Semantic Segmentation

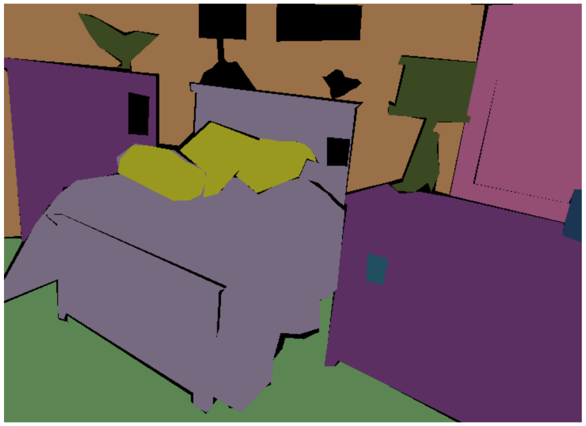
Prepared for CSC2541: Visual Perception for Autonomous Driving

Stefania Raimondo

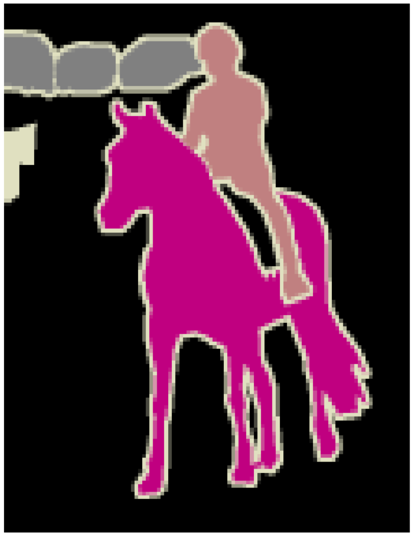
March 15, 2015

Semantic segmentation

• Pixel-level classification



(Badrinarayanan, Kendall, & Cipolla, 2015)



(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

Previously…

- Hand-engineered features, various classifiers

- Deep Convolutional Neural Nets
  - Success at other *high-level* vision tasks (abstract representations)

- DCNN hurdles for low-level tasks:
  - Signal down-sampling ---> reduced signal resolution
  - Spatial invariance ---> limits spatial accuracy

  - In general – hard to train

#1 “DeepLab” (2014)

Semantic image segmentation with deep convolutional nets and fully connected crfs – Chen et al., 2014

Idea

Overcome the two hurdles of DCNNs using the “atrous” algorithm (downsampling issue) and CRFs (spatial insensitivity)

Do not rely on front-end segmentation systems

DeepLab

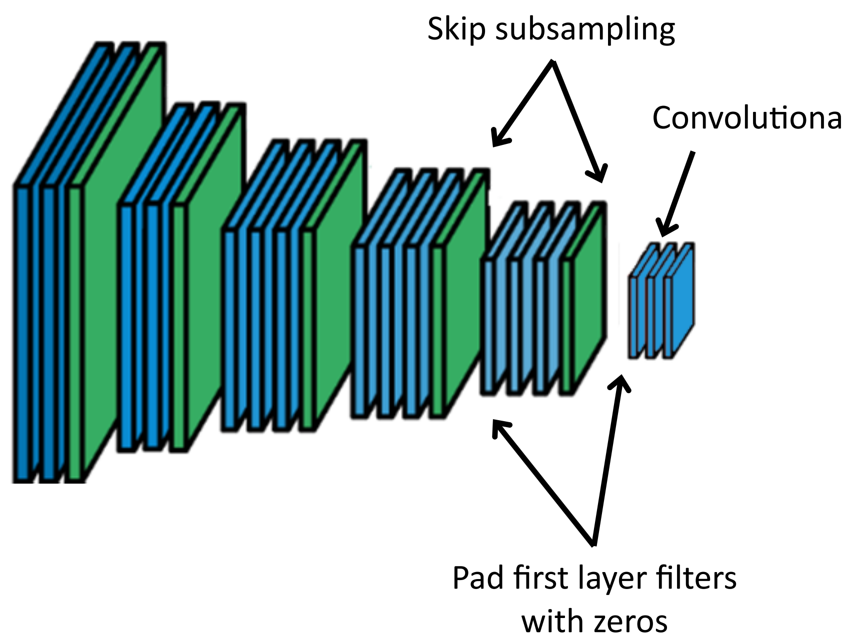
- Deeper (more max-pooling)…
  - … increased invariance and large receptive fields
  - … loss of spatial accuracy

- Previous solutions:
  - Segmentation – 2 stage approaches
  - Harness information from multiple layers

