Advanced topics in deep learning: segmentation and pose estimation

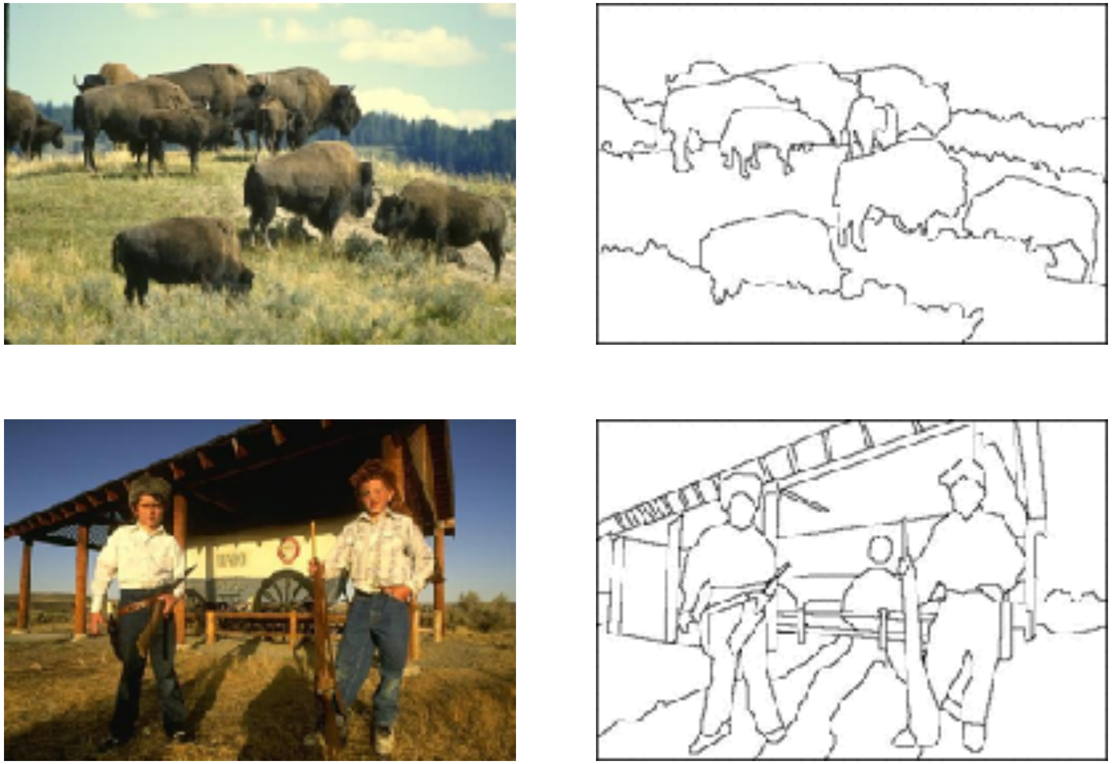
COS 429: Computer Vision

PRINCETON UNIVERSITY

Semantic segmentation

Recall: contour/boundary detection

- Separate image into coherent "regions"

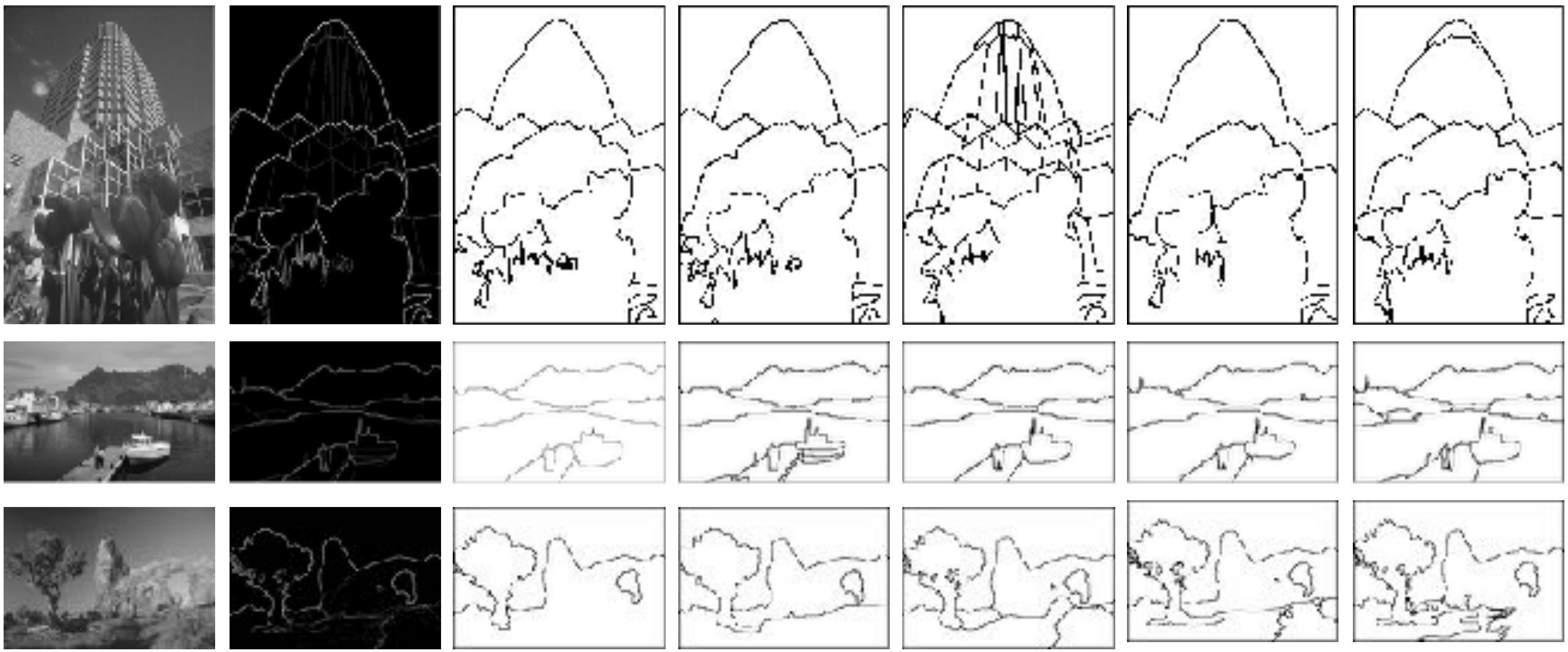


Berkeley segmentation database:
http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

Lazebnik

Human agreement

Berkeley segmentation dataset



A Measure for Objective Evaluation of Image Segmentation Algorithms

R. Unnikrishnan C. Pantofaru M. Hebert

CVPR 2005

Recall: unsupervised/superpixel segmentation

Efficient Graph-Based Image Segmentation
P. Felzenszwalb, D. Huttenlocher
International Journal of Computer Vision, Vol. 59, No. 2, September 2004

http://cs.brown.edu/~pff/segment/

Example Results

Segmentation parameters: sigma = 0.5, K = 500, min = 50.

Segmentation parameters: sigma = 0.5, K = 1000, min = 100.

Generating object proposals

Segmentation as Selective Search for Object Recognition.
Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, Arnold W. M. Smeulders
ICCV 2011

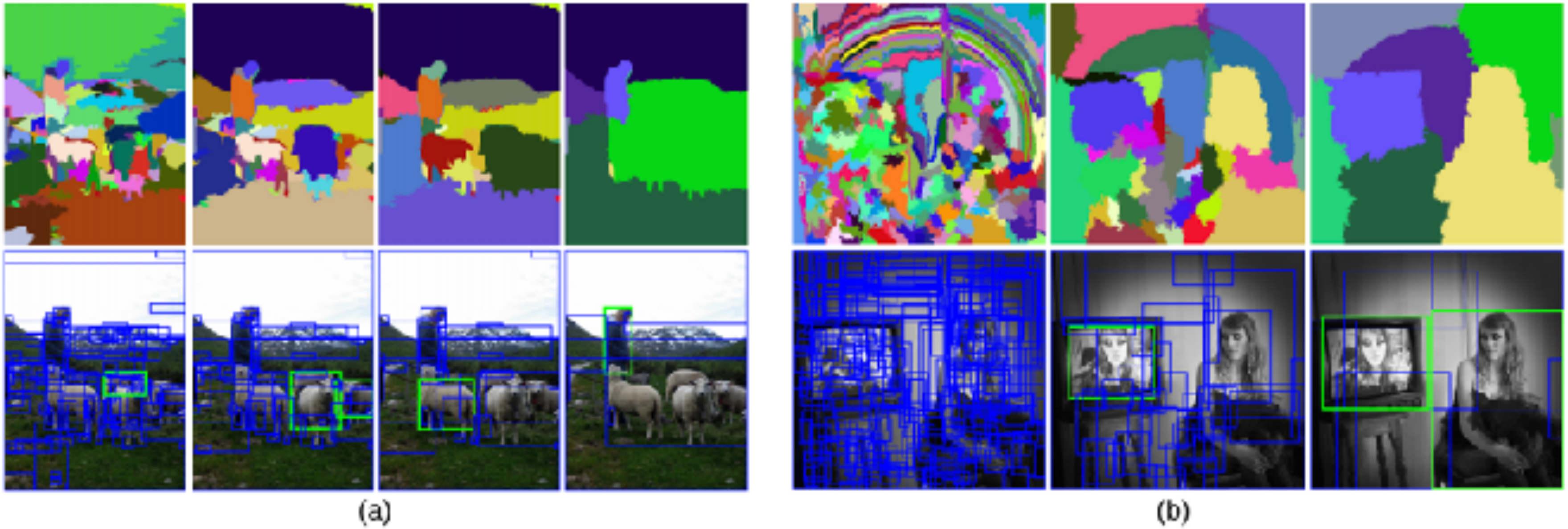


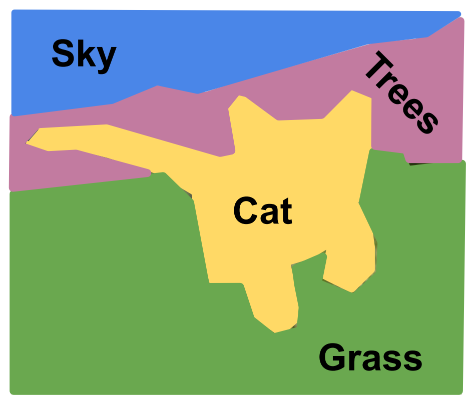
Figure 2: Two examples of our selective search showing the necessity of different scales. On the left we find many objects at different scales. On the right we necessarily find the objects at different scales as the girl is contained by the tv.

