al. [11] generates 0.15 false positives per frame or approximately 4 false positive per second. Similarly, the system proposed in [10], which is evaluated on a significantly larger dataset, generates 0.10 false positives per frame. Clearly, such systems that rely on a subset of polyp characteristics are not clinically viable—a limitation that this paper aims to overcome.

3 Proposed Method

Our computer-aided polyp detection system is designed based on our algorithms [7, 8, 10], consisting of 2 stages where the first stage utilizes geometric features to reliably generate polyp candidates and the second stage employs a comprehensive set of deep features to effectively remove false positives. Figure 2 shows a schematic overview of the suggested method.

3.1 Stage 1: Candidate Generation

Our unique polyp candidate generation method exploits the following two properties: (1) polyps have distinct appearance across their boundaries, (2) polyps,

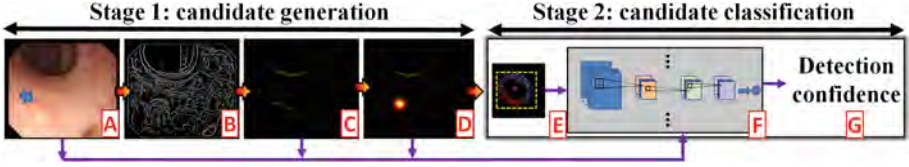


Fig. 2. Our system consists of 2 stages: candidate generation and classification. Given a colonoscopy frame (A), we first obtain a crude set of edge pixels (B). We then refine this edge map using a classification scheme where the goal is to remove as many non-polyp boundary pixels as possible (C). The geometric features of the retained edges are then captured through a voting scheme, generating a voting map whose maximum indicates the location of a polyp candidate (D). In the second stage, a bounding box is estimated for each generated candidate (E) and then a set of convolution neural networks—each specialized in one type of features—are applied in the vicinity of the candidate (F). Finally, the CNNs are aggregated to generate a confidence value (G) for the given polyp candidate.

irrespective of their morphology and varying levels of protrusion, feature at least one curvilinear head at their boundaries. We capture the first property with our image characterization and edge classification schemes, and capture the second property with our voting scheme.

Constructing Edge Maps. Given a colonoscopy image, we use Canny’s method to extract edges from each input channel. The extracted edges are then put together in one edge map. Next, for each edge in the constructed edge map, we determine edge orientation. The estimated orientations are later used for extracting oriented patches around the edge pixels.

Image Characterization. Our patch descriptor begins with extracting an oriented patch around each edge pixel. The patch is extracted so that the containing boundary is placed vertically in the middle of the patch. This representation allows us to capture desired information across the edges independent of their orientations. Our method then proceeds with forming 8×16 sub-patches all over the extracted patch. Each sub-patch has 50 % overlap with the neighboring sub-patches. For a compact representation, we compress each sub-patch into a 1D signal S by averaging intensity values along each column. We then apply a 1D discrete cosine transform (DCT) to the resulting signal:

$$C_k = \frac{2}{n} w(k) \sum_{i=0}^{n-1} S[i] \cos\left(\frac{2i+1}{2n} \pi k\right) \quad (1)$$

where

$$w(k) = 1/\sqrt{2}, k = 0 \text{ and } w(k) = 1, 1 \leq k \leq n-1.$$

With the DCT, the essential information of the intensity signal can be summarized in a few coefficients. We discard the DC component C_0 because the average patch intensity is not a robust feature—it is affected by a constant change in patch intensities. However, the next 3 DCT coefficients $C_1 - C_3$ are more reliable and provide interesting insight about the intensity signal. C_1 measures whether the average patch intensity along the horizontal axis is monotonically decreasing (increasing) or not, C_2 measures the similarity of the intensity signal against a valley (ridge), and finally C_3 checks for the existence of both a valley and a ridge in the signal. The higher order coefficients $C_4 - C_{15}$ may not be reliable for feature extraction because of their susceptibility to noise and other degradation factors in the images.

The selected DCT coefficients $C_1 - C_3$ are still undesirably proportional to linear illumination scaling. We therefore apply a normalization treatment. Mathematically,

$$C_i = \frac{C_i}{\sqrt{C_1^2 + C_2^2 + C_3^2}}, i = 1, 2, 3.$$

Note that we use the norm of the selected coefficients for normalization rather than the norm of entire DCT coefficients. By doing so, we can avoid the expensive computation of all the DCT components. The final descriptor for a given patch is obtained by concatenating the normalized coefficients selected from each sub-patch.

The suggested patch descriptor has 4 advantages. First, our descriptor is fast because compressing each sub-patch into a 1D signal eliminates the need for expensive 2D DCT and that only a few DCT coefficients are computed from each intensity signal. Second, due to the normalization treatment applied to the DCT coefficients, our descriptor achieves invariance to linear illumination changes, which is essential to deal with varying lighting conditions (see Fig. 1). Third, our descriptor is rotation invariant because the patches are extracted along the dominant orientation of the containing boundary. Fourth, our descriptor handles small positional changes by selecting and averaging overlapping sub-patches in both horizontal and vertical directions.

Edge Classification. Our classification scheme has 2 layers. In the first layer, we learn a discriminative model to distinguish between the boundaries of the structures of interest and the boundaries of other structures in colonoscopy images. The structures of interest consists of *p*olyps, *v*essels, *l*umen areas, and *s*pecular reflections. Specifically, we collect a stratified set of $N_1 = 100,000$ oriented patches around the boundaries of structures of interest and r random structures in the training images, $S^1 = \{(p_i, y_i) | y_i \in \{p, v, l, s, r\}, i = 1, 2, \dots, N_1\}$. Once patches are extracted, we train a five-class random forest classifier with 100 fully grown trees. The resulting probabilistic outputs can be viewed as the similarities between the input patches and the predefined structures of interest. Basically, the first layer receives low-level image features from our patch descriptor and then produces mid-level semantic features.

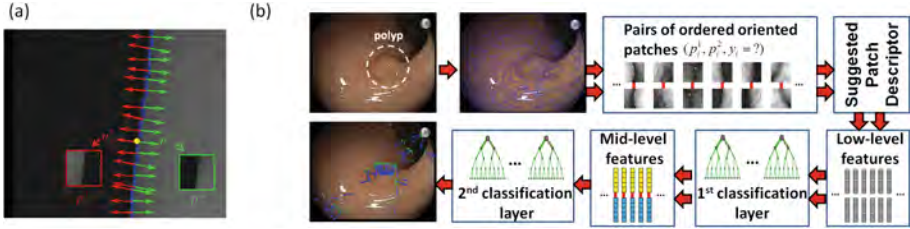


Fig. 3. (a) A pair of image patches $\{p_i^1, p_i^2\}$ extracted from an edge pixel. The green and red arrows show the two possible normal directions $\{n_i^1, n_i^2\}$ for a number of selected edges on the displayed boundary. The normal directions are used for patch alignment. (b) The suggested edge classification scheme given a test image. The edges that have passed the classification stage are shown in green. The inferred normal directions are visualized with the blue arrows for a subset of the retained edges (Color figure online).

In the second layer, we train a 3-class random forest classifier with 100 fully grown trees. Specifically, we collect $N_2 = 100,000$ pairs of oriented patches, of which half are randomly selected from the polyp boundaries and the rest are selected from random non-polyp edge segments. For an edge pixel at angle θ , one can obtain two oriented image patches $\{p_i^1, p_i^2\}$ by interpolating the image along the two possible normal directions $\{n_i^1, n_i^2\}$. As shown in Fig. 3(a), for an edge pixel on the boundary of a polyp, only one of the normal directions points to the polyp region. Our classification scheme operates on each pair of patches with two objectives: (1) to classify the underlying edge into polyp and non-polyp categories, and (2) to determine the desired normal direction among n_i^1 and n_i^2 such that it points towards the polyp location. Henceforth, we refer to the desired normal direction as “voting direction”.

Once image pairs are collected, we order the patches $\{p_i^1, p_i^2\}$ within each pair according to the angles of their corresponding normal vectors, $\angle n_i^1 < \angle n_i^2$. In this way, the patches are represented in a consistent order. Each pair of ordered patches is then assigned a label $y_i \in \{0, 1, 2\}$, where “0” indicates that the underlying edge does not lie on a polyp boundary, “1” indicates that the edge lies on a polyp boundary and that n_i^1 is the voting direction, and “2” indicates that the edge lies on a polyp boundary but n_i^2 shows the voting direction. Mathematically, $S^2 = \{(p_i^1, p_i^2, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, \dots, N_2\}$. To generate semantic features, each pair of ordered patches undergoes the image characterization followed by the first classification layer. The resulting mid-level features are then concatenated to form a feature vector f_i . This process is repeated for N_2 pairs of ordered patches, resulting in a labeled feature set, $\{(f_i, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, \dots, N_2\}$, which is needed to train the second classifier. We train a 3-class classifier to learn both edge labels and the voting directions (embedded in y_i). Figure 3(b) illustrates how the suggested edge classification scheme operates given a test image.

Candidate Localization. Our voting scheme is designed to generate polyp candidates in regions surrounded by curvy boundaries. The rationale is such

boundaries can represent the heads of polyps. In our voting scheme, each edge that has passed the classification stage, casts a vote along its voting direction (inferred by the edge classifier). The vote cast by the voter v at a receiver pixel $r = [x, y]$ is computed as

$$M_v(x, y) = \begin{cases} C_v \exp(-\frac{\|\vec{v}\vec{r}\|^2}{\sigma_F}) \cos(\angle \vec{n}^* \vec{v}\vec{r}), & \text{if } \angle \vec{n}^* \vec{v}\vec{r} < \pi/2 \\ 0, & \text{if } \angle \vec{n}^* \vec{v}\vec{r} \geq \pi/2 \end{cases} \quad (2)$$

where the exponential and cosinusoidal functions enable smooth vote propagation, which we will later use to estimate a bounding box around each generated candidate. In Eq. 2, C_{v_i} is the classification confidence, $\vec{v}\vec{r}$ is the vector connecting the voter and receiver, σ_F controls the size of the voting field, and $\angle \vec{n}^* \vec{v}\vec{r}$ is the angle between the voting direction \vec{n}^* and $\vec{v}\vec{r}$. Figure 4(a) shows the voting field for an edge pixel lying at 135 degree. As seen, due to the condition set on $\angle \vec{n}^* \vec{v}\vec{r}$, the votes are cast only in the region pointed by the voting direction.

It is essential for our voting scheme to prevent vote accumulation in the regions that are surrounded by low curvature boundaries. For this purpose, our voting scheme first groups the voters into 4 categories according to their voting directions, $V^k = \{v_i | \frac{k\pi}{4} < \text{mod}(\angle n_i^*, \pi) < \frac{(k+1)\pi}{4}\}$, $k = 0 \dots 3$. Our voting scheme then proceeds by accumulating votes of each category in a separate voting map. To produce the final voting map, we multiply the accumulated votes generated in each category. A polyp candidate is then generated where the final voting map has the maximum vote accumulation (MVA). Mathematically,

$$MVA = \arg \max_{x, y} \prod_{k=0}^3 \sum_{v \in V^k} M_v(x, y). \quad (3)$$

Comparing Fig. 4(b) and (c) clarifies how the suggested edge grouping mitigates vote accumulation between parallel lines, assigning higher temperature to only regions surrounded by curvy boundaries. Another important characteristic of our voting scheme is the utilization of voting directions. As shown in Fig. 4(d), casting votes along both possible normal directions can result in mislocalized candidates; however, incorporating voting directions allows for more accurate candidate localization (Fig. 4(e)).

3.2 Stage 2: Candidate Classification

Our candidate classification method begins with estimating a bounding box around each polyp candidate followed by a novel score fusion framework based on convolutional neural networks (CNNs) [4] to assign a confidence value to each generated candidate.

Bounding Box Estimation. To measure the extent of the polyp region, we estimate a narrow band around each candidate, so that it contains the voters that have contributed to vote accumulation at the candidate location. In other

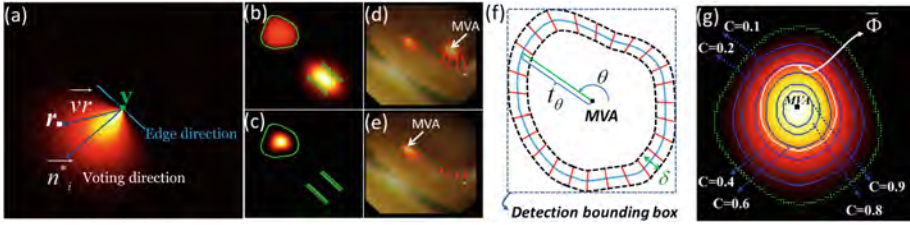


Fig. 4. (a) The generated voting map for an edge pixel lying at 135 degree. (b) Without edge grouping, all the votes are accumulated in one voting map, which results in undesirable vote accumulation between the parallel lines. (c) With the suggested edge grouping, higher temperature is assigned to only within the curvy boundaries. (d) Casting votes along both possible normal directions can result in a candidate placed outside the polyp region. (e) Casting votes only along the inferred voting directions results in a successful candidate localization. (f) A narrow band is used for estimating a bounding box around candidates. (g) A synthetic shape and its corresponding voting map. The isocontours and the corresponding representative isocontour are shown in blue and white, respectively (Color figure online).

words, the desired narrow band will enclose the polyp boundary and thus can be used to estimate a bounding box around the candidate location. As shown in Fig. 4(f), the narrow band B consists of a set of radial lines ℓ_θ parameterized as $\ell_\theta : MVA + t[\cos(\theta), \sin(\theta)]^T, t \in [t_\theta - \frac{\delta}{2}, t_\theta + \frac{\delta}{2}]$, where δ is the bandwidth, and t_θ is the distance between the candidate location and the corresponding point on the band skeleton at angle θ . Once the band is formed, the bounding box is localized so that it fully contains the narrow band around the candidate location (see Fig. 4(f)). The bounding box will be later used for data augmentation where we extract patches in multiple scales around the polyp candidates.

To estimate the unknown δ and t_θ for a given candidate, we use the isocontours of the corresponding voting map. The isocontour Φ_c of the voting map V is defined as $\Phi_c = \{(x, y) | V(x, y) = c \times M\}$ where M denotes the maximum of the voting map and c is a constant between 0 and 1. As shown in Fig. 4(g), the isocontours of the voting map, particularly those located farther away from the candidate, have the desirable feature of following the shape of the actual boundary from which the votes have been cast at the candidate location. Therefore, one can estimate the narrow band's parameters from the isocontours such that the band encloses the object's boundary. However, in practice, the shape of far isocontours are undesirably influenced by other nearby voters in the scene. We therefore obtain the representative isocontour $\bar{\Phi}$ by computing the median shape of the isocontours of the voting map (see Fig. 4(g)). We have experimented with different sets of isocontours and found out that as long as their parameter c is uniformly selected between 0 and 1, the resulting representative isocontour serves the desired purpose.

Let d_{iso}^i denotes the distance between the i^{th} point on the representative isocontour $\bar{\Phi}$ and the candidate location. We use d_{iso}^i to predict d_{obj}^i , the distance

between the corresponding point on the object boundary and the candidate location. For this purpose, we employ a second order polynomial regression model

$$d_{obj}^i = b_0 + b_1(d_{iso}^i) + b_2(d_{iso}^i)^2, \quad (4)$$

where b_0 , b_1 , and b_2 are the regression coefficients that are estimated using a least square approach. Once the model is constructed, we take the output of the model d_{obj} at angle θ with respect to MVA as t_θ and the corresponding prediction interval as the bandwidth δ .

Probability Assignment. We propose a score fusion framework based on convolutional neural networks (CNNs) that can learn and integrate color, texture, shape, and temporal information of polyps in multiple scales for more accurate candidate classification. We choose to use CNNs because of their superior performance in major object detection challenges. The attractive feature of CNNs is that they jointly learn a multi-scale set of image features and a discriminative classifier during a supervised training process. While CNNs are known to learn discriminate patterns from raw pixel values, it turns out that preprocessing and careful selection of the input patches can have a significant impact on the performance of the subsequent CNNs. Specifically, we have found out that partial illumination invariance achieved by histogram equalizing the input patches significantly improves the performance of the subsequent CNNs and that curse of dimensionality caused by patches with more than 3 channels results in CNNs with inferior performance.

Considering these observations, we propose a 3-way image presentation that is motivated by the three major types of polyp features suggested in the literature: (1) for color and texture features, we collect histogram-equalized color patches P_C around each polyp candidate; (2) for temporal features, we form 3-channel patches P_T by stacking histogram-equalized gray channel of the current frame and that of the previous 2 frames; (3) for shape in context, we form 3-channel patches P_S by stacking the gray channel of the current frame and the corresponding refined edge channel and voting channel produced in the candidate generation stage (see Fig. 2).

We collect the three sets of patches P_C , P_T , and P_S from candidate locations in the training videos, label each individual patch depending on whether the underlying candidate is a true or false positive, and then train a CNN for each set of the patches. Figure 5(a) shows the test stage of the suggested score fusion framework. Given a new polyp candidate, we collect the three sets of patches in multiple scales and orientations around the candidate location, apply each of the trained CNNs on the corresponding patches, and take the maximum response for each CNN, resulting in three probabilistic scores. The final classification confidence is computed by averaging the resulting three scores.



Fig. 5. (a) The test stage of the suggested score fusion framework. (b) Network layout used for training the deep convolution networks.

4 Experiments

For evaluation, we have used 40 short colonoscopy videos. We have randomly halved the database at video level into the training set containing 3800 frames with polyps and 15100 frames without polyps, and the test set containing 5700 frames with polyps and 13200 frames without polyps. Each colonoscopy frame in our database comes with a binary ground truth image. For performance evaluation, we consider a detection as a true (false) positive if it falls inside (outside) the white region of the ground truth image.

Our candidate generation stage yielded a sensitivity of 73.6% and 0.8 false positives/frame. For candidate classification, we used Krizhevsky’s GPU implementation [4] of CNNs. With data augmentation, we collected 400,000 32×32 patches for P_C , P_T , and P_S where half of the patches were extracted around false positive candidates and the rest around true positive candidates. Specifically, for a candidate with an $N \times N$ bounding box, we extracted patches at three scales $sN \times sN$ with $s \in \{1, 1.2, 1.4\}$ and then resized them to 32×32 patches. Furthermore, we performed data augmentation [4] by extracting patches at multiple orientations and translation in each given scale. We have used the layout shown in Fig. 5(b) for all the CNNs used in this paper.

Figure 6(a) shows FROC analysis of the suggested system. As seen, our system based on the suggested score fusion approach shows a relatively stable performance over a wide range of voting fields. For comparison, we have also reported the performance of our system based on individual CNNs trained using color patches (P_C), temporal patches (P_T), and shape in context patches (P_S). We have also experimented with the channel fusion approach where color, shape, and temporal patches are stacked for each polyp candidate followed by training one CNN for the resulting 9-channel training patches. To avoid clutter in the figure, only their best performance curves obtained by $\sigma_F = 70$ are shown. As seen in Fig. 6(a), the proposed score fusion framework yields the highest performance, achieving 50% sensitivity at 0.002 FPs/frame, outperforming [10] with 0.10 FPs/frame at the same sensitivity.

FROC analysis is widely used for evaluating computer-aided detection systems designed for static datasets such as CT scans and mammograms. However, for temporal or sequence-based datasets such as colonoscopy videos, it has the

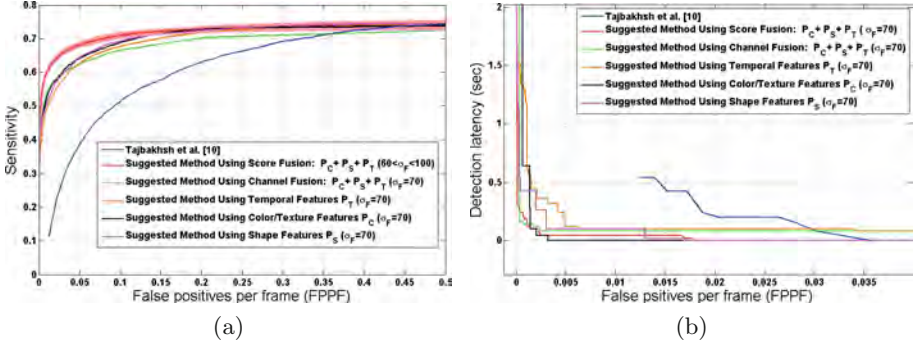


Fig. 6. (a) Analysis of FROC. (b) Analysis of polyp detection latency.

drawback of excluding the factor of time. While it is desirable for a polyp detection system to detect as many polyp instances as possible, it is also important to measure how quickly a polyp is detected after it appears in the video. We therefore employ a new performance curve [8] that measures the polyp detection latency with respect to the number of false positives. Briefly, if t_1 denotes the arrival frame of the polyp, t_2 denotes the frame in which the polyp is detected, and fps is the frame rate of the video, the detection latency is then computed as $\Delta T = (t_2 - t_1)/fps$. As with FROC, we change a threshold on the detection confidences and then at each operating point measure the median polyp detection latency of the test positive shots and the number of false positives in the entire test set. As seen in Fig. 6(b), different variations of our system yield significantly less number of false positives than our previous work [10] at nearly all operating points.

On a desktop computer with a 2.4 GHz quad core Intel and an Nvidia GeForce GTX 760 video card, our system processes each image at 2.65s, which is significantly faster than [11] with run-time of 7.1s and [2] with run-time of 19s. We should note that a very large fraction of the computation time (2.6s) is caused by the candidate generation stage and that the candidate classification based on CNNs is extremely fast because CNNs are only applied to the candidate location in each frame. We expect a significant speedup of our system using parallel computing optimization.

5 Conclusion

We proposed a new computer-aided polyp detection system for colonoscopy videos. Our system was based on context-aware shape features to generate a set of candidates and convolutional neural networks to reduce the generated false positives. We evaluated our system using the widely-used FROC analysis, achieving 50 % sensitivity at 0.002 FPs/frame, outperforming state-of-the-art systems [10, 11], which generate 0.15 FPs/frame and 0.10 FPs/frame at 50 % sensitivity, respectively. We also evaluated our system using a latency analysis,

demonstrating a significantly lower polyp detection latency than [10] particularly in low false positive rates.

Acknowledgment. This research has been supported by an ASU-Mayo Clinic research grant.

References

1. Alexandre, L.A., Nobre, N., Casteleiro, J.: Color and position versus texture features for endoscopic polyp detection. In: International Conference on BioMedical Engineering and Informatics, BMEI 2008, vol. 2, pp. 38–42. IEEE (2008)
2. Bernal, J., Snchez, J., Vilario, F.: Towards automatic polyp detection with a polyp appearance model. *Pattern Recogn.* **45**(9), 3166–3182 (2012)
3. Karkanis, S.A., Iakovidis, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M.: Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Trans. Inform. Technol. Biomed.* **7**(3), 141–152 (2003)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012), <https://code.google.com/p/cuda-convnet/>
5. Leufkens, A., van Oijen, M., Vleggaar, F., Siersema, P.: Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* **44**(05), 470–475 (2012)
6. Park, S.Y., Sargent, D., Spofford, I., Vosburgh, K., A-Rahim, Y.: A colon video analysis framework for polyp detection. *IEEE Trans. Biomed. Eng.* **59**(5), 1408–1418 (2012)
7. Tajbakhsh, N., Chi, C., Gurudu, S.R., Liang, J.: Automatic polyp detection from learned boundaries. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 97–100. IEEE (2014)
8. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on.* pp. 79–83. IEEE (2015)
9. Tajbakhsh, N., Gurudu, S.R., Liang, J.: A classification-enhanced vote accumulation scheme for detecting colonic polyps. In: Yoshida, H., Warfield, S., Vannier, M.W. (eds.) *Abdominal Imaging 2013. LNCS*, vol. 8198, pp. 53–62. Springer, Heidelberg (2013)
10. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automatic polyp detection using global geometric constraints and local intensity variation patterns. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) *MICCAI 2014, Part II. LNCS*, vol. 8674, pp. 179–187. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-10470-6_23
11. Wang, Y., Tavanapong, W., Wong, J., Oh, J., de Groen, P.: Part-based multi-derivative edge cross-section profiles for polyp detection in colonoscopy. *IEEE J. Biomed. Health Inform.* **PP**(99), 1–1 (2013)

Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge

Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilango Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debar, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, and Aymeric Histace*

Manuscript received October 30, 2016; revised January 27, 2017; accepted January 31, 2017. Date of publication February 2, 2017; date of current version June 1, 2017. This work was supported in part by ASU-Mayo Clinic partnerships, in part by the Spanish Government through the Funded Project iVENDIS under Project DPI2015-65286-R, in part by FSEED, in part by the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya under Grant 2014-SGR-1470 and Grant 2014-SGR-135, in part by par SATT IdInnov (France) through the Project Smart Videocolonoscopy under Grant 186, and in part by the European Union through the ERC starting grant COMBIOSCOPY under the New Horizon Framework Programme under Grant ERC-2015-StG-37960. (Jorge Bernal and Nima Tajkbaksh share first co-authorship. Aymeric Histace and Jianming Liang share last co-authorship) Asterisk indicates corresponding author.

J. Bernal and F. J. Sánchez are with the Computer Science Department, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain, and also with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.

N. Tajkbaksh and J. Liang are with Arizona State University, Tempe, AZ 85281 USA.

B. J. Matuszewski is with the School of Engineering, University of Central Lancashire, Preston PR1 2HE, U.K.

H. Chen and L. Yu are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

Q. Angermann and O. Romain are with ETIS, ENSEA, CNRS, University of Cergy-Pontoise, F95000 Cergy, France.

B. Rustad is with Oslo University Hospital, 0379 Oslo, Norway, and also with OmniVision, University of Oslo, 0313 Oslo, Norway.

I. Balasingham is with Oslo University Hospital, 0379 Oslo, Norway.

K. Pogorelov is with the Media Performance Group, Simula Research Laboratory, and University of Oslo, 0313 Oslo, Norway.

S. Choi is with Seoul National University, Seoul 08826, South Korea.

Q. Debar is with the University of Nice-Sophia Antipolis, 06000 Nice, France.

L. Maier-Hein is with the Junior Group Computer-assisted Interventions, German Cancer Research Center, 69120 Heidelberg, Germany.

S. Speidel is with the Institute for Anthropomatics, Karlsruhe Institute of Technology, 76021 Karlsruhe, Germany.

D. Stoyanov and P. Brandao are with the Centre for Medical Image Computing and Department of Computer Science, University College London, London WC1E 6BT, U.K.

H. Córdova, C. Sánchez-Montes, and G. Fernández-Esparrach are with the Endoscopy Unit, Gastroenterology Department, Hospital Clínic, IDIBAPS, CIBEREHD, University of Barcelona, Barcelona, Spain.

S. R. Gurudu is with the Division of Gastroenterology and Hepatology, Mayo Clinic, Scottsdale, AZ 85259 USA.

X. Dray is with ETIS, ENSEA, CNRS, University of Cergy-Pontoise, F95000 Cergy, France, and also with the Lariboisière Hospital-APHP, 75000 Paris, France.

*A. Histace is with ETIS, ENSEA, CNRS, University of Cergy-Pontoise, F95000c Cergy, France.

Digital Object Identifier 10.1109/TMI.2017.2664042

Abstract— Colonoscopy is the gold standard for colon cancer screening though some polyps are still missed, thus preventing early disease detection and treatment. Several computational systems have been proposed to assist polyp detection during colonoscopy but so far without consistent evaluation. The lack of publicly available annotated databases has made it difficult to compare methods and to assess if they achieve performance levels acceptable for clinical use. The Automatic Polyp Detection sub-challenge, conducted as part of the Endoscopic Vision Challenge (<http://endovis.grand-challenge.org>) at the international conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2015, was an effort to address this need. In this paper, we report the results of this comparative evaluation of polyp detection methods, as well as describe additional experiments to further explore differences between methods. We define performance metrics and provide evaluation databases that allow comparison of multiple methodologies. Results show that convolutional neural networks are the state of the art. Nevertheless, it is also demonstrated that combining different methodologies can lead to an improved overall performance.

Index Terms— Endoscopic vision, polyp detection, hand-crafted features, machine learning, validation framework.

I. INTRODUCTION

THIS paper introduces the results and main conclusions of the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy, conducted as part of the *Endoscopic Vision Challenge* (<http://endovis.grand-challenge.org>). More precisely, we present a validation study comparing the performance of different polyp detection methods covering different methodologies proposed by participating teams, providing an insight analysis of their detection yield. In this section, we introduce both clinical and technical contexts.