- DeepLab Alternative: CRF

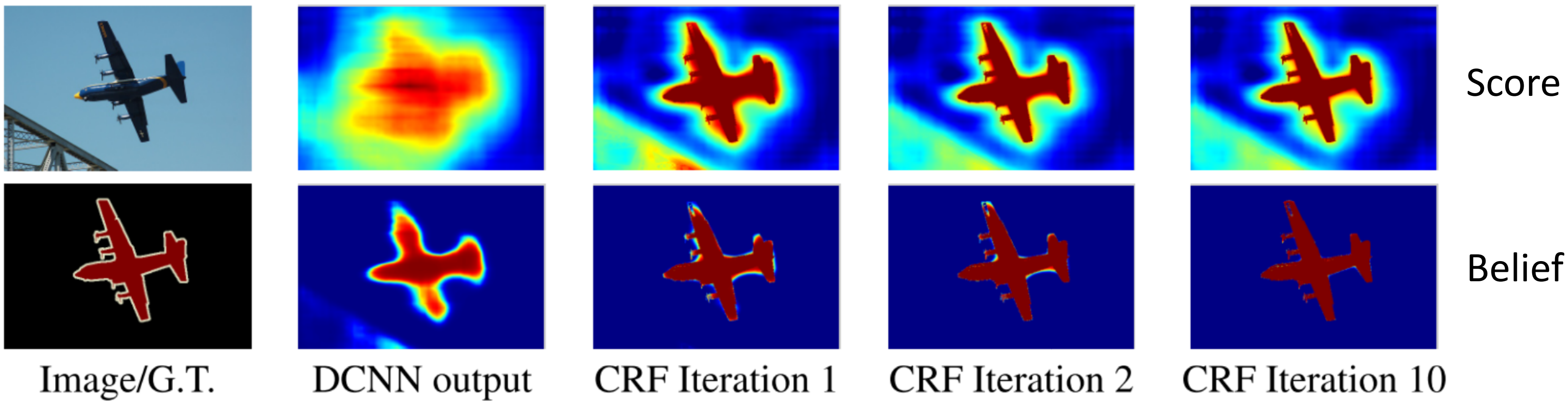
DeepLab DCNN

- Modify ImageNet pre-trained VGG-16 (Simonyan & Zisserman, 2014)
  - fully convolutional
  - dense features
  - Upsampling by bilinear interpolation



(image modified from Badrinarayanan, Kendall, & Cipolla, 2015)

DeepLab CRF



(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

DeepLab CRF

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j)$$ ⟵ Fully connected model

From DCNN label probabilities

Gaussian, pairwise

$$w_1 \exp\Big(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\Big) + w_2 \exp\Big(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\Big)$$

Differences in position and intensity

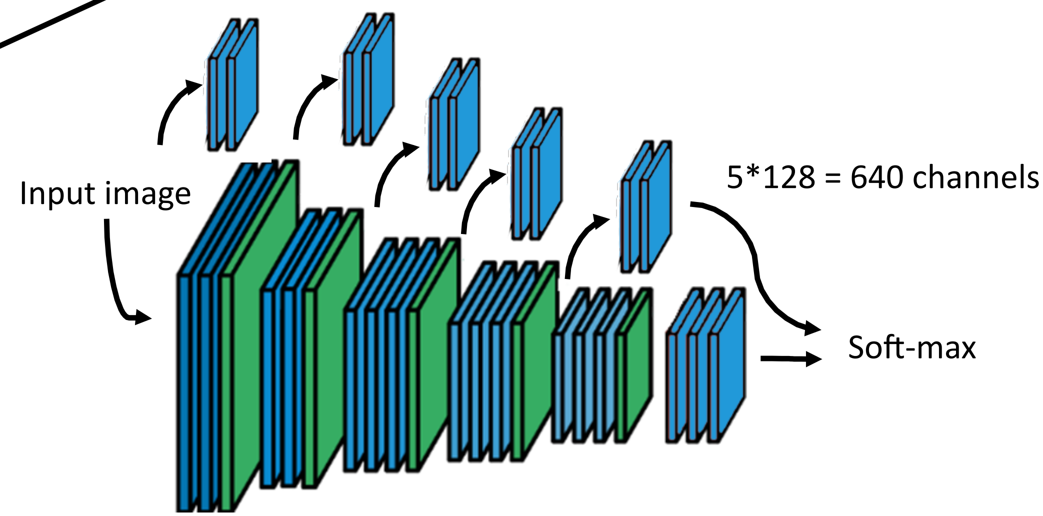
Just position

(equations from Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

DeepLab Variations

- DeepLab (no CRF)
- DeepLab-CRF
- DeepLab-MSc (CRF)
- DeepLab-7x7 (CRF)
- DeepLab-4x4 (CRF)
- DeepLab-LargeFOV (CRF/ MSc)

Multi-scale Prediction



(image modified from Badrinarayanan, Kendall, & Cipolla, 2015)

DeepLab Results

- PASCAL VOC 2012
  - 20 classes + background
  - ~1.5k images for testing/training/validation
  - ~10.5k extra training annotations
  - Performance: IOU averaged across classes

- Most results/experiments provided on 'val' set

Method	mean IOU (%)
DeepLab	59.80
DeepLab-CRF	63.74
DeepLab-MSc	61.30
DeepLab-MSc-CRF	65.21
DeepLab-7x7	64.38
DeepLab-CRF-7x7	67.64
DeepLab-LargeFOV	62.25
DeepLab-CRF-LargeFOV	67.64
DeepLab-MSc-LargeFOV	64.21
DeepLab-MSc-CRF-LargeFOV	68.70

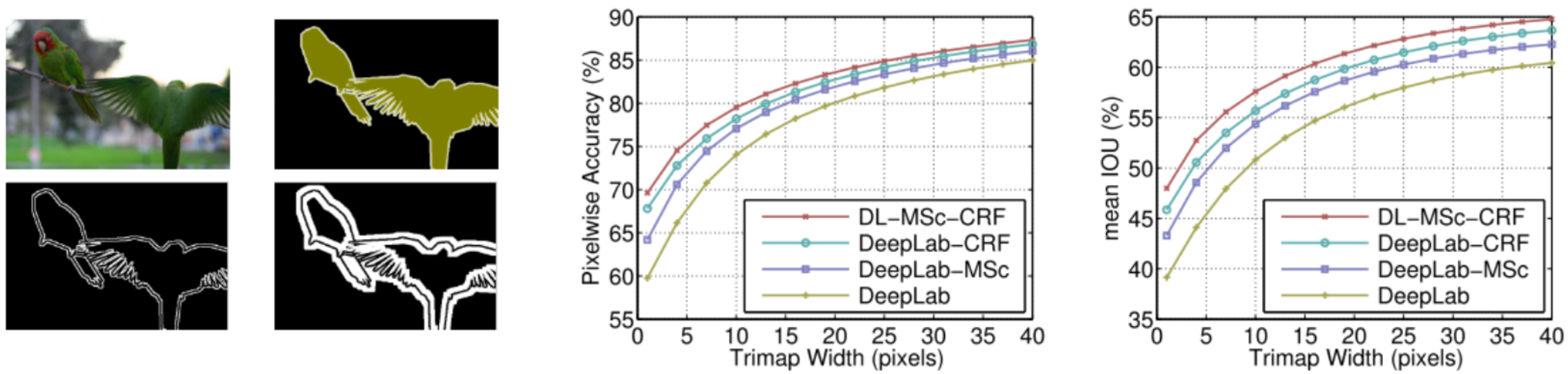
(a)

