CSC 2547: Machine Learning for Vision as Inverse Graphics

Anthony Bonner
www.cs.toronto.edu/~bonner

Scene Understanding

- Much more than just classification.
- Needs a rich 3-dimensional representation of the world.
- Objects, shape, position, orientation, appearance, category, composition, …
- Relationships between objects.
(part-of, next-to, on-top-of, …)
- Illumination, camera angle, …

Inverse Graphics

- Computer graphics represents the world this way internally.
- Inverse problems:
  - Graphics generates a 2D image from a 3D representation.
  - Scene understanding generates a 3D representation from a 2D image.

Paper Presentations

- Each week will focus on one or two topics, as listed on the course web page (soon).
- You can vote for your choice of topic/week (soon).
- I will assign you to a week (soon).
- Papers on each topic will be listed on the course web page.
- If you have a particular paper you would like to add to the list, please let me know.

Paper Presentations

- Goal: high quality, accessible tutorials.
- 7 weeks and 40 students = 6 students per week and 20 minutes per student (including questions).
- 2-week planning cycle:
  - 2 weeks before your presentation, meet me after class to discuss and assign papers.
  - The following week, meet the TA for a practice presentation (required).
  - Present in class under strict time constraints.

Team Presentatations

- Papers may be presented in teams of two or more with longer presentations (20 minutes per team member).
- Unless a paper is particularly difficult or long, a team will be expected to cover more than one paper (one paper per team member).
- A team may cover one of the listed papers and one or more of its references (but see me first).

Tentative Topics

- Discriminative and generative approaches
- Capsule networks
- Point Nets and 3D point clouds
- Group symmetries and equivariance
- Visual attention and transformers
- CNNs for 3D
- Part-whole relationships
- Contrastive and semi-supervised learning
- Adversarial learning

Discriminative Approaches

- Train a single neural net.
- Image is the input
- Scene representation is the output.
- Supervised learning.