A. Clinical Context

Colorectal cancer (CRC) is the third largest cause of cancer deaths in the United States among men and women, and it is expected to have resulted in about 49,196 deaths in 2016 in the USA [1]. CRC arises from adenomatous polyps (or adenomas), that are growths of glandular tissue originating from the colonic mucosa. Though adenomas are initially benign, they

might become malignant over time and spread to adjacent and distant organs such as lymph nodes, liver or lungs, being ultimately responsible for complications and death [2].

CRC prevention is first based on the detection of at-risk patients: those with symptoms (such as hematochezia and anemia), those with positive screening tests (such as a fecal occult blood test or a fecal immunochemical test), and those with a past history of adenoma or with a family history of advanced adenoma or CRC. In these groups of patients, a colonoscopy is proposed to detect polyps before any malignant transformation or at an early cancer stage. This stage refers to the most superficial colon layers, with no deep invasion, and it is associated with a 5-year survival rate over 90% [1], [3]. If any polyp found is characterized as a likely adenoma, its removal should be considered to confirm the diagnosis, to set its histological stage and to confirm its complete removal, giving clinicians clues to determine the need and timing of the next colonoscopy [4].

Though colonoscopy is the gold standard for colon screening, other alternatives, such as CT colonography [5] or wireless capsule endoscopy (WCE) [6], are also used to search for polyps. They are less invasive to patients and do not present perforation risk. Though, as colonoscopy, they require bowel preparation. Nevertheless in these cases, if a polyp is found, a colonoscopy must be considered to remove the suspicious lesion. These alternatives have specific limitations that may affect the outcome of the screening. For instance, CT colonography has a low small lesions (5 mm or less) detection rate due to resolution constraints [7] and it implies using ionising radiation. WCE allows to detect all kind of lesions but their observation depends on whether they are recorded during the progress of the camera through the gastrointestinal tract or not. Moreover, its diagnostic yield is highly dependent on the cleanliness of the colon (whereas colonoscopy has some in-situ lavage capabilities). Last but not least, the analysis of the information provided by WCE can be highly time-consuming, as the recorded videos can last up to 8 hours [8].

Colonoscopy presents some drawbacks, polyp miss-rate being the most important among these. Colonoscopy rarely misses polyps bigger than 10 mm, but the miss-rate increases significantly with smaller sized and/or flat polyps [9], [10]. It has also to be noted that colonoscopies are seldom recorded, so a new procedure must be performed to revisit explored areas.

The outcome of the colonoscopy exploration depends on: 1) bowel preparation [11]; 2) specific choice of endoscope and video processor, affecting image quality and preventing the use of certain image enhancing tools; 3) clinicians' skills, as both endoscopist's experience and his/her actual concentration during the intervention may influence the degree of procedure completion (reaching the cecum or not) and the percentage of the colon that has been explored [12], [13] and 4) patient-specific issues, as due to colon movements and the appearance of folds and angulations during the exploration, some parts of the colon which may potentially present polyps may not be reached [9]. Moreover, patients' personal and family history can increase the risk of having a polyp and, in this case, the exploration should be even more thorough.

B. Technical Strategies to Improve Polyp Detection Rate

Apart from the continuous improvement of clinicians' skills through training programs and practice [14], technical efforts are being undertaken to improve colonoscopy's outcome. We clustered them into two groups: improvement of devices and the development of computational support systems.

Amongst the device improvements, the following should be highlighted: 1) increase in image resolution and, consequently, textural information; 2) the use of wide-angle cameras showing more colon wall surface; 3) the development of zooming and magnification techniques [15] and 4) the development of new imaging methodologies such as autofluorescence imaging [16] or virtual chromoendoscopy (Olympus' Narrow Band Imaging [17], Fujinon's FICE [18] or Pentax's i-Scan [19]). This last group of techniques modify how the scene is observed by improving the contrast of endoluminal scene elements, which may help in lesion detection and also with in-vivo lesion diagnosis due to the enhanced visualization of lesion tissues [20]. These advances have fostered the cooperation between clinicians and computer scientists in the development and validation of computer-aided support systems for colonoscopy, aimed to help clinicians in all stages of CRC diagnosis. A significant part of this effort has been focused on computer assisted polyp detection. As it is indicated in [21], cooperation between technologists and clinicians is essential to develop clinically useful solutions, with both these groups understanding challenges and limitations in their respective domains.

Automatic polyp detection in colonoscopy videos has been an active research topic during the last 20 years and several approaches have been proposed. We present a review of the most relevant methods in Section II but, to the best of our knowledge, none of them has been adopted for a routine patient treatment. There might be several reasons behind this. First of all, in order for a given method to be clinically useful, it has to meet real time constraints; e.g. for videos acquired at 25 frames per second (fps) the maximum time available to process each image frame should be under 40ms. Secondly, some of them are built from a theoretical model of a polyp appearance [14], [22] and therefore limited to only certain polyp morphologies, which may not translate to the actual scene where polyp appearance varies greatly. Thirdly, the majority of methods are mainly focused on the polyps and they do not consider the presence of other elements such as folds, blood vessels or the lumen that can affect methods' performance [14]. Last but not least, some of these methods have been only trained and tested on selected good quality still image frames. The lack of temporal coherence and the great variability in polyp appearance due to camera progression and visibility conditions might impact their performance in the full sequences analysis, as they might cause instability in their response against similar stimuli.

Computational methods also have to deal with additional colonoscopy-specific challenges. For instance, they should consider the impact of image artifacts generated due to scene illumination (specular highlights, overexposed regions) or to specific configuration of the videoprocessor attached to the colonoscopy, which might overlay information over the

scene view. These artifacts, apart from altering the view of the scene, might not be stable within consecutive frames and therefore methods should both compensate their impact on the individual frame polyp detection and tracking in the full sequence analysis. Additionally, though an effort is made to ensure an adequate bowel preparation, some particles may still appear which, in some cases, could lead to false detections when isolated or to occlusion leading to miss detection or localization errors. As mentioned before, these methods have to cope with a great degree of variability in polyp appearance which depends on illumination conditions, camera position and on clinician skills when progressing through the colon. Finally, available methods have been typically validated on small and restricted databases, under specific endoscope device conditions (brand and resolution), in some cases even covering only one specific polyp type, shape or morphology hindering their actual performance in a more generic setting.

C. Motivation of the Comparison Study

Unfortunately, the lack of a common validation framework, which is a frequent problem in medical and endoscopy image analysis [21], has limited the effectiveness of the comparison between existing approaches, making it difficult to determine which of them could have actual advantage in clinical use. To cope with this, efforts have been made on publishing fully annotated databases [14], [22] and on organizing challenges as part of international conferences (ISBI, MICCAI), which offer a basis to discuss validation strategies.

Considering this and taking inspiration from recent works on quantitative comparative methods' analysis in areas such as laparoscopic 3D Surface Reconstruction [23] or liver segmentation [24], we present in this paper a complete validation study of polyp detection methods performed as part of the 2015 MICCAI sub-challenge on Automatic Polyp Detection. This sub-challenge was organized jointly by three research teams: 1) Computer Vision Center/Universitat Autònoma de Barcelona and Hospital Clinic from Barcelona, Spain (CVC-CLINIC); 2) ETIS Lab (ENSEA/CNRS/University of Cergy-Pontoise) and Lariboisière Hospital-APHP at Paris, France (ETIS-LARIB), and 3) Arizona State University and Mayo Clinic, USA (ASU-Mayo).

The objective of this paper is to present a comparative study of polyp detection methods under a newly proposed validation framework. This validation framework was firstly introduced as part of MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy and we present in this paper the results of the mentioned sub-challenge. Beyond this, we also propose additional experiments to assess even more in-depth the performance of an automatic polyp detection method. These new experiments are focused on exploring the actual clinical applicability of a given method by assessing up to what extent they are affected by some of the technical and clinical challenges reported in the literature or whether they incorporate temporal coherence features or not. Finally we also go beyond the individual analysis of methods and propose combination strategies in order to study whether a combination method may lead to improved individual performance.

The remainder of the paper is structured as follows: In Section II we present the methods proposed by each of the participating teams in the challenge, including them in the context of existing published methods. In Section III we describe the complete validation framework. Results from the comparative study are presented in Section IV. Section V provides an in-depth analysis of the results and discusses some topics related to challenge organization. Finally, the concluding remarks are drawn in Section VI.

II. AUTOMATIC POLYP DETECTION METHODS

A. Historical Review of Computational Polyp Detection Methods

After analyzing approaches reported in the literature, we propose to cluster methods into three groups: 1) **hand-crafted**; 2) **end-to-end learning** and 3) **hybrid** approaches. This taxonomy represents the different historical trends of polyp detection methods, as in early 2000s, the majority of the methods used a given texture descriptor to guide a classification method but, subsequently, some researchers decided to go for hand-crafted features, aiming at a real time implementation. As technology evolved and the computational capabilities increased, techniques such as neural networks that were developed in the past and abandoned due to excessive computational cost have now resurfaced.

Regarding hand-crafted methods, the majority are based on exploiting low-level image processing methods to obtain candidate polyp boundaries (using Hessian filters in the work of Iwahori *et al.* [25], intensity valleys in the work of Bernal *et al.* [14] or Hough transform in the work of Silva *et al.* [26]) and then use resulting information to define cues unique to polyps. For instance, the work of Zhu *et al.* [27] analyzes curvatures of detected boundaries whereas the method of Kang and Doraiswami [28] is focused on searching ellipsoidal shapes typically associated with polyps. Finally, the method of Hwang *et al.* [29] combines curvature analysis and shape fitting in their strategy.

Concerning end-to-end learning, texture and color information were formerly used as descriptors such as in the work of Karkanis *et al.* [30] which proposed the use of color wavelets, the work of Ameling *et al.* [31] that exploits the use of co-occurrence matrices or the work of Gross *et al.* [32], which proposed the use of local binary patterns. Active learning methodologies have also been introduced as in the work of Angermann *et al.* [33] to reinforce the tradeoff between performance and computation time. Some of the most recent methods use deep learning tools to aid in polyp detection tasks, as in the work of Park and Sargent [34] or in the work of Ribeiro *et al.* [35]. In these very recent developments, differences among methods are based on the selection of a specific network architecture and databases used for training.

Finally, there are several hybrid methods which combine both methodologies for polyp detection, such as in the works of Tajbakhsh *et al.* [22], which combines edge detection and feature extraction to boost detection accuracy, the work of Bae and Yoon [36], that propose a system based on imbalanced learning and discriminative feature learning; the work of

Silva *et al.* [26], which uses hand-crafted features to filter non-informative image regions and the work of Ševo *et al.* [37], which combines edge density and convolutional networks.

As mentioned in Section I, the great majority of the methods are tested on private databases though we can observe that more recent publications such as the work of Park and Sargent [34] or the work of Ribeiro *et al.* [35] have started to use publicly available databases such as the ones used in the MICCAI 2015 Sub-challenge on Automatic Polyp Detection. Related to this, apart from new proposals, some of the referenced methods have been adopted by participants, such as the works of Bernal *et al.* [14], Silva *et al.* [26] or the work of Tajbakhsh *et al.* [22]. We provide in the next subsection a brief description of participating methods highlighting their most relevant contributions to the field. We grouped the methods following the taxonomy defined earlier in this subsection.

B. MICCAI 2015 Polyp Detection Sub-Challenge Methods

1) Hand-Crafted Features:

- **CVC-CLINIC:** This method [14] is based on a model of appearance considering polyps as protruding surfaces, being their boundaries defined from intensity valleys detection. Their proposal includes a pre-preprocessing stage to mitigate the impact of other valley-rich structures (blood vessels, specular highlights). To build final energy maps highlighting polyp presence, four different constraints (continuity, completeness, concavity, and robustness against spurious structures) are imposed to candidate boundaries to differentiate polyps from other structures.

2) End-to-End Learning:

- **CUMED:** The architecture of the proposed network contains two sections including a downsampling path and an upsampling path [38]. The former contains convolutional and max-pooling layers while the latter contains convolutional and upsampling layers, increasing the resolutions of feature maps and output prediction masks. To alleviate the problem of vanishing gradients and encourage the back-propagation of gradient flow in deep neural networks, the auxiliary classifiers are injected to train the network. Furthermore, they can serve as regularization to reduce over-fitting and improve the discriminative capability of features in intermediate layers [39], [40].
The classification layer, after fusing multi-level contextual information, produces the detection results. Network training is formulated as a pixel-wise classification problem with respect to ground-truth masks. The highlight of this approach is that it explores multi-level feature representations with fully CNNs in an end-to-end way, taking an image as input and directly providing the score map. In addition, feature-rich hierarchies from a large scale auxiliary dataset are transferred into the model to reduce over-fitting and further boost detection performance [41].
- **UNS-UCLAN:** This method, inspired by reported works [42]–[44], uses three CNNs trained at different image scales, namely 1, 0.5, and 0.25, of the original

training images. For all the scales the CNNs use the same architecture, but they are trained independently on the RGB images at their corresponding scale. After this initial training phase, the last fully connected part of each CNN is removed and the outputs from the 'convolutional part' of all the three networks are fed as input to a single Multi-Layer Perceptron (MLP) network. This additional network is trained independently from the three CNNs. In this approach CNNs are used as feature extraction engines operating at different spatial scales, and the MLP performs the classification based on these features. The method's output is the polyp incidence probability map, which is then processed to locate dominant probability peaks, as peaks locations and probability values are returned as the final output of the system. The training was performed exclusively on the CVC-CLINIC database.

- **OUS:** This method is based on the popular AlexNet model [44] for CNNs and its slight modification CaffeNet, which is pre-trained on the ILSVRC 2012 [45] dataset. Computations are achieved using the Caffe library [46]. The original model is modified to take input patches of size 96×96 , and the kernel size of the two first pooling layers is decreased from 3 to 2, while the last pooling layer is removed. The output layer is modified to give two outputs, polyp or non-polyp. In order to increase the training examples, data augmentation is performed in the form of random mirroring, rotation, up- and down-scaling, cropping, and brightness adjustment. Final polyp presence or absence was determined by using a sliding-window strategy, with three scalings for still frame analysis and two for full video sequence analysis.
- **SNU:** This methodology proposes a two-step approach: detection and localization. For both steps, CNNs were used. Starting from GoogleNet (pre-trained on the ImageNet dataset), a CNN fine-tuning was performed. Input image is resized to 224×224 pixels prior training and data augmentation (rotation and scaling) is also performed. Training set images are augmented by using several degrees of random rotation and scaling. Detection is considered as a simple binary classification task whereas, for localization, CNN are applied on polyp-positive images which are then segmented into a uniform-sized 8×8 grid (64 grids per image). Then, for each image, one grid is overlaid in black and then CNNs are applied thereafter to perform the binary classification task. The 64 overlaid grid images are then sorted by classification score to calculate final polyps' position.

3) Hybrid Approaches:

- **PLS:** The proposed full localization scheme consists of two parts, detection and localization. Regarding detection, two sets of images, one containing polyps, and the other without polyps, are used for training. Global image features [47] are used as they are easy and fast to calculate. Based on similarity scores between input frame and training ones and results ranks, the detection

TABLE I
SUMMARY OF INFORMATION FROM THE TEAMS THAT TOOK PART IN MICCAI 2015 CHALLENGE ON AUTOMATIC POLYP DETECTION

Team acronym	Full team details	Methodology	Published	Still-frame analysis	Video analysis	Training (seconds)	Testing (seconds)	System tested
ASU	Arizona State University (USA)	Hybrid	Yes [22]	No	Yes	N/A	2.7	2.4 GHz Intel quad core processor and an NVIDIA GeForce GTX 760 video card
CUMED	Department of Computer Science and Engineering, Chinese University of Hong Kong (China)	End-to-end learning (CNNs)	No	Yes	Yes	10800	0.2	A standard PC with a 2.50 GHz Intel(R) Xeon(R) E5-1620 CPU and a NVIDIA GeForce GTX Titan X GPU
CVC-CLINIC	Computer Vision Center and Universitat Autònoma de Barcelona (Spain)	Hand-crafted	Yes [14]	Yes	Yes	N/A	10	Intel core i7-4790 at 3.6GHz
ETIS-LARIB	ETIS, ENSEA, University of Cergy-Pontoise, CNRS, Cergy (France)	Hybrid	Yes [26]	Yes	No	196	2.14	Intel i5 4200U 2.30 GHz
OUS	Oslo University Hospital, OUS Norway, University of Oslo (Norway)	End-to-end learning (CNNs)	No	Yes	Yes	86400	5	Intel i5, 4 cores at 2.8 GHz, 4 GB RAM. Graphic card with 4 GB memory used for training
PLS	Polyp Localize and Spot Team, Media Performance Group, Simula Research Laboratory and University of Oslo (Norway)	Hybrid	No	Yes	Yes	0.33 per image	0.145	2 Intel(R) Xeon(R) CPU E5-2650 at 2.00GHz CPU, 64 GB of RAM, NVIDIA Corporation GK110, GeForce GTX TITAN
SNU	Seoul National University, Seoul (South Korea)	End-to-end learning (CNNs)	No	Yes	Yes	360	0.8-1	NVIDIA TITAN X GPU
UNS-UCLAN	School of Engineering, University of Central Lancashire, Preston (UK) and University of Nice-Sophia Antipolis, Nice (France)	End-to-end learning (CNNs)	No	Yes	No	18000	5	i7-5930K @ 3.5GHz (6 cores), 64 GB RAM, NVIDIA GeForce GTX TITAN X

subsystem decides in real-time to which class (polyp or no polyp) the input frame belongs to.

The localization scheme is implemented as a sequence of preprocessing filters (RGB to YCbCr color space conversion, removal of borders and sub-images, flare masking and low-pass filtering) and uses the polyp's physical shape to find its exact position, approximating polyps by elliptical shape regions presenting local features that differentiate them from surrounding tissues. The final decision regarding polyp location is taken by means of the maximum values in the energy map computed using the elliptical shape of the polyp's usual appearance. Finally, the method outputs four possible locations per frame.

- **ETIS-LARIB:** This method [26] is inspired by the psycho-visual methodology used by clinicians when performing an endoscopic examination. First, a detection of the Regions of Interests (ROI) that may contain a polyp is performed using shape and size image features. This first pre-selection allows a first and fast scanning of the image. Due to being circular/elliptical shapes associated to polyps, a Hough transform was used for this first filtering stage. Once ROIs are detected, a second analysis, based on texture is achieved in order to remove those ROIs with no actual polyp content. To achieve this, an ad-hoc classifier based on a boosting-based learning process using texture features computed from co-occurrence matrices (standard Haralick features) is proposed.
- **ASU:** This method [22] consists of two stages. In the first stage, a set of polyp candidates is generated using geometric features. Specifically, given a colonoscopy frame,

a crude set of edge pixels is first obtained. This edge map is then refined using a classification and feature extraction scheme [48]. The goal of the edge classification scheme is to remove as many non-polyp boundary pixels as possible from the initial edge map. The geometry of the retained edges is then used in a voting scheme that localizes polyps candidates as objects with curved boundaries in the refined edge maps. The voting scheme further estimates a bounding box for each generated candidate based on the generated voting map. In the second stage, an ensemble of CNNs -each of them specialized in one type of features- is applied to each candidate bounding box [49]. Finally, the outputs of the CNNs are averaged to generate a confidence score for a given polyp candidate.

Table I shows a summary of the different methods participating at MICCAI 2015 Challenge on Automatic Polyp Detection. As each method was tested under different conditions, computation times are given to complete the information on the training and testing processes.

III. VALIDATION STUDY

We introduce in this section the complete validation study proposed to assess and compare the performance of different polyp detection methods.

A. Definitions and General Performance Metrics

We define **Polyp detection** as the capability of a given method to determine polyp presence in a colonoscopy frame (**Polyp presence detection**) and, once this is determined, it

TABLE II
PERFORMANCE METRICS FOR POLYP DETECTION

Metric	Abbreviation	Calculation
Precision	Prec	$Prec = \frac{TP}{TP+FP}$
Recall	Rec	$Rec = \frac{TP}{TP+FN}$
Specificity	Spec	$Spec = \frac{TN}{FP+FN}$
F1-measure	F1	$F1 = \frac{2 \times Prec \times Rec}{Prec + Rec}$
F2-measure	F2	$F2 = \frac{5 \times Prec \times Rec}{4 \times Prec + Rec}$

is able to provide the location of the polyp within the image (**Polyp localization**). Consequently, a good polyp detection method should select images (video frames) containing polyps and ignore all others and it should indicate the position of all polyps present in an image. There are some terms defined next which are key to set performance metrics. As we deal with images from real patients examinations, we will find two different cases: images with polyps and images without polyps.

In the first case, if detection output is within the polyp, the method is said to be providing a **True Positive (TP)** or correct alarm. It has to be noted that only one TP will be considered per polyp, no matter how many detections fall within the polyp. Any detection that falls outside the polyp is considered a **False Positive (FP)** or false alarm. The absence of alarm in images with a polyp is considered a **False Negative (FN)**, counting one per each polyp in the image that has not been detected. Regarding images without polyps, we define as a **True Negative (TN)** whenever the method does not provide any output for this particular image. Any detection provided for frames without a polyp counts as a **False Positive (FP)**. Considering these definitions, we propose the use of the frame-based performance metrics presented in Table II.

B. Databases

Three different databases are used in the context of the validation study presented in this paper. Two publicly available databases were proposed for still frame analysis, **CVC-CLINIC** and **ETIS-LARIB**. CVC-CLINIC [14] contains 612 Standard Definition (SD) frames and comprises 31 different polyps from 31 sequences. ETIS-LARIB database contains 196 High Definition (HD) frames and comprises 44 different polyps from 34 sequences. More details on these databases are presented in Table IV. It has to be noted that all images contain at least a polyp; both databases were built to cover as many different polyp appearances as possible. Ground truth consisting of a polyp mask was generated using the same procedure for both databases: Images were annotated by expert videoendoscopists from the corresponding associated clinical institution. These experts (one per hospital) were asked to outline the boundaries of any polyps present in the image. These boundaries are used to generate a binary mask representing the actual polyp area within the image, also to be used for validation purposes. Examples from these two databases are shown in the first two columns of Fig. 1.

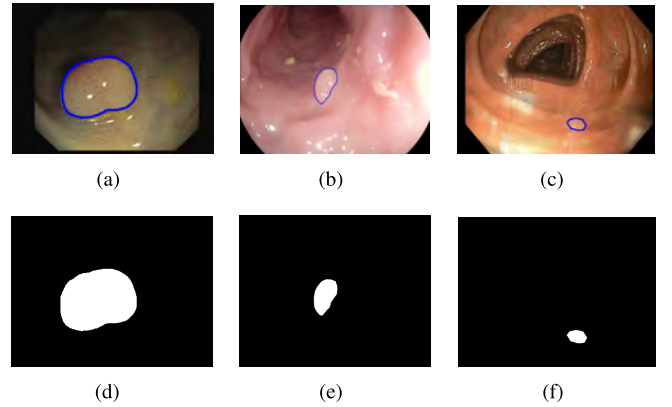


Fig. 1. Illustration of the content of the CVC-CLINIC (first column), ETIS-LARIB (second column) and ASU-Mayo Clinic (third column) databases. The first column shows the original images with the corresponding reference polyp contour shown as a blue line and the second contains binary masks representing the ground truth.

The **ASU-Mayo Clinic Colonoscopy Video © Database** [22] comprises a set of short and long colonoscopy videos, collected at the Department of Gastroenterology at Mayo Clinic, Arizona. This database consists of 38 different, fully annotated videos. The videos were selected to display maximum variation in colonoscopy procedures including different resolutions and examination strategies (careful vs. fast inspection) and also include frames containing biopsy instruments or device information. Ground truth consisting of binary masks (polyp frames) and black frames (non-polyp frames) were created by volunteer students at Arizona State University and have been reviewed and corrected by a trained expert. Table III outlines information about the videos in that database, including for each video duration in seconds (Length), number of frames with polyps and the total number of frames (Polyp/Total) and the image resolution (Res). An example from this database is shown in the third column of Fig. 1.

C. Statistical Analysis

In order to account for statistically significant differences in performance between methods, we propose first to perform a Saphiro-Wilk test to find out whether the available data follows a normal distribution or not. In the first case (normal distribution) statistically significant differences across methods will be assessed using an analysis of variance (ANOVA) to detect differences regarding proposed metrics. In the second case (no normal distribution), the Kruskal-Wallis test will be used. All tests are done at a confidence level $1 - \alpha = 0.95$.

Considering the scope of the analysis presented in the paper, the metric that will be used to compare different methods will be F1-score, as it presents a balance between missed polyps and false alarms. We perform a statistical study of this metric only in videos with polyp (and potentially non-polyp) frames, as the number of samples in the still-frame analysis is not big enough to provide with statistically relevant conclusions and the analysis in videos with no polyps would cause the F1 score to be zero for all methods. We also perform statistical analysis of detection latency but, for the sake of a proper statistical

TABLE III
CONTENT OF ASU-MAYO CLINIC COLONOSCOPY VIDEO © DATABASE

Training database						Testing database					
Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]	Video	Length	Polyp/Total [Res]
1	22	0/682 [712 × 480]	11	10	245/324 [1920 × 1080]	1	19	0/599 [712 × 480]	11	15	338/452 [1920 × 1080]
2	27	0/838 [712 × 480]	12	30	910/910 [1920 × 1080]	2	20	0/625 [712 × 480]	12	4	134/0134 [1920 × 1080]
3	25	0/769 [712 × 480]	13	17	374/519 [1920 × 1080]	3	20	0/628 [712 × 480]	13	10	312/312 [1920 × 1080]
4	23	0/712 [712 × 480]	14	16	391/501 [856 × 480]	4	20	0/607 [712 × 480]	14	60	0/1815 [712 × 480]
5	61	0/1843 [712 × 480]	15	36	1106/1200 [856 × 480]	5	30	693/918 [856 × 480]	15	59	0/1795 [712 × 480]
6	64	0/1925 [712 × 480]	16	11	209/339 [1920 × 1080]	6	40	1218/1218 [856 × 480]	16	54	0/1627 [712 × 480]
7	51	0/1550 [712 × 480]	17	13	234/418 [856 × 480]	7	18	445/555 [712 × 480]	17	60	0/1807 [712 × 480]
8	58	0/1740 [712 × 480]	18	18	189/259 [1920 × 1080]	8	14	335/446 [856 × 480]	18	61	0/1835 [712 × 480]
9	60	0/1802 [712 × 480]	19	20	235/616 [1920 × 1080]	9	13	290/396 [1920 × 1080]			
10	54	0/1639 [712 × 480]	20	13	385/410 [856 × 480]	10	60	548/1805 [1920 × 1080]			

TABLE IV

SUMMARY OF CONTENT OF STILL FRAME VALIDATION DATABASES.
SD STANDS FOR STANDARD DEFINITION, HD STANDS
FOR HIGH DEFINITION

Database	Purpose	Institution	Content	Device
CVC-CLINIC	Training	Hospital Clinic, Barcelona, Spain	612 SD frames (388 × 284) from 31 sequences	Olympus Q160AL and Q165L, Exera II videoprocessor
ETIS-LARIB	Testing	Lariboisière Hospital, Paris, France	196 HD frames (1225 × 966) from 34 sequences	Pentax 90i series, EPKi 7000 videoprocessor

comparison, this analysis is only done for those teams which detect the polyp in all sequences.

D. MICCAI 2015 Sub-Challenge Validation Study

Two different scenarios were presented to the participants of the challenge: (i) still frame analysis and (ii) full video analysis. In the following, we present specific information of the two presented scenarios, including validation databases and performance metrics used in each of them.

1) *Still Frame Analysis*: The objective of this analysis was to explore localization capabilities of a polyp detection method. We aim to test how different methods perform in challenging high-definition (HD), high-quality images showing great variability in polyp appearances. In this case, each image contains at least one polyp and images have been selected in order to have shots in which polyp appearance can be mistaken with other elements of the scene (folds, vessels).

Two different databases were used in this study: **CVC-CLINIC** is used for the **training** stage whereas **ETIS-LARIB** is used during the **testing** stage. Participating methods are compared using performance metrics exposed in Table II. Additionally, in case a given method provides confidence values a Precision-Recall curve is also provided otherwise the operating point will represent its performance.

2) *Video Analysis*: In this second scenario we aim to explore full polyp detection capabilities (localization and presence detection) of a given method in full sequences from actual colonoscopy procedures. In this case, polyp detection methods have to deal, apart from appearance variability, with potential polyp presence or absence in each image and, moreover, with variability in image quality (blurring, bowel preparation). Additionally, the videos in the second scenario may contain

images with extra-endoluminal elements such as device information or surgical instruments. We also have to consider that, as in real procedures, nor all the sequences or all the frames contain a polyp.