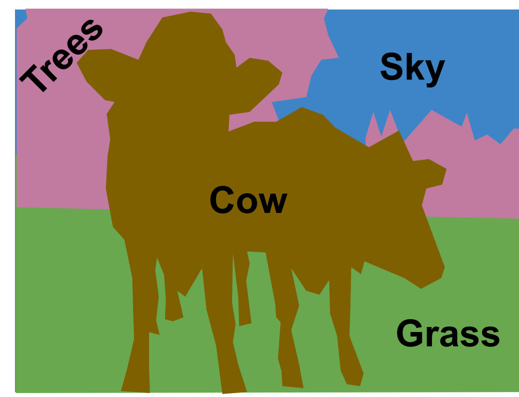
https://www.koen.me/research/selectivesearch/

Semantic segmentation

Label each pixel in the image with a category label

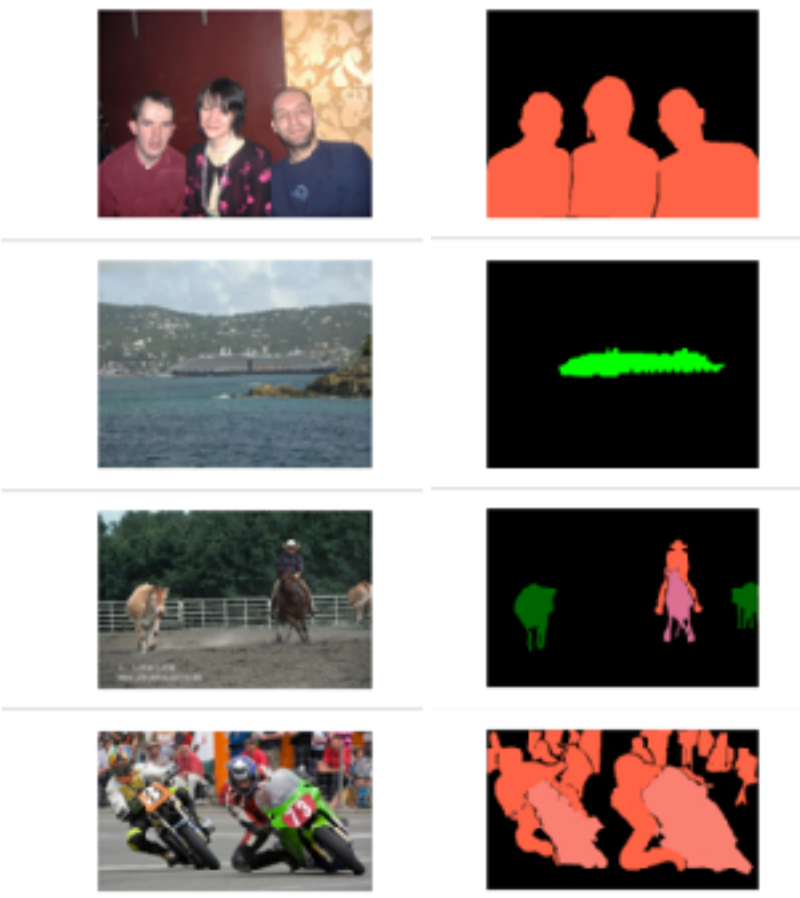
Don't differentiate instances, only care about pixels





Semantic segmentation

PASCAL VOC (20 objects)



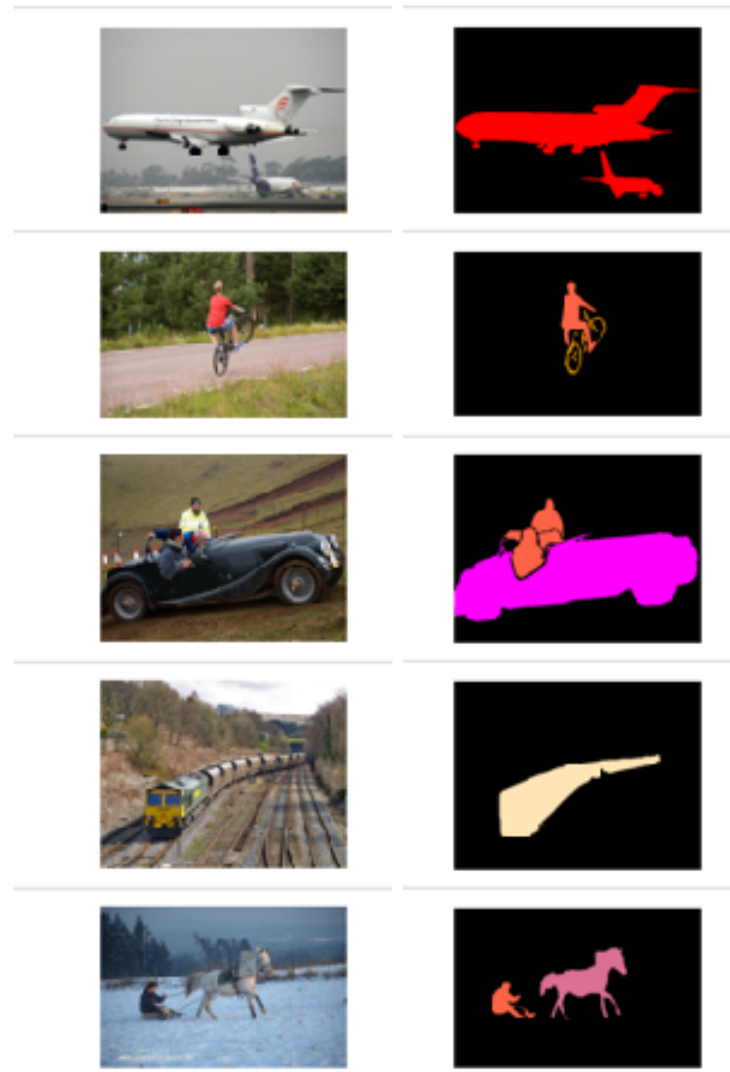
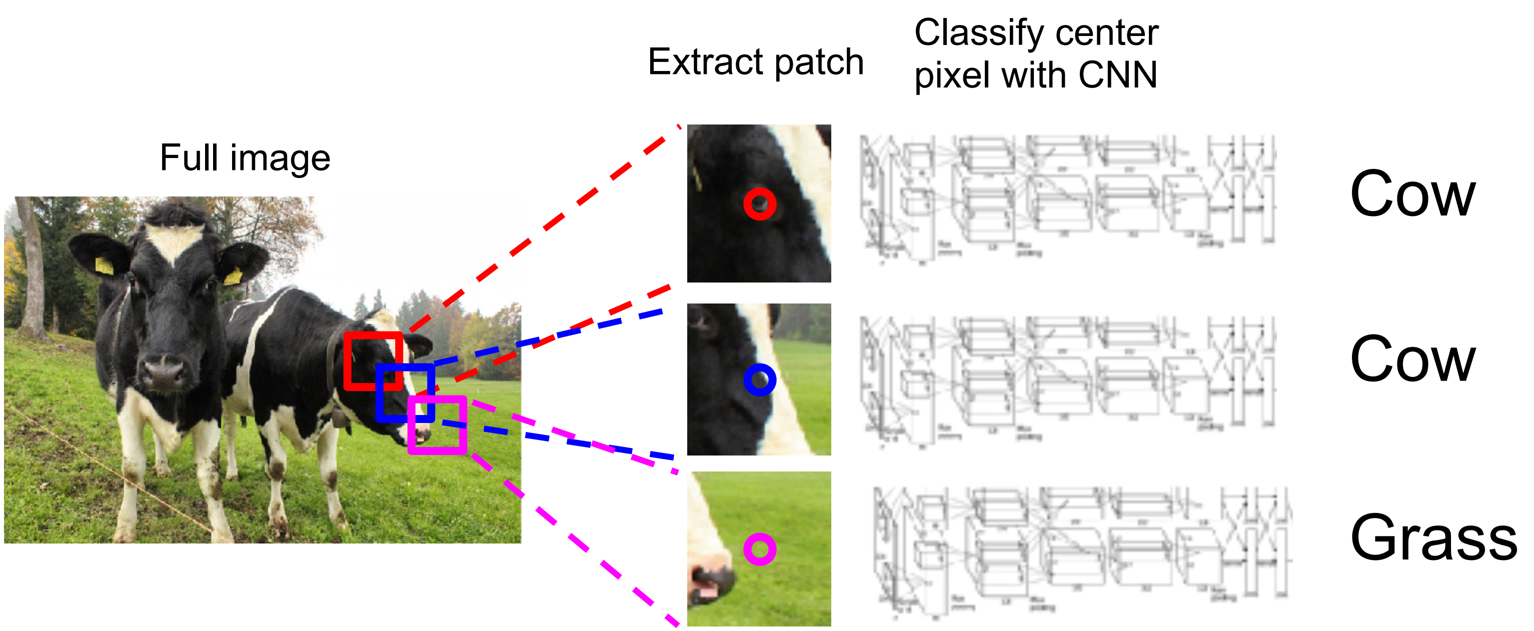