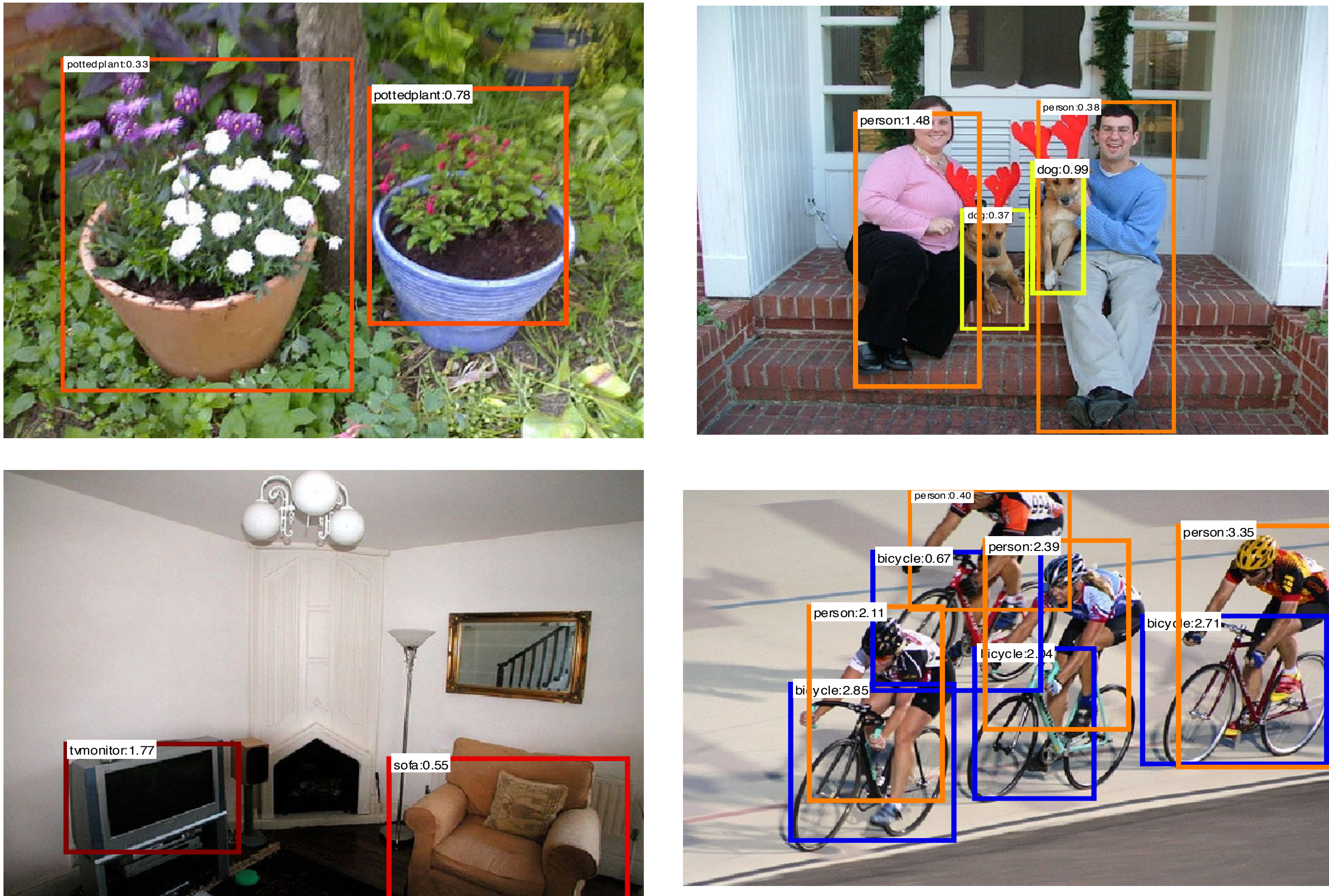
Discriminative Approaches

- Problem: need a labeled scene representation for each training image.
- Use simulated data:
  - Generate many scenes
  - Use a graphics program to generate images of the scene.
- The machine-vision community has many labeled benchmarks of real data.

Human Pose Estimation

From Tompson et al, *Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation,* arXiv 2014.