The **ASU-Mayo Clinic Colonoscopy Video © Database** was used in this experiment. Apart from using common performance metrics exposed in Table II, we proposed an additional performance metric to assess whether how fast a given detection method reacts to polyp presence. In this context, **Detection Latency (DL)**: $DL = first_detection - first_appearance$ represents the delay in frames between the first appearance of the polyp in the video sequence ($first_appearance$) and the first actual detection of the polyp by a method ($first_detection$). Considering this, a clinically useful support system should have a DL close to zero. Finally we also provide with Receiver Operating Curves (ROC), though again, each method's representation depends on whether they provide detections' confidence values or not.

From a general organization perspective, all teams taking part in the challenge were to use the same data for both their training and testing stages. Participants were provided with labelled training data on June 15th whereas unlabelled testing data (still frames and full sequences) was released on July 24th. In order to take part in the challenge, each participating team was asked to provide a unique CSV file for the analysis of the ETIS-LARIB database and/or one CSV file for each of the 18 testing videos in the ASU-Mayo database, depending on the sub-category the team would take part in. Each row in the CSV file represents a detection candidate region. Additionally, teams could also provide a confidence value (value between 0 and 1) for the performance curves drawing purposes, though this was not mandatory. Finally, though 8 different teams took part in the challenge, not all of them participated in all categories. ASU did not take part in the still-frame analysis sub-category whereas ETIS-LARIB and UNS-UCLAN did not take part in the video analysis one.

E. Additional Validation Experiments

1) *Combination of Methods*: In this study we propose to go beyond the analysis of individual methods by providing quantitative elements on how potential combinations of some of the presented approaches may lead to an improved performance. Inspired by [50], we have studied two options of fusing the

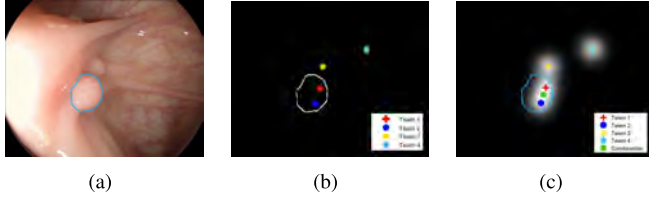


Fig. 2. Synthetical examples of different ways to perform a combination of methods: (a) original image; (b) result of combination by union, and (c) result of combination by saliency map creation. Outputs from different teams are represented by different colors and shapes. In all images, the contour of the polyp is represented as a blue curve.

methods, namely: 1) **combination by union** and 2) **saliency map creation**.

The first one consists of adding, for a particular frame, the outputs from all submitted methods. Saliency maps creation proposes a combination of the output of the methods in order to generate heat maps which aim to represent those areas in the image where most of the methods coincide in their decision regarding polyp location, following the methodology proposed in [14]. We show in Fig. 2 a graphical comparison between both strategies.

In this case, we treat the output of each method as a 'fixation' or vote, and we create saliency maps from this set of discrete fixations/votes. These fixation points are interpolated by a Gaussian function to build up the final saliency map for a given image as follows:

$$s(x, y) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2\pi\sigma_s^2} \cdot \exp\left(-\frac{(x - x_n^f)^2 + (y - y_n^f)^2}{2\sigma_s^2}\right), \quad (1)$$

where: x and y denote, respectively, the horizontal and vertical positions of a given image pixel; x_n^f and y_n^f represent the horizontal and vertical coordinates of a detection point (fixation); N indicates the total number of detected points and finally σ_s denotes the standard deviation of the Gaussian function, determined as proposed in [14]. We determine the location of the global maximum of the saliency map as the final output of the combination of methods for a particular frame.

Two versions of saliency map creation have been implemented: (**saliency by union**) calculates the saliency maps for each frame considering all the methods that have provided any output whereas (**saliency voting**) only calculates the saliency maps if the majority of the teams in the studied combination provide an output for the specific frame.

We provide results for each combination strategy in the two challenge scenarios (still frame analysis and video analysis) using the same frame-based performance metrics.

2) Impact of Image Challenges on Method's Performance: This experiment aims to study the impact of some of the technical and clinical challenges reported in Section I over the performance of a polyp detection method. In order to study this, we proposed clinicians and computer scientists from the contributing teams to define the main image challenges present in colonoscopy frames that were to be studied. The following ones were selected: 1) Presence of overlay information in images (including patient information

TABLE V
BREAKDOWN OF CLINICAL AND TECHNICAL CHALLENGES
WITHIN ASU-MAYO TEST DATABASE

Challenge	Number of frames	Challenge	Number of frames
Overlay information	517[02.94%]	Massive specular highlights presence	8638[49.15%]
Overexposed regions	8270[47.05%]	Intestinal content	10013[56.97%]
Visible luminal region	4963[28.24%]	Images showing an incomplete polyp	3286[18.69%]
Specular highlights within polyp	93[19.88%]	Low visibility	12788[72.67%]

and camera shots); 2) Presence of specular highlights; 3) Appearance of overexposed regions; 4) Occurrence of intestinal content (fecal particles, bubbles); 5) Presence of the luminal region; 6) Lack of visibility of the whole polyp in the image; 7) Presence of specular highlights within the polyp region and 8) Images with low visibility (due to blurring or excessive intestinal content). Fig. 3 shows examples of each of the challenges.

A graphical user interface was built for experts to label individually each frame from the testing videos of **ASU-Mayo Clinic Colonoscopy Video © Database** according to the mentioned image challenges. For the sake of statistical representativeness of the results, we did not perform the same experiment for **ETIS-LARIB** database due to its smaller size. As some of them may lead to subjective interpretations we collected three different annotations per frame and the final decision of a frame for each challenge was taken by majority voting from the three experts. We present statistics about the presence of the different image challenges in Table V.

We can observe how roughly half of the frames contain a high number of specular highlights, some degree of intestinal content and overexposed regions. Regarding polyp frames, which equate to a 25% (4313) of the frames, we can observe that about a 30% of them (1360) do not show completely the polyp and that nearly all of them (3959) present specularities within the polyp region. Finally, it is interesting to mention that more than a 70% of the images were considered of low visibility quality, which indicates how the methods are tested in clear challenging conditions.

Once we have final annotations, we broke down the methods performance analysis into two groups: frames with polyps and frames without polyps. For the first case, we analyze differences between performance for frames with and without a specific image challenge regarding Precision, Recall and F1-score whereas for the second the same kind of analysis was done regarding Specificity score.

3) Impact of Polyp Morphology on Methods' Performance: This experiment assesses whether methods' performance depends on the polyp morphology. This analysis examines if the methods perform differently for polyps with different associated morphological type. Such analysis could be useful to check whether existing methods are able to cope with different morphologies as well as determining which method to choose if a given morphology is predicted before the examination. In order to study these potential differences in

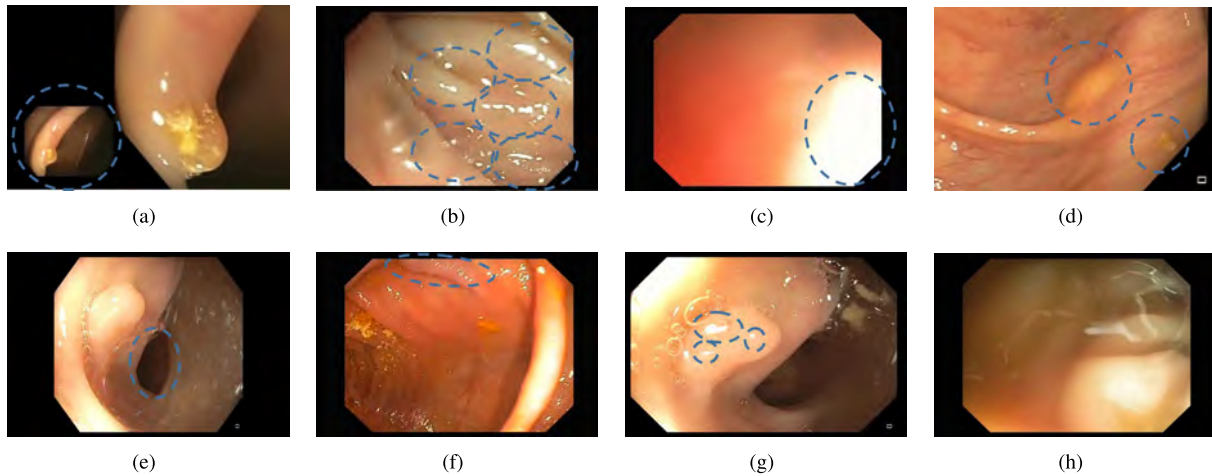


Fig. 3. Examples of the 8 technical and clinical challenges selected for the study: (a) Presence of overlay information; (b) High presence of specular highlights; (c) Overexposed regions; (d) Intestinal content; (e) Luminal region; (f) Polyp cannot be seen completely in the image; (g) Specular highlights within the polyp and (h) Impact of low visibility quality.

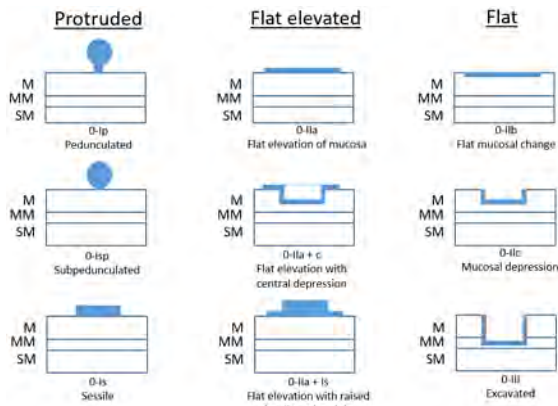


Fig. 4. Graphical representation of Paris classification of endoscopic polyps. M stands for mucosa, MM for muscularis mucosa and SM for submucosa.

performance, we propose to categorize each of the polyps that appear in the testing databases using the Paris classification criteria [51]. We show graphical examples of each type in Fig. 4.

To account for differences in performance related to polyp morphology we will use Precision, Recall and F1 scores as defined in Table II. We also study differences in latency score for the case of video sequences analysis.

4) *Temporal Coherence on Method's Response*: One capability that a computational method should have when dealing with video analysis is temporal stability in its response. That is, if a given method detects a polyp in a given frame and considering normal camera movement, its output for the following frame should provide a relevant detection. As we can observe in the example shown in Fig. 5, none of the methods presented in the challenge incorporated per se temporal stability capabilities in their methodologies but we consider that it is important to assess up to what extent they provide this kind of stability. Moreover, and as a consequence of this stable temporal output, the method should provide with correct detection in the majority of the frames in which the polyp appears.

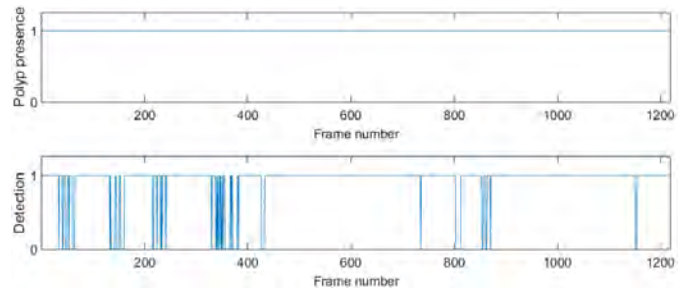


Fig. 5. Example of non-temporal coherence of polyp detection methods. The example represents the performance of the CVC-CLINIC method for the testing video 6 of ASU-Mayo Clinic Colonoscopy Video © Database. Image at the top shows ground polyp presence per frame (1 is polyp, 0 is no polyp) whereas bottom image shows detection score (1 correct, 0 no correct).

In order to study this we perform two evaluations. Regarding detection stability in consecutive frames, for each testing video from **ASU-Mayo Clinic Colonoscopy Video © Database** that contained a polyp, we extracted the pairs of consecutive frames containing a polyp. We analysed methods' output for each pair of consecutive polyp frames and calculated as metric the percentage of these pairs in which the method provided correct output - detection inside the polyp mask - for both frames. With respect of overall detection stability in a sequence, we study Recall scores over the different sequences, analyzing mean and standard deviation values to account for intra and inter-sequence stability on detection performance.

5) *Analysis of the Direct Output of the Methods*: As mentioned in Section III-D, participating teams were only asked to provide CSV files indicating detection output for the testing frames (x and y position). This file is created from the output of the different methods and we propose here to analyze this actual output. As a first study, we asked the teams participating in the still frame analysis challenge sub-category to provide their actual output for each frame of **ETIS-LARIB** database.

In this context, we foresee the output of a method to be interpreted as a likelihood or heat map, in which brighter

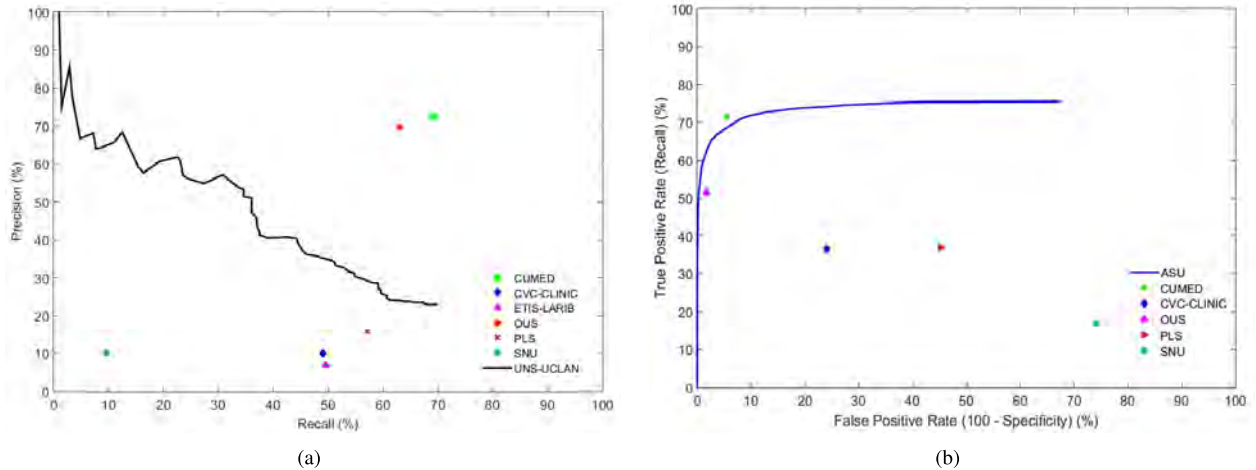


Fig. 6. Performance curves: (a) Precision-Recall curve for the analysis of the ETIS-LARIB database and (b) Receiver Operating Characteristic (ROC curves) for the analysis of the ASU-Mayo database. For the ROC curve, SNU operating point is calculated from the videos the team provided results for. Methods are represented with a line in cases where the confidence value has been provided for each detection, otherwise the operating point is used.

TABLE VI
SUMMARY OF STILL FRAME ANALYSIS RESULTS

	TP	FP	FN	Prec	Rec	F1	F2
Individual team analysis							
CUMED	144	55	64	72.3	69.2	70.7	69.8
CVC-CLINIC	102	920	106	10.0	49.0	16.5	27.5
ETIS-LARIB	103	1373	105	6.9	49.5	12.2	22.3
OUS	131	57	77	69.7	63.0	66.1	64.2
PLS	119	630	89	15.8	57.2	24.9	37.6
SNU	20	176	188	10.2	9.6	9.9	9.7
UNS-UCLAN	110	226	98	32.73	52.8	40.4	47.1
Combination of teams analysis							
Best concatenation result (all teams)	188	3410	20	5.2	90.4	9.9	21.2
Best saliency result (CUMED, OUS)	159	38	49	80.7	76.4	78.5	77.2
Best saliency voting (CUMED, OUS)	159	38	49	80.7	76.4	78.5	77.2

(hotter) areas of the image represent parts of the image more likely to contain a polyp. By analyzing these maps, we could observe up to what extent method's attention is only focused on the polyp. To measure this, we propose Concentration Ratio (CR) to compare these maps as proposed in [14]; CR measures, for each frame, the rate of total energy in the image (calculated as the sum of the each pixel's value from the energy map image) falling within the polyp. High CR values are interpreted as a method actually focusing on the polyp, being lower values related to sparser energy maps.

IV. RESULTS

In this section, we present the performance achieved by each method in the several experimental studies proposed in the paper, including those part of the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy.

A. MICCAI 2015 Challenge Results

We present main still frame analysis results in both Fig. 6 (a) and in Table V. CUMED offers the best performance in all metrics evaluated, being the team which detected the most

polyps (144) frames along with providing the lowest number of false alarms (55). The comparison between the performance of CNN-based approaches shows the importance of specific network configuration, as relevant differences in both number of detected polyp frames and false alarms can be observed - for instance, the number of detected polyp frames falls into a range between 131 (OUS) and 20 (SNU) -. Finally we can observe a performance gap between end-to-end learning and hybrid/hand-crafted methods, which provide less correct detections and significantly more false alarms. It has to be noted that as PLS provides four locations per image, the number of false alarms for this method is inherently higher than for other methods.

Considering full video analysis, the study shows superior performance by CNN-based methods -see Table VII and in Fig. 6 (b)-, with CUMED being the method providing a higher number of polyp frames detected (3081). In this case, it has to be noted that CUMED does not outperform all other methods in all considered metrics, as ASU provides a better balance between true and false alarms (higher F1-score) at the cost of detecting less polyp frames (2636 vs. 3081).

We present in Table VII a complete breakdown of video analysis results, dividing them into 3 groups according to the degree of polyp presence in the sequences: 1) videos containing frames with and without polyps; 2) videos containing only frames with polyps, and 3) videos containing non-polyp frames. In all cases, results again show a superior performance of CUMED in terms of total number of polyp frames detected. Deepening the analysis, we observe a decrease in the difference in performance observed in global analysis between hand-crafted methods (CVC-CLINIC) and CNN-based methods when videos with only polyp frames are analyzed. This can be related to those methods being designed to highlight polyp-like structures in the image (localization) but not for determining specific polyp presence. The analysis of sequences without polyp frames shows that PLS offers the best performance, which is possibly due to the presence of a specific polyp presence module in this approach.

TABLE VII
SUMMARY OF THE MOST RELEVANT RESULTS REGARDING VIDEO SEQUENCE ANALYSIS

All videos									
	TP	FP	TN	FN	Prec [%]	Rec [%]	Spec [%]	F1 [%]	F2 [%]
<i>Individual team analysis</i>									
ASU	2636	184	13149	1677	93.5	61.1	98.6	73.9	65.7
CUMED	3081	769	13010	1232	80.0	71.4	94.4	75.5	73.0
CVC-CLINIC	1578	3456	10927	2735	31.3	36.6	75.9	33.8	35.4
OUS	2222	229	13245	2091	90.6	51.5	98.3	65.7	56.4
PLS	1594	10103	12258	2719	13.6	36.9	54.8	19.9	27.5
SNU(Only videos with polyps)	721	3285	1140	3592	17.9	16.7	25.7	17.3	16.9
<i>Combination of teams analysis</i>									
Best concatenation result (all teams)	3576	14741	10064	737	19.5	82.9	40.6	31.6	50.3
Best saliency result (all teams)	3294	4070	10064	1019	44.7	76.4	71.2	56.4	66.9
Best saliency voting result (ASU,CUMED)	3316	557	12915	997	85.6	76.8	95.9	81.0	78.4
Videos with frames with and without polyp									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
<i>Individual team analysis</i>									
ASU	1218	92	1864	1431	92.9	45.9	95.3	61.5	51.1
CUMED	1439	600	1692	1210	70.6	54.3	73.8	61.4	57.0
CVC-CLINIC	195	1343	1430	2454	12.7	7.4	51.6	9.3	8.0
OUS	651	55	1914	1998	92.2	24.6	97.2	38.8	28.8
PLS	328	6953	920	2321	4.5	12.4	11.7	6.6	9.2
SNU	282	2085	1140	2367	11.9	10.6	35.3	11.2	10.9
<i>Combination of teams analysis</i>									
Best combination by union (all teams)	1949	11128	493	700	14.9	73.6	4.2	24.8	41.2
Best saliency by union (all teams)	1588	2557	493	1061	38.3	59.9	16.2	46.7	53.9
Best saliency voting result (CUMED,ASU)	1698	439	1649	951	79.4	64.0	79.0	70.9	66.7
Videos with only polyp frames									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
<i>Individual team analysis</i>									
ASU	1418	40	no	246	97.2	85.2	N/A	90.8	87.4
CUMED	1642	149	no	22	91.7	98.7	N/A	95.0	97.2
CVC-CLINIC	1383	272	no	281	83.5	83.1	N/A	83.3	83.2
OUS	1571	167	no	93	90.4	94.4	N/A	92.3	93.6
PLS	1266	3150	no	398	28.7	76.1	N/A	41.6	57.2
SNU	439	1200	no	1225	26.8	26.4	N/A	26.6	26.5
<i>Combination of teams analysis</i>									
Best combination by union (all teams)	1664	4978	N/A	0	25.0	100.0	N/A	40.0	62.6
Best saliency by union (all teams)	1662	2	N/A	2	99.8	99.8	N/A	99.8	99.8
Best saliency voting (all teams)	1662	2	N/A	2	99.9	99.9	N/A	99.9	99.9
Videos without polyp frames									
	TP	FP	TN	FN	Prec	Rec	Spec	F1	F2
<i>Individual team analysis</i>									
ASU	N/A	52	11286	N/A	N/A	N/A	99.5	N/A	N/A
CUMED	N/A	20	11318	N/A	N/A	N/A	99.8	N/A	N/A
CVC-CLINIC	N/A	1841	9497	N/A	N/A	N/A	83.8	N/A	N/A
OUS	N/A	7	11331	N/A	N/A	N/A	99.9	N/A	N/A
PLS	N/A	0	11338	N/A	N/A	N/A	100.0	N/A	N/A
<i>Combination of teams analysis</i>									
Best combination by union (all teams)	N/A	1920	9454	N/A	N/A	N/A	83.2	N/A	N/A
Best saliency by union (PLS,OUS)	N/A	20	11318	N/A	N/A	N/A	99.8	N/A	N/A
Best saliency voting (CUMED,PLS,OUS)	N/A	0	11338	N/A	N/A	N/A	100.0	N/A	N/A

As mentioned in Section III-C, a statistical analysis is performed to account for differences in performance between methods. Results of the Saphiro-Wilk test over the F1 results for each video and method indicates a normal data distribution ($p\text{-value} > 0.05$). Considering this, we perform a subsequent ANOVA analysis and multicomparison test to compare the different methods. The ANOVA test detects significant differences across F1 values ($p\text{-value} = 5.4e^{-10}$), which are explored in the multicomparison test shown in Fig. 8. Results of this test show the superior performance of ASU, providing

CUMED with a comparable performance different from the rest in a statistically significant way. CUMED and OUS also show performances comparable to each other. Finally CVC, PLS, and SNU also present comparable performances.

We present in Fig. 7 detection latency results. We can observe how there are only two teams (ASU and CUMED) which present latency scores for all the videos. We perform a statistical analysis to account for the differences between them. The result of the Saphiro-Wilk test indicates a non-normal data distribution ($p\text{-value} < 0.05$) and, consequently, the

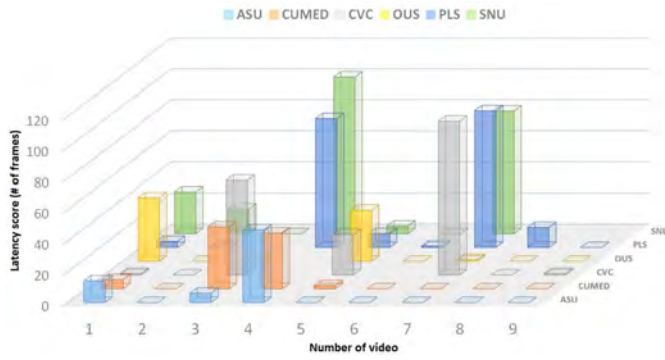


Fig. 7. Detection latency for polyp-containing videos.

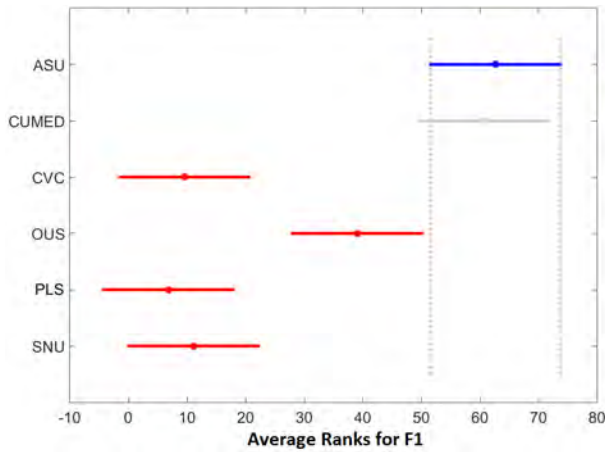


Fig. 8. Multicomparison test for the analysis of the F1 score in videos showing frames with and without polyps. Each method is represented as a horizontal line whose center is located in corresponding method's mean F1 score and whose width corresponds to the variance calculated according anova1 fit model. The best ranked group is represented by a blue horizontal line, comparable methods are shown in grey, and methods that are different in a statistically significant way are shown in red.

Kruskal-Wallis test is performed to account for statistically significant differences. In this case the test confirms the null hypothesis that both data samples come from the same distribution $p\text{-value} = 0.76$, which can be observed in Fig. 8. Concerning the rest of the methods, we can observe that they do not detect the polyps in all the videos which is also a cause of the difference in performances shown in Table VII.

B. Additional Validation Experiments

1) *Combination of Methods*: We have included in Table VI and in Table VII the best performance achieved after applying each of the proposed method combination strategies. The most important though logical conclusion extracted is that a combination of methods leads to better detection results. As expected, any combination of methods leads to an increase of the total number of detected polyps. This shows that different methods detect different polyps and that even those with lower performance can contribute positively to the overall detection.

We can observe in Fig. 9 (d-f) that if we do not include all teams in the combination, the number of correct polyp frame detections could be affected. We can also observe in Fig. 9 (a-c) that the combination of the two best methods in

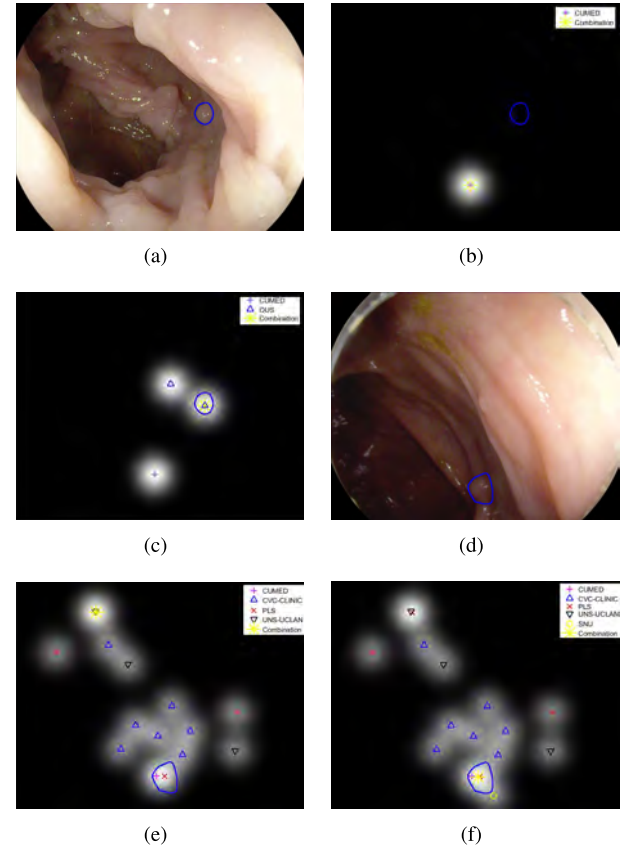


Fig. 9. Examples of the benefits of using a saliency-map-based approach. The first row shows the impact of combining the two best methods that surpass their individual performances: (a) original image; (b) saliency map with the position of detection points superimposed (best method, CUMED); (c) saliency map with the position of detection points superimposed (two best method, CUMED and OUS). The second row shows the positive impact of the worst performing method: (d) original image; (e) saliency map with the position of detection points superimposed (all method); (f) saliency map with the position of detection points superimposed (all method but SNU). The polyp contour is represented in blue. Each method is represented by a different color and shape.

each category surpasses the individual methods' performance, which indicates the potential of saliency map methods to build up more reliable systems. It is clear that the combination by union strategy increases the total number of detected polyp frames at the cost of vastly increasing the number of false alarms and, consequently, other strategies should be explored to achieve a clinically useful system.