Method	mean IOU (%)
MSRA-CFM	61.8
FCN-8s	62.2
TTI-Zoomout-16	64.4
DeepLab-CRF	66.4
DeepLab-MSc-CRF	67.1
DeepLab-CRF-7x7	70.3
DeepLab-CRF-LargeFOV	70.3
DeepLab-MSc-CRF-LargeFOV	71.6

(b)

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

DeepLab Results

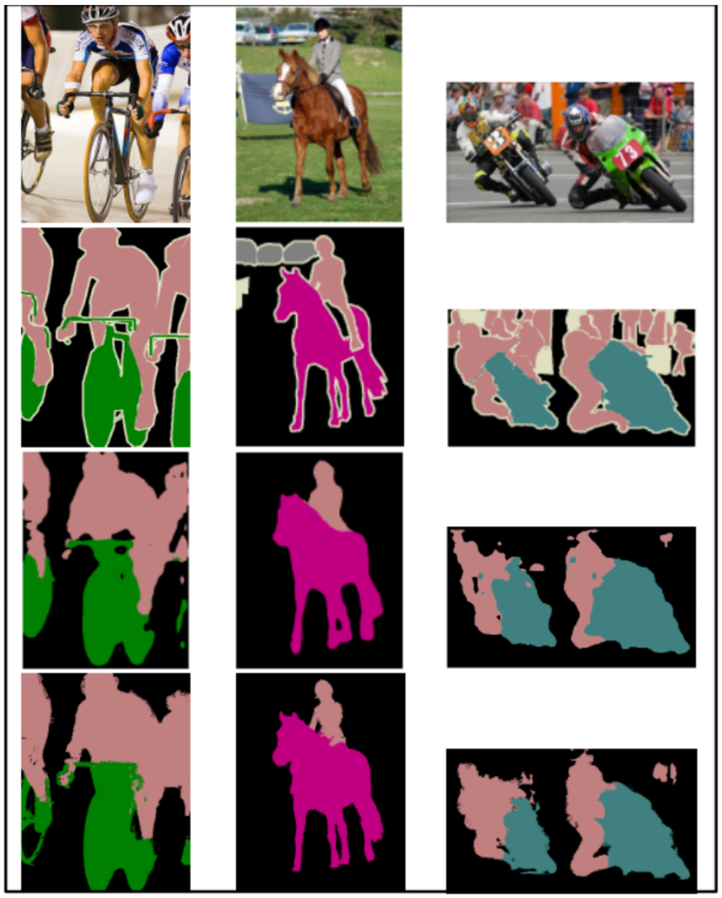


(a) (b) (c)

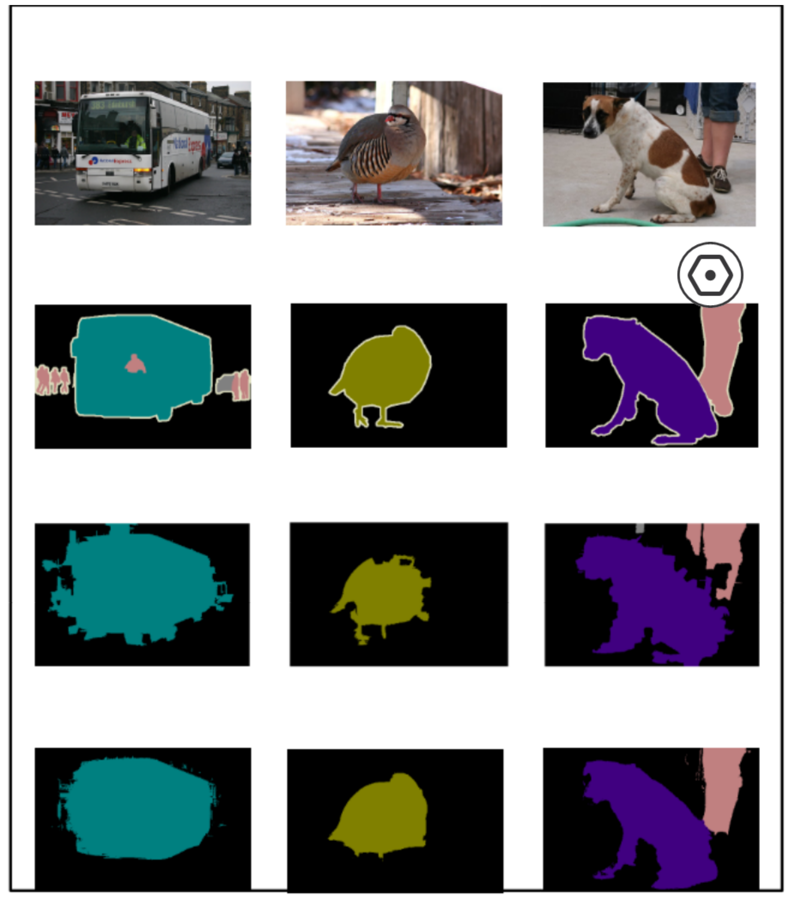
(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

DeepLabResults

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)



(a) FCN-8s vs. DeepLab-CRF



(b) TTI-Zoomout-16 vs. DeepLab-CRF

(Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014)