Figure from http://vision.stanford.edu/whats_the_point/

Semantic segmentation idea: sliding window

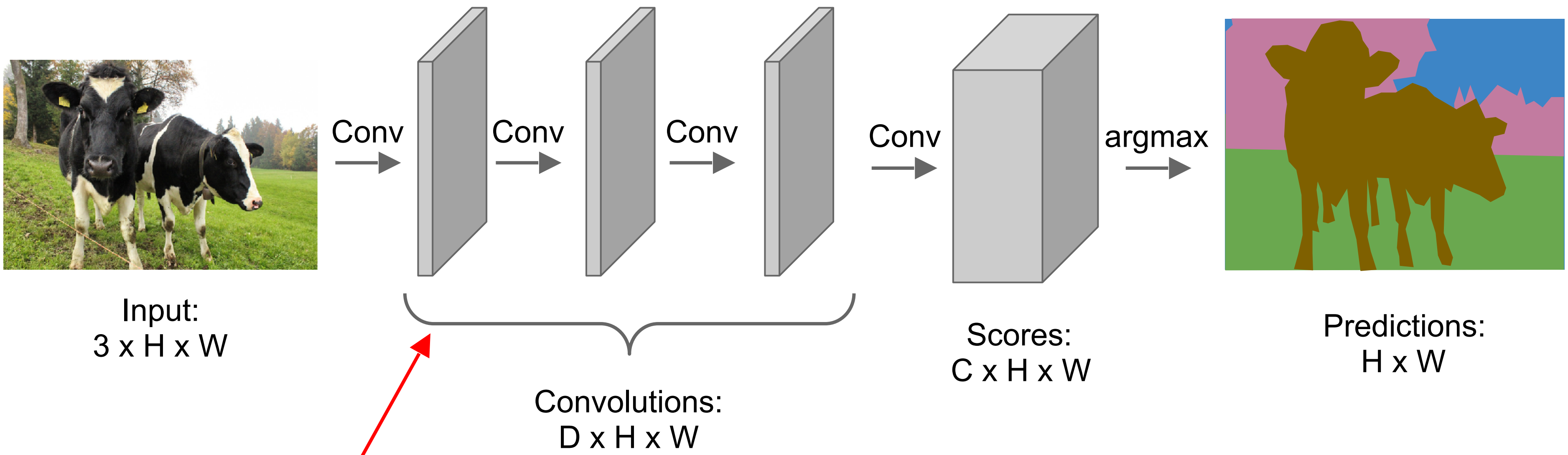


Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic segmentation idea: fully convolutional

Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



Problem: convolutions at
original image resolution will
be very expensive ...

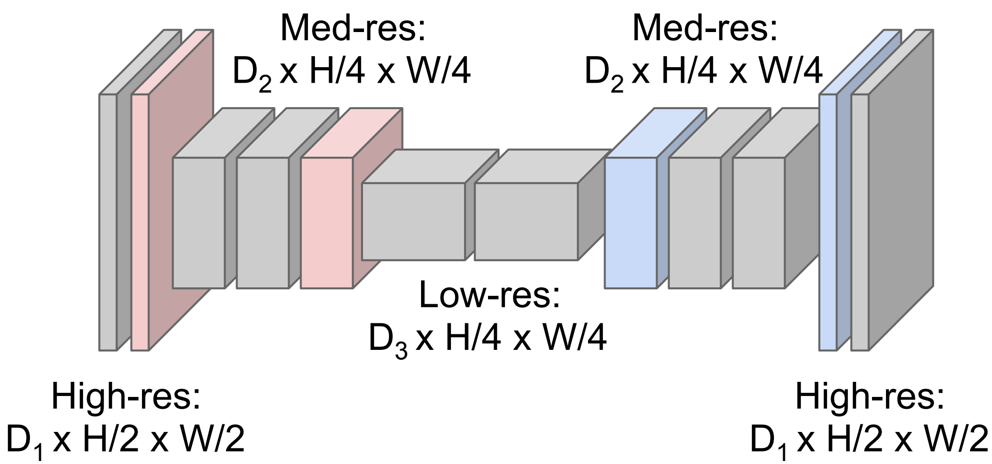
Semantic segmentation idea: fully convolutional

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???

Input:
3 x H x W



Predictions:
H x W

Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015
Noh et al, “Learning Deconvolution Network for Semantic Segmentation”, ICCV 2015

In-Network upsampling: "Unpooling"

Nearest Neighbor

1	2
3	4

Input: 2 x 2

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Output: 4 x 4

"Bed of Nails"

1	2
3	4

Input: 2 x 2

1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

Output: 4 x 4

In-Network upsampling: "Max Unpooling"

Max Pooling
Remember which element was max!

1	2	6	3
3	5	2	1
1	2	2	1
7	3	4	8

Input: 4 x 4

5	6
7	8

Output: 2 x 2

→ … →

Rest of the network

Max Unpooling
Use positions from pooling layer

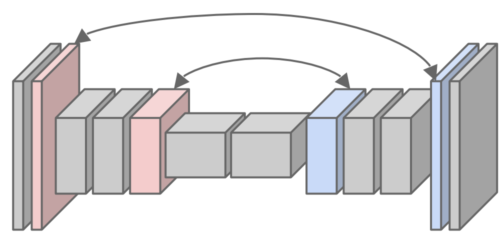
1	2
3	4

Input: 2 x 2

0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

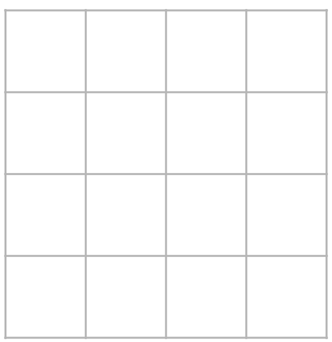
Output: 4 x 4

Corresponding pairs of downsampling and upsampling layers

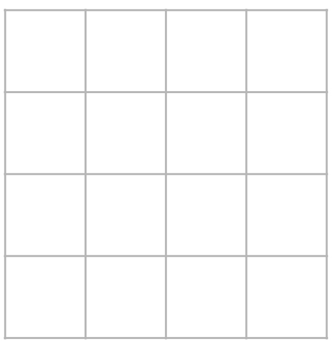


Learnable Upsampling: Transpose Convolution

Recall: Typical 3 x 3 convolution, stride 1 pad 1



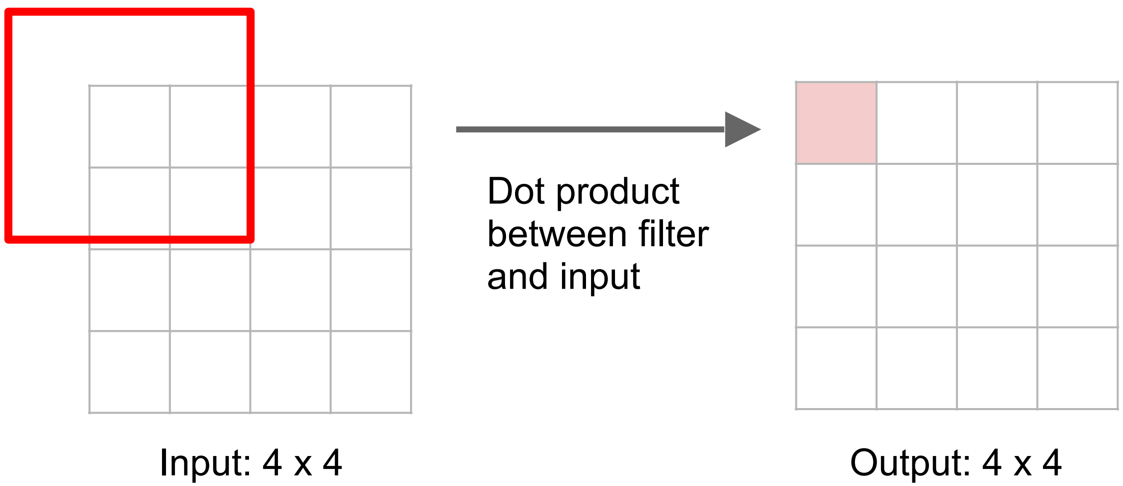
Input: 4 x 4



Output: 4 x 4

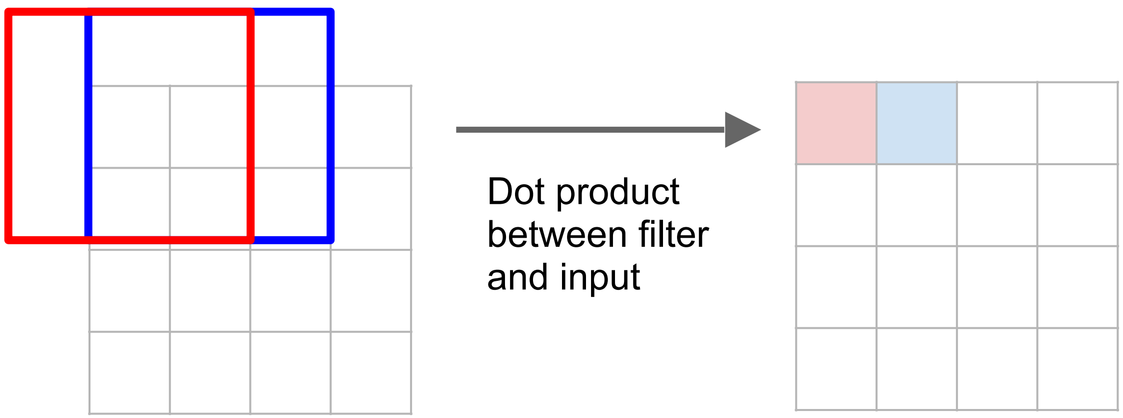
Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 1 pad 1



Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 1 pad 1

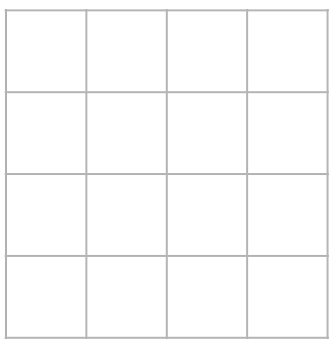


Input: 4 x 4

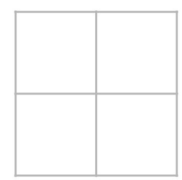
Output: 4 x 4

Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1



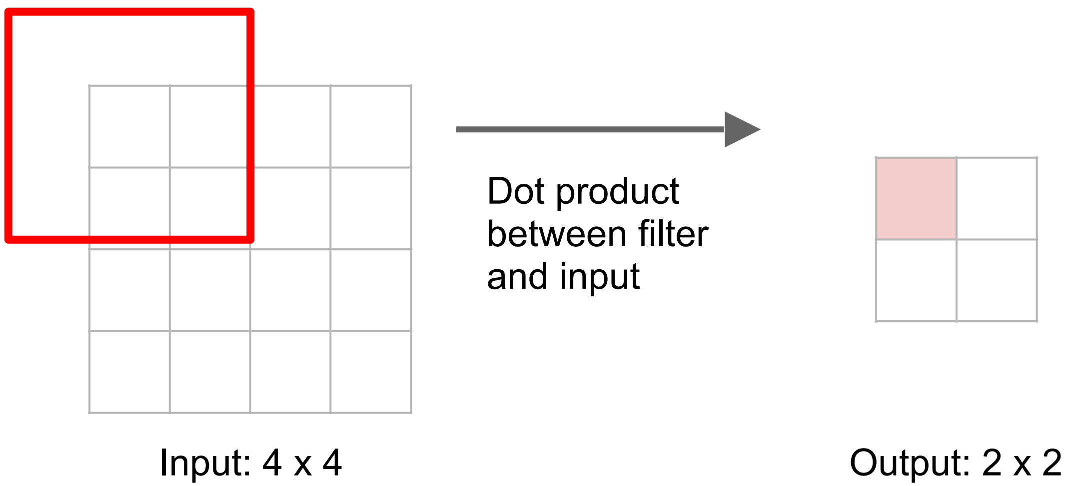
Input: 4 x 4



Output: 2 x 2

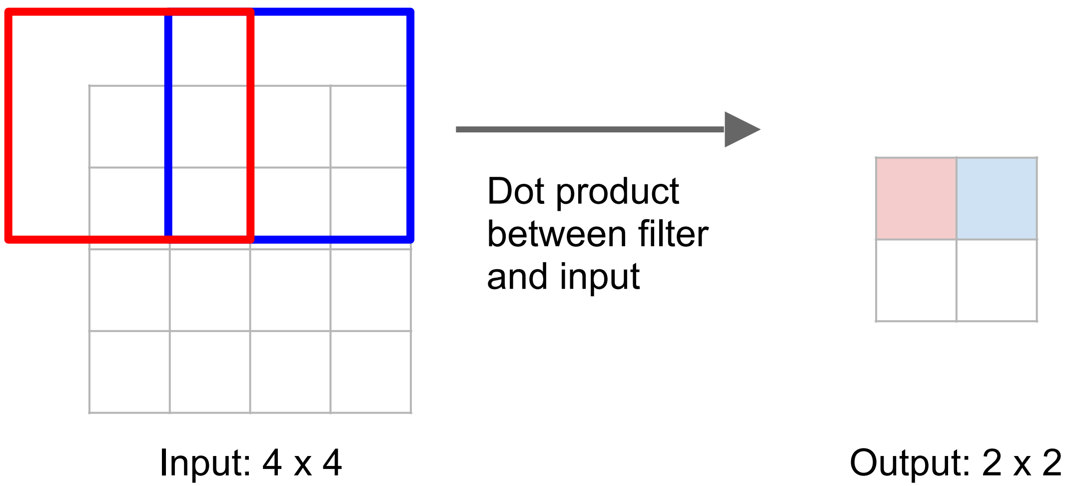
Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1



Learnable Upsampling: Transpose Convolution

Recall: Normal 3 x 3 convolution, stride 2 pad 1

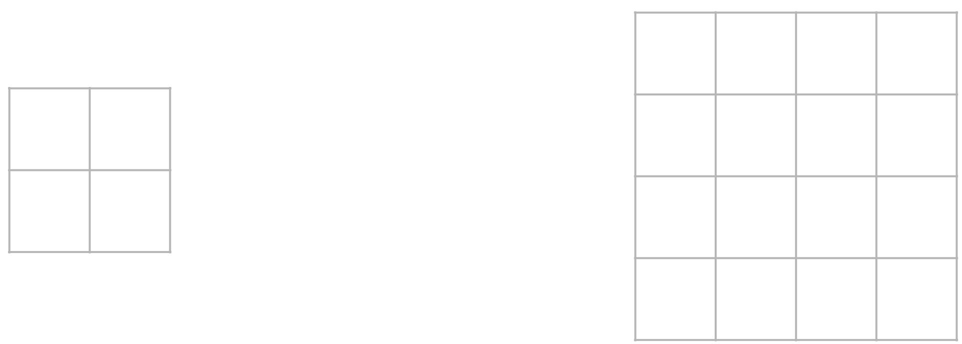


Filter moves 2 pixels in the input for every one pixel in the output

Stride gives ratio between movement in input and output

Learnable Upsampling: Transpose Convolution

3 x 3 **transpose** convolution, stride 2 pad 1

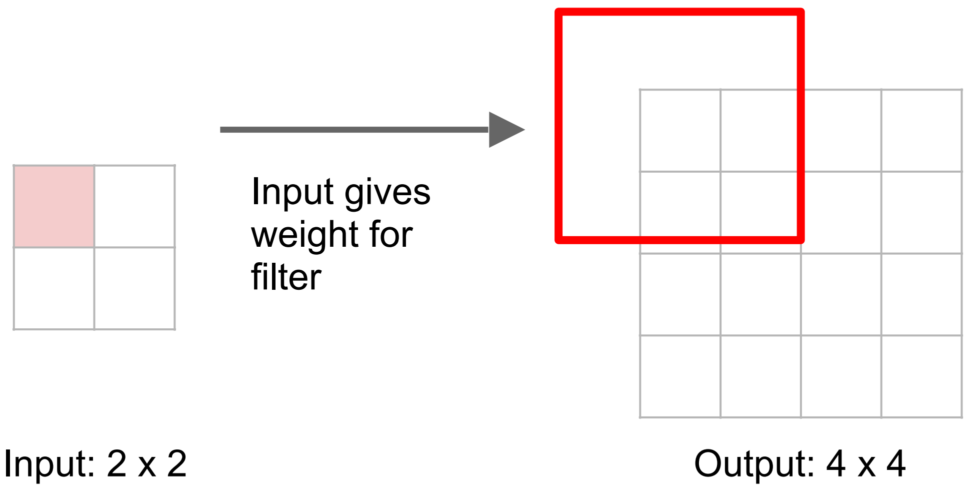


Input: 2 x 2

Output: 4 x 4

Learnable Upsampling: Transpose Convolution

3 x 3 **transpose** convolution, stride 2 pad 1



Learnable Upsampling: Transpose Convolution

3 x 3 **transpose** convolution, stride 2 pad 1

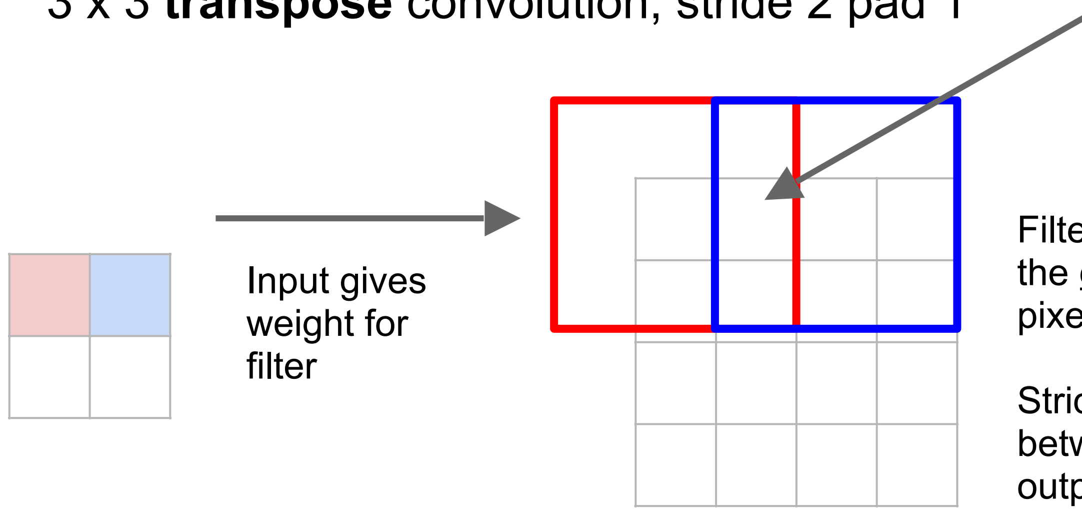
Other names:

-Deconvolution

-Upconvolution

-Fractionally strided convolution

-Backward strided convolution



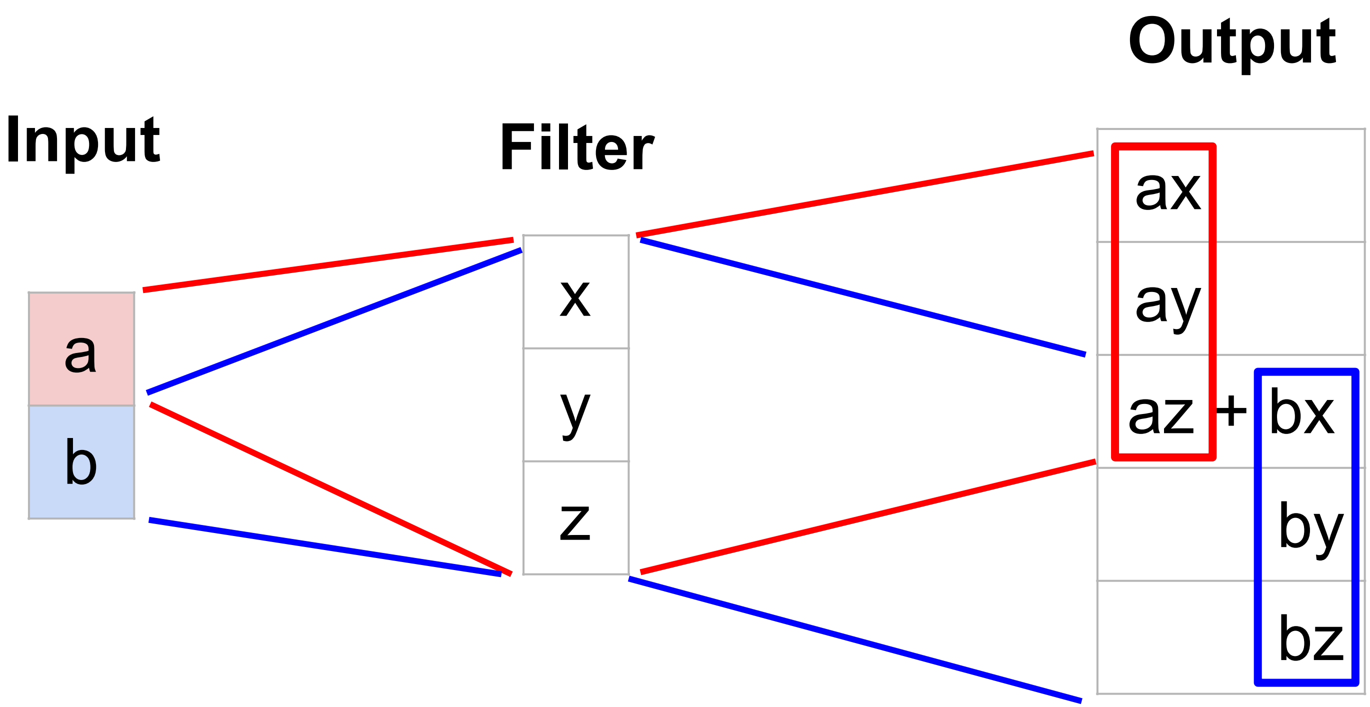
Input: 2 x 2

Output: 4 x 4

Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

Transpose Convolution: 1D example



Output contains copies of the filter weighted by the input, summing at where at overlaps in the output

Need to crop one pixel from output to make output exactly 2x input

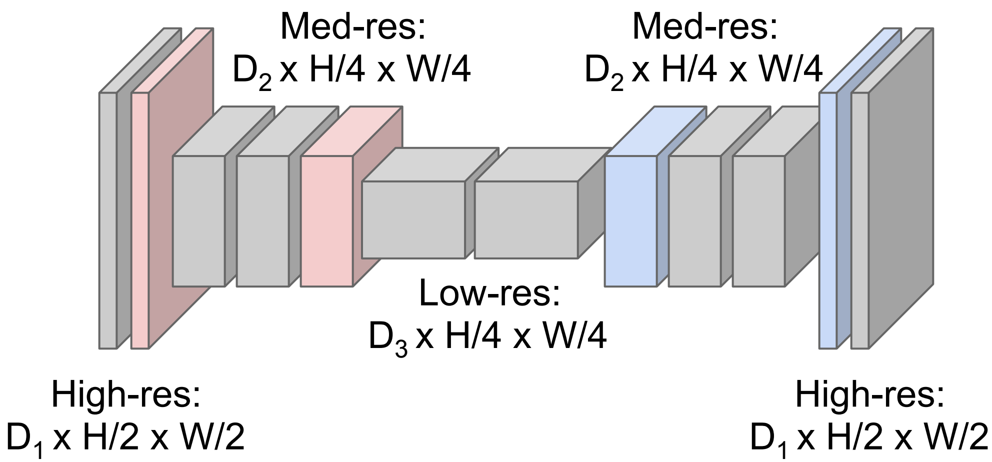
Semantic segmentation idea: fully convolutional

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: unpooling or strided transpose convolution

Input:
3 x H x W

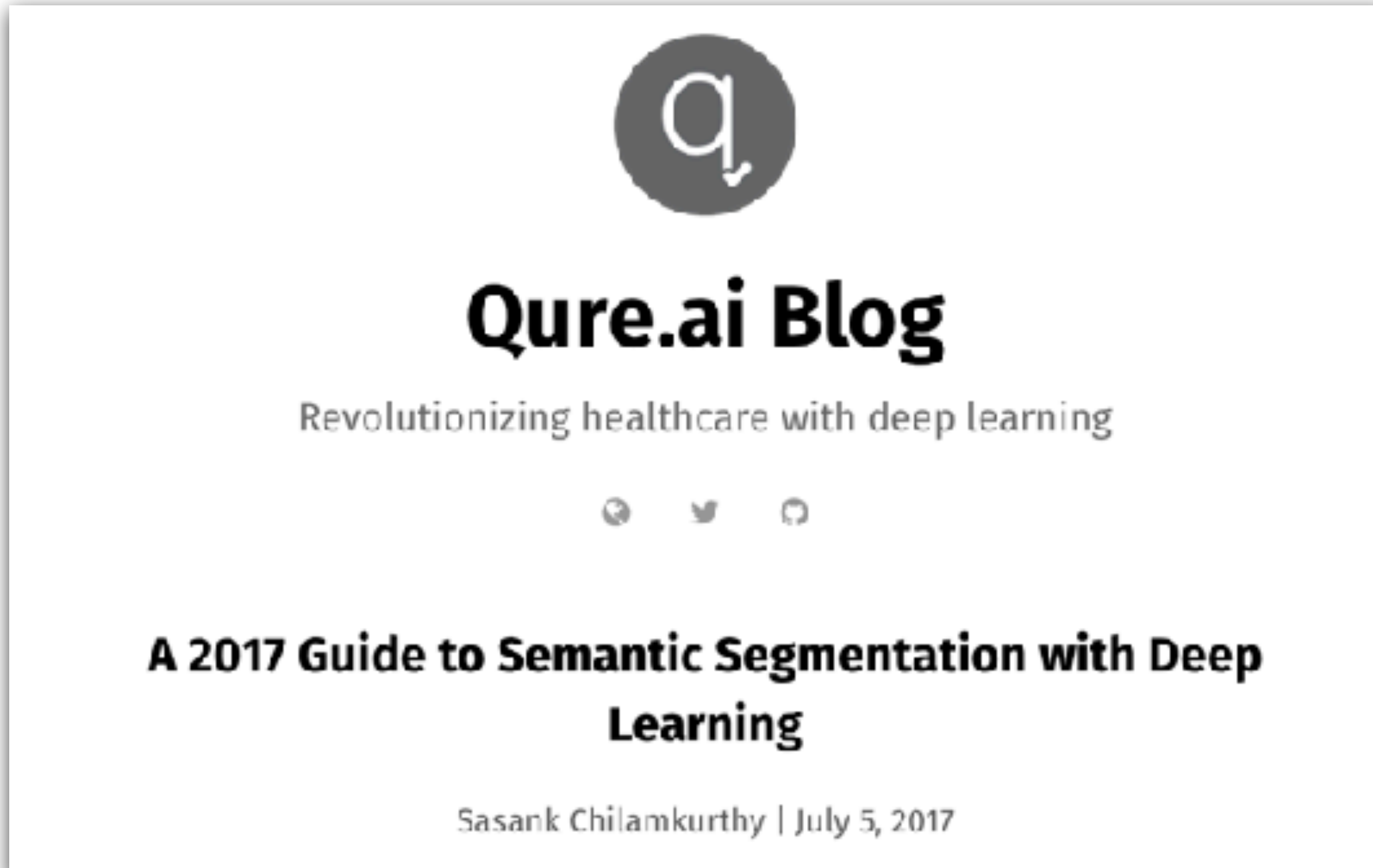


Predictions:
H x W

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Semantic segmentation literature

http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review


Qure.ai Blog

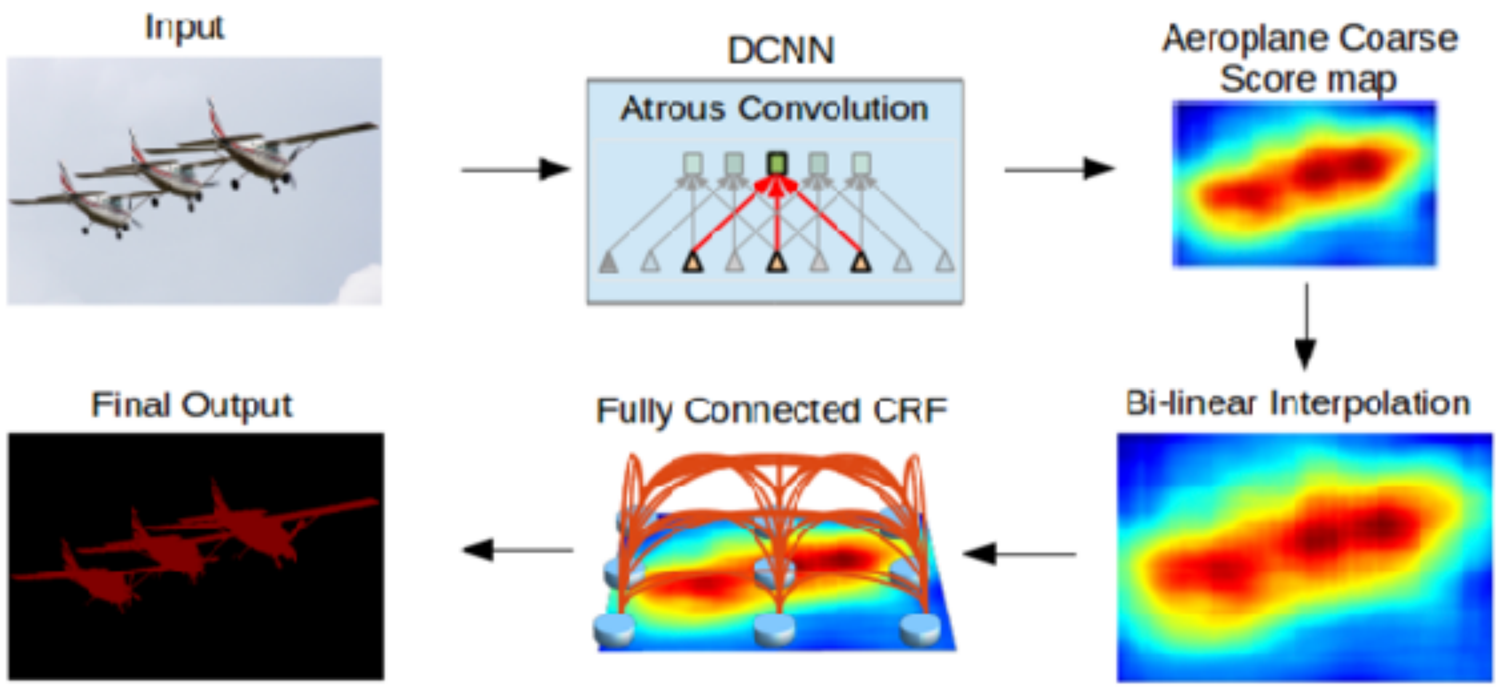
Revolutionizing healthcare with deep learning