Considering this, we observe that the use of saliency maps leads to a better balance between correct and false alarms. Regarding the two saliency-map sub-strategies, the voting strategy leads to a slightly better performance for the case of still frame analysis, specifically observed in the reduction of FP. This can be explained as being due to those poorly performing methods providing outputs for almost all frames. Once their contribution is not considered as majority is not achieved for a particular frame, these false alarms vanish.

Taking into account these results, if we consider a potential combination of methods as the solution for polyp detection, we would propose saliency maps with a voting sub-strategy as the strategy that leads to a better compromise between correct detections and false alarms, though other potential

TABLE VIII
IMPACT OF CLINICAL AND TECHNICAL CHALLENGES ON INDIVIDUAL METHODS' PERFORMANCE

Clinical and Technical Challenges								
Team	Overlay information	High specular highlights presence	Overexposed regions	Intestinal content	Luminal region	Incomplete polyp	Specular highlights inside polyp	Low visibility
<i>Polyp frames: Recall [differences in %]</i>								
ASU	-14.3	48.2	-10.2	7.7	-11.10	-0.2	43.5	-11.9
CUMED	-3.9	42.1	-24.7	8.3	-17.1	-1.3	23.0	-30.1
CVC-CLINIC	-28.8	24.7	-20.7	28.9	-35.5	3.4	26.9	-11.6
OUS	-27.3	49.6	-31.3	18.8	-21.4	15.9	41.0	-14.1
PLS	-12.0	21.3	-3.7	28.4	-16.1	1.0	35.3	-13.4
SNU	-18.5	13.9	-2.1	2.5	-10.7	-2.5	14.8	5.6
<i>Polyp frames: F1-score [differences in %]</i>								
ASU	-11.1	50.8	-8.4	5.5	-8.2	0.3	43.5	-9.3
CUMED	-18.6	32.1	-28.0	8.3	-16.9	1.1	16.3	-18.0
CVC-CLINIC	-31.6	23.3	-23.4	30.3	-40.8	1.0	27.9	-13.1
OUS	-25.6	58.9	-30.4	15.0	-16.1	13.0	44.2	-11.9
PLS	-7.4	20.2	-5.1	18.7	-15.1	2.5	25.9	-4.0
SNU	-20.6	15.3	-1.9	3.2	-12.9	-4.1	15.6	6.6
<i>Non-polyp frames: Specificity [differences in %]</i>								
ASU	0.8	-0.4	0.5	-0.7	-0.5	N/A	N/A	0.3
CUMED	-7.9	-1.2	2.6	0.1	-3.6	N/A	N/A	4.8
CVC-CLINIC	-63.8	-26.0	12.5	-24.5	2.1	N/A	N/A	32.7
OUS	0.1	-0.2	-0.1	-0.1	-0.1	N/A	N/A	0.1
PLS	13.7	-3.9	28.2	-17.0	-32.2	N/A	N/A	42.5
SNU	40.8	5.6	-15.7	-14.6	17.9	N/A	N/A	20.9

combinations can be explored. For instance, we can think of a system which also includes the specific detection modules that some approaches have presented here (PLS, SNU), with polyp localization within a given image being then obtained using CNN-based approaches (CUMED, OUS).

2) Impact of Image Challenges on Individual Methods' Performance: We present in Table VIII a summary of the results of the experiment assessing the impact of several image challenges on individual methods' performance. With respect to polyp frames, the first conclusion to be extracted is that low visibility images and the presence of specular highlights within the polyp affect all methods in the same way. We interpret the impact of image quality as being both mucosa wall and its elements, such as polyps, better visually defined in good visibility images hence helping in polyp detection. We associate the positive impact of specular highlights inside the polyp to polyps appearing commonly as protruding elements in the scene and, as a consequence of this, specularities appear in their surface as their reflect light back to the camera [52].

There are some image challenges that generally seem to make polyp frames detection difficult such as the presence of overlay information and overexposed regions, with the latter being more prevalent in the explored images. The clear view of the luminal region also negatively affects detection capabilities, which we interpret as result of lumen presenting strong boundaries and contrast in comparison with the mucosa, which is a feature that polyps also exhibit. Surprisingly, the presence of intestinal content affects positively Recall and F1-score; this could be explained by the fact that this image challenge appears clearly different from polyps (weak contours, different color). Finally, the degree of completeness of the polyp seems

to present a low impact on the performance of the methods, specially regarding F1-score.

Regarding non-polyp frames, we can observe that the presence of overlay information and overexposed regions helps methods to discard frames without a polyp. Intestinal content leads to more false alarms, as does the presence of the luminal region and the presence of specular highlights; the three of them may falsely indicate the presence of a polyp, as they may also present contrast to mucosa or an indication of protrudness. We can also observe how methods tend to provide a higher number of false alarms for good quality images, which we interpret as a result of structures likely to be confused with polyps being better visually defined.

With respect to individual methods, we observe that those including boundary information (ASU, CVC-CLINIC and SNU) in their methodologies are specially damaged by the presence of structures with strong boundaries such as lumen or overlay information. End-to-end learning approaches are less affected in non-polyp frames analysis and they benefit from the presence of specular highlights in polyp frames.

3) Impact of Polyp Morphology on Methods' Performance: We present in Table X and in Table IX results of the study on the impact on polyp morphology on method's performance. It has to be noted that we only provide results for the morphologies that appear in each particular database, as described in Section III. We can observe that for both still frames and video sequences analysis, methods' performance do depend on polyp morphology. With respect to still frame analysis, we can observe that Recall scores are higher for sessile polyps (including sub-types 0-ISp and 0-IIa+Is) than for those less elevated (including flat) ones. We associate this to appearing sessile polyps more different to

TABLE IX

IMPACT OF POLYP MORPHOLOGY IN METHODS' PERFORMANCE:
VIDEO SEQUENCE ANALYSIS. ONLY FRAMES CONTAINING
A POLYP ARE CONSIDERED FOR METRICS CALCULATION

	0-Is (4 polyps, 2212 images)				0-IIa (5 polyps, 2101 images)			
	Prec	Rec	F1	Lat	Prec	Rec	F1	Lat
ASU	97.4	73.7	83.9	0.5	97.2	47.9	64.1	13.4
CUMED	92.1	86.8	89.4	0.0	77.3	55.3	64.4	16.8
CVC-CLINIC	81.9	62.5	70.9	0.3	19.3	9.3	12.5	46.7
OUS	90.9	77.6	83.7	0.0	92.1	24.1	38.2	18.7
PLS	24.6	62.6	35.3	3.5	10.2	9.9	10.1	46.0
SNU	26.6	23.3	24.9	79.0	15.9	9.7	12.1	42.4

the mucosa and hence attracting the attention of the different methods. We can also observe how CVC-CLINIC and ETIS-LARIB, despite offering worse overall performance, are able to detect all kind of polyps though they obtain worse Precision scores.

Concerning video sequences, differences regarding degree of polyp elevation follow the same trend; in this case we can observe big differences in Recall for all methods but, in this case, Precision is not greatly affected but for the case of CVC-CLINIC, which is logical due to its big dependence on boundary presence to guide polyp detection; boundaries between mucosa and the polyp are less distinguishable for the case of slightly elevated polyps. Finally, this positive increase in Recall score associated to sessile polyps also has an impact in latency score; all teams achieve smaller latency scores for those videos containing polyps of this morphological type (videos 2, 6, 8 and 9) in Fig. 7.

4) *Temporal Coherence on Method's Response*: We present results of our temporal coherence study on Table XI(a). For both consecutive frame and within sequence detection stability, we can observe that results follow the same trend than the analysis of individual frames, being CUMED and ASU the teams which present a higher degree of temporal coherence, despite none of them including temporal information as part of their methodology. We can also observe how CUMED and ASU are able to correctly detect polyp frames in more than half of the polyp frames that each sequence contain, which can be associated to them being more capable to cope with polyp appearance variability within a same sequence.

5) *Analysis of the Direct Output of the Methods*: We present mean and standard deviation values of CR in Table XI(b). As we can observe, CUMED achieves the higher mean CR value across all frames from ETIS-LARIB database, concentrating around half of the total energy of the image inside the polyp. It has to be noted that, in this case, differences between methods can be associated to several reasons. First of all, it is clear that methods detecting correctly more polyps will be prone to concentrate more energy inside them hence the superior performance of CUMED, which was the best performing team in still frame analysis. Second, we also have to consider how the actual output of the method looks like, as it can have an impact in the specific metric considered.

We observe in Fig. 10 how some methods do not provide probabilistic energy maps but binary masks approximating the polyp region. Due to these regions having pre-determined shapes, two problems may appear. First, it is highly unlikely

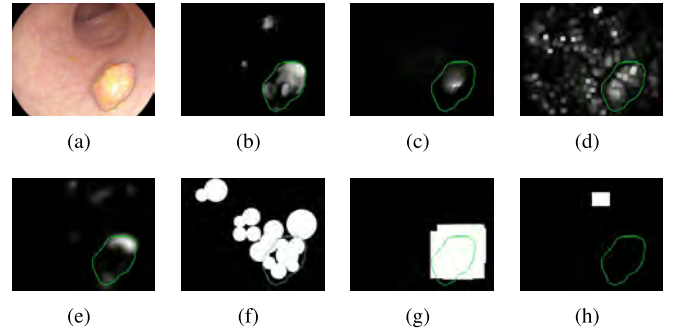


Fig. 10. Comparison of energy maps provided by each method: (a) original image (b) CUMED (c) CVC-CLINIC (d) PLS (e) UNS-UCLAN (f) ETIS-LARIB (g) OUS,s and (h) SNU. In each of the images, a green line represents the reference polyp mask.

that actual polyps fit those shapes, so that any pixel-wise metric value can be damaged by the shape choice. Second, if methods' evaluation is based on the calculation of detection scores from single-position values and this position is calculated as the centroid of the pre-determined shape in case of large regions partially covering the polyp, it may happen that the detection position falls outside the polyp when in fact part of the polyp region was covered by method's output.

Consequently, we think it is not fair to include those methods (ETIS-LARIB, OUS and SNU) in a CR-based comparison. We did statistically compare the different energy map-based methods. Preliminary results shown in Fig. 11 indicate again a superior performance of CUMED over the rest. Regarding the statistical significance of the differences, the Saphiro-Wilk test over CR values indicates a normal distribution of data. Therefore ANOVA and multicomparison tests are performed to study potential differences across methods. The ANOVA test detects significant differences across CR values ($p - value = 2.51e^{-61}$), which are explored in the multicomparison test shown in Fig. 11. CUMED's performance is statistically different from the rest of the approaches, which present comparable performances between them.

V. DISCUSSION

A. Impact of the Methodology Used on Method's Performance

The main result of this comparative study is that methods including some degree of machine learning outperform classic hand-crafted methods, specially regarding specificity scores in non-polyp videos. This correlates with the trend actually observed on computer vision research; methods traditionally were hybrid, using hand-crafted features and machine learning to classify a given input image according to the specific problem to solve. There is an extensive amount of hand-crafted features defined within the computer vision community, covering from general ones such as HOG or SIFT features to others more domain-specific, such as the ones presented by CVC-Clinic team. Designing hand-crafted features to solve specific problems can be complicated and highly time consuming, as well as limiting the widespread use of a new developed technology.

TABLE X
IMPACT OF POLYP MORPHOLOGY IN METHODS' PERFORMANCE: STILL FRAME ANALYSIS

Team	0-Is (27p,127im)			0-IIa (9p,45im)			0-IIb (4p,6im)			0-IIa+c (2p,6 im)			0-Isp (1p,6im)			0-IIa+Is (1p,6im)		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
CUMED	79.5	79.5	79.5	57.4	60.0	58.7	33.3	11.1	16.6	42.8	50.0	46.1	100.0	100.0	100.0	83.3	83.3	83.3
CVC-CLINIC	7.7	44.8	13.1	10.4	48.9	17.1	35.7	27.8	31.2	42.8	100.0	60.0	66.6	100.0	80.0	18.2	100.0	30.7
ETIS-LARIB	5.4	50.4	9.7	13.9	46.6	21.4	6.7	22.2	10.4	75.0	50.0	60.0	11.6	83.3	20.4	18.2	100.0	30.7
OUS	67.3	76.4	71.6	66.6	40.0	50.0	0.0	0.0	0.0	100.0	66.6	80.0	100.0	100.0	100.0	85.7	100.0	92.3
PLS	15.3	58.3	24.2	15.0	60.0	24.1	0.0	0.0	0.0	33.3	100.0	50.0	28.6	100.0	44.4	25.0	100.0	40.0
SNU	11.8	11.8	11.8	4.4	4.4	4.4	0.0	0.0	0.0	16.6	16.6	16.6	0.0	0.0	0.0	33.3	33.3	33.3
UNS-UCLAN	24.2	74.0	36.5	15.3	55.5	24.0	0.0	0.0	0.0	75.0	100.0	85.7	75.0	100.0	85.7	33.3	100.0	50.0

TABLE XI

TEMPORAL COHERENCE AND CONCENTRATION RATIO RESULTS. FOR EACH METHOD, MEAN AND STANDARD DEVIATION VALUES OF CORRESPONDING METRIC ARE PROVIDED. (a) TEMPORAL COHERENCE. (b) CONCENTRATION RATIO

(a) Temporal Coherence

Method	Consecutive frames	Within sequence
ASU	54.8 \pm 21.4	61.2 \pm 21.9
CUMED	63.7 \pm 23.9	67.7 \pm 23.2
CVC-CLINIC	34.1 \pm 37.6	30.3 \pm 37.0
OUS	48.7 \pm 30.8	47.2 \pm 34.1
PLS	27.4 \pm 26.1	31.7 \pm 30.4
SNU	10.2 \pm 06.3	11.7 \pm 11.7

(b) Concentration Ratio

Method	Value
CUMED	48.5 \pm 26.7
CVC-CLINIC	17.9 \pm 24.1
PLS	11.9 \pm 14.4
UNS-UCLAN	18.7 \pm 18.8

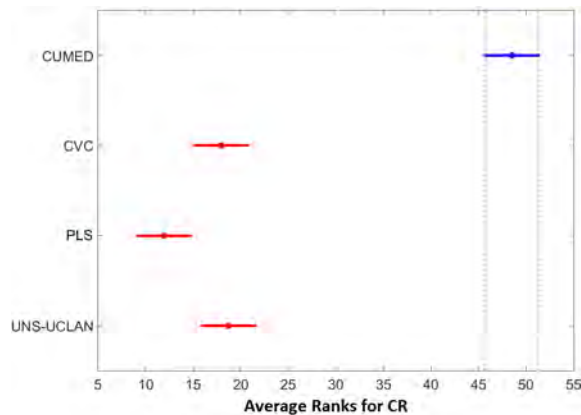


Fig. 11. The multicomparison test for the analysis of the CR score in the ETIS-LARIB database. Each method is represented as a horizontal line centered in corresponding method's mean CR score and whose width corresponds to the variance calculated according anova1 fit model. The best ranked group is represented by a blue line, comparable methods are shown in grey, and statistically significant different methods are shown in red.

CNNs allow to learn jointly problem-specific features and the classifier to differentiate among classes. Their great power comes from the ability of learning problem-specific features in an increasing depth of complexity and abstraction. As it has been shown in this paper, we can observe a superior performance of CNN-based approaches over hand-crafted methods. We can also observe differences in performance between CNN-based methods which shows how obtaining good performance of these networks depends strongly on defining the proper architecture and having quality data to feed the network. Regarding the design of the network, there are several details to take into account which go from pure architecture decisions (number of layers, number and size of

the filters of each layer or activation functions) to how the training is done (choice of optimization method, setting a learning rate, data preprocessing). A proper selection of these parameters may lead to a boost in performance achieved.

Apart from differences related to parameter tuning, we can also observe that one important difference between CNN-based proposals rely on the type of data used to define the network. CUMED uses only colonoscopy data whereas OUS and SNU networks are pre-trained over general image databases. CNNs are structured in layers and each of them captures a different kind of features from the data; first layers capture basic image features such as boundaries whereas deeper ones capture more meaningful and abstract features built over the previous ones. Considering this, features learnt on the first layers might work well in several domains but those learnt in the last layers are more application dependent. One big requirement to use CNNs is to have a large amount of labelled data that may not be available for the case of medical imaging analysis. One widespread solution (used in SNU methodology) is to train the network on a very big database such as ImageNet (with over 10^6 images and 10^3 classes) and then fine tune the network to adapt for a more specific domain. The problem relies on ImageNet containing images potentially very different that the ones that the polyp detection system will have to deal with and, as results show, this may limit the use of those pre-trained networks. In this sense, we can observe how methods using colonoscopy data from end to end (CUMED) obtain better performance than those trained in general datasets such as OUS or SNU, which indicate that efforts should be made to build up domain-specific networks in order to obtained desired performance levels.

Concerning a general comparison among methods regardless their methodology, we can also observe from Table VII that recall scores improve if we only consider frames with polyps. As non-polyp frames are included in the study, performance of hand-crafted and some of the hybrid methods decrease with respect to all metrics. We explain this as being due to some of the methods being specifically tuned for polyp-like structures detection, but not on specific polyp presence or absence; this can be observed in the high number of false positives that these approaches provide as they seem to find these polyp-like structures in frames without a polyp. We link this to non-polyp frames containing structures which guide polyp detection methods, such as, boundary information which also appears associated in other endoluminal structures such as folds or vessels. We also observe strong differences in the performance of hand-crafted methods when dealing with polyp

frames in the two proposed scenarios (still frames and videos). This could be related to the fact that the ETIS-LARIB database presents a high number of lateral polyp views, deviating from the model of appearance which the hand-crafted method is based on [14].

B. Impact of Clinical and Technical Image Challenges on Method's Performance

We have presented in this paper a preliminary study on whether image challenges defined and reported by clinicians and technicians do impact the performance of an automatic polyp detection method. Results exposed in Table VIII show that all of them, in a certain degree, should be tackled in order the automatic system to efficiently assist clinicians during the procedure. The most straightforward conclusion from this experiment is that image quality matters, as methods' performance decrease when only bad quality images are considered. The presence of extra-endoluminal structures such as overlay information or overexposed regions do also affect negatively the performance of automatic methods. This indicates that efforts should be made during the exploration in order a computational support system to efficiently assist clinicians. We can also observe that results do improve if luminal region is not present in the image; this correlates with actual exploration guidelines in which a thorough inspection of the mucosa is prescribed in order to efficiently detect polyps.

It is also interesting to observe how there are some cases in that image challenges considered for both technical and clinical domains do not suppose an actual technical challenge. For instance, we would expect that the presence of intestinal content or the observation of specular highlights over the polyp would impact negatively the performance of an automatic method; results show that studied methods are indeed positively affected by their presence. We associate this to these image challenges appearing clearly different from polyps.

Moreover, and also related to this, there are some image challenges which may provide unexpected results and which would need to be better defined to avoid potential subjectivity. Though we have gathered three observations per frame to mitigate this, some of the image challenges should be defined appropriately to avoid discrepancies between observers. For instance, the presence of intestinal content, image quality or, specially, the high specular highlights presence should be redefined as, for the former, we should also consider its size and type (solid, bubbles) and, for the latter, we may consider not only the number but their size and position in the image.

Apart from the image challenge experiment, we have also performed another one to assess the impact of polyp morphology on methods' performance. Even considering that this experiment is limited to the actual morphologies that the databases contain, differences can already be observed in a way such methods obtain better performance as the polyp protrudes more from the mucosa. With respect to polyps with flat morphology cited by clinicians as the most difficult ones to the detect [9], [10] we observe, for the case of still frames analysis in which they are present, that there are methods that are already able to detect them, despite its low presence in training and testing databases.

TABLE XII

SUMMARY OF INDIVIDUAL METHOD'S PERFORMANCE. FOR LATENCY AND TEMPORAL COHERENCE MEAN AND STANDARD DEVIATION VALUES FROM THE ANALYSIS OF THE 9 VIDEOS WITH POLYPS ARE PROVIDED

Method	DP [%]	Latency [frames]	Rec [%]	F1 [%]	TempC [%]
ASU	100.0	7.4 ± 15.6	61.1	73.9	54.8 ± 21.4
CUMED	100.0	9.3 ± 16.4	71.4	75.9	63.7 ± 23.9
CVC-CLINIC	77.8	26.8 ± 39.0	36.6	33.7	34.1 ± 37.6
OUS	88.9	9.4 ± 17.2	51.5	65.7	48.7 ± 30.8
PLS	88.9	24.7 ± 37.8	36.9	19.9	27.4 ± 26.1
SNU	77.8	32.5 ± 40.9	10.6	11.2	10.2 ± 06.3

C. Towards Clinical Applicability

One of the objectives of the challenge and, consequently, of this paper, was to assess if any of the participating methods is close to clinical applicability. In order to assess this, we have proposed several studies to observe certain features that a given clinically applicable computational polyp detection method should have. The main feature that a clinically applicable system should have is that it should detect all polyps regardless their appearance (high detection rate (DR), measured as the percentage of polyps detected in at least one frame out of the total of polyps present in the testing videos). This detection should also be fast enough to be of an actual help; speed here is not only considered in terms of computation time but also in response to a stimuli as a computational method should react to polyp presence as soon as it appears in the image (associated to a low latency score). The actual response of a given method should be stable over time (high temporal coherence) in order to provide an smooth assistance to clinicians in polyp search. Finally, and considering the scope of application of the methods, the number of false alarms should be kept low (high F1 score associated to an also high Recall value) as the contrary would suppose indicating the clinicians to further explore uninteresting regions of the image.

Considering these criteria, we present in Table XII a summary of the main results presented in this paper for the video analysis challenge. Columns are ordered according to the, under our point of view, relevance of the specific criteria. As it can be seen, there are only two methods (CUMED, ASU) that may be actually considered for a potential clinical use as they do detect all polyps. Concerning the rest of criteria, both do perform similarly: ASU presents a lower latency which could be compensated by CUMED's higher temporal coherence degree. Concerning frame-based metrics we can observe that ASU leads to a better balance between true and false alarms though CUMED detects polyps in more frames.

It has to be noted that we have not included in Table XII information regarding computation time for comparison purposes as they have not been tested under the same configuration and, consequently, provided times may vary in an actual final deployed system. Nevertheless, a clinically useful method should operate under real-time constraints. Considering that videos are recorded on 25 or 30 frames per second, processing of a new frame should not take more than 40 ms (33 ms for NTSC systems) in order not to suppose a delay in overall procedure time. All methods studied in this paper

have computation times higher than these threshold values and, consequently, do not comply with real-time constraints though the processing of all frames might not be needed considering due to the small variation between consecutive frames due to usual smooth camera movement.

D. Possible Improvements in Validation Framework

During the analysis of the performance of each of the methods, we have discovered several aspects to be considered for future iterations of this study.

The first one deals with the variability of the image quality provided in the training and testing stages. In this study, the databases used for validation come from three different sources presenting differences in image size or acquisition system, as we have source data from both OLYMPUS and PENTAX devices. This was done on purpose, as it is impossible to predict under which specific scenario a given system can be potentially used, as there is no standard regarding resolution or manufacturer and a given method should perform similarly regardless of the specific conditions. But it is true that these variabilities may have affected the performance of the different methods, as training was done using images with resolutions different from the ones used for testing. These differences in resolution can imply to have a greater level of texture detail which can impact the performance of systems trained with SD images (i.e. edge detection could be greatly affected by the presence of small texture details inside the polyp).

Also related to database content, and after observing that polyp morphology can impact methods' performance, an effort could also be made on enlarging the databases to cover those types that are not currently present. It is important to mention that performance of learning-based approaches for certain morphologies could be affected by the lack of frames of this particular type in the training set. In our experiment, this only happens for still frames analysis as CVC-CLINIC database does not contain polyps of types 0-IIa+c and 0-IIa+Is which are indeed present in ETIS-LARIB database. Nevertheless it has to be noted that these types are only present in 12 frames out of the 196 frames of the database and, consequently, global performance should not be greatly affected by this issue. Finally, not all types are represented in the database (for instance, proposed databases have no examples of types 0-Ip, 0-IIc or 0-III); it would surely be helpful to study how computational methods deal with those polyp types reported as the ones with higher associated miss-rate [9], [10].

The second one of improvement deals with how actual results are calculated. The majority of results presented in this paper have been calculated from the CSV files provided by participants in the challenge. Though they are useful to represent the actual performance of the method, we think it is also necessary to analyze how these files are generated (the actual output of the method they come from) in order to have a deep understanding on how a given method performs. In this sense, we proposed a preliminary study comparing the amount of actual image energy that is kept inside the polyp. As it was shown in Fig. 10, there are big differences in how the actual output of the methods is calculated, inherent in each teams' methodology. Therefore, if we wanted to present a fair

comparison between methods over their direct output, specific guidelines should be given to participants in order to gather comparable data.

Finally, we think that Precision-Recall and ROC curves should be used for methods' comparison as well. In order to provide these curves for all teams, confidence values should have been provided; in this case, only one team per sub-category (UNS-UCLAN in still-frame analysis and ASU-Mayo for full video analysis) provided this information whereas the rest only provided what we assume are results obtained using the best configuration of each particular method. Nevertheless, we have presented both curves in Fig. 6 along with quantitative data in both Table VI and Table VII.

VI. CONCLUSIONS

We present in this paper a complete validation study comparing the performance of different polyp detection methods. Eight different teams took part in this challenge, ranging from methodologies based on hand-crafted methods to trending techniques such as CNNs. We propose the use of uniform performance metrics and common, publicly available, fully annotated databases to objectively assess their performance.

The analysis of the results obtained by each method shows a superior performance by methods using machine learning as part of their methodologies, obtaining promising performance in both still frames and full-sequence sets. The global analysis of methods' performance shows that some of them are close to be clinically applicable as they are able to detect polyps in all sequences with a small reaction time. We have also shown how there is a clear link between clinical and technical challenges and that mitigating them is key to improve methods' performance. As it was expected, our preliminary study proves that image quality and careful mucosa inspection do have a positive impact in methods' performance.

A deep analysis of the results shows that, as different methods detect different polyps, there is room for improvement by combining some of the methods into a new solution. Going along this line, we have performed a first study on how to combine some of the methods in order to improve detection performance. Preliminary results show how the combination of the best methods can be used to exceed best individual scores, indicating the potential of creating clinically useful systems integrating capabilities from several individual methods.

Beyond presenting challenge results, this study shows areas in which methods might focus to increase their performance, such as the ability to work equally under different conditions, the necessity of include spatial and temporal coherence in full sequences analysis or by considering the presence of other elements of the scene to help in polyp detection task. More importantly, this study also shows how efforts should be made between clinicians and computer scientists to build up image acquisition protocols that can help to better observe (clinicians) and analyze (computational methods) the endoluminal scene. Finally and concerning availability of data to test methods, the study shows that granting access to large available labelled data is needed for a comprehensive validation of a polyp detection method and that this might lead to a boost in performance of end-to-end learning methods.

We believe efforts should also be made to create and use data from new imaging technologies such as magnification endoscopy or virtual chromoendoscopy, due to increased visualization performance already observed by clinicians.

After analyzing the complete validation study, we have detected several areas in which the study can be extended to provide with an even deeper comparative analysis of the performance of polyp detection methods. More precisely, future studies should tackle some of the issues detected such as the variability in source data resolution and size and should aim to cover all different polyp morphological types. Moreover, a consensus should be reached on how the information provided by each method is to be interpreted to allow a comparison beyond simple detection positions. This may result in, apart from a more complete analysis, a deeper understanding on how each method works and in which scenarios each of them show the most benefit, thinking of potential optimized combinations of them to finally build up a clinically useful method.

ACKNOWLEDGEMENTS

The authors would like to thank EndoVis challenge organizers for their continuous help and guidance through both challenge and paper preparation. The idea of organizing a competition for polyp detection in colonoscopy was first conceived by Dr. Jianming Liang, and the foundational framework was established by Drs. Tajbakhsh and Liang before the first challenge at ISBI-2015. The associated ground truth images in the ASU-Mayo Clinic Colonoscopy Video © Database were created by Saiswathi Javangula, Ireen Khan, Kamran Bodushev, Sarah Fallah-Adl, and Tracy Phan. The ASU-Mayo Clinic Colonoscopy Video © Database is copyrighted and its use is granted to the work for the challenge on polyp detection in colonoscopy as reported this IEEE TMI paper. For any other uses, a prior agreement must be obtained from Arizona State University.