DeepLab Summary + Future Work…

- State-of-the-art on Pascal Segmentation
- Fast – 8fps DCNN, 0.5s CRF
- Step away from relying on segmentation
  - …but still requires post-processing of NN output
  - …but still not trained end-to-end

- Future work:
  - End-to-end training of CNN + CRF
  - Apply to video, depth maps…
  - Training with weakly supervised annotations

#2 SegNet (2015)

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation – Badrinarayanan et al. 2015

Idea:

Performance boosting support algorithms should not be used to hide deficiencies in core network performance

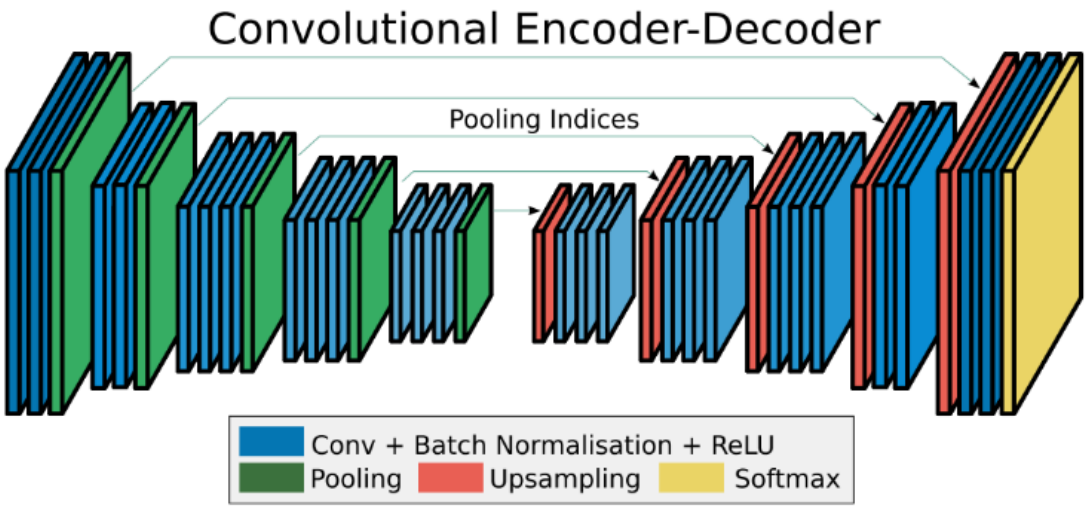
Should train end-to-end

Focus on decoding architecture

SegNet Contributions

- Efficient architecture (memory + computation time)
  - Upsampling reusing max-pooling indices
- Reasonable results without performance boosting addition
- Comparison to FCN

SegNet Architecture



(Badrinarayanan, Kendall, & Cipolla, 2015)

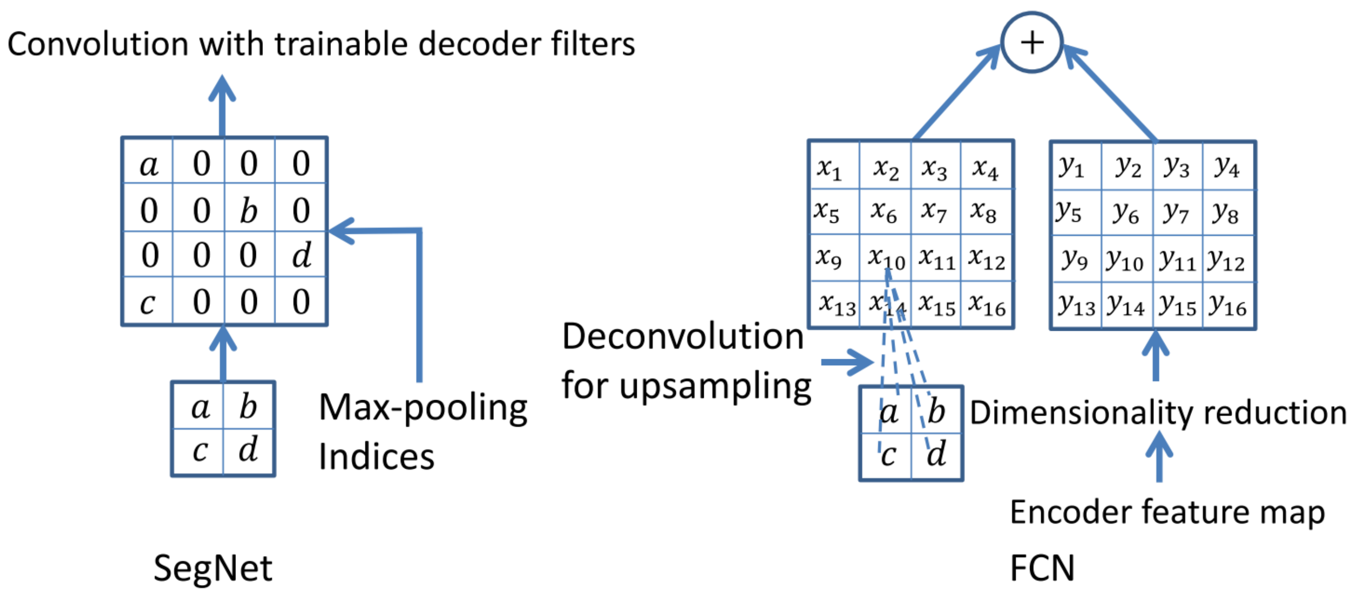
SegNet Evaluation

- Pascal VOC 2012
  - Lots of background – favors methods using weakly labelled data
  - Same objects with different backgrounds
- CamVid – for variants
  - 11 classes, day and dusk, ~300 testing/training images
- SUN RGB-D
  - 37 indoor scene classes, ~5000 training/testing images