Object Detection and Localization

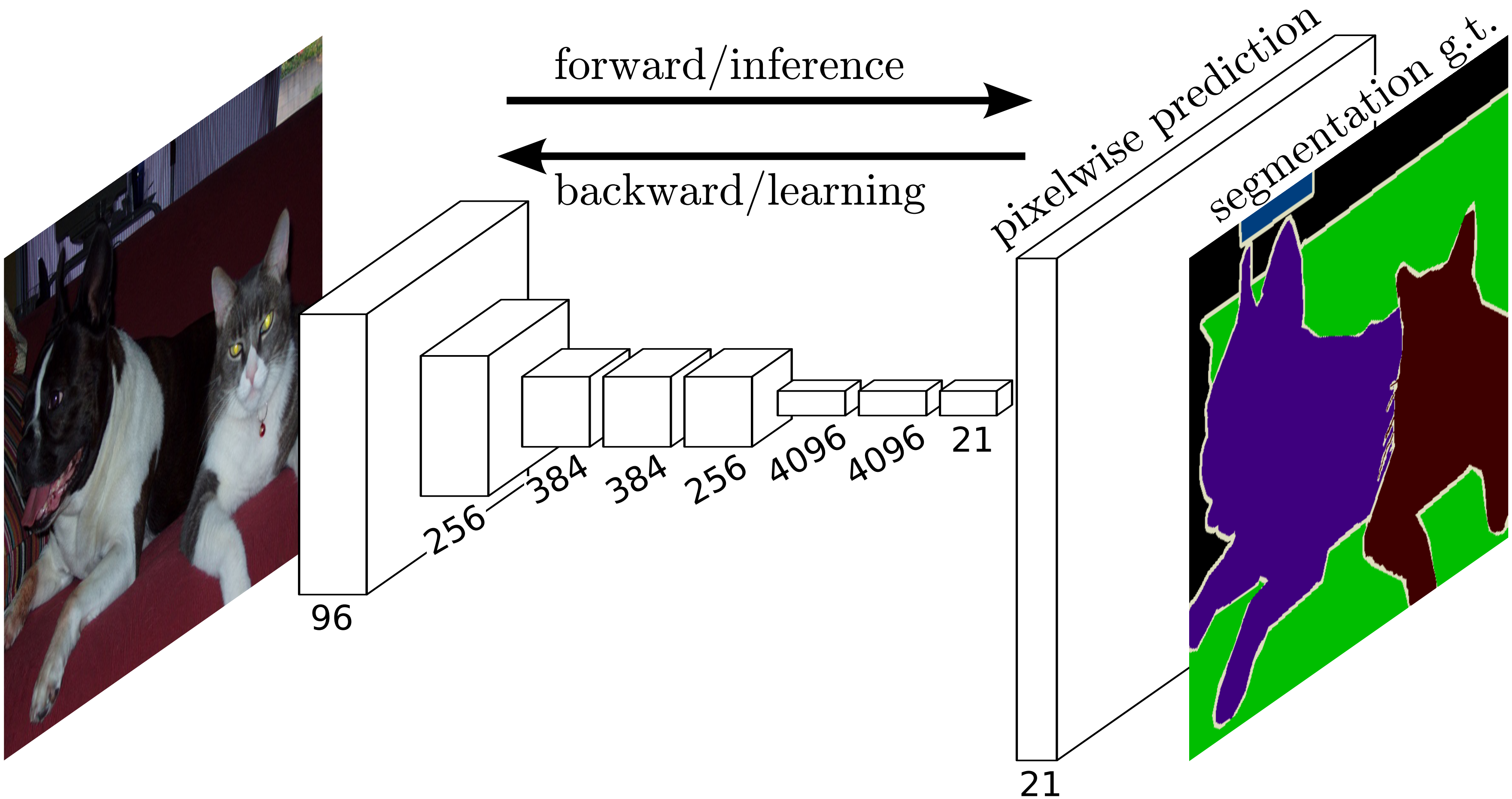


From He et al, *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, arXiv 2015

Image Transformation

- Simplest case:
  - Train a single neural net.
  - Image as input
  - Transformed image as output
- More complex cases:
  - Train two or more feed-forward neural nets.
  - Two or more images as input (one per neural net).
  - Combine outputs into a transformed image.

Semantic Segmentation



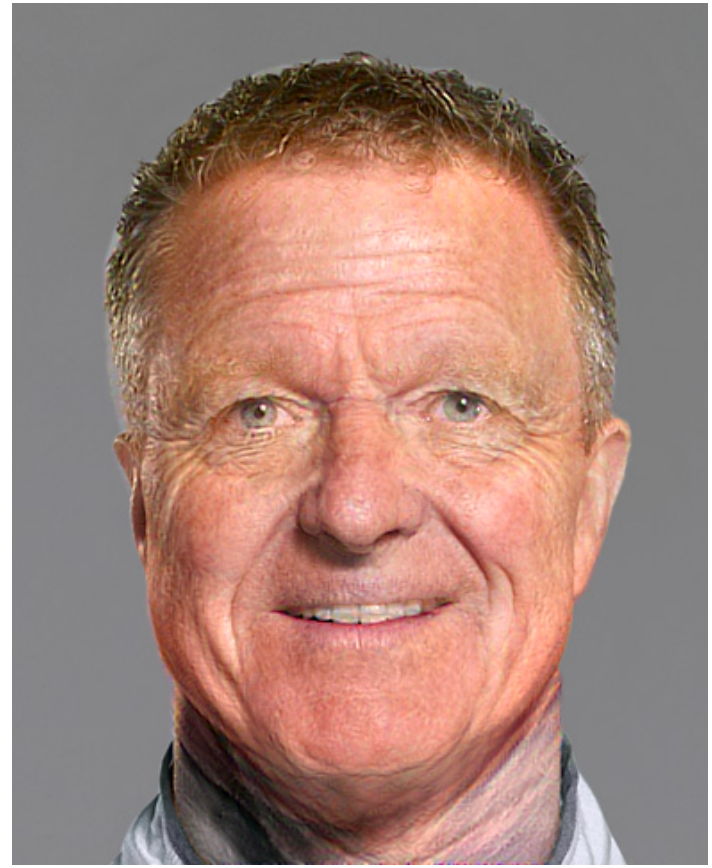
From Long et al, *Fully Convolutional Networks for Semantic Segmentation*, CVPR 2015