REFERENCES

- [1] *Colorectal Cancer*, American Cancer Society, Atlanta, GA, USA, Jan. 2016.
- [2] M. Gschwanter and S. E. A. Kriwanek, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, 2002.
- [3] A. Jemal *et al.*, "Cancer statistics, 2008," *CA, Cancer J. Clin.*, vol. 58, no. 2, pp. 71–96, 2008.
- [4] R. Jover *et al.*, "Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy (ESGE) guideline," *Endoscopy*, vol. 45, no. 10, pp. 842–851, 2013.
- [5] D. Burling *et al.*, "CT colonography standards," *Clin. Radiol.*, vol. 65, no. 6, pp. 474–480, 2010.
- [6] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, p. 417, May 2000.
- [7] C. D. Johnson *et al.*, "Accuracy of CT colonography for detection of large adenomas and cancers," *New England J. Med.*, vol. 359, no. 12, pp. 1207–1217, 2008.
- [8] M. J. Farnbacher *et al.*, "Quickview video preview software of colon capsule endoscopy: Reliability in presenting colorectal polyps as compared to normal mode reading," *Scandin. J. Gastroenterol.*, vol. 49, no. 3, pp. 339–346, 2014.
- [9] A. Leufkens *et al.*, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.
- [10] J. C. Van Rijn *et al.*, "Polyp miss rate determined by tandem colonoscopy: A systematic review," *Amer. J. Gastroenterol.*, vol. 101, no. 2, pp. 343–350, 2006.
- [11] B. Lebowitz *et al.*, "The impact of suboptimal bowel preparation on adenoma miss rates and the factors associated with early repeat colonoscopy," *Gastrointestinal Endoscopy*, vol. 73, no. 6, pp. 1207–1214, 2011.
- [12] S.-H. Lee *et al.*, "An adequate level of training for technical competence in screening and diagnostic colonoscopy: A prospective multicenter evaluation of the learning curve," *Gastrointestinal Endoscopy*, vol. 67, no. 4, pp. 683–689, 2008.
- [13] M. A. Armin *et al.*, "Visibility map: A new method in evaluation quality of optical colonoscopy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 396–404.
- [14] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.
- [15] M. J. Bruno, "Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis," *Gut*, vol. 52, no. 4, pp. iv7–iv11, 2003.
- [16] H. Inomata *et al.*, "Efficacy of a novel auto-fluorescence imaging system with computer-assisted color analysis for assessment of colorectal lesions," *World J. Gastroenterol.*, vol. 19, no. 41, pp. 7146–7153, 2013.
- [17] H. Machida *et al.*, "Narrow-band imaging in the diagnosis of colorectal mucosal lesions: A pilot study," *Endoscopy*, vol. 36, no. 12, pp. 1094–1098, 2004.
- [18] R. Coriat *et al.*, "Computed virtual chromoendoscopy system (FICE): A new tool for upper endoscopy?" *Gastroentérol. Clin. Biol.*, vol. 32, no. 4, pp. 363–369, 2008.
- [19] A. Hoffman *et al.*, "High definition colonoscopy combined with i-Scan is superior in the detection of colorectal neoplasias compared with standard video colonoscopy: A prospective randomized controlled trial," *Endoscopy*, vol. 42, no. 10, pp. 827–833, 2010.
- [20] N. Gupta *et al.*, "Accuracy of *in vivo* optical diagnosis of colon polyp histology by narrow-band imaging in predicting colonoscopy surveillance intervals," *Gastrointestinal Endoscopy*, vol. 75, no. 3, pp. 494–502, 2012.
- [21] D. K. Iakovidis and A. Koulouzidis, "Software for enhanced video capsule endoscopy: Challenges for essential progress," *Nature Rev. Gastroenterol. Hepatol.*, vol. 12, no. 3, pp. 172–186, 2015.
- [22] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.
- [23] L. Maier-Hein *et al.*, "Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1913–1930, Oct. 2014.
- [24] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.
- [25] I. Yuji *et al.*, "Automatic polyp detection in endoscope images using a Hessian filter," in *Proc. MVA*, 2013, pp. 21–24, 2013.
- [26] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, 2014.
- [27] H. Zhu, Y. Fan, and Z. Liang, "Improved curvature estimation for shape analysis in computer-aided detection of colonic polyps," in *Proc. Int. Workshop Comput. Challenges Clin. Opportunities Virtual Colonoscopy Abdominal Imag.*, 2010, pp. 9–14.
- [28] J. Kang and R. Doraiswami, "Real-time image processing system for endoscopic applications," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, vol. 3, May 2003, pp. 1469–1472.
- [29] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Oct. 2007, pp. II-465–II-468.
- [30] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.
- [31] S. Ameling *et al.*, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin*. Berlin, Germany: Springer, 2009, pp. 346–350.
- [32] S. Gross *et al.*, "A comparison of blood vessel features and local binary patterns for colorectal polyp classification," *Proc. SPIE*, vol. 6918, Feb. 2009.
- [33] Q. Angermann, A. Histace, and O. Romain, "Active learning for real time detection of polyps in videocolonoscopy," *Procedia Comput. Sci.*, vol. 90, pp. 182–187, Jul. 2016.
- [34] S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," *Proc. SPIE*, vol. 9785, Mar. 2016.

- [35] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *Proc. IEEE 29th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2016, pp. 253–258.
- [36] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2379–2393, Nov. 2015.
- [37] I. Ševo *et al.*, "Edge density based automatic detection of inflammation in colonoscopy videos," *Comput. Biol. Med.*, vol. 72, pp. 138–150, May 2016.
- [38] H. Chen, X. J. Qi, J. Z. Cheng, and P. A. Heng, "Deep contextual networks for neuronal structure segmentation," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1167–1173.
- [39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. (2014). "Deeply-supervised nets." [Online]. Available: <https://arxiv.org/abs/1409.5185>
- [40] H. Chen, X. Qi, L. Yu, and P.-A. Heng. (2016). "DCAN: Deep contour-aware networks for accurate gland segmentation." [Online]. Available: <https://arxiv.org/abs/1604.02677>
- [41] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.
- [42] S. Park, M. Lee, and N. Kwak, "Polyp detection in colonoscopy videos using deeply-learned hierarchical features," Dept. Transdisciplinary Sci., Seoul Nat. Univ., Suwon, South Korea, Tech. Rep. n°0b, 2015.
- [43] S. Demyanov. *A Convolutional Neural Network Toolbox*. [Online]. Available: <https://github.com/sdemyanov/ConvNet>, 2013
- [44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [45] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [46] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678. [Online]. Available: <http://doi.acm.org/10.1145/2647868.2654889>
- [47] M. Riegler, M. Larson, M. Lux, and C. Kofler, "How 'how' reflects what's what: Content-based exploitation of how users frame social images," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 397–406.
- [48] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection using global geometric constraints and local intensity variation patterns," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Berlin, Germany: Springer, 2014, pp. 179–187.
- [49] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 79–83.
- [50] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.
- [51] Participants in the Paris Workshop, "The Paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon: November 30 to December 1, 2002," *Gastrointest Endoscopy*, vol. 58, no. 6, pp. S3–S43, 2003.
- [52] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

United Snakes

Jianming Liang^{a,*}, Tim McInerney^{b,c}, Demetri Terzopoulos^{d,c}

^a *Computer Aided Diagnosis and Therapy, Siemens Medical Solutions USA, Inc., Malvern, PA 19355, USA*

^b *Department of Computer Science, Ryerson University, Toronto, Ont., Canada M5B 2K3*

^c *Department of Computer Science, University of Toronto, Toronto, Ont., Canada M5S 3H5*

^d *Department of Computer Science, University of California at Los Angeles, Los Angeles, CA 90095, USA*

Received 21 September 2005; accepted 27 September 2005

Available online 28 November 2005

Abstract

Since their debut in 1987, snakes (active contour models) have become a standard image analysis technique with several variants now in common use. We present a framework called “United Snakes”, which has two key features. First, it unifies the most popular snake variants, including finite difference, B-spline, and Hermite polynomial snakes in a consistent finite element formulation, thus expanding the range of object modeling capabilities within a uniform snake construction process. Second, it embodies the idea that the heretofore presumed competing technique known as “live wire” or “intelligent scissors” is in fact complementary to snakes and that the two techniques can advantageously be combined by introducing an effective hard constraint mechanism. The United Snakes framework amplifies the efficiency and reproducibility of the component techniques, and it offers more flexible interactive control while further minimizing user interactions. We apply United Snakes to several different medical image analysis tasks, including the segmentation of neuronal dendrites in EM images, dynamic chest image analysis, the quantification of growth plates in MR images and the isolation of the breast region in mammograms, demonstrating the generality, accuracy and robustness of the tool.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Snakes; Active contours; Live wire; Intelligent scissors; Finite elements; Interactive image analysis; Neuronal dendrite segmentation; Dynamic chest image analysis; Growth plate quantification; Breast region isolation

1. Introduction

Snakes (active contour models) quickly gained popularity following their debut in 1987 (Kass et al., 1988). They have proven to be especially useful in medical image analysis (McInerney and Terzopoulos, 1996; Singh et al., 1998) and for tracking moving objects in video (Terzopoulos and Szeliski, 1992; Blake and Isard, 1998), among other applications. Variants such as finite element snakes (Cohen and Cohen, 1993), B-snakes (Menet et al., 1990; Blake and Isard, 1998), and Fourier snakes (Staib and Duncan, 1992) have been proposed in an effort to improve aspects of the original finite difference implementation (e.g., to decrease initialization sensitivity, increase robustness

against noise, improve selectivity for certain classes of objects, etc.). No formulation has yet emerged as the “gold standard”. Rather, the primary variants seem well-suited to different applications with particular image modalities and processing scenarios.

Given the broad array of choices for the user, there is a need for a portable and reusable snakes implementation which unites the best features of the variants while maintaining the simplicity and elegance of the original formulation. To this end, our first contribution in this paper is to unify the most important snakes variants, including finite difference, B-spline, and Hermite polynomial snakes, in a comprehensive finite element formulation, where a particular type of snake can be derived by simply changing the finite element shape functions at the user level.

Subsequent to snakes, a related technique, known as “live wire” or “intelligent scissors” (Mortensen et al.,

* Corresponding author. Tel.: +1 610 448 1460.

E-mail address: jianming.liang@computer.org (J. Liang).

1995; Falcão et al., 1996; Barrett and Mortensen, 1997; Falcão and Udupa, 1997; Mortensen and Barrett, 1998; Falcão et al., 1998, 2000) emerged as an effective interactive boundary tracing tool. Based on dynamic programming (Falcão et al., 1998) or Dijkstra's graph search algorithm (Mortensen and Barrett, 1998), it was originally developed as an interactive 2D extension to earlier optimal boundary tracking methods. Live wire features several similarities with snakes, but it is generally considered in the literature as a competing technique. Our second contribution in this paper is the idea that live wire and snakes are in fact complementary techniques that can be advantageously combined via a simple yet effective method for imposing hard constraints on snakes. An advantage of this combination is the efficient handling of large images – a potential obstacle for live wire alone.

We call our software implementation *United Snakes* (Liang et al., 1999a,b), because it unites several snake variants with live wire to offer a general purpose tool for interactive image segmentation that provides more flexible control while reducing user interaction. United Snakes is implemented in the highly portable Java programming language. We have applied United Snakes to several different medical image analysis tasks including the segmentation of neuronal dendrites in EM images, dynamic chest image analysis, the quantification of growth plates in MR images and the isolation of the breast region in mammograms, demonstrating the generality, accuracy, robustness, and ease of use of the tool.

In the remainder of this paper, we first describe our finite element framework in Section 2 and show how several snake variants can be integrated within it. Section 3 describes the live wire technique. We justify the idea of combining snakes with live wire in Section 4 and develop a hard constraint mechanism in Section 5 that makes this combination possible. Section 6 presents results utilizing the United Snakes system in medical image segmentation scenarios. We conclude in Section 7 and propose future extensions of United Snakes.

2. Finite element unification of snakes

A snake is a time-varying parametric contour $\mathbf{v}(s, t) = (x(s, t), y(s, t))^T$ in the image plane $(x, y) \in \mathbb{R}^2$, where x and y are coordinate functions of parameter s and time t . The shape of the contour subject to an image $I(x, y)$ is dictated by an energy functional $\mathcal{E}(\mathbf{v}) = \mathcal{S}(\mathbf{v}) + \mathcal{P}(\mathbf{v})$. The first term is the internal deformation energy defined as

$$\mathcal{S}(\mathbf{v}) = \frac{1}{2} \int_0^L \alpha(s) \left| \frac{\partial \mathbf{v}}{\partial s} \right|^2 + \beta(s) \left| \frac{\partial^2 \mathbf{v}}{\partial s^2} \right|^2 ds, \quad (1)$$

where $\alpha(s)$ controls the “tension” of the contour and $\beta(s)$ regulates its “rigidity”. The second term is an external image energy

$$\mathcal{P}(\mathbf{v}) = \int_0^L P_I(\mathbf{v}) ds, \quad (2)$$

which couples the snake to the image via a scalar potential function $P_I(x, y)$ typically computed from $I(x, y)$ through image processing. The Euler–Lagrange equations of motion for a dynamic snake are

$$\mu \frac{\partial^2 \mathbf{v}}{\partial t^2} + \gamma \frac{\partial \mathbf{v}}{\partial t} - \frac{\partial}{\partial s} \left(\alpha \frac{\partial \mathbf{v}}{\partial s} \right) + \frac{\partial^2}{\partial s^2} \left(\beta \frac{\partial^2 \mathbf{v}}{\partial s^2} \right) = \mathbf{q}(\mathbf{v}). \quad (3)$$

The first two terms represent inertial forces due to the mass density $\mu(s)$ and damping forces due to the dissipation density $\gamma(s)$. The next two terms represent the internal stretching and bending deformation forces. On the right-hand side are the external forces $\mathbf{q}(\mathbf{v}) = -\nabla P_I(\mathbf{v}) + \mathbf{f}(s, t)$, where the image forces are the negative gradient of the image potential function. The user may guide the dynamic snake via time-varying interaction forces $\mathbf{f}(s, t)$ (usually applied through an input device such as a mouse), driving the snake out of one energy minimizing equilibrium and into another. Viewed as a dynamical system, the snake may also be used to track moving objects in a time-varying (video) image $I(x, y, t)$.

2.1. Finite element formulation

In a finite element formulation (Zienkiewicz and Taylor, 1989), the parametric domain is partitioned into finite sub-domains, so that the snake contour is divided into “snake elements”. Each element e is represented geometrically using shape functions $\mathbf{N}(s)$ and nodal variables $\mathbf{u}^e(t)$. The nodal variables of all the elements are assembled into the snake nodal variable vector $\mathbf{u}(t)$. This leads to a discrete form of the equations of motion (3) as a system of second-order ordinary differential equations in $\mathbf{u}(t)$:

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{g}, \quad (4)$$

where \mathbf{M} is the mass matrix, \mathbf{C} is the damping matrix, \mathbf{K} is the stiffness matrix, and \mathbf{g} is the external force vector, which are assembled from corresponding element sub-matrices that depend on the shape functions \mathbf{N} (Appendix A details the finite element formulation).

By using different shape functions and thereby generating different stiffness matrices, the behavior of the resulting snake can be adapted to specific tasks. For example, snakes that use B-spline shape functions are typically characterized by a low number of degrees of freedom, typically use polynomial basis functions of degree 2 or higher, and are inherently very smooth. Therefore, these “B-snakes” (Menet et al., 1990; Blake and Isard, 1998) can be effective in segmentation or tracking tasks involving noisy images where the target object boundaries may exhibit significant gaps in the images. On the other hand, object boundaries with many fine details or rapid curvature variations may best be segmented by a snake that uses simpler shape functions and more degrees of freedom, such as a finite difference snake (Kass et al., 1988). Various contour representations are reviewed in Gavrilu (1996). The unification of these different shape functions in a single framework expands the range of object modeling capabilities,

and the range of segmentation and tracking scenarios that can be handled by a single tool.

The following sections address Hermitian shape functions, B-spline shape functions, and “shape functions” for finite difference snakes. Since the two coordinate functions $x(s)$ and $y(s)$ of the snake $\mathbf{v}(s)$ are independent, we shall discuss the shape functions in terms of only one component $x(s)$; the shape functions for $y(s)$ assume an identical form.

2.2. Hermitian shape functions

In the case of Hermitian snakes, $x(s)$ ($0 \leq s \leq l$, where l is the element parametric length) is approximated with a cubic polynomial function, parameterized by position x and slope θ at the endpoints $s = 0$ and $s = l$ of an element. We can show that $x(s) = \mathbf{N}_h \mathbf{u}^{e_i}$, where $\mathbf{u}^{e_i} = [x_i, \theta_i, x_{i+1}, \theta_{i+1}]^T$ are the nodal variables of element e_i and $\mathbf{N}_h = \mathbf{sH}$ are the Hermitian shape functions, with $\mathbf{s} = [1, s, s^2, s^3]$ and the *Hermitian shape matrix* is

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3/l^2 & -2/l & 3/l^2 & -1/l \\ 2/l^3 & 1/l^2 & -2/l^3 & 1/l^2 \end{bmatrix}. \quad (5)$$

It is reasonable to assume that the mass density $\mu(s)$, the dissipation density $\gamma(s)$, the tension function $\alpha(s)$ and rigidity function $\beta(s)$ are constant within the element. Hence, for element e_i , the mass matrix is

$$\mathbf{M}^{e_i} = \mu_i 420l \begin{bmatrix} 156 & 22l & 54 & -13l \\ 22l & 4l^2 & 13l & -3l^2 \\ 54 & 13l & 156 & -22l \\ -13l & -3l^2 & -22l & 4l^2 \end{bmatrix} \quad (6)$$

the damping matrix is

$$\mathbf{C}^{e_i} = \gamma_i 420l \begin{bmatrix} 156 & 22l & 54 & -13l \\ 22l & 4l^2 & 13l & -3l^2 \\ 54 & 13l & 156 & -22l \\ -13l & -3l^2 & -22l & 4l^2 \end{bmatrix} \quad (7)$$

and the stiffness matrices associated with the tension and rigidity components are, respectively,

$$\mathbf{K}_\alpha^{e_i} = \frac{\alpha_i}{30l} \begin{bmatrix} 36 & 3l & -36 & 3l \\ 3l & 4l^2 & -3l & -l^2 \\ -36 & -3l & 36 & -3l \\ 3l & -l^2 & -3l & 4l^2 \end{bmatrix}, \quad (8)$$

$$\mathbf{K}_\beta^{e_i} = \frac{\beta_i}{l^3} \begin{bmatrix} 12 & 6l & -12 & 6l \\ 6l & 4l^2 & -6l & 2l^2 \\ -12 & -6l & 12 & -6l \\ 6l & 2l^2 & -6l & 4l^2 \end{bmatrix}. \quad (9)$$

An analytic form of the external forces $\mathbf{q}(\mathbf{v})$ in (3) is generally not available. Therefore, Gauss–Legendre quadrature

(Kwon and Bang, 1997) may be employed to approximate the value of the integral for the element external force vector \mathbf{F}^e . For element e_i we have

$$\mathbf{F}_x^{e_i} = \int_0^l \mathbf{N}_h^T \mathbf{q}_x(\mathbf{v}(s)) ds = \sum_j \rho_j \mathbf{N}_h(\xi_j)^T \mathbf{q}_x(\mathbf{v}(\xi_j)), \quad (10)$$

where the subscript x indicates the association with coordinate function $x(s)$, and where ξ_j and ρ_j are the j th Gaussian integration point and its corresponding weighting coefficient, respectively. $\mathbf{F}_y^{e_i}$ is derived in a similar fashion.

To make the global matrix assembly process identical for all shape functions, we introduce *assembling matrices*. Suppose that we have a snake with n elements and N nodes ($N = n$ if the snake is closed and $N = n + 1$ if it is open). For the i th element e_i of the snake ($0 \leq i \leq n - 1$), the assembling matrices are $\mathbf{G}_M = \mathbf{G}_C = \mathbf{G}_\alpha^{e_i} = \mathbf{G}_\beta^{e_i} = \mathbf{G}_F^{e_i} = \mathbf{G}^{e_i}$, where

$$(\mathbf{G}^{e_i})_{jk} = \begin{cases} 1 & \text{if } (j + di) \bmod (dN) = k, \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

are $(2d) \times (dN)$ matrices, with d the number of degrees of freedom of each node in an element (here $d = 2$). Hence, \mathbf{K}_α , \mathbf{K}_β and \mathbf{F} may be assembled as follows:

$$\mathbf{M} = \sum_{i=0}^{n-1} (\mathbf{G}_M^{e_i})^T \mathbf{M}^{e_i} (\mathbf{G}_M^{e_i}), \quad (12)$$

$$\mathbf{C} = \sum_{i=0}^{n-1} (\mathbf{G}_C^{e_i})^T \mathbf{C}^{e_i} (\mathbf{G}_C^{e_i}), \quad (13)$$

$$\mathbf{K}_\alpha = \sum_{i=0}^{n-1} (\mathbf{G}_\alpha^{e_i})^T \mathbf{K}_\alpha^{e_i} (\mathbf{G}_\alpha^{e_i}), \quad (14)$$

$$\mathbf{K}_\beta = \sum_{i=0}^{n-1} (\mathbf{G}_\beta^{e_i})^T \mathbf{K}_\beta^{e_i} (\mathbf{G}_\beta^{e_i}), \quad (15)$$

$$\mathbf{F} = \sum_{i=0}^{n-1} (\mathbf{G}_F^{e_i})^T \mathbf{F}^{e_i}. \quad (16)$$

In our implementation, we set the element parametric length l to 1. Only the shape matrix and the assembling matrices are determined by specific shape functions. Therefore, in the following section we shall focus only on the derivation of the shape matrix and the assembling matrices for B-spline shape functions, and briefly mention other kinds of shape functions which are suitable for snakes.

2.3. B-spline shape functions

For B-spline shape functions, the $x(s)$ coordinate function of $\mathbf{v}(s)$ is constructed as a weighted sum of N_B basis functions $B_n(s)$, for $n = 0, \dots, N_B - 1$, as follows: $x(s) = \mathbf{B}(s)^T \mathbf{Q}^x$, where $\mathbf{B}(s) = [\mathbf{B}_0(s), \dots, \mathbf{B}_{N_B-1}(s)]^T$, $\mathbf{Q}^x = [x_0, \dots, x_{N_B-1}]^T$ and x_i are the weights applied to the respective basis functions $B_n(s)$.

A B-spline span serves as an element in our finite element formulation (hence “span” and “element” are interchangeable terms). Consequently, we shall determine the

nodal variables, the shape matrix, and the assembling matrix associated with a span. When all spans are of unit length, the knot multiplicities at the breakpoints are m_0, \dots, m_L (L is the number of spans and the total number of knots $N_B = \sum_{i=0}^L m_i$), the knot values k_i are determined by $k_i = l$, such that $0 \leq (i - \sum_{j=0}^l m_j) < m_{l+1}$. Furthermore, the n th polynomial $B_{n,d}^\sigma$ in span σ can be computed as follows:

$$B_{n,1}^\sigma(s) = \begin{cases} 1 & \text{if } k_n \leq \sigma < k_{n+1}, \\ 0 & \text{otherwise,} \end{cases} \quad (17)$$

$$B_{n,d}^\sigma(s) = \frac{(s + \sigma - k_n)B_{n,d-1}^\sigma(s)}{k_{n+d-1} - k_n} + \frac{(k_{n+d} - s - \sigma)B_{n+1,d-1}^\sigma(s)}{k_{n+d} - k_{n+1}}. \quad (18)$$

For span σ , the index b_σ for the first basis function whose support includes the span can be determined as $b_\sigma = [(\sum_{i=0}^\sigma m_i) - d] \bmod N_B$. Therefore,

$$I = [b_\sigma, (b_\sigma + 1) \bmod N_B, \dots, (b_\sigma + d - 1) \bmod N_B]$$

are the indices of the nodal variables and also those of the d polynomials $B_{n,d}^\sigma$.¹ Now, the shape matrix for span σ can be constructed by collecting the coefficients of each of the d polynomials $B_{n,d}^\sigma$ as its columns. For example, the shape matrix of a regular cubic B-spline is

$$\mathbf{H} = \begin{bmatrix} 1/6 & 2/3 & 1/6 & 0 \\ -1/2 & 0 & 1/2 & 0 \\ 1/2 & -1 & 1/2 & 0 \\ -1/6 & 1/2 & -1/2 & 1/6 \end{bmatrix} \quad (19)$$

and the element matrices for element e_i are

$$\mathbf{M}^{e_i} = \frac{\mu_i}{5040} \begin{bmatrix} 20 & 129 & 60 & 1 \\ 129 & 1188 & 933 & 60 \\ 60 & 933 & 1188 & 129 \\ 1 & 60 & 129 & 20 \end{bmatrix}, \quad (20)$$

$$\mathbf{C}^{e_i} = \frac{\gamma_i}{5040} \begin{bmatrix} 20 & 129 & 60 & 1 \\ 129 & 1188 & 933 & 60 \\ 60 & 933 & 1188 & 129 \\ 1 & 60 & 129 & 20 \end{bmatrix}, \quad (21)$$

$$\mathbf{K}_\alpha^{e_i} = \frac{\alpha_i}{120} \begin{bmatrix} 6 & 7 & -12 & -1 \\ 7 & 34 & -29 & -12 \\ -12 & -29 & 34 & 7 \\ -1 & -12 & 7 & 6 \end{bmatrix}, \quad (22)$$

$$\mathbf{K}_\beta^{e_i} = \frac{\beta_i}{6} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -3 & 6 & -3 & 0 \\ 0 & -3 & 6 & -3 \\ 1 & 0 & -3 & 2 \end{bmatrix}. \quad (23)$$

¹ In an open B-spline snake, d knots are introduced at the two ends. As a result, the index for the first basis function in the first span is zero (i.e., $b_0 = 0$) and the index of the last basis function in the last span is $N_B - 1$. For a closed B-spline snake, the index needs to be wrapped properly (Blake and Isard, 1998).

The assembling matrix \mathbf{G}^{e_i} can be defined as

$$(\mathbf{G}^{e_i})_{jk} = \begin{cases} 1 & \text{if } (j + b_\sigma) \bmod N_B = k, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

In a similar fashion as above, we may construct other kinds of shape functions; for instance, NURBS shape functions (Terzopoulos and Qin, 1994), Catmull-Rom shape functions, Bézier shape functions, and Fourier shape functions (Staib and Duncan, 1992). The latter are global shape functions over the whole snake, thus the associated assembling matrix becomes an identity matrix.

2.4. Finite difference snakes in element form

Despite the differences between finite element snakes and finite difference snakes, the finite difference snakes can also be constructed in the finite element fashion, using the Dirac delta function $\delta(s)$ as the shape function. The construction primitives are as follows. For a snake with n nodes, \mathbf{M}^{e_i} is a 1×1 matrix and its corresponding assembling matrix $\mathbf{G}_M^{e_i}$ is a $1 \times n$ matrix:

$$\mathbf{M}^{e_i} = \mu_i [1]^T [1] = \mu_i [1], \quad (25)$$

$$(\mathbf{G}_M^{e_i})_{0,k} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise,} \end{cases} \quad (26)$$

where $0 \leq i \leq n - 1$ for both open and closed snakes. \mathbf{C}^{e_i} is also a 1×1 matrix with a $1 \times n$ assembling matrix $\mathbf{G}_C^{e_i}$:

$$\mathbf{C}^{e_i} = \gamma_i [1]^T [1] = \gamma_i [1], \quad (27)$$

$$(\mathbf{G}_C^{e_i})_{0,k} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise,} \end{cases} \quad (28)$$

where $0 \leq i \leq n - 1$ for both open and closed snakes. $\mathbf{K}_\alpha^{e_i}$ is a 2×2 matrix and its corresponding assembling matrix $\mathbf{G}_\alpha^{e_i}$ is a $2 \times n$ matrix:

$$\mathbf{K}_\alpha^{e_i} = \alpha_i \begin{bmatrix} -1 & 1 \end{bmatrix}^T \begin{bmatrix} -1 & 1 \end{bmatrix} = \alpha_i \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad (29)$$

$$(\mathbf{G}_\alpha^{e_i})_{jk} = \begin{cases} 1 & \text{if } (j + i) \bmod n = k, \\ 0 & \text{otherwise,} \end{cases} \quad (30)$$

where $0 \leq i \leq n - 2$ for an open snake and $0 \leq i \leq n - 1$ for a closed snake. $\mathbf{K}_\beta^{e_i}$ is a 3×3 matrix and with it is associated a $3 \times n$ assembling matrix $\mathbf{G}_\beta^{e_i}$:

$$\mathbf{K}_\beta^{e_i} = \beta_i \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & -2 & 1 \end{bmatrix} = \beta_i \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}, \quad (31)$$

$$(\mathbf{G}_\beta^{e_i})_{jk} = \begin{cases} 1 & \text{if } (j + i) \bmod n = k, \\ 0 & \text{otherwise,} \end{cases} \quad (32)$$

where $0 \leq i \leq n - 3$ for an open snake and $0 \leq i \leq n - 1$ for a closed snake. The $1 \times n$ assembling matrix $\mathbf{G}_F^{e_i}$ is defined as

$$(\mathbf{G}_F^{e_i})_{0,k} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise,} \end{cases} \quad (33)$$

where $0 \leq i \leq n - 1$ for both open and closed snakes.