SegNet Evaluation

- Measures
  - Global accuracy (G)- % pixels correctly classified
  - Class average accuracy (C) – mean accuracy over all classes
  - Mean intersection over union (I/U)
    - Penalizes false positives, not optimized for

SegNet Decoders



(Badrinarayanan, Kendall, & Cipolla, 2015)

SegNet Decoder Evaluation

- SegNet-Basic
  - 4 enc. + 4 dec., all max-pooling, no RELU, 7x7 kernel
- SegNet-Basic-SingleChannelDecoder
  - decoder only convolve their corresponding upsampled feature map

- FCN Basic
  - fully convolutional decoding technique
- FCN Basic-NoAddition
  - skips the addition step (space)

- Bilinear-Interpolation – no learning for upsampling

SegNet Decoder Evaluation

				Median frequency balancing						Natural frequency balancing					
		Encoder	Infer	Test			Train			Test			Train		
Variant	Params (M)	storage (MB)	time (ms)	G	C	I/U	G	C	I/U	G	C	I/U	G	C	I/U
Fixed upsampling															
Bilinear-Interpolation	0.625	0	24.2	77.9	61.1	43.3	89.1	90.2	82.7	82.7	52.5	43.8	93.5	74.1	59.9
Upsampling using max-pooling indices															
SegNet-Basic	1.425	1x	52.6	82.7	62.0	47.7	94.7	96. 2	92.7	84.0	54.6	46.3	96.1	83.9	73.3
SegNet-Basic-EncoderAddition	1.425	64x	53.0	83.4	**63.6**	48.5	94.3	95.8	92.0	**84.2**	56.5	**47.7**	95.3	80.9	68.9
SegNet-Basic-SingleChannelDecoder	0.625	1x	33.1	81.2	60.7	46.1	93.2	94.8	90.3	83.5	53.9	45.2	92.6	68.4	52.8
Learning to upsample (bilinear initialisation)															
FCN-Basic	0.65	11x	24.2	81.7	62.4	47.3	92.8	93.6	88.1	83.9	55.6	45.0	92.0	66.8	50.7
FCN-Basic-NoAddition	0.65	n/a	23.8	80.5	58.6	44.1	92.5	93.0	87.2	82.3	53.9	44.2	93.1	72.8	57.6
FCN-Basic-NoDimReduction	1.625	64x	44.8	**84.1**	63.4	**50.1**	**95.1**	**96.5**	**93.2**	83.5	**57.3**	47.0	**97.2**	**91.7**	**84.8**
FCN-Basic-NoAddition-NoDimReduction	1.625	0	43.9	80.5	61.6	45.9	92.5	94.6	89.9	83.7	54.8	45.5	95.0	80.2	67.8

(Badrinarayanan, Kendall, & Cipolla, 2015)

SegNet Evaluation

- SUN RGB-D Results – complex indoor scenes
- CamVid - out-door road scenes
- Pascal VOC 2012 – few classes, varying backgrounds

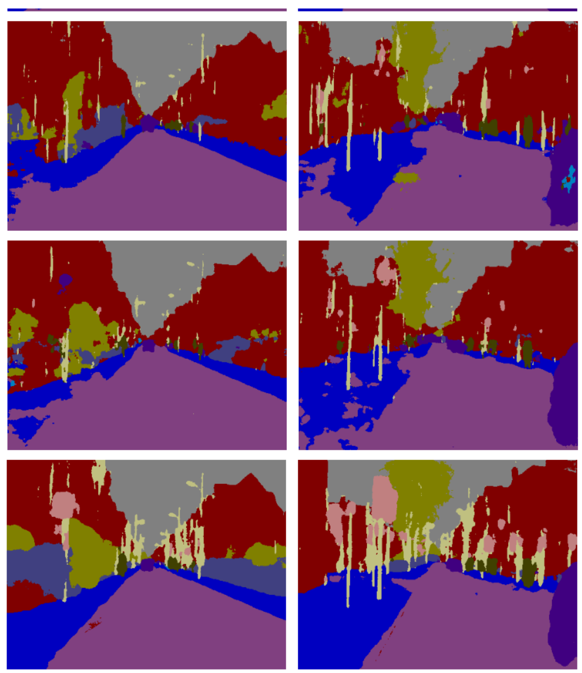
- Demo
  - http://mi.eng.cam.ac.uk/projects/segnet/#demo

SegNet CamVid Results

SegNet-Basic with only local contrast normalized RGB as input (median freq. balancing)

SegNet with only local contrast normalized RGB as input (pre-trained encoder, median freq. balancing)

SegNet with only local contrast normalized RGB as input (pretrained encoder , median freq. balancing + large training set)



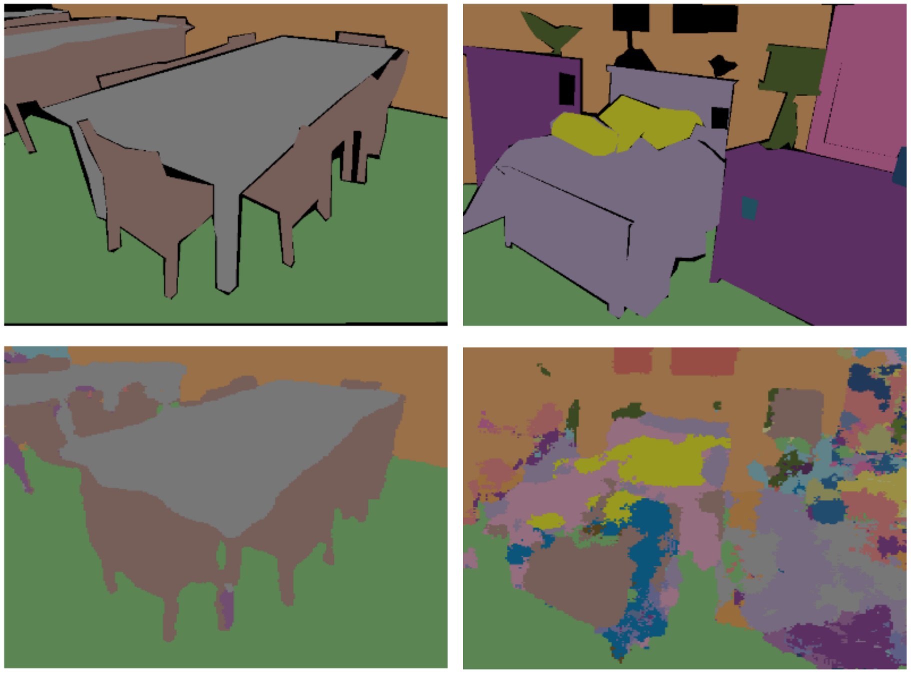
(Badrinarayanan, Kendall, & Cipolla, 2015)