A 2017 Guide to Semantic Segmentation with Deep Learning

Sasank Chilamkurthy | July 5, 2017

Semantic segmentation literature

http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review



DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs
*Liang-Chieh Chen**, George Papandreou*, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille

Cute aside

Photographic Image Synthesis with Cascaded Refinement Networks

Qifeng Chen and *Vladlen Koltun*

International Conference on Computer Vision (ICCV) 2017 (Selected for full oral presentation)

Abstract

We present an approach to synthesizing photographic images conditioned on semantic layouts. Given a semantic label map, our approach produces an image with photographic appearance that conforms to the input layout. The approach thus functions as a rendering engine that takes a two-dimensional semantic specification of the scene and produces a corresponding photographic image. Unlike recent and contemporaneous work, our approach does not rely on adversarial training. We show that photographic images can be synthesized from semantic layouts by a single feedforward network with appropriate structure, trained end-to-end with a direct regression objective. The presented approach scales seamlessly to high resolutions; we demonstrate this by synthesizing photographic images at 2-megapixel resolution, the full resolution of our training data. Extensive perceptual experiments on datasets of outdoor and indoor scenes demonstrate that images synthesized by the presented approach are considerably more realistic than alternative approaches.

http://vladlen.info/publications/photographic-image-synthesis-cascaded-refinement-networks/

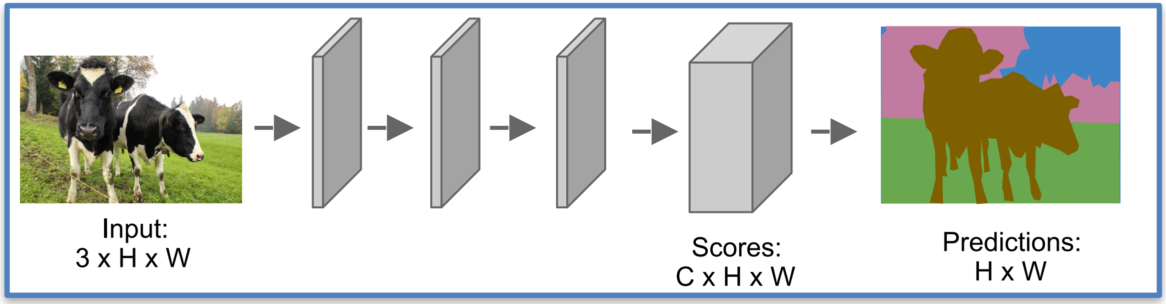
Instance segmentation

Instance segmentation task

- Masks for each individual object instance
- Sometimes called “object detection” now
- Consider two approaches:
  - Start from a semantic segmentation model
  - Start from an object detection model

MSCOCO

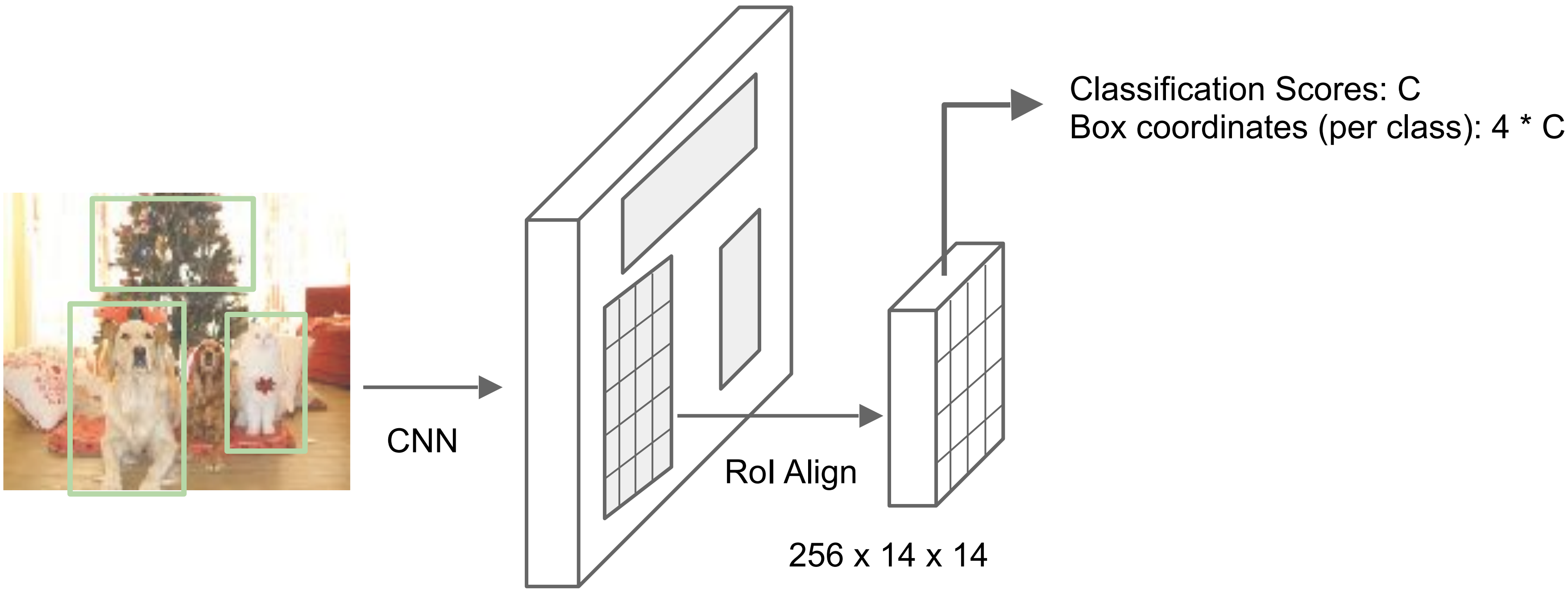
Attempt #1: Starting from semantic segmentation



Issue: don't know the number of instances

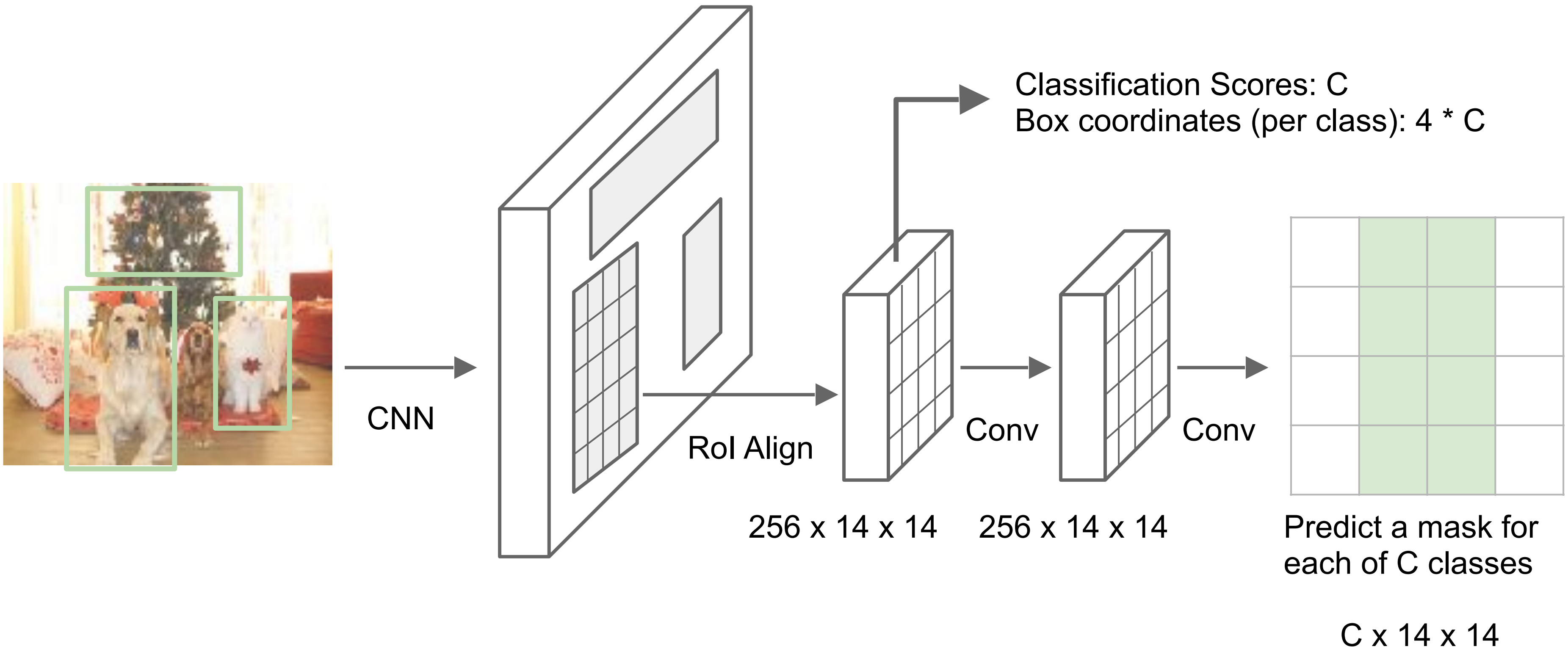
(we'll come back to this)