With the above formulation of finite difference snakes, we have a uniform finite element construction for a variety of snake representations, which leads to a relatively straightforward United Snakes implementation in an object-oriented programming language, such as Java.

3. Live wire

Live wire (or intelligent scissors) is a recently proposed interactive boundary tracing technique (Mortensen et al., 1995; Falcão et al., 1996, 1997; Falcão and Udupa, 1997; Mortensen and Barrett, 1998; Falcão et al., 1998; Falcão et al., 2000). Although it shares some similarities with snakes – it was originally developed as an interactive 2-D extension to previous stage-wise optimal boundary tracking methods – it is generally considered in the literature as a competing technique to snakes. Like snakes, the idea behind the live wire technique is to allow image segmentation to occur with minimal user interaction, while at the same time allowing the user to exercise control over the segmentation process. However, live wire realizes the idea differently from snakes.

Live wire is very easy to use. The user begins by placing an initial *seed* point near the boundary of the object of interest. As the cursor, or *free* point is moved around the image, the current calculated boundary, called the *live wire* or *trace*, from the seed point to the free point is dynamically displayed. If the displayed trace is acceptable and the user clicks the mouse, the free point is collected as an additional seed point, and this trace will be frozen and will become part of the extracted object boundary. The resulting live wire boundaries are piecewise optimal (i.e., optimal between seed points), while the snake gives an optimal solution over the entire contour.

The genesis of live wire has its origin in the early collaboration between Udupa (University of Pennsylvania) and Barrett (Brigham Young University) (Mortensen and Barrett, 1998; Falcão et al., 1998). Their two research groups have since independently developed different live wire systems. They share two essential components: a local cost function that assigns lower cost to image features of interest, such as edges, and an expansion process that forms optimal boundaries for objects of interest based on the cost function and seed points provided interactively by the user. However, they employ different underlying graph models with different local cost functions. In (Mortensen and Barrett, 1998), each pixel represents a graph node, and directed, weighted edges are created between each pixel and its eight adjacent neighbors. In (Falcão et al., 1998), the graph nodes are pixel corners and they are connected by oriented, weighted edge cracks, called boundary elements (*bels* for short). In both cases, when the image is large, a corresponding large underlying graph may have to be maintained and live wire performance will be compromised. To improve the efficiency of live wire, the two groups have developed extensions known as live lane (Falcão et al., 1998) and

toboggan-based intelligent scissors (Mortensen and Barrett, 1998; Mortensen, 2000), respectively.

Live-wire-like user interaction techniques have been proposed in the snakes literature. In (Cohen and Kimmel, 1997), Cohen and Kimmel compute the global minimal path between two points using Sethian's fast marching algorithm (Sethian, 1997), which has sub-pixel accuracy² and may eliminate metrication errors of graph search algorithms. A minimal path between two points is also obtained in (Grzeszczuk and Levin, 1994) based on simulated annealing. In a technique called "static" snakes, proposed in (Neuenschwander et al., 1994), the user initially specifies two end snake points and then the snake takes image information into account progressively from the two end points to its center, resulting in a minimal path between the two points. A similar technique has also been proposed in (Hyche et al., 1992). Dubuisson-Jolly and Gupta have formulated tracking an active contour with shape constraints in an image sequence as a shortest path problem (Dubuisson-Jolly and Gupta, 2001).

3.1. Trace formation

Boundary finding in live wire can be formulated as a directed graph search for an optimal (minimum cost) path using Dijkstra's algorithm in the underlying graph model. First, the graph is initialized with the local costs as described in the next section. Once the user selects a seed point (node), it will be used as the starting point for a recursive expansion process. In the expansion process, the local cost at the seed point is summed into its neighboring nodes. The neighboring node with the minimum cumulative cost is then further expanded and the process produces a dynamic "wavefront". The wavefront expands in the order of minimum cumulative cost. Consequently, it propagates preferentially in directions of highest interest (i.e., along image edges).

For any dynamically selected goal node (i.e., the free point) within the wavefront, the optimal path back to the seed point which forms a live wire trace can be displayed in real time. When the cursor (the free point) moves, the old live wire trace is erased and a new one computed and displayed in real time. The expansion process aims to compute an optimal path from a selected seed point to *every* other point in the image and lets the user choose among paths interactively, based on the current cursor position.

Live wire may be implemented very efficiently in multi-threaded programming languages, such as Java, because the expansion process and the user interface can execute in separate, parallel threads. Since the free point is generally near the target object boundary, the expansion process will most likely have already advanced beyond that point and the live wire trace can be displayed immediately. That

² Toboggan-based live wire (Mortensen and Barrett, 1999) obtains sub-pixel localization by fitting an edge model to the tobogganed region boundaries.

is, the live wire trace can typically be displayed before the expansion process has finished sweeping over the entire image. Therefore, our implementation (Liang et al., 1999a,b) is equivalent to the interleaved computation proposed in (Mortensen et al., 1995; Mortensen and Barrett, 1998) or live wire-on-the-fly introduced in (Falcão et al., 2000) in terms of computation cost, and the multi-threaded Java implementation is more elegant in software design and in supporting user interactions.

3.2. Local cost functions

Many local cost functions can be defined. In (Mortensen et al., 1995), the local cost $l(\mathbf{p}, \mathbf{q})$ on the directed link from \mathbf{p} to a neighboring pixel \mathbf{q} is defined as a weighted sum of six local component costs created from various edge features:

$$l(\mathbf{p}, \mathbf{q}) = \omega_Z f_Z(\mathbf{q}) + \omega_G f_G(\mathbf{q}) + \omega_D f_D(\mathbf{p}, \mathbf{q}) + \omega_P f_P(\mathbf{q}) + \omega_I f_I(\mathbf{q}) + \omega_O f_O(\mathbf{q}), \quad (34)$$

where $f_Z(\mathbf{q})$ is the Laplacian zero-crossing function at \mathbf{q} , $f_G(\mathbf{q})$ is the gradient magnitude at \mathbf{q} , $f_D(\mathbf{p}, \mathbf{q})$ is the gradient direction from \mathbf{p} to \mathbf{q} , $f_P(\mathbf{q})$ is the edge pixel value at \mathbf{q} , $f_I(\mathbf{q})$ and $f_O(\mathbf{q})$ are the “inside” and “outside” pixel values at \mathbf{q} , respectively, while ω_Z , ω_G , ω_D , ω_P , ω_I and ω_O are their corresponding weights.

The Laplacian zero-crossing function $f_Z(\mathbf{q})$ is a binary function defined as

$$f_Z(\mathbf{q}) = \begin{cases} 0 & \text{if } I_L(\mathbf{q}) = 0, \\ 1 & \text{otherwise,} \end{cases} \quad (35)$$

where $I_L(\mathbf{q})$ is the Laplacian of the image I at pixel \mathbf{q} . The gradient magnitude serves to establish a direct connection between edge strength and cost. The function f_G is defined as an inverse linear ramp function of the gradient magnitude G

$$f_G = \frac{\max(G') - G'}{\max(G')} = 1 - \frac{G'}{\max(G')}, \quad (36)$$

where $G' = G - \min(G)$. When calculating $l(\mathbf{p}, \mathbf{q})$, the function $f_G(\mathbf{q})$ is further scaled by 1 if \mathbf{q} is a diagonal neighbor to \mathbf{p} and by $1/\sqrt{2}$ if \mathbf{q} is a horizontal or vertical neighbor.

The gradient direction $f_D(\mathbf{p}, \mathbf{q})$ adds a smoothness constraint to the boundary by associating a higher cost for sharp changes in boundary direction. With $\mathbf{D}'(\mathbf{p})$ defined as the unit vector normal to the gradient direction $\mathbf{D}(\mathbf{p})$ at pixel \mathbf{p} (i.e., $\mathbf{D}(\mathbf{p}) = [I_x(\mathbf{p}), I_y(\mathbf{p})]$ and $\mathbf{D}'(\mathbf{p}) = [I_y(\mathbf{p}), -I_x(\mathbf{p})]$), the formulation of the gradient direction cost is

$$f_D(\mathbf{p}, \mathbf{q}) = \frac{2}{3\pi} \{ \arccos[d_p(\mathbf{p}, \mathbf{q})] + \arccos[d_q(\mathbf{p}, \mathbf{q})] \}, \quad (37)$$

where $d_p(\mathbf{p}, \mathbf{q}) = \mathbf{D}'(\mathbf{p}) \cdot \mathbf{L}(\mathbf{p}, \mathbf{q})$ and $d_q(\mathbf{p}, \mathbf{q}) = \mathbf{L}(\mathbf{p}, \mathbf{q}) \cdot \mathbf{D}'(\mathbf{q})$ are vector dot products and

$$\mathbf{L}(\mathbf{p}, \mathbf{q}) = \frac{1}{\|\mathbf{p} - \mathbf{q}\|} \begin{cases} \mathbf{q} - \mathbf{p} & \text{if } \mathbf{D}'(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) \geq 0, \\ \mathbf{p} - \mathbf{q} & \text{if } \mathbf{D}'(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) < 0 \end{cases} \quad (38)$$

is the normalized bidirectional link or unit edge vector between pixels \mathbf{p} and \mathbf{q} .

Along with the gradient magnitude f_G , pixel value features (f_P , f_I and f_O) are used in on-the-fly training to increase the live wire dynamic adaptation (Mortensen and Barrett, 1998). With the typical gray-scale image pixel value range $[0, 255]$, they are defined as

$$f_P(\mathbf{q}) = \frac{1}{255} I(\mathbf{p}), \quad (39)$$

$$f_I(\mathbf{q}) = \frac{1}{255} I(\mathbf{p} + k \cdot \mathbf{D}(\mathbf{p})), \quad (40)$$

$$f_O(\mathbf{q}) = \frac{1}{255} I(\mathbf{p} - k \cdot \mathbf{D}(\mathbf{p})), \quad (41)$$

where $\mathbf{D}(\mathbf{p})$ is the unit vector of the gradient direction as defined above, and k is a constant distance value for determining the inside and outside features.

In (Falcão et al., 1998), the local cost assigned to each boundary element (bel) \mathbf{b} is a linear combination of the costs with its eight possible features f_i :

$$l(\mathbf{b}) = \frac{\sum_{i=1}^8 w_i c_{f_i}(f_i(\mathbf{b}))}{\sum_{i=1}^8 w_i}, \quad (42)$$

where w_i is the associated weight with feature f_i , and where c_{f_i} , called the *feature transform* function of feature f_i , converts feature value $f_i(\mathbf{b})$ into a cost value. The eight features of a bel \mathbf{b} include the intensity values on positive and negative sides of \mathbf{b} (f_1 and f_2), four different gradient magnitude approximations (f_3, f_4, f_5, f_6), orientation-sensitive gradient magnitude (f_7) and boundary distance (f_8). Each feature value ($f_i, 1 \leq i \leq 8$) may be converted into a cost value with any of the following six feature transforms: linear (c_1), inverted linear (c_2), Gaussian (c_3), inverted Gaussian (c_4), modified hyperbolic (c_5), and inverted modified hyperbolic (c_6). Training methods have been developed for optimum selection of the bel features and automatic selection of the parameters with their feature transforms, based on the typical segments painted by the user along the desired object boundary.

4. Combining snakes and live wire

Excluding user interaction, an accurate initialization is generally needed in order for a snake to lock onto image features of interest in all but the simplest images. Therefore, researchers have been actively investigating techniques to mitigate the sensitivity of snakes to their initialization. Among these techniques are the use of an inflation force (Terzopoulos et al., 1988; Cohen and Cohen, 1993), a chamfer distance map (Cohen and Cohen, 1993) and gradient vector flow (Xu and Prince, 1998). These techniques can work well if the image feature map is relatively clean. However, most clinical images are noisy, contain many uninteresting edges, or texture is present. Hence, these more automatic techniques can fail. For this reason, we explore an alternative direction – instead of attempting

to mitigate initialization sensitivity, we seek to increase the efficiency of interactive initialization. In particular, we enable the user to instantiate (i.e., construct, initialize, and activate) snakes quickly and with minimal effort by exploiting the strengths of the live wire technique.

In this section, we first justify the complementarity of snakes and live wire, and then we show that the combination of snakes and live wire also provides an efficient mechanism for handling large images.

4.1. Snakes and live wire are complementary

There are numerous ways to define the local cost functions in live wire, as long as sufficiently low cost values are assigned to the desired object boundaries. Therefore, various techniques developed for computing snake potentials (Kass et al., 1988; Blake and Isard, 1998) can be used for the generation of local cost maps in live wire. For instance, the chamfer distance map (Cohen and Cohen, 1993) and gradient vector flow (Xu and Prince, 1998) are readily usable. Similarly, the local cost map computed for live wire may be treated as an image potential in snakes. Therefore, in United Snakes, snakes and live wire may share the same image potential (local cost map).

In general, live wire provides user-friendly control during the segmentation process. The user may freely move the free point on the image plane, and the corresponding live wire trace is interactively displayed in real time. However, once the free point is collected as an additional seed point, the trace is frozen and it becomes a part of the extracted object boundary. At this point, the user has no further control over the trace between seed points other than *backtracking*. Therefore, when the shape of the object boundary is complex or when object boundaries are noisy and unclear, the user may need to backtrack to produce acceptable traces. Consequently, many seed points may be needed to guide the live wire to an accurate result. Furthermore, it is not uncommon for the user to make small errors when placing seed points using a mouse or other input device, forcing the user to repeat the placement. By contrast, a snake may be dynamically adjusted or refined as it deforms at any time and at any point on the snake via intuitive mouse-controlled forces. However, the best performance of the snake is often achieved when user-specified constraint points are utilized. The constraint points do not “lock in” the solution – they too may be changed dynamically, allowing further refinement of the extracted object boundary.

Live wire seeks a global minimal path between two points. Therefore, when a section of the desired object boundary has a weak edge relative to a nearby strong but uninteresting edge, the live wire snaps to the strong edge rather than the desired weaker boundary. In order to mitigate this problem, Falcão et al. (1998) have developed training techniques for optimum feature selection and automatic parameter selection, and Mortensen and Barrett have proposed *on-the-fly training* (Mortensen and Barrett,

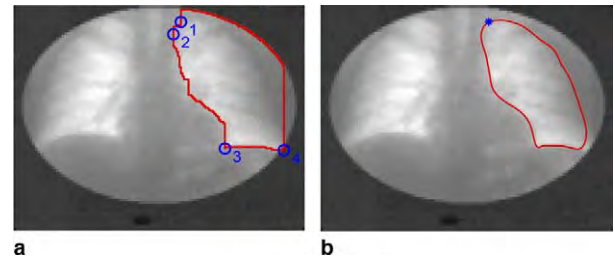


Fig. 1. (a) Delineation of the lung in X-ray fluoroscopy images using live wire (seed points are shown). (b) A Hermite snake instantiated from live wire traces with the first seed point imposed as a hard constraint. It is interactively pulled out of the strong edge with spring forces and then locks onto the lung boundary.

1998). Basically, these techniques (dynamically) update the cost map to filter out the image features which do not have edge characteristics similar to the sample boundary specified by the user. In other words, these methods rely on the assumption that the edge property is relatively consistent along the object boundary. Training is most effective for those objects with relatively consistent boundary properties and may be counter-productive for objects with sudden and/or dramatic changes in their boundary properties (Mortensen and Barrett, 1998). For example, in the lung image of Fig. 1(a), the live wire snaps to the strong edges of the elliptical viewport rather than the desired lung boundary. In this case, training is ineffective since the edge property of the lung boundary varies considerably over its extent and is also disturbed by the ribs (not obvious to the eye). Consequently, it is difficult to specify a typical segment of the lung boundary. Nevertheless, in situations where training can be effective, snakes can also take advantage of it by utilizing the trained cost map. Moreover, in United Snakes, the user has more control, using spring forces to pull the snake out of one minimum into another without training, as shown in Fig. 1(b).

The underlying graph search makes it possible for live wire to bridge boundary gaps and pass through noisy areas. For instance, in MR wrist images from Falcão et al. (2000) (Fig. 2), live wire bridges the gap along the vessel boundary in Fig. 2(a) and passes through a noisy region in Fig. 2(b). Even if there are no image features at all between two points, live wire can still provide a minimal path – a straight line. However, live wire is inherently

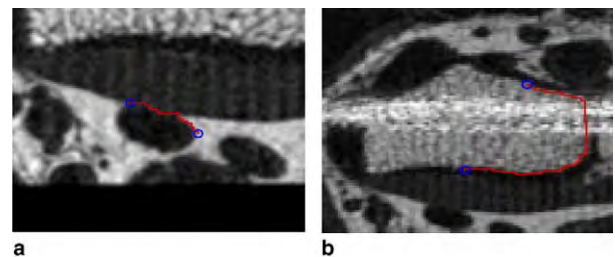


Fig. 2. Live wire bridges the gap along the vessel boundary (a) and passes through a noisy region (b) in an MR wrist image.

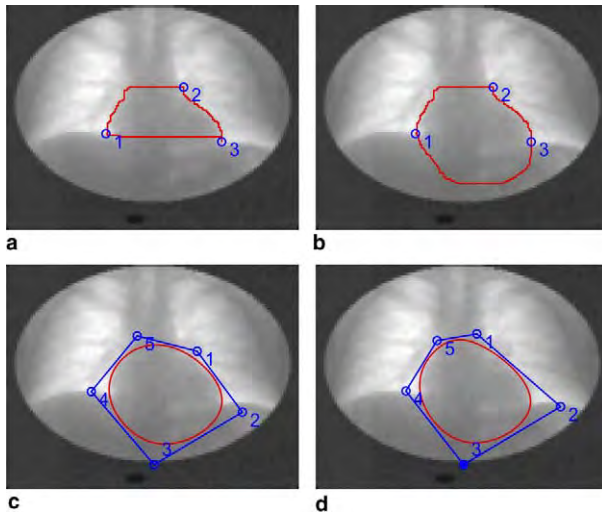


Fig. 3. (a) Delineation of the heart in X-ray fluoroscopy images using live wire (seed points are shown). (b) The unacceptable segment replaced by manual drawing. Alternately, the user may place multiple seed points and let live wire generate a piecewise linear path between adjacent seed points to approximate the missing cardiac boundary. (c) Initial B-spline snake and control polygon instantiated from live wire traces in (b). (d) Resulting segmentation after a few iterations with control point 3 as a hard constraint, which effectively bridges the gaps along the cardiac boundary.

image-based, rather than model-based. Fundamentally, it is not designed to bridge gaps in a manner that is consistent with the image features bordering the gaps and the smoothness of the traces cannot be guaranteed. For instance, in Fig. 3(a), part of the live wire trace from seed point 1 to seed point 2 is a straight line where the cardiac boundary is missing, and the live wire technique does not generate an acceptable cardiac boundary from seed point 3 to seed point 1. The user may place multiple seed points and let live wire generate a piecewise linear path between adjacent seed points to approximate the missing cardiac boundary. In this case, we have found that it is convenient to draw a rough curve manually between the points (Fig. 3(b)). Snakes, on the other hand, are model-based and were designed to adhere to image edges and interpolate between edge features in regions of sparse and noisy data (i.e., fill in the gaps). For example, a B-snake instantiated from the live wire traces is more effective in bridging the gaps along the cardiac boundary, as shown in Fig. 3(d).

In summary, it is desirable to enable the user to exercise more control over the live wire traces between seed points,

impose smoothness on live wire traces, and bridge complicated gaps along object boundaries. This is what snakes are good at doing. Snakes adhere to edges with sub-pixel accuracy and they may also be adjusted interactively as parametric curves with intuitively familiar physical behaviors. Furthermore, snakes have the power to track moving objects, while live wire does not.

However, the efficient performance of interactive snakes is linked to fast, reasonably accurate initialization and user-specified constraints. Even with a few seed points, live wire can quickly give much better results than casual manual tracing. Hence, the resulting live wire boundary can serve well to instantiate a snake. The live wire seed points reflect the user's prior knowledge of the object boundary. They can therefore serve as either hard or soft point constraints for the snake, depending on the user's confidence in the accuracies of these seed points.

Because a live wire-traced initial object boundary is more accurate than a hand-drawn boundary, and with the further incorporation of the seed points as snake constraints, the snake will very quickly lock onto the desired object boundary. If necessary, the user may then correct mistakes inherited from the live wire-generated boundary by applying mouse-controlled spring forces to the snake. Because the user still has the opportunity to correct the deficiencies of the trace as the snake is evolving, the number of seed points needed by live wire to generate the object boundary can be further reduced. Consequently, a satisfactory initial object boundary can be generated very quickly using live wire. Other hard or soft constraints may be added during the snake deformation process as well. Because constrained values may be changed dynamically, the user may adjust the seed points to further refine the object boundary as the snake deforms.

To illustrate their performance, we apply United Snakes to an angiogram (Fig. 4) and a vertebra image (Fig. 5), to which Mortensen and Barrett applied their live wire algorithm in (Mortensen and Barrett, 1998). With only a few seed points, United Snakes generate the boundaries shown in Figs. 4 and 5(c), which are comparable to the ideal boundaries used as references in (Mortensen and Barrett, 1998).

As further evidence that the United Snakes framework improves upon the robustness and accuracy of its component techniques, Fig. 6 shows a synthetic image of a known curve degraded by strong Gaussian white noise (variance

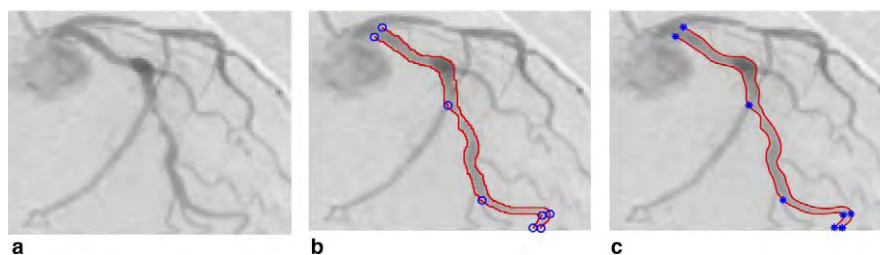


Fig. 4. Segmenting a vessel in an angiogram. (a) The image used in (Mortensen and Barrett, 1998). (b) Live wire segmentation. (c) United Snakes generate boundaries comparable to ideal boundaries in (Mortensen and Barrett, 1998).

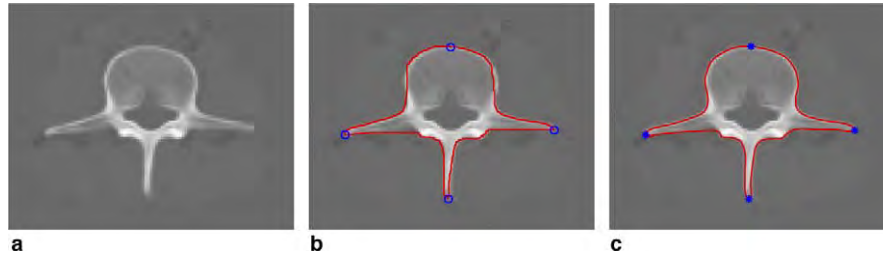


Fig. 5. Segmenting the outer boundary of a vertebra. (a) The image used in (Mortensen and Barrett, 1998). In United Snakes, we only expect a coarse object boundary from live wire. To illustrate this point, referring to Eq. (35), we have set $\omega_G = 0.50$, $\omega_Z = 0.5$, and turned all the other parameters off (i.e., $\omega_D = \omega_P = \omega_I = \omega_O = 0$), resulting in the live wire segmentation (b). From it, United Snakes generate a boundary (c) which is comparable to the *ideal* boundary in (Mortensen and Barrett, 1998).

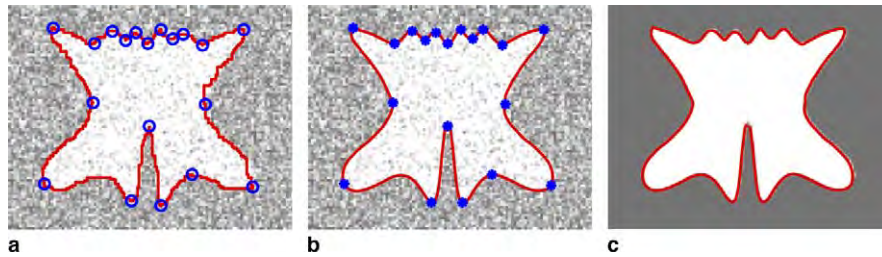


Fig. 6. Performance of United Snakes demonstrated using a noisy synthetic image. This image was designed to challenge the snakes and live wire with the high curvature points as well as the small wave details. (a) A live wire is sensitive to noise (the required seed points are shown). (b) United Snakes are robust against noise. (c) The segmented boundary accurately conforms to the ideal boundary.

0.25). Given its image-based nature, the live wire is sensitive to noise as shown in Fig. 6(a). A snake instantiated by the live wire gives a better result (Fig. 6(b)). Fig. 6(c) shows that the United Snakes result is very close to the boundary in the ideal image, despite the strong noise. This performance is a consequence of the imposed hard constraints, without which the snake would slip away from high curvature points.

4.2. Handling large images

Large images are now common in clinical settings because high resolution is often needed to make accurate diagnoses. For instance, in the mammogram analysis task (see Section 6.4), we need to handle images with a typical resolution of 3500×6500 pixels. However, due to the nature of its underlying graph-based algorithm, the basic live wire algorithm is unable to handle large images efficiently. To support user interaction, live wire aims to compute an optimal path from the last seed to *every* other point in the image. Even with our efficient multi-threaded Java implementation of Dijkstra's algorithm (e.g., with bucket sort (Mortensen and Barrett, 1998) or Dia's method (Falcão et al., 2000)), the performance of live wire will be significantly compromised when working with large images. The reason is that the lower bound of its computational complexity is $O(m)$, where m is the number of image pixels involved in the computation of an optimal path from the seed to the free point; that is, all the pixels within the wavefront (i.e., the expansion process). In the worst case, m is the total number of pixels in the image.

For effective user interaction, the thread responsible for computing an optimal path from the seed point to every other point in the image should not stop until the user has selected the current free point as a seed. This ensures the user may move with more freedom in the image plane to select optimal paths and quickly generate an acceptable object boundary with a minimal number of seed points. That is, the user should be able to place two neighboring seed points as far apart as desired. However, when the desired object boundary is not clear/sharp (e.g., chest images, mammograms, etc.) or has many branches (e.g., a retinal angiogram), the wavefront will spread too widely and include many pixels for any path of reasonable length. As a result, the memory required to maintain the auxiliary information in Dijkstra's algorithm will increase dramatically for large images. Furthermore, as demonstrated also in (Falcão et al., 2000), when the image size changes from 128×128 to 1024×1024 , the live wire performance will be reduced by a factor of 400, and the ultra fast live wire on the fly may still be 40 times slower.

The combination of live wire and snakes in United Snakes provides a new mechanism for handling large images. The computational complexity of snakes is $O(n)$ in each iteration, where n is the number of snake nodal variables. In United Snakes, we typically require live wire to generate only a coarse boundary with a few seed points. Therefore, we can construct a *truncated* pyramid of images, and let the live wire work at the top of the pyramid with a small image size (for example, 128×128 or 256×256), thus efficiently supporting user interaction. The snake “descends” the image pyramid from coarse to fine levels of

resolution, tolerating any live wire errors introduced at the top of the pyramid, and accurately locks onto the desired object boundary. The original large image is still displayed to the user and thus the seed points can generally be accurately specified or dynamically adjusted if necessary. The extra memory needed to maintain the pyramid is offset by the reduced memory necessary for the auxiliary information in Dijkstra's algorithm. In practice, we do not have to maintain a pyramid for the original image, but only for the image potential. Assuming the pyramid has n levels and the original image occupies M amount of memory, the extra memory required for the pyramid then is

$$\left(\frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots + \frac{1}{4^{(n-1)}}\right)M = \left(1 - \frac{1}{4^{(n-1)}}\right)\frac{M}{3}, \quad (43)$$

while the reduced auxiliary memory (e.g., only for the cumulative cost map) in the live wire implementation is $(1 - \frac{1}{4^{(n-1)}})M$.