SegNet CamVid Results

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	Mean I/U
SfM+Appearance [26]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a
Boosting [27]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a
Dense Depth Maps [30]	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a
Structured Random Forests [29]	n/a											51.4	72.5	n/a
Neural Decision Forests [60]	n/a											56.1	82.1	n/a
Local Label Descriptors [61]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a
Super Parsing [31]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a
SegNet-Basic	81.3	72.0	93.0	81.3	14.8	93.3	62.4	31.5	36.3	73.7	42.6	62.0	82.7	47.7
SegNet-Basic (layer-wise training [12])	75.0	84.6	91.2	82.7	36.9	93.3	55.0	37.5	44.8	74.1	16.0	62.9	84.3	n/a
SegNet	**88.8**	87.3	92.4	82.1	20.5	97.2	57.1	49.3	27.5	84.4	30.7	65.2	**88.5**	55.6
SegNet (3.5K dataset training)	73.9	**90.6**	90.1	**86.4**	**69.8**	94.5	**86.8**	**67.9**	**74.0**	**94.7**	**52.9**	**80.1**	86.7	**60.4**
CRF based approaches														
Boosting + pairwise CRF [27]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a
Boosting+Higher order [27]	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a
Boosting+Detectors+CRF [28]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a

(Badrinarayanan, Kendall, & Cipolla, 2015)

SegNet SUN RGB-D Results



(Badrinarayanan, Kendall, & Cipolla, 2015)

SegNet Pascal Results

Method	Encoder size (M)	Decoder size (M)	Total size (M)	Class avg. acc.	Inference 500 × 500 pixels	Inference 224 × 224 pixels
DeepLab [14] (validation set)	n/a	n/a	$<$ 134.5	58	n/a	n/a
FCN-8 [2] (multi-stage training)	134	**0.5**	134.5	62.2	210ms	n/a
Hypercolumns [43] (object proposals)	n/a	n/a	$>$ 134.5	62.6	n/a	n/a
DeconvNet [9] (object proposals)	138.35	138.35	276.7	69.6	n/a	92ms (× 50)
CRF-RNN [10] (multi-stage training)	n/a	n/a	$>$ 134.5	**69.6**	n/a	n/a
SegNet	**14.725**	14.725	**29.45**	59.1	**94ms**	**28ms**

(Badrinarayanan, Kendall, & Cipolla, 2015)

-10%

SegNet Summary + Future Work

- Reasonable results w/out support methods
- Comparison of FCN decoding method
- Some failures:
  - Lacks smoothness on large objects
  - Cannot handle clutter

- Future work:
  - Estimate labelling uncertainty
  - Real-time application
  - Dropout during training and testing

#3 Joint Seg + 3D Reconstruction

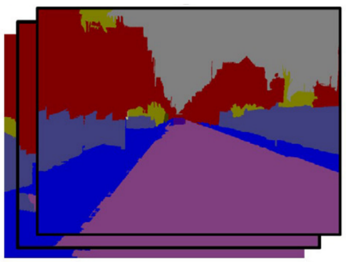
Joint Semantic Segmentation and 3D Reconstruction from Monocular Video – Kundu et al. 2014

Idea

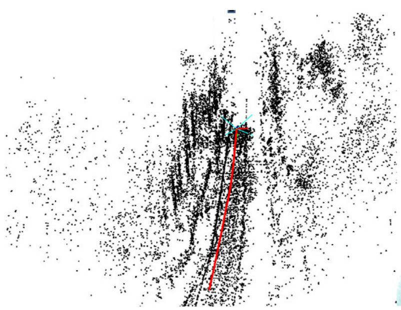
Structural and semantic information is necessary for some applications and can benefit from each other

Method

2D image segmentation

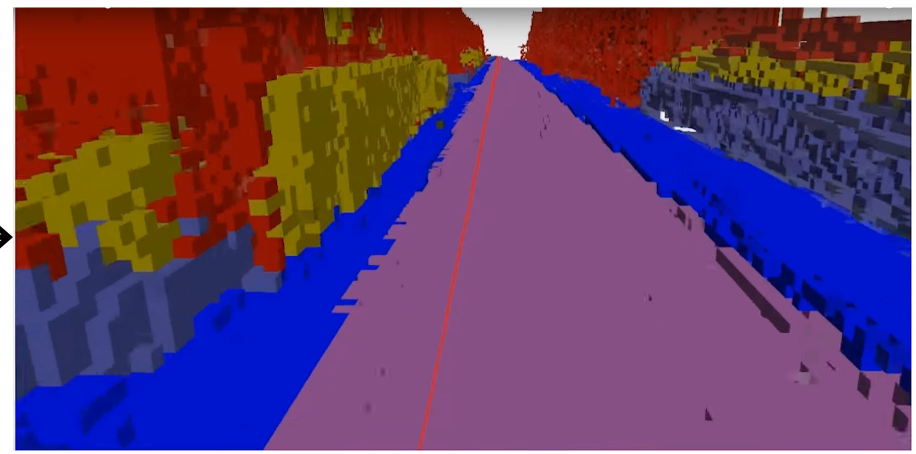


SLAM: trajectory + point cloud



CRF

3D labelled voxel representation



(images from Kundu, Li, Dellaert, Li, & Rehg, 2014)