Starting from detection model: Faster RCNN



Ren et al, “Faster R-CNN”, NIPS 2015

Mask R-CNN



He et al, "Mask R-CNN", ICCV 2017

Mask R-CNN

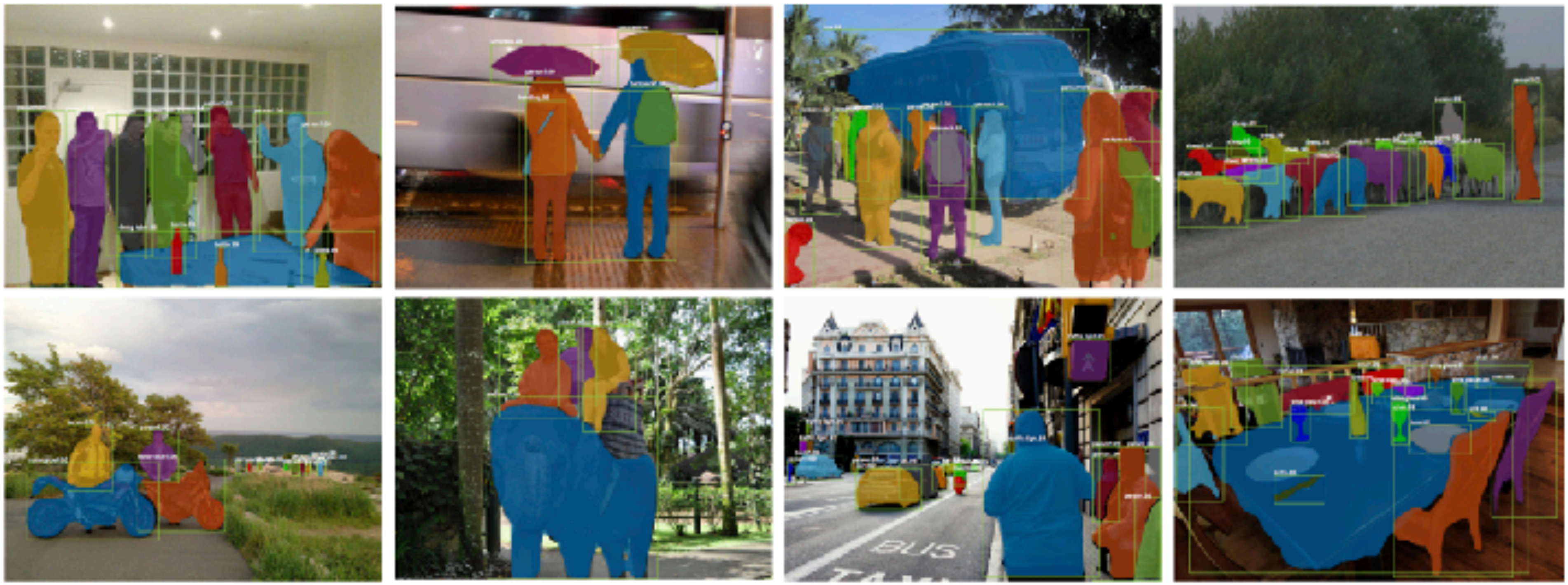


Figure 2. **Mask R-CNN** results on the COCO test set. These results are based on ResNet-101 [19], achieving a *mask* AP of 35.7 and running at 5 fps. Masks are shown in color, and bounding box, category, and confidences are also shown.

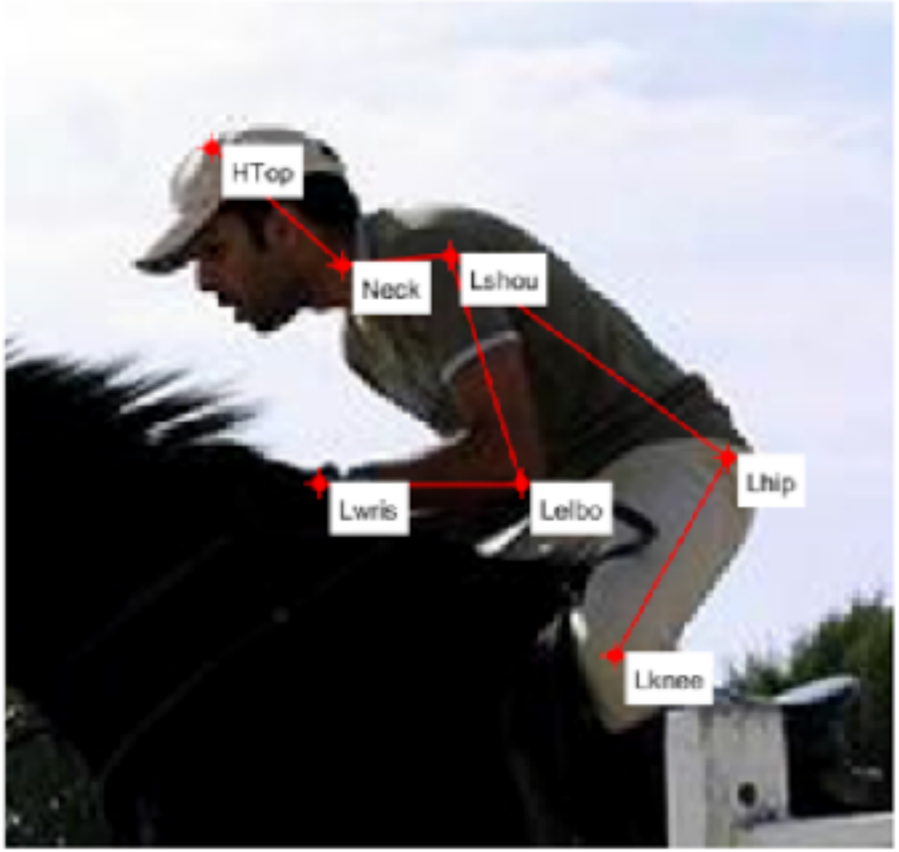
He et al, “Mask R-CNN”, ICCV 2017

Mask R-CNN also does pose…

He et al, “Mask R-CNN”, ICCV 2017

Human pose estimation

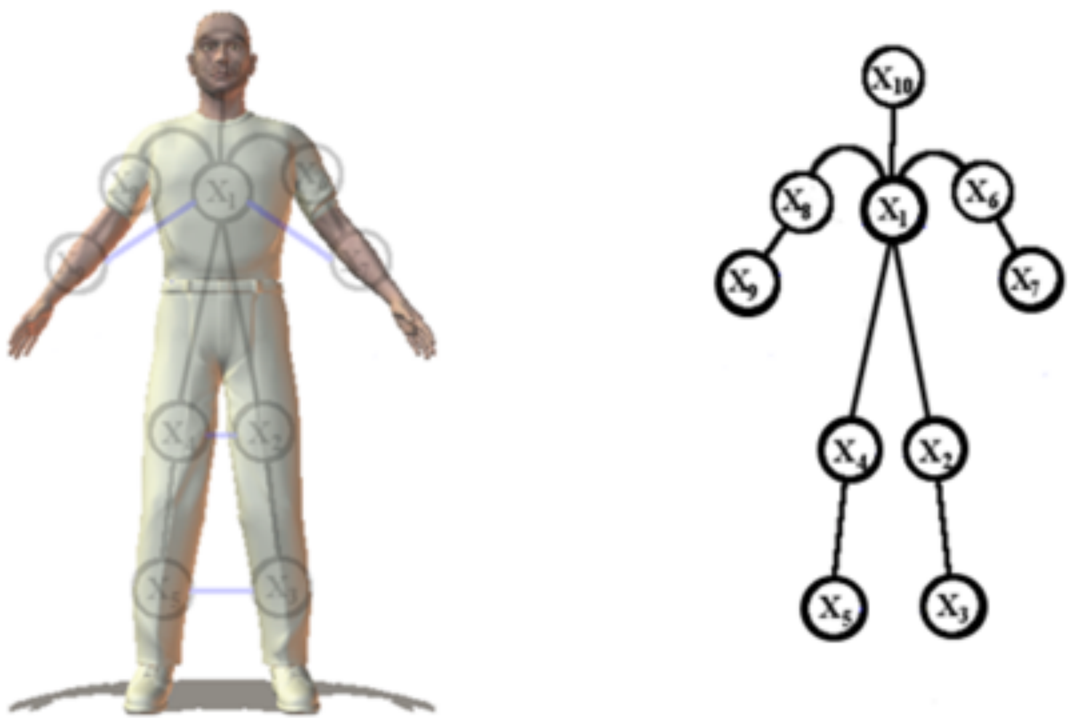
Human pose estimation task



Fangting Xia, Peng Wang, Xianjie Chen, Alan Yuille, Joint Multi-Person Pose Estimation and Semantic Part Segmentation in a Single Image. In *CVPR*, 2017

https://sites.google.com/view/pasd/dataset?authuser=0

Pictorial structures model



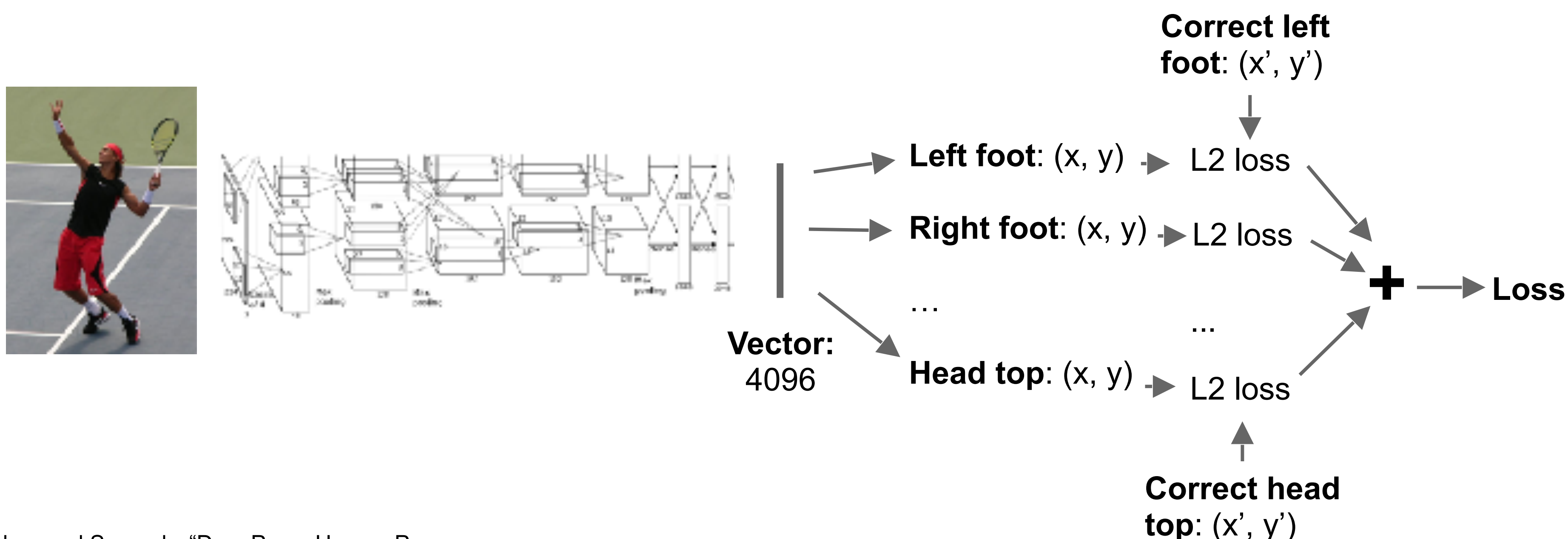
P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. IJCV 2005.