As a demonstration, Fig. 7(a) shows a retinal angiogram with pixel resolution of 256×256 obtained by down-sampling the original large image with resolution 1024×1024 . Suppose we would like to trace the vessel starting from point S to one of the target points 0–9 (see Fig. 7(b)). Once point S is selected as the first seed, the corresponding branch should ideally be instantly available once the user points to any of the 10 branch end-points. This real-time user interaction is achievable by live wire when the image size is under 300×300 on modest PCs (Fig. 7(c)). For larger images, however, real-time user interaction becomes increasingly difficult to achieve using live wire alone. Table 1 shows the time needed for the wavefront to reach the 10 targets as well as the time needed to sweep over the entire image at different resolutions on an 866 MHz Pentium PC with SUN JDK1.3. From the table, we can see that the time required at 256×256 resolution is approximately 1/4 of

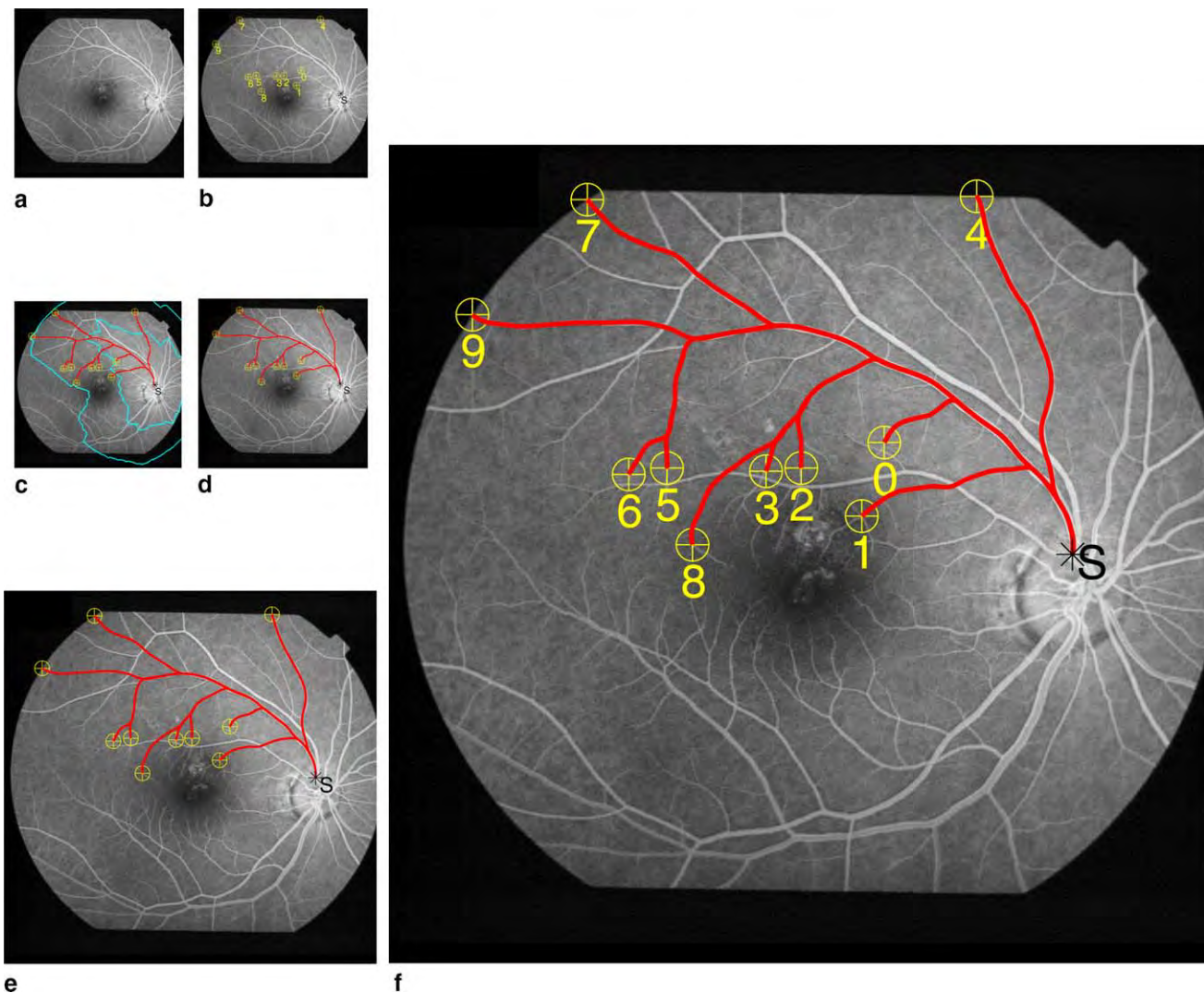


Fig. 7. (a) A retinal angiogram with pixel resolution 256×256 obtained by down-sampling the original 1024×1024 image. (b) The seed point S and 10 marked target points. (c) The superimposed live wire traces shown only for the first and last target points. (d) The superimposed snakes dynamically instantiated from the live wire traces in (c) descend the truncated pyramid reaching the intermediate level with resolution 512×512 (e) and the original large image (f). This mechanism supports real-time user interaction: once point "S" is selected as a seed, the corresponding vessel branch is instantly available when the user points to a new position (such as, the 10 targets) on the original large image.

Table 1

The time (in milliseconds) needed for the wavefront to reach the 10 targets shown in Fig. 7, as well as the time needed to sweep over the entire image at different resolutions on an 866 MHz Pentium PC with SUN JDK1.3

Targets	0	1	2	3	4	5	6	7	8	9	Entire image
256 × 256	65	109	118	131	131	170	205	211	225	240	405
512 × 512	374	526	545	604	805	875	979	1025	1081	1248	2178
1024 × 1024	1562	2293	2393	2703	3479	3683	4425	4543	4688	5481	9802

The time required at 256 × 256 resolution is approximately 1/4 of that at 512 × 512 resolution, which is roughly 1/4 of that at 1024 × 1024 resolution. The time needed for a snake sliding down is $O(n)$ in each iteration, where n is the number of snake nodal variables. We use five iterations at each level of the pyramid. The longest snake (from point S to point 9 in Fig. 7) in this experiment has 100 nodal variables. When the live wire trace is available at the top level (256 × 256), it takes 15 iterations or 77 ms for the snake to descend the pyramid. So, the total time needed is 317 (77 + 240) ms, which is much less than 5481 ms when working directly at the resolution of 1024 × 1024.

that at 512 × 512 resolution, which is roughly 1/4 of that at 1024 × 1024 resolution. This can be justified by the observation that reducing an image by a factor of 2 in linear dimension while maintaining its aspect ratio reduces its area in pixels to 1/4, and that the complexity of the wavefront computation is proportional to the latter. Thus, with a three-level pyramid, we can make the algorithm approximately 16 times faster and, with four levels, it becomes approximately 64 times faster.

In United Snakes, snakes that are dynamically instantiated from live wire traces at the top of the truncated image pyramid can easily descend the pyramid, reaching the original large image (Fig. 7(f)) via intermediate level(s) (Fig. 7(e)), resulting in real-time user interaction on the original large image. Thus, United Snakes with the image pyramid scheme yields real-time response – a critical factor in any interactive segmentation scheme – with sub-pixel accuracy in original large images.

5. Hard constraints

Our combination of snakes and live wire relies on an efficient constraint mechanism. A constraint on a snake may be either soft or hard. Hard constraints generally compel the snake to pass through certain positions or take certain shapes, whereas soft constraints merely encourage a snake to do so. Two kinds of soft constraints, springs and volcanos, were described in the original snakes paper (Kass et al., 1988) and they are incorporated into our finite element formulation. Hard constraints have been used to prevent snake nodes from clustering in dynamic programming snakes (Amini et al., 1990). Generic hard constraints are discussed in (Fua and Brechbühler, 1997). In this section, we propose a convenient mechanism, called *pins*, as a simple yet effective way to impose hard constraints on snakes for the integration of snakes and live wire.

Suppose that we wish to guarantee that the snake node i sticks at position (x_i^c, y_i^c) in the Hermitian parameterization. Recall that in the Hermitian parameterization, the polynomial shape of each element is parameterized by the position and slope of $x(s)$ and $y(s)$ at the two nodes (position and slope variables occupy alternating positions in the nodal variable vector \mathbf{u}). Therefore, the snake stiffness matrix \mathbf{K} may be updated with

$$\mathbf{K}_{2i,j} = \begin{cases} 1 & \text{if } 2i = j, \\ 0 & \text{otherwise,} \end{cases} \quad (44)$$

where $0 \leq j \leq 2(N-1)$ and N is the number of snake nodes. The system force vector \mathbf{F} is updated as

$$\mathbf{F}_{2i}^x = x_i^c, \quad \mathbf{F}_{2i}^y = y_i^c, \quad (45)$$

where x and y indicate coordinate function $x(s)$ and $y(s)$, respectively. It is then guaranteed that the snake node i is always at position (x_i^c, y_i^c) .

A drawback of this simple technique, however, is that the updated system stiffness matrix is no longer symmetric. Consequently, we are unable to store the stiffness matrix economically using skyline storage, nor factorize it into LDL^T form (see Appendix A). Nevertheless, since the position of node i is given, a constant force may be derived from the stiffness matrix for each degree of freedom and subtracted from its corresponding position in the system force vector so that we can restore the symmetry of the stiffness matrix while keeping the system in balance. In our implementation, we store column $2i$ of \mathbf{K} into a vector \mathbf{k}^{2i} ; i.e., $\mathbf{k}_j^{2i} = \mathbf{K}_{j,2i}$, for $0 \leq j \leq 2(N-1)$, before \mathbf{K} is made symmetric with

$$\mathbf{K}_{j,2i} = \begin{cases} 1 & \text{if } 2i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (46)$$

To keep the system in balance, the system force vector \mathbf{F} is updated with

$$\mathbf{F}_j^x = \mathbf{F}_j^x - x_i^c \mathbf{k}_j^{2i}, \quad \mathbf{F}_j^y = \mathbf{F}_j^y - y_i^c \mathbf{k}_j^{2i} \quad (47)$$

for $0 \leq j \neq 2i \leq 2(N-1)$. We can constrain the slope in the same way. If we constrain two node variables of an element in both position and slope, this element will be frozen. Its two neighboring elements will also be influenced by the constraint. The constraints on a B-snake are imposed on the nodes of its control polygon. Imposing hard constraints in this manner also lessens computational cost, in terms of both memory and time, since the number of entries in the skyline storage of the stiffness matrix is reduced. Consequently, the LDL^T factorization and forward/backward substitutions can be performed more efficiently (see Appendix A). It is also possible to apply more general constraints to any point on the snake as is described in (Terzopoulos and Qin, 1994).

In the formulation above, the updated stiffness matrix only indicates which degrees of freedom of the snake are constrained, it does not contain any constraint values. These are recorded in the system force vector. As a result, the constraint values may be updated dynamically during snake deformation. Hence, the user can move the constraint points around the image plane to refine the object boundary as the snake is deforming. This property is very useful when integrating snakes with live wire. While a snake is deforming, additional hard constraints may be imposed on the snake to restrict its deformation. Because these constraints are unknown before the snake is instantiated, they may be incorporated on-the-fly using reaction forces on the system force vector without changing the stiffness matrix. However, small time steps are required to ensure the stability of the snake. In our implementation, we create a new snake from the current snake plus the hard constraints, since the LDL^T factorization is fast.

Hard constraints play very important roles in capturing the intricate details and bridging gaps along object boundaries in image segmentation. For instance, to segment the bladder from an MR image of a female abdomen shown in Fig. 8, neither the live wire (Fig. 8(a)) nor its corresponding, instantiated dynamic snake (Fig. 8(b)) would be able to capture the intricate details indicated by the rectangle, which require additional seed points (Fig. 8(c)). With all the seed points imposed as hard constraints, the corresponding snake accurately captures the details without

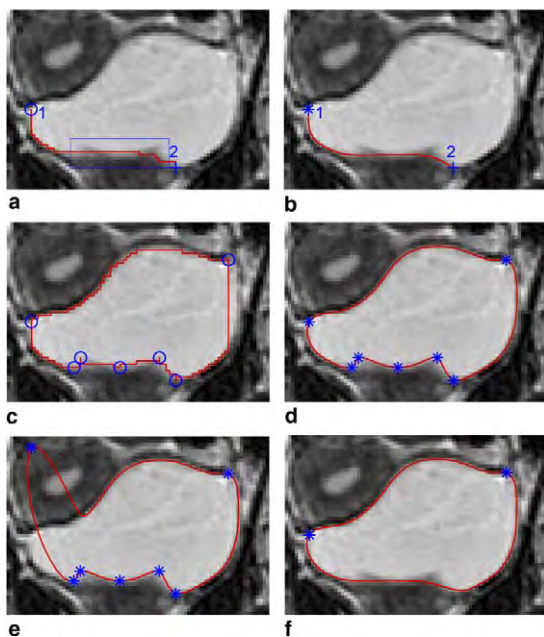


Fig. 8. Segmenting the bladder in an MR image of a female abdomen. Neither the live wire (a) nor its corresponding dynamic snake (b) would be able to capture the intricate details indicated by the rectangle without additional live wire seed points (c). (d) With all the seed points naturally imposed as hard constraints, the corresponding snake accurately captures the fine, intricate details without any further user intervention. (e) Hard constraint point 1 is deliberately moved far away from the desired bladder boundary to illustrate the adjustment capability. (f) Releasing the hard constraints will lose the details.

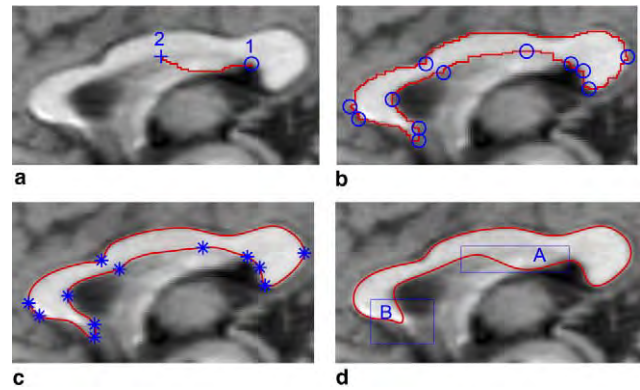


Fig. 9. Segmenting the corpus callosum in an MR head image of a human volunteer. (a) The live wire snaps to the nearby strong edge. (b) An additional seed bridges the missing boundary. (c) The corresponding snake with the seed points naturally imposed as hard constraints nicely captures the desired object. (d) When releasing the hard constraints, the strong image force in the region of region A will gradually drive the snake away from the desired boundary, while, in region B, the snake will become insufficiently peaked due to its internal energy.

any further user intervention as shown in Fig. 8(d). Hard constraint points may be adjusted to refine the object boundary. For illustration purposes, hard constraint point 1 has been deliberately moved far away from the desired bladder boundary in Fig. 8(e). Releasing the hard constraints will lose the details as shown in Fig. 8(f).

The desired object boundary might be unclear or even missing in many clinical images. For example, in segmenting the corpus callosum in an MR head image of a human volunteer, the live wire snaps to the nearby strong edge 9(a), and additional seed points are required to bridge the missing boundary 9(b). These seed points can be naturally imposed as hard constraints on the corresponding snake, which nicely captures the desired object in Fig. 9(c). When releasing the hard constraints, the strong image force in region A (see Fig. 9(d)) will gradually drive the snake away from the desired location, while the snake will become insufficiently peaked in region B due to its internal energy.

5.1. Static vs. dynamic constraint integration

We have argued that a hard constraint mechanism is crucial in practical image segmentation. Live wire generally requires seed points at the critical, complicated locations where the desired boundary is twisted, unclear, weak or even missing. These seed points, interactively provided by the user to guide the live wire, capture the user's expert prior knowledge about the desired object boundary, and they can naturally be imposed as hard constraints on the snake that is then instantiated from the complete live wire trace. We refer to this form of livewire-snake integration as *static* integration – once the live wire result is used to instantiate a snake, the segmentation process continues using only the constrained, user-controlled snake.

A more *dynamic* constraint integration “mode” is often useful – once the live wire trace between the last seed point

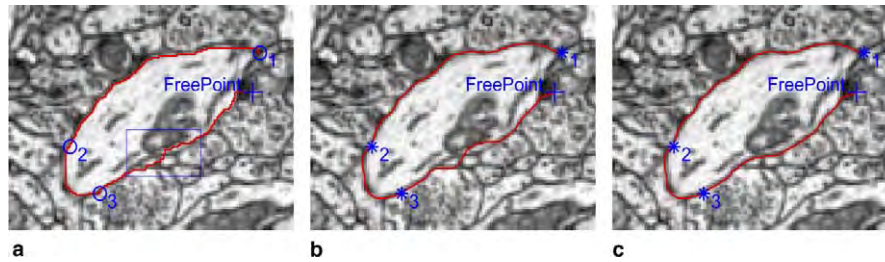


Fig. 10. Using United Snakes in dynamic mode to segment neuronal EM images. (a) Live wire boundary showing three seed points and free point (rectangle indicates a problem area). (b) Open snakes dynamically generated from the live wire traces and constrained by seed and free points. (c) Third snake corrected in the problem area using the mouse.

and the free point is formed, a corresponding open snake with constraints at the seed point and the free point is instantiated and set in motion. When the free point is chosen and collected as a seed point, this open snake is merged with the snake (if any) instantiated from previous live wire traces. All seed points are automatically applied as constraints. Fig. 10 illustrates this process where “+” indicates the current free point. The live wire and snake results are shown separately in the neuronal EM images in Figs. 10(a) and (b), respectively. Since the snake is automatically set in motion, the user may use the mouse-controlled springs to adjust it in any problematic areas along the snake trace (Fig. 10(c)).

6. Applying United Snakes

In this section, we apply United Snakes to several different medical image analysis tasks, demonstrating the generality, accuracy, robustness, and ease of use of the tool.

6.1. Segmenting neuronal dendrites in EM images

A neuronal dendrite is the receiving unit of a nerve cell. The area of contact between the dendrites of different cells is called a synapse and is located on the dendritic spines. In humans, morphological changes in dendritic spines are seen with aging and with diseases that affect the nervous system, such as dementia, brain tumors and epilepsy (Carlsson et al., 1994). Detailed anatomical models of dendritic spines and their synapses will provide new insights into their function, thus providing better opportunities to understand the underlying causes and effects of these diseases. To build such models, the dendrite must be segmented from the surrounding tissue in positive electron microscopy (see Carlsson et al., 1994 for a detailed description of how snakes are used in reconstruction of 3D nerve cell models from serial microscopy). Here, we are interested in localizing nerve cell membranes, which appear dark in positive microscopy.

In the United Snakes system, the user begins an image segmentation task using a live wire. An initial seed point is placed near the boundary of the object of interest. As the cursor, or free point, is moved around, the live wire, or trace, is interactively displayed from the seed point to

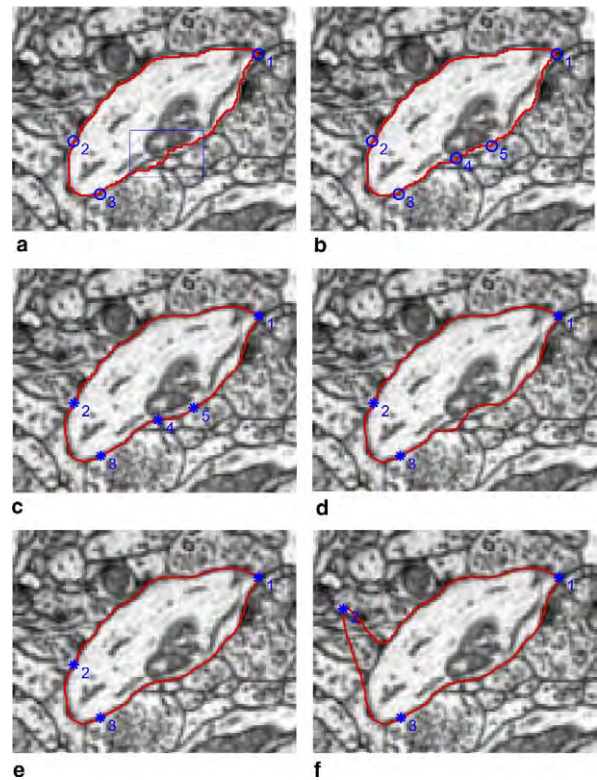


Fig. 11. Using United Snakes in static mode to segment neuronal EM images. (a) Approximate live wire boundary using just three seed points (rectangle indicates a problem area). (b) Additional seed points can improve live wire's accuracy. (c) Instantiated from the live wire traces in (b), the snake tolerates live wire errors and locks on cell boundary without further user interaction. (d) Instantiated from the live wire traces in (a), the snake “sticks” in the problem area, but it is easily adjusted (e) using the mouse. (f) Snake adjustment capability illustrated by moving constraint point 2.

the free point. If the displayed trace is acceptable, the free point is collected as an additional seed point. For example, we can capture an approximate cell boundary in Fig. 11(a) with just three seed points.

The live wire tends to stick to the object boundary using the seed points as a guide. The trace between the two adjacent seed points is frozen. The user has no further control over these traces other than backtracking. In order to generate a more accurate result in the area indicated by a rectangle, more seed points may be placed as in Fig. 11(b).

Although the live wire boundary is somewhat jagged and exhibits some small errors, it is in general as accurate as manual tracing, but more efficient and reproducible.

Next, we instantiate a snake from the live wire-generated boundary and use the seed points to constrain it. The user may select a shape function for the snake which is suitable for the object boundary. In our cell segmentation example, if the live wire result with five seed points is used to instantiate a finite difference snake, it is able to tolerate the live wire errors and very quickly and accurately lock onto the cell boundary without any further user interaction (Fig. 11(c)). Using the live wire result with three seed points, the snake becomes “stuck” in the problematic area (Fig. 11(d)) due to the live wire-generated boundary errors. However, this situation can be easily remedied using the mouse spring (Fig. 11(e)). Furthermore, as the snake is deforming, the hard constraints may be adjusted to refine the snake boundary. In Fig. 11(f), for example, constraint point 2 is moved to illustrate this snake boundary adjustment capability. By contrast, it is not nearly as easy to adjust a seed point in the live wire algorithm.

In summary, the information from live wire including the user guidance and expert prior knowledge is fully utilized by the snake; the snake very quickly locks onto the image features of interest with reasonable tolerance to mistakes in the live wire traces.

6.2. Dynamic chest image analysis

The aim of the dynamic chest image analysis task is to show focal and general abnormalities of lung ventilation and perfusion based on a sequence of digital chest fluoroscopy frames collected over a short time period (typically about 4 s) (Liang, 2000; Liang et al., 1997a,b, 1998, 2001, 2003). The project uses only plain X-ray fluoroscopy for the ventilation and perfusion studies; the radiation dose to patients is low and, unlike a nuclear medicine scan, no preparation is required before the examination and radioactive isotopes are unnecessary. The information gleaned from these images is helpful in several aspects of cardiothoracic radiology. Diseases directly related to the parameters being measured include pulmonary embolism, pulmonary emphysema, cardiac failure, congenital heart disease and other diseases (tumors, obstructive lesions or infections) which may change pulmonary ventilation and/or perfusion. The shapes and motions of the lung and heart give an indispensable clue to ventilation and perfusion examinations. Therefore, an essential first step for ventilation and perfusion analysis is the delineation of the lungs and the heart from each frame in a chest image sequence. The United Snakes system has been used for this purpose. Typically most of the user interactions to initialize and edit the snake are applied to the first image of the sequence only. The resulting snake is then propagated and deformed through the remaining frames of the image sequence.

We employ the dynamic integration mode, which was described in Section 5.1, to delineate the lung boundaries

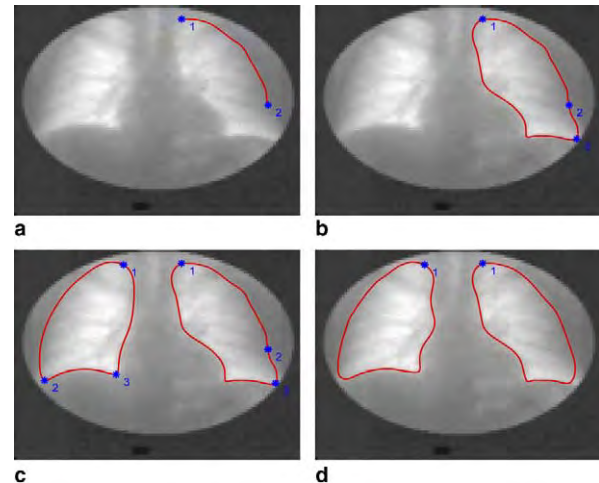


Fig. 12. Dynamic lung delineation with United Snakes. (a) An automatically instantiated Hermite snake. (b) United Snakes result for the left lung with three seed points. (c) The final result for both lungs. (d) Releasing hard constraints except those at both lung apices for lung motion tracking.

interactively for the first image in the sequence. Fig. 12(a) shows the first dynamically instantiated Hermite snake with two end points (seeds) applied as hard constraints. Three seed points are sufficient for delineating the left lung (Fig. 12(b)), similarly, for the right lung shown in Fig. 12(c).

In the dynamic integration method, all seed points are automatically applied as hard constraints. Although hard constraints can be dynamically adjusted, for motion tracking it is not convenient to perform hard constraint adjustments in each frame. Therefore, the United Snakes system allows the user to add or release hard constraints dynamically. The edge information at the lung apex is very weak and there is no observable motion in quiet breath. Consequently, it is desirable to maintain a hard constraint there. All other hard constraints are released for lung motion tracking. Fig. 12(d) shows a Hermite snake with the first seed imposed as hard constraint for each lung.

We apply the snake motion tracking mechanism on the entire image sequence, resulting in the registration of the lung from one frame to another. Since the first seed is applied as hard constraint, the snake can firmly stick at the apex, although the edge information there is rather weak. Fig. 13 illustrates the tracking result for every fifth image.

In the case of the heart, we first employ the static integration method for heart boundary tracing with the B-spline shape function in the first image. A least squares approximation to the initial curve shown in Fig. 3(b) with a cubic B-spline with 5 knots can be used as an initialization to a B-snake in Fig. 3(c). A hard constraint may be further imposed on control polygon node 3 to effectively bridge the gap along the heart boundary. The resulting B-snake for the first frame (Fig. 3(d)) is then used to track the heart motion through the whole image sequence (see

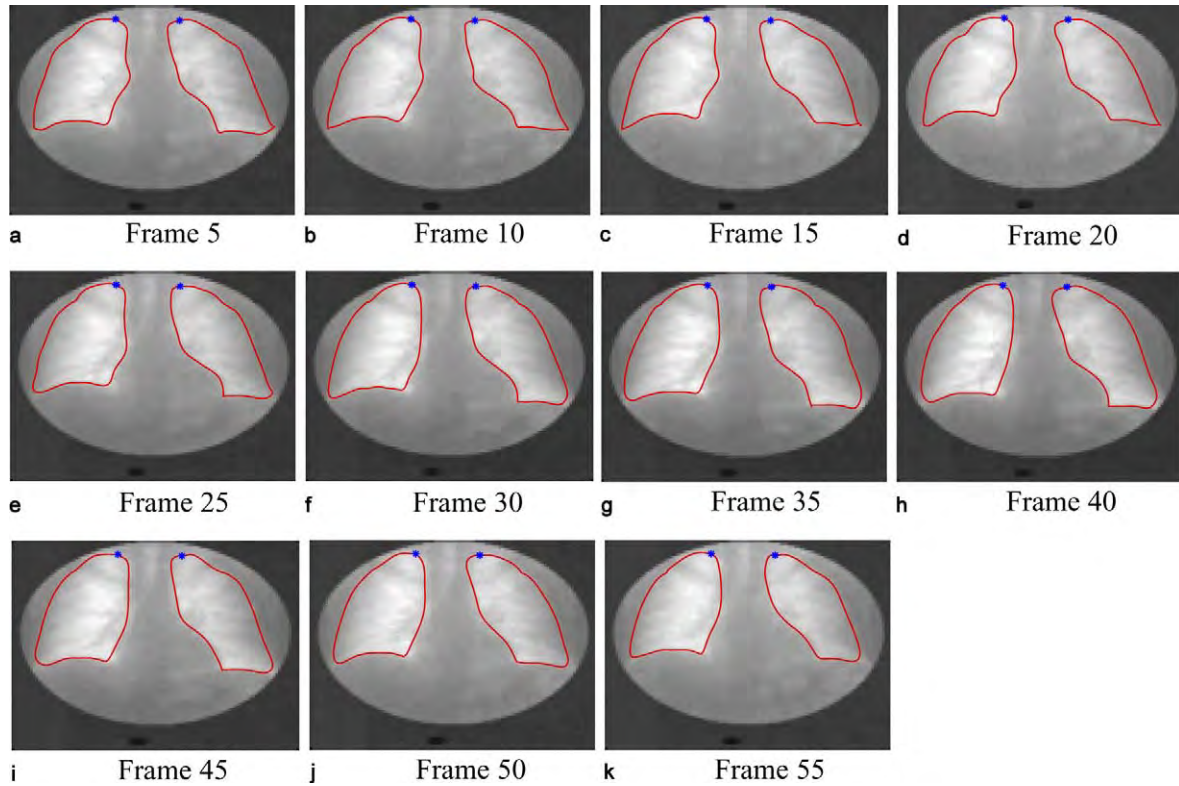


Fig. 13. Lung motion tracking result for every fifth frame.