Contributions

- Method for simultaneous 3D structure and semantics
  - …but not the first to use 3D features to improve 2D segmentation
  - …and not the first to use 2D segmentation to improve 3D depth estimation

- Benefits of this approach
  - Temporally consistent
  - Monocular
  - Does not require dense depth maps
  - Efficient for real-time applications

- First 3D reconstruction of monocular Camvid

CRF Model

$m_i \in \mathcal{M}$ ⟵ voxel i's semantic label $\mathcal{L}_{\mathcal{M}} = \{Free, Road, Car, \ldots\}$

$\mathcal{D} = \{\boldsymbol{z}^r_{1:P}, \boldsymbol{z}^s_{1:Q}, \boldsymbol{g}_{1:T}\}$ ⟵ input data/measurements

camera trajectory per image

with-depth

- pixel + pose
- 2D label
- depth

semantic-only

- pixel + pose
- 2D label

CRF Model

$\mathcal{M}^* = \arg\max_{\mathcal{M}} P(\mathcal{M}|\mathcal{D})$ ⟵ probability of voxel assignments given measurements

$$P(\mathcal{M}|\mathcal{D}) = P(\mathcal{M}) \prod_{p=1}^{P} P\left(z_p^r \middle| \boldsymbol{m}_p, g_p\right) \prod_{q=1}^{Q} P\left(z_q^s \middle| \boldsymbol{m}_q, g_q\right)$$

Do not assume voxels are independent

prior

width-depth

semantic-only

$$= P(\mathcal{M}) \prod_{p=1}^{P} P\left(\boldsymbol{m}_p \middle| z_p^r, g_p\right) \prod_{q=1}^{Q} P\left(\boldsymbol{m}_q \middle| z_q^s, g_q\right)$$

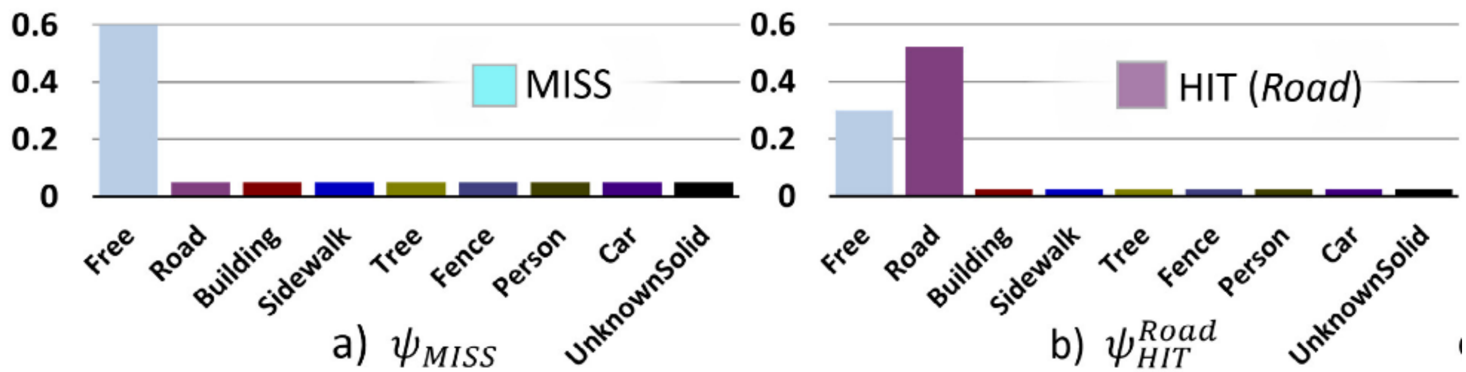
map is independent of camera trajectory

measurements are independent given the map

$$P(\mathcal{M}|\mathcal{D}) = \frac{1}{Z(\mathcal{D})} \prod_{i} \psi_u^i(m_i) \prod_{i,j \in \mathcal{N}} \psi_p(m_i, m_j) \prod_{R \in \mathcal{R}} \psi_h(\mathbf{m}_R)$$

Unary potentials

$$\psi_u^i(m_i) = [\psi_{\text{MISS}}(m_i)]^{N_M} \prod_{l \in \mathcal{L}_{\mathcal{I}} \setminus Sky} [\psi_{\text{HIT}}^l(m_i)]^{N_{Hl}}$$

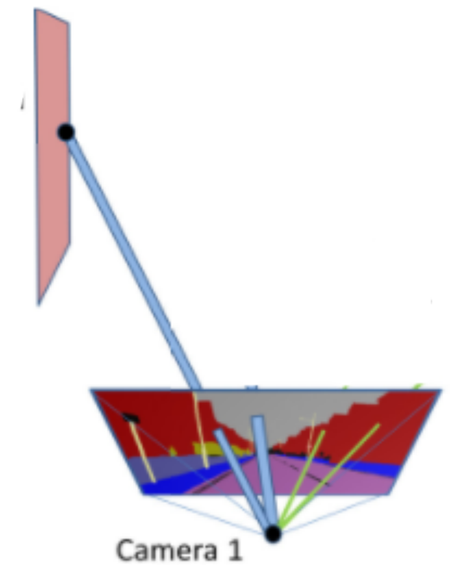


(Kundu, Li, Dellaert, Li, & Rehg, 2014)