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. CVPR 2009

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixture-of-parts. CVPR 2011.

Figure from Sigal et al. “Human pose estimation” https://cs.brown.edu/~ls/Publications/SigalEncyclopediaCVdraft.pdf

Regression-based model



Toshev and Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks", CVPR 2014

Model based on keypoint heatmaps

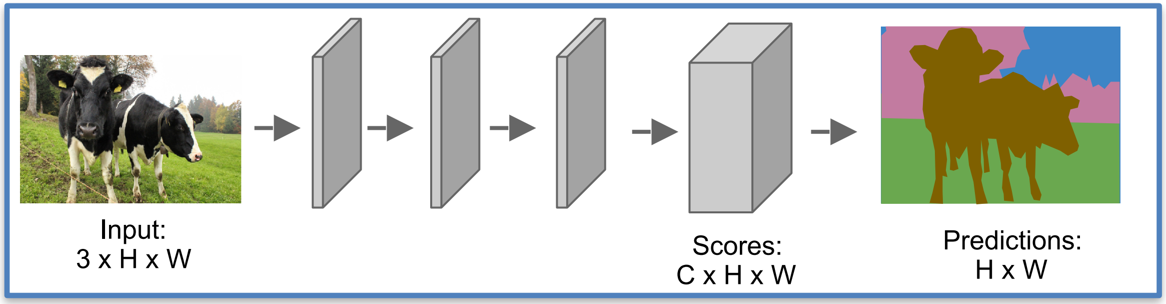
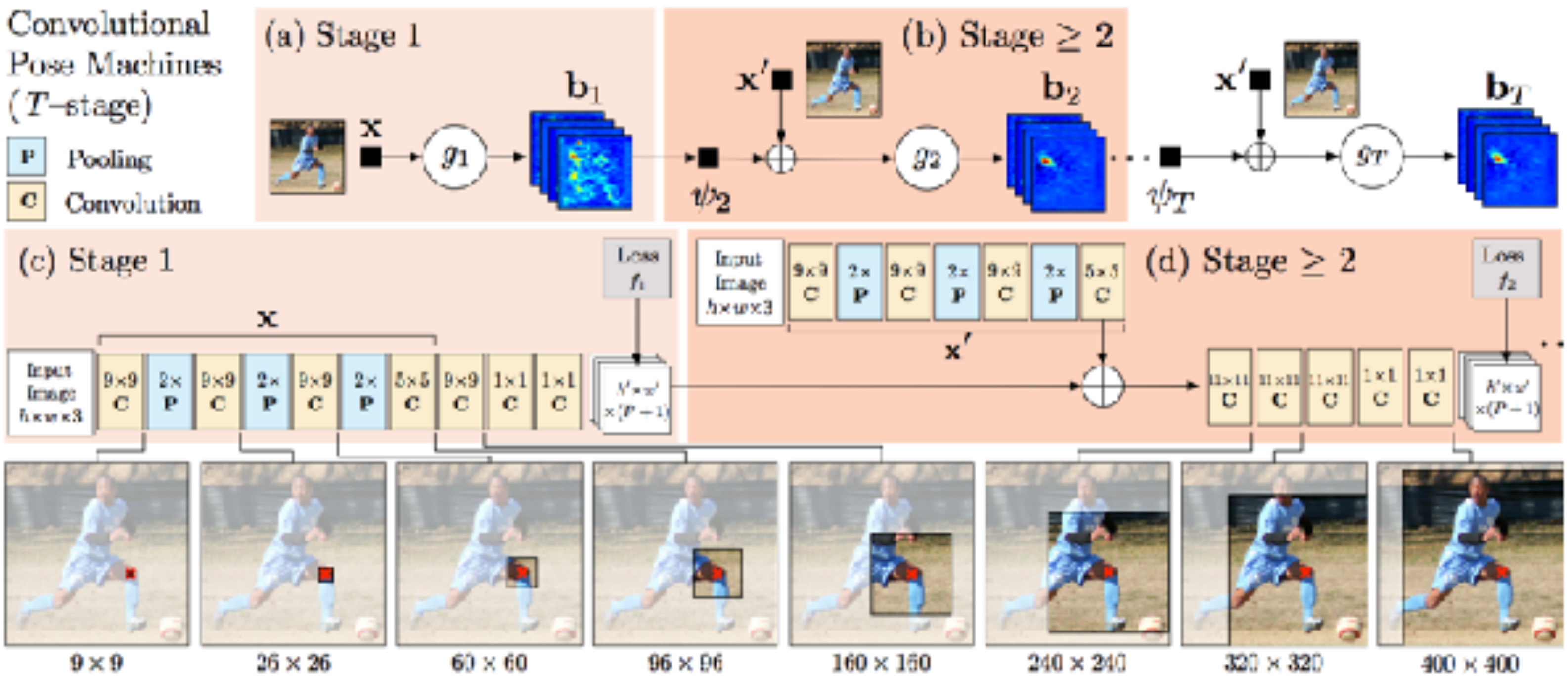


Figure from Newell et al. Stacked Hourglass Networks for Human Pose Estimation. ECCV 2016

Model based on keypoint heatmaps



(e) Effective Receptive Field

Wei et al. “Convolutional Pose Machines” CVPR 2016

cf also

Carriera et al. “Human Pose Estimation with Iterative Error Feedback” CVPR 2016

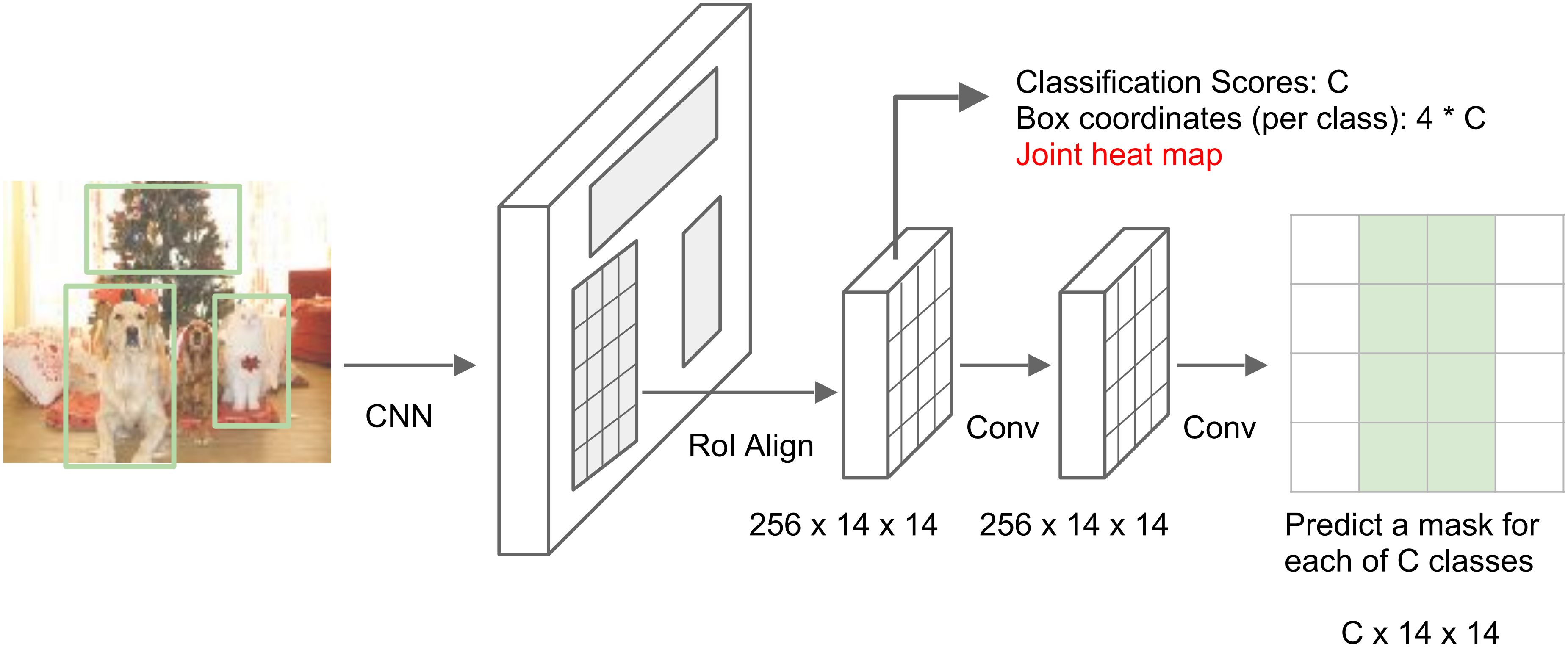
Newell et al. Stacked Hourglass Networks for Human Pose Estimation. ECCV 2016

Xia et al. “Joint Multi-Person Pose