Fig. 14). Since, in this patient orientation, there is no significant motion with the missing cardiac boundary, it is desirable to apply a hard constraint on the control polygon

node 3. In the case of cardiac motion tracking, the hard constraint is not only effective for single images but also for the entire image sequence.

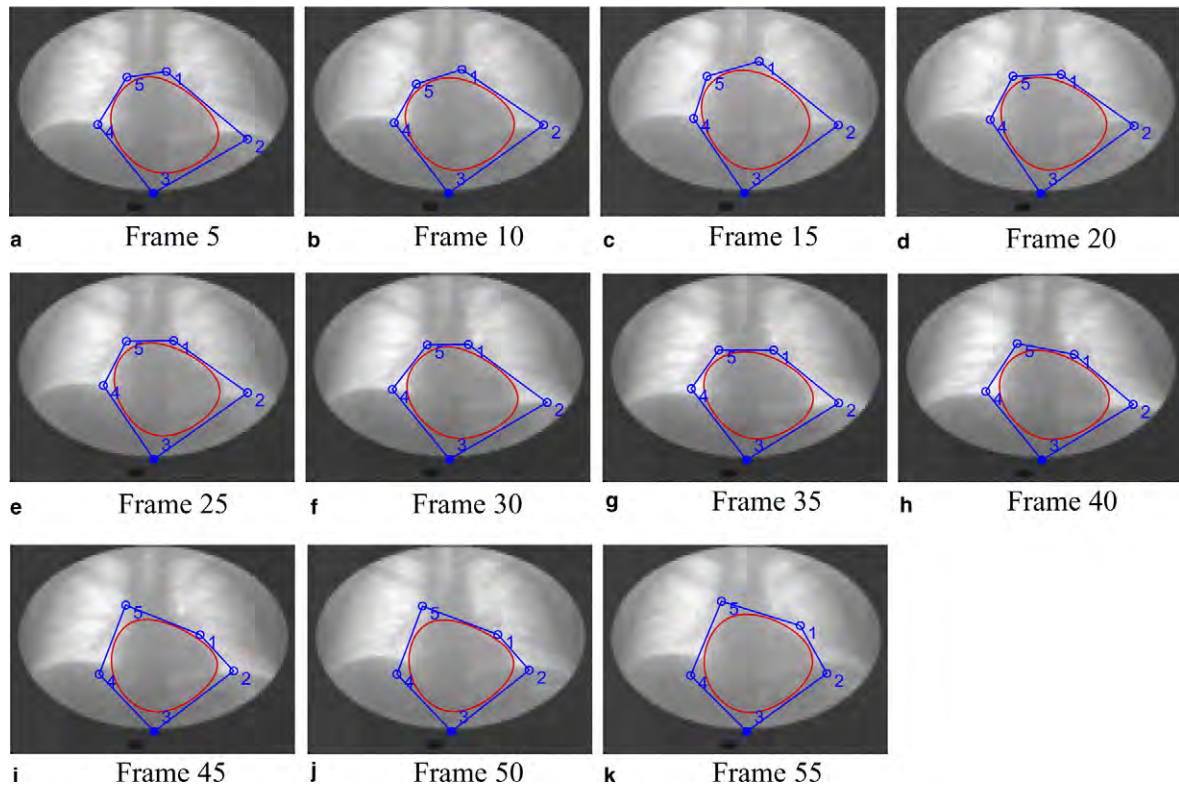


Fig. 14. Cardiac motion tracking result for every fifth frame.



Fig. 15. Quantifying growth plates in MR images. (a) An MR growth plate image. (b) The live wire results. (c) The United Snakes results.

6.3. Quantifying growth plates in MR images

The aim of the growth plate image analysis task is to determine the right time for surgery for patients with abnormal growth of the legs. To this end, the four tiny (essentially horizontal) lines in the image (Fig. 15(a)) must be detected to quantify the growth plate.

In this scenario, it is difficult for the user to trace an initial contour for a snake manually because of the small size

of the lines and the small distance between each pair of lines. However, live wire can be used to generate quickly an acceptable snake initialization with just two or three seed points as shown in Fig. 15(b). In the final results shown in Fig. 15(c), two hard boundary conditions are applied on each of four finite difference snakes.

6.4. Isolating the breast region in mammograms

The goal of the mammogram project is to use pattern recognition techniques to detect abnormalities in the breast tissue. The mammograms we are handling are very large with a typical resolution of 3500×6500 pixels, requiring about 30 MB of disk space. For effective and efficient abnormality detection, it is essential to isolate the breast region from the background (Ojala et al., 2000, 2001). For instance, the original mammogram in Fig. 16 has a resolution of 3691×6466 . In United Snakes, we can achieve the real-time interactive segmentation of the breast region on the original mammogram with only two or three seed points using the truncated image pyramid technique proposed in Section 4.2.

7. Discussion and conclusion

It is concluded in (Falcão et al., 1998) that the main goals of research in interactive segmentation methods are (i) to provide as complete control as possible to the user of the segmentation process while it is being executed and (ii) to minimize the user's intervention and the total user time required for segmentation. The entire segmentation process may be thought of as consisting of two tasks: *recognition* and *delineation*. Recognition determines roughly where the object (boundary) is, while delineation defines precisely the spatial extent of the object region/boundary in the image. For practical applications, we have found that an additional task – *refinement* – is essential. The errors in reproducibility occur mostly in the vicinity of seed points (Mortensen and Barrett, 1999). In United Snakes, both live wire traces and hard constraint points can be interactively adjusted for refinement. Furthermore, dynamically instantiated snakes can tolerate the live wire errors and thus reduce the number of the seed points which are interactively given by the user. In other words, United Snakes provides more complete control to the user while further

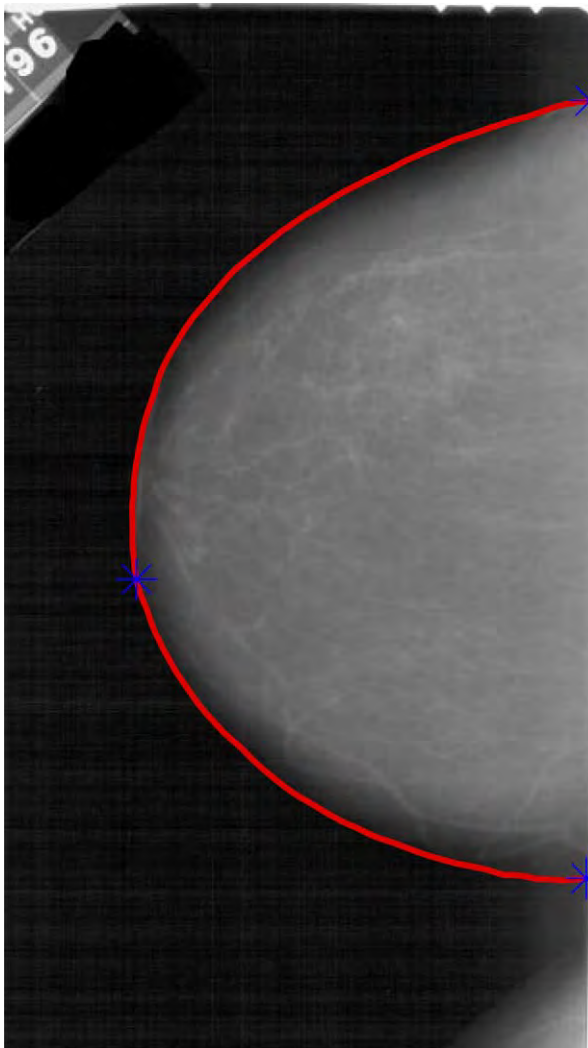


Fig. 16. Real-time isolation of the breast region in mammograms from a 3691×6466 pixel image using only three seed points.

minimizing the user's intervention in the interactive segmentation process.

In summary, our United Snakes framework unites several snake variants with live wire to provide a general purpose tool for interactive medical image segmentation and tracking. The union of these techniques amplifies the efficiency, flexibility and reproducibility of the component techniques. The United Snakes technique offers more control for relatively less user interaction. As it quickly locks onto the image features of interest with reasonable tolerance to errors in live wire, the snake fully exploits the user guidance and expert prior knowledge captured by the initial live wire trace and the seed points. We have demonstrated the generality, accuracy and robustness of United Snakes in applications ranging from the segmentation of neuronal dendrites in EM images, to the analysis of dynamic chest images, to the quantification of growth plates, to the isolation of the breast region in mammograms, among other examples. We believe that United Snakes are in several ways superior to live wire or snakes alone.

We the creators of the United Snakes, in order to form a more perfect union of snake technologies, plan to incorporate within our framework, affine cell image decomposition methods for snake topological adaptability (McInerney and Terzopoulos, 2000), advanced snake motion tracking mechanisms (Terzopoulos and Szeliski, 1992; Blake and Isard, 1998), generic hard constraint mechanisms (Fua and Brechbühler, 1997; Fua, 1997), automatic learning and adaptation of shape functions to specific images, and other snake techniques. We anticipate that such efforts will further enhance the effectiveness of this image segmentation tool.

Acknowledgments

This work was carried out at the Turku Centre for Computer Science, Turku, Finland, and in the Department of Computer Science, University of Toronto, Toronto, Canada. J. Liang thanks the Turku Centre for Computer Science and the Science Foundation of Instrumentarium Corporation (now GE Healthcare) for their financial support, as well as the Faculty of Mathematics and Natural Sciences of the University of Turku for a faculty research award that supported the dynamic chest image analysis project. The authors gratefully acknowledge the insightful comments and suggestions of Prof. Alexandre X. Falcão, Prof. Pascal Fua, Prof. Darius M. Gavrilă, Prof. Timo Järvi, Prof. Ron Kimmel, Dr. Aaro Kiuru, Prof. Martti Kormano, Prof. Eric Mortensen, Dr. Erkki Svedström, Prof. Jayaram K. Udupa, and the anonymous referees. The chest images were acquired by Dr. Raimo Virkki and provided by Dr. Aaro Kiuru. The growth plate image was provided by Prof. Martti Kormano and Dr. Matti Sauna-aho. The mammogram was provided by Prof. Olli Navalainen. The cell image was obtained from Dr. Kristen Harris of the Harvard Medical School. The retinal angiogram was obtained from Dr. Piotr Jasiobedzki. The angiogram and spine images were

provided courtesy of Eric Mortensen of Brigham Young University. The bladder image and the corpus callosum images were provided courtesy of Johannes Hug of Swiss Federal Institute of Technology, Switzerland. The MR wrist images were provided courtesy of Dr. Alexandre X. Falcão of State University of Campinas, Brazil.

Appendix A. Finite element Snakes formulation

The two coordinate functions $x(s, t)$ and $y(s, t)$ of the snake $v(s, t)$ are independent, we shall develop the finite element formulation and the corresponding matrix equations in terms of only one component $x(s, t)$. An identical form will be assumed for component $y(s, t)$. We apply Galerkin's method to the Euler–Lagrange equation for $x(s, t)$:

$$\mu \frac{\partial^2 x}{\partial t^2} + \gamma \frac{\partial x}{\partial t} - \frac{\partial}{\partial s} \left(\alpha \frac{\partial x}{\partial s} \right) + \frac{\partial^2}{\partial s^2} \left(\beta \frac{\partial^2 x}{\partial s^2} \right) - q(x) = 0, \quad (\text{A.1})$$

which expresses the necessary condition for the snake at equilibrium. The average weighted residual is

$$I = \int_0^L \left(\mu \frac{\partial^2 x}{\partial t^2} + \gamma \frac{\partial x}{\partial t} - \frac{\partial}{\partial s} \left(\alpha \frac{\partial x}{\partial s} \right) + \frac{\partial^2}{\partial s^2} \left(\beta \frac{\partial^2 x}{\partial s^2} \right) - q(x) \right) \times w(s) ds = 0, \quad (\text{A.2})$$

where $w(s)$ is an arbitrary test function. By performing integrations by parts once for the third term and twice for the fourth term of (A.2), we arrive at the weak formulation of the snake model:

$$\int_0^L w \mu \frac{\partial^2 x}{\partial t^2} ds + \int_0^L w \gamma \frac{\partial x}{\partial t} ds + \int_0^L \left(\frac{\partial w}{\partial s} \alpha \frac{\partial x}{\partial s} \right) ds + \int_0^L \left(\frac{\partial^2 w}{\partial s^2} \beta \frac{\partial^2 x}{\partial s^2} \right) ds - \int_0^L w q ds + b = 0, \quad (\text{A.3})$$

where

$$b = \left[-w \alpha \frac{\partial x}{\partial s} + w \frac{\partial}{\partial s} \left(\beta \frac{\partial^2 x}{\partial s^2} \right) - \frac{\partial w}{\partial s} \beta \frac{\partial^2 x}{\partial s^2} \right]_0^L \quad (\text{A.4})$$

are the boundary conditions at the two boundary points, $s = 0$ and $s = L$. We approximate $x(s, t)$ as

$$x(s, t) = \mathbf{N}(s) \mathbf{u}(t), \quad (\text{A.5})$$

where $\mathbf{N}(s) = [N_1(s), N_2(s), \dots, N_n(s)]$ are the shape functions and $\mathbf{u}(t) = [u_1(t), u_2(t), \dots, u_n(t)]^T$ are the n nodal variables (degrees of freedom) of the snake model, implying the derivatives of $x(s, t)$ are

$$\frac{\partial^2 x}{\partial t^2} = \mathbf{N} \ddot{\mathbf{u}}, \quad \frac{\partial x}{\partial t} = \mathbf{N} \dot{\mathbf{u}}, \quad \frac{\partial x}{\partial s} = \frac{\partial \mathbf{N}}{\partial s} \mathbf{u}, \quad \frac{\partial^2 x}{\partial s^2} = \frac{\partial^2 \mathbf{N}}{\partial s^2} \mathbf{u}. \quad (\text{A.6})$$

In Galerkin's method, the arbitrary test function w takes the form

$$w = \mathbf{N} \mathbf{c}, \quad (\text{A.7})$$

where \mathbf{N} are the same shape functions as in (A.5), and \mathbf{c} is an arbitrary vector. As w is a scalar, we have

$$w = w^T = \mathbf{c}^T \mathbf{N}^T. \quad (\text{A.8})$$

Substituting (A.5)–(A.8) into (A.3) yields the snake equations of motion

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} - \mathbf{F} + \mathbf{P} = \mathbf{0}, \quad (\text{A.9})$$

where \mathbf{M} is the mass matrix, \mathbf{C} is the damping matrix, \mathbf{K} is the stiffness matrix, \mathbf{F} is the force vector, and \mathbf{P} is the boundary forces, defined as follows:

$$\mathbf{M} = \int_0^L \mathbf{N}^T \mu \mathbf{N} \, ds, \quad (\text{A.10})$$

$$\mathbf{C} = \int_0^L \mathbf{N}^T \gamma \mathbf{N} \, ds, \quad (\text{A.11})$$

$$\mathbf{K} = \mathbf{K}_x + \mathbf{K}_\beta, \quad (\text{A.12})$$

$$\mathbf{K}_x = \int_0^L \left(\frac{\partial \mathbf{N}}{\partial s} \right)^T \alpha \left(\frac{\partial \mathbf{N}}{\partial s} \right) ds, \quad (\text{A.13})$$

$$\mathbf{K}_\beta = \int_0^L \left(\frac{\partial^2 \mathbf{N}}{\partial s^2} \right)^T \beta \left(\frac{\partial^2 \mathbf{N}}{\partial s^2} \right) ds, \quad (\text{A.14})$$

$$\mathbf{F} = \int_0^L \mathbf{N}^T q \, ds, \quad (\text{A.15})$$

$$\mathbf{P} = \left[-\mathbf{N}^T \alpha \frac{\partial \mathbf{N}}{\partial s} + \mathbf{N}^T \frac{\partial}{\partial s} \left(\beta \frac{\partial^2 \mathbf{N}}{\partial s^2} \right) - \left(\frac{\partial \mathbf{N}}{\partial s} \right)^T \beta \frac{\partial^2 \mathbf{N}}{\partial s^2} \right]_0^L \mathbf{u}. \quad (\text{A.16})$$

Eq. (A.9) gives the finite element formulation for the whole snake. To achieve acceptable accuracy in the finite element approximation, the integration domain should be discretized into a number of small subdomains, resulting in the finite element mesh. That is, the snake contour is divided into small segments (elements), each of which can still be considered a snake. Applying (A.9) to an element e , we have $\mathbf{M}^e \ddot{\mathbf{u}}^e + \mathbf{C}^e \dot{\mathbf{u}}^e + \mathbf{K}^e \mathbf{u}^e - \mathbf{F}^e + \mathbf{P}^e = \mathbf{0}$, where \mathbf{M}^e is the element mass matrix, \mathbf{C}^e is the element damping matrix, \mathbf{K}^e is the element stiffness matrix, \mathbf{F}^e the element force vector, and \mathbf{P}^e the element boundary forces applied to the boundary points of the element. Assembling the element matrices results in the system matrix motion equation (4). In a closed snake, the boundary forces will cancel each other. In an open snake, the boundary conditions may be assumed to be zero at the two ends. However, for generality and clarity, we introduce \mathbf{g} for the external force vector.

To solve the motion equation (4), we replace the time derivatives of \mathbf{u} with the backward finite differences

$$\ddot{\mathbf{u}} = (\mathbf{u}^{(t+\Delta t)} - 2\mathbf{u}^{(t)} + \mathbf{u}^{(t-\Delta t)})/(\Delta t)^2, \quad \dot{\mathbf{u}} = (\mathbf{u}^{(t+\Delta t)} - \mathbf{u}^{(t)})/\Delta t,$$

where the superscripts denote the quantity evaluated at the time given in the parentheses and the time step is Δt . This yields the update formula

$$\mathbf{A}\mathbf{u}^{(t+\Delta t)} = \mathbf{b}\mathbf{u}^{(t)} + \mathbf{c}\mathbf{u}^{(t-\Delta t)} + \mathbf{g}, \quad (\text{A.17})$$

where $\mathbf{A} = \mathbf{M}/(\Delta t)^2 + \mathbf{C}/\Delta t + \mathbf{K}$ and $\mathbf{b} = 2\mathbf{M}/(\Delta t)^2 + \mathbf{C}/\Delta t$ and $\mathbf{c} = -\mathbf{M}/(\Delta t)^2$. Because \mathbf{A} is symmetric and banded, it can be economically saved in skyline storage, and efficiently factorized uniquely into the form $\mathbf{A} = \mathbf{LDL}^T$, where \mathbf{L} is a lower triangular matrix and \mathbf{D} is a diagonal

matrix (Bathe and Wilson, 1976). The solution $\mathbf{u}^{(t+\Delta t)}$ to (A.17) is obtained by first solving $\mathbf{L}\mathbf{s} = \mathbf{b}\mathbf{u}^{(t)} + \mathbf{c}\mathbf{u}^{(t-\Delta t)}$ with forward substitution, then $\mathbf{L}^T\mathbf{u} = \mathbf{D}^{-1}\mathbf{s}$ with backward substitution. Since \mathbf{A} is constant, only a single factorization is necessary. Therefore, at each time step only the forward/backward substitutions are performed to integrate the snake equations of motion forward through time.

References

- Amini, A., Weymouth, T., Jain, R., 1990. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (9), 855–867.
- Barrett, W., Mortensen, E., 1997. Interactive live-wire boundary extraction. *Medical Image Analysis* 1 (4), 331–341.
- Bathe, K.-J., Wilson, E.L., 1976. *Numerical Methods in Finite Element Analysis*. Prentice-Hall, Englewood Cliffs, NJ.
- Blake, A., Isard, M., 1998. *Active Contours*. Springer, Berlin.
- Carlbom, I., Terzopoulos, D., Harris, K., 1994. Computer-assisted registration, segmentation, and 3D reconstruction from images of neuronal tissue sections. *IEEE Transactions on Medical Imaging* 13 (2), 351–362.
- Cohen, L., Cohen, I., 1993. Finite element methods for active contour models and balloons for 2D and 3D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (11), 1131–1147.
- Cohen, L., Kimmel, R., 1997. Global minimum for active contour models: a minimal path approach. *International Journal of Computer Vision* 24 (1), 57–78.
- Dubuisson-Jolly, M.P., Gupta, A., 2001. Tracking deformable templates using a shortest path algorithm. *Computer Vision and Image Understanding* 81 (1), 26–45.
- Falcão, A.X., Udupa, J.K., 1997. Segmentation of 3D objects using live wire. In: *SPIE on Medical Imaging 1997*, vol. 3034, Newport Beach, CA, pp. 228–239.
- Falcão, A.X., Udupa, J.K., Miyazawa, F.K., 2000. An ultra-fast user-steered segmentation paradigm: live-wire-on-the-fly. *IEEE Transactions on Medical Imaging* 19 (1), 55–62.
- Falcão, A.X., Udupa, J.K., Samarasekera, S., Hirsch, B.E., 1996. User-steered image boundary segmentation. In: *Proceedings of SPIE on Medical Imaging*, vol. 2710, Newport Beach, CA, pp. 278–288.
- Falcão, A.X., Udupa, J.K., Samarasekera, S., Sharma, S., 1998. User-steered image segmentation paradigms: live wire and live lane. *Graphical Models and Image Processing* 60, 233–260.
- Fua, P., 1997. Model-based optimization: an approach to fast, accurate, and consistent site modeling from imagery. In: Firschein, O., Strat, T.M. (Eds.), *RADIUS: Image Understanding for Intelligence Imagery*. Morgan Kaufmann, Los Altos, CA.
- Fua, P., Brechbühler, C., 1997. Imposing hard constraints on deformable models through optimization in orthogonal subspaces. *Computer Vision and Image Understanding* 65, 148–162.
- Gavrila, D.M., 1996. Hermite deformable contours. In: *Proceedings of the International Conference on Pattern Recognition*, Vienna, Austria, pp. 130–135.
- Grzeszczuk, R., Levin, D., 1994. Brownian strings: segmenting images with stochastically deformable contours. In: Robb, R. (Ed.), *Proceedings of the Third Conference on Visualization in Biomedical Computing (VBC'94)*, SPIE Proceedings, vol. 2359. SPIE, pp. 72–89.
- Hyche, M.E., Ezquerro, N.F., Mullick, R., 1992. Spatiotemporal detection of arterial structure using active contours. In: *Proceedings of the Second Conference on Visualization in Biomedical Computing (SPIE vol. 1808)*, Chapel Hill, NC, October, pp. 52–62.
- Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: active contour models. *International Journal of Computer Vision* 1 (4), 321–331.
- Kwon, Y.W., Bang, H., 1997. *The Finite Element Method Using Matlab*. CRC Mechanical Engineering Series. CRC Press, Boca Raton, FL.

- Liang, J., 2000. Dynamic chest image analysis: new model-based methods for dynamic pulmonary imaging and other applications. Turku Centre for Computer Science, Turku, Finland, December [TUCS Dissertation No. 31] (available at <http://www.cs.toronto.edu/~liang/phddissertation.pdf>).
- Liang, J., Haapanen, A., Järvi, T., Kiuru, A., Kormano, M., Svedström, E., Virkki, R., 1998. Dynamic chest image analysis: model-based pulmonary perfusion analysis with pyramid images. In: Hoffman, E.A., (Ed.), *Medical Imaging 1998: Physiology and Function from Multidimensional Images*, San Diego, CA, pp. 63–72.
- Liang, J., Järvi, T., Kiuru, A., Kormano, M., Svedström, E., 2001. Dynamic chest image analysis: evaluation of model-based perfusion analysis with pyramid images. In: *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Istanbul, Turkey, October, pp. 415–420 (invited paper).
- Liang, J., Järvi, T., Kiuru, A., Kormano, M., Svedström, E., 2003. Dynamic chest image analysis: model-based perfusion analysis in dynamic pulmonary imaging. *EURASIP Journal on Applied Signal Processing* (5), 437–448 (special issue on Advances in Modality-Oriented Medical Image Processing).
- Liang, J., Järvi, T., Kiuru, A., Kormano, M., Svedström, E., Virkki, R., 1997. Dynamic chest image analysis: model-based ventilation study with pyramid images. In: Hoffman, E.A. (Ed.), *Medical Imaging 1997: Physiology and Function from Multidimensional Images*, Newport Beach, CA, pp. 81–92.
- Liang, J., McInerney, T., Terzopoulos, D., 1999a. United snakes. In: *Proceedings of the Seventh International Conference on Computer Vision (ICCV'99)*, Kerkyra (Corfu), Greece, September. IEEE Computer Society Press, Silver Spring, MD, pp. 933–940.
- Liang, J., McInerney, T., Terzopoulos, D., 1999b. Interactive medical image segmentation with united snakes. In: *Proceedings of the Second International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 99)*, Cambridge, England, September. Springer, Berlin, pp. 116–127.
- Liang, J., Virkki, R., Järvi, T., Kiuru, A., Kormano, M., Svedström, E., 1997. Dynamic chest image analysis: evaluation of model-based ventilation study with pyramid images. In: Zurawski, R., Liu, Z.-Q. (Eds.), *IEEE First International Conference on Intelligent Processing Systems*, Beijing, China, pp. 989–993.
- McInerney, T., Terzopoulos, D., 2000. Topology adaptive snakes. *Medical Image Analysis* 4, 73–91.
- McInerney, T., Terzopoulos, D., 1996. Deformable models in medical image analysis: a survey. *Medical Image Analysis* 1 (2), 91–108.
- Menet, S., Saint-Marc, P., Medioni, G., 1990. B-snakes: implementation and application to stereo. In: *Proceedings DARPA*, pp. 720–726.
- Mortensen, E.N., 2000. Simultaneous multi-frame subpixel boundary definition using toboggan-based intelligent scissors for image and movie editing. Ph.D. Thesis, Department of Computer Science, Brigham Young University, Provo, UT.
- Mortensen, E.N., Barrett, W.A., 1995. Intelligent scissors for image composition. In: *Proceedings of Computer Graphics (SIGGRAPH'95)*, Los Angeles, CA, August, pp. 191–198.
- Mortensen, E.N., Barrett, W.A., 1998. Interactive segmentation with intelligent scissors. *Graphical Models and Image Processing* 60, 349–384.
- Mortensen, E.N., Barrett, W.A., 1999. Toboggan-based intelligent scissors with a four parameter edge model. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, June, pp. 452–458.
- Neuenschwander, W., Fua, P., Székely, G., Kübler, O., 1994. Initializing snakes. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'94)*. IEEE Computer Society Press, Silver Spring, MD, pp. 658–663.
- Ojala, T., Liang, J., Näppi, J., Nevalainen, O., 2000. Interactive segmentation of the breast region from digitized mammograms. In: *Proceedings of the IASTED International Conference on Signal Processing and Communications (SPC 2000)*, Marbella, Spain, September, pp. 132–136.
- Ojala, T., Näppi, J., Nevalainen, O., 2001. Accurate segmentation of the breast region from digitized mammograms. *Computerized Medical Imaging and Graphics* 25 (1).
- Sethian, J.A., 1997. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences of the United States of America* 93 (4), 1591–1595.
- Singh, A., Goldgof, D., Terzopoulos, D., 1998. *Deformable Models in Medical Image Analysis*. IEEE Computer Society Press, Silver Spring, MD.
- Staib, L., Duncan, J., 1992. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (11), 1061–1075.
- Terzopoulos, D., Qin, H., 1994. Dynamic NURBS with geometric constraints for interactive sculpting. *ACM Transactions on Graphics* 13 (2), 103–136.
- Terzopoulos, D., Szeliski, R., 1992. Tracking with Kalman snakes. In: Blake, A., Yuille, A. (Eds.), *Active Vision*. MIT Press, Cambridge, MA, pp. 3–20.
- Terzopoulos, D., Witkin, A., Kass, M., 1988. Constraints on deformable models: recovering 3D shape and nonrigid motion. *Artificial Intelligence* 36 (1), 91–123.
- Xu, C., Prince, J.L., 1998. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing* 7 (3), 359–369.
- Zienkiewicz, O., Taylor, R., 1989. *The Finite Element Method*. McGraw-Hill, New York, NY.