Updating Potentials

1) With-depth measurement
2) Semantic-only measurements

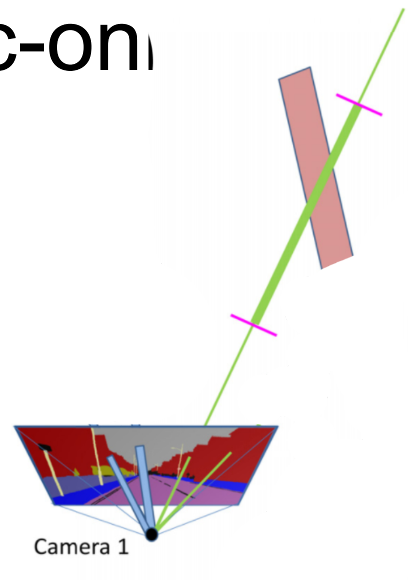
$$\psi_u^i(m_i) = [\psi_{\text{MISS}}(m_i)]^{N_M} \prod_{l \in \mathcal{L}_{\mathcal{I}} \setminus Sky} [\psi_{\text{HIT}}^l(m_i)]^{N_{Hl}}$$



(image modified from Kundu, Li, Dellaert, Li, & Rehg, 2014)

Updating Potentials (Semantic-on

- Depth statistics (per grid cell)
- For low-depth-uncertainty categories:
  - Same as with-depth, add unary factors
- For high-depth-uncertainty categories:
  - Add a higher order factor
  - Joins voxels along the ray bwtn min-max depth



(image modified from Kundu, Li, Dellaert, Li, & Rehg, 2014)

$$\psi_h(\mathbf{m}_R) = \begin{cases} \alpha & \text{if atleast one of } \mathbf{m}_R \text{ is } \neg \mathit{Free} \\ \beta & \text{if all of } \mathbf{m}_R \text{ is } \mathit{Free} \end{cases}$$

Updating Potentials

- Pairwise potentials
  - Neighbors in each direction are treated differently
    - Ex. road more likely in the horizontal direction
  - Lower cost for free neighbor

Implementation details…

- Octree data structure
  - Unused voxels are uninitialized
  - Minimal storage/computation
  - Pairwise/higher order are static across all voxels
  - Only store factor values, not the measruements
- Clamping
  - High probability (0.98) voxels are treated like evidence
  - 3D support for clamped voxels
    - Extra hit unaries for neighbors
    - Including free-space boundaries
- Improving SLAM
  - Reject matches if they lie on different semantic categories
  - Bundle adjustment (minimize re-projection errors)

Results

- Camvid, Leuven, KITTI
  - Fast forward-moving datasets

- Video
  - http://www.cc.gatech.edu/~akundu7/projects/JointSegRec/
- Qualitative 3D reconstruction results
- Quantitative 2D segmentation results
  - Label accuracy
  - Temporal consistency (entropy)

Results

- 2D segmentation results…

CamVid seq05VD	*Building*		*Road*		*Car*		*Sidewalk*		*Sky*		*Tree*		*Fence*		*All*	
	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)
Ours	**0.0**	98.30	**0.0**	97.77	**0.0**	95.75	**0.0**	**98.33**	NA	99.27	**0.0**	**83.63**	**0.0**	73.74	**0.0**	**95.51**
[20]	0.114	**98.52**	0.024	95.99	0.231	89.41	0.177	96.53	NA	**99.81**	0.168	83.02	0.299	**75.59**	0.095	94.58
[24]	0.114	94.78	0.016	98.85	0.106	99.69	0.184	94.11	NA	99.21	0.173	80.34	0.249	39.06	0.084	92.41
[31]	0.025	95.01	0.004	**98.97**	0.046	**99.87**	0.062	73.17	NA	99.26	0.037	74.08	0.107	4.38	0.019	87.88

Leuven	*Building*		*Road*		*Car*		*Sidewalk*		*Sky*		*Bike*		*Pedestrian*		*All*	
	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)
Ours	**0.0**	**96.51**	**0.0**	**99.40**	**0.0**	**91.78**	**0.0**	66.97	NA	**95.30**	**0.0**	83.82	**0.0**	NA	**0.0**	**95.74**
[19]	0.046	95.84	0.116	98.75	0.150	91.42	0.429	**74.89**	NA	93.29	0.264	**84.68**	0.686	**61.76**	0.094	95.24

Kitti seq05	*Building*		*Road*		*Car*		*Sidewalk*		*Sky*		*Tree*		*Fence*		*All*	
	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)	H(bits)	Acc(%)
Ours	**0.0**	**98.90**	**0.0**	**98.72**	**0.0**	96.95	**0.0**	**98.35**	NA	99.37	**0.0**	96.45	**0.0**	**96.34**	**0.0**	**97.20**
[20]	0.165	97.47	0.113	87.85	0.203	**98.14**	0.158	96.00	NA	**99.75**	0.129	**97.47**	0.220	91.55	0.163	95.15

(Kundu, Li, Dellaert, Li, & Rehg, 2014)

3D+Seg Summary

- CRF based method
- Dense reconstruction
- Temporally consistent
- “Tractable for large outdoor environments”

- Future work:
  - Real-time application
  - Incorporating multi-camera information (already done)

Overall Summary

- Semantic Segmentation – pixel-level
  - Need dense output
  - Need to preserve spatial details

- CRF (and other methods) can boost accuracy of NN models
- DCNN models may be able to stand on their own
  - If given sufficient training data and proper architecture
- Use 3D information to augment 2D segmentation

References

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv:1511.00561 [cs]*. Retrieved from http://arxiv.org/abs/1511.00561

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv:1412.7062 [cs]*. Retrieved from http://arxiv.org/abs/1412.7062

Kundu, A., Li, Y., Dellaert, F., Li, F., & Rehg, J. (2014). Joint Semantic Segmentation and 3D Reconstruction from Monocular Video. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (Vol. 8694, pp. 703–718). Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-10599-4_45