Artistic Style Transfer

From Gatys et al, *A Neural Algorithm of Artistic Style*, arXiv 2015

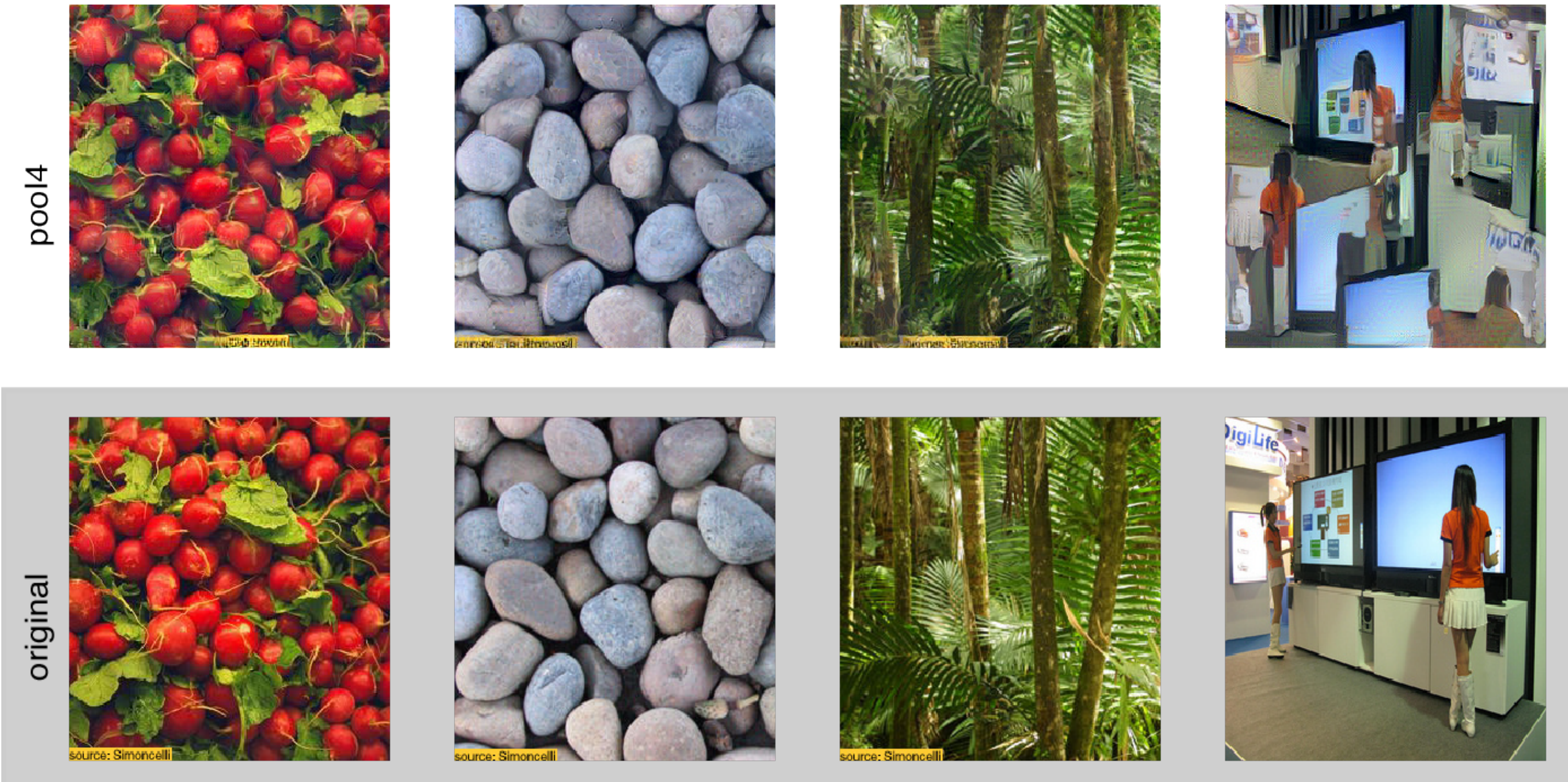
Feature Interpolation


Input


Older

From Upchurch et al, *Deep Feature Interpolation for Image Content Changes*, arXiv 2017

Texture Synthesis



From Gatys et al, *Texture Synthesis Using Convolutional Neural Networks*, NIPS 2015

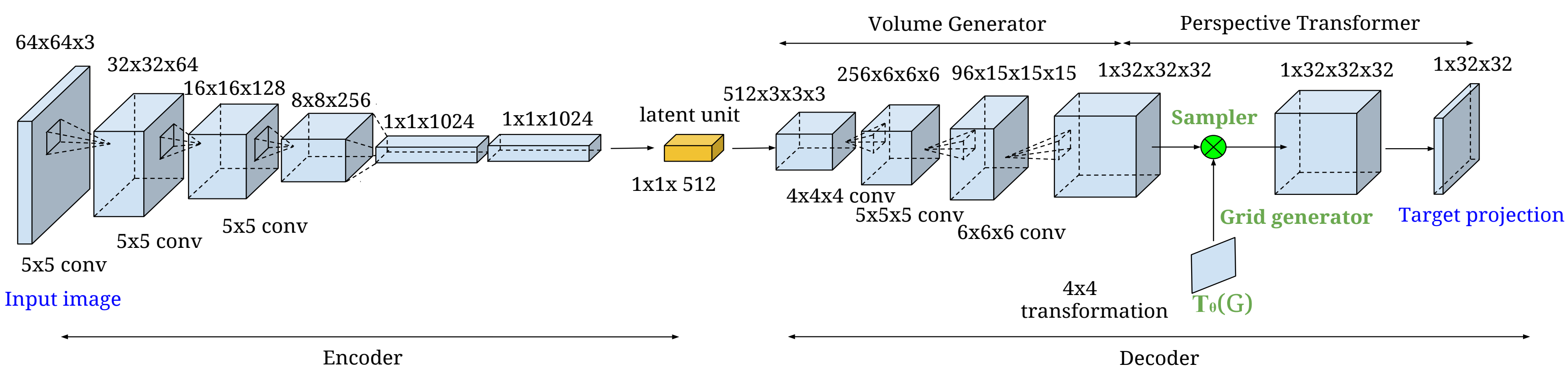
Generative Approaches

- Given a scene, s, a graphics program, G, produces an image, G(s).
- Given an image, x, find s such that $G(s) \approx x$
- More generally, find $P(s|x)$,.
- $P(s|x)$ is high when $G(s)$ is close to x.

Variational Approximations

- Finding $P(s|x)$ is intractable in general.
- Use variational approximations.
- Variational auto-encoders work very well.
- G can be a neural net that we learn (unsupervised).
- Computationally intensive.

Variational Autoencoders



From Yan et al, *Perspective Transformer Nets*, arXiv 2017

Learning 3D Shape

From Yan et al, *Perspective Transformer Nets*, arXiv 2017

Making Visual Analogies

- Given images A, B, C, generate image D so that D is to C as B is to A.

From Reed et al, *Deep Visual Analogy-Making*, NIPS 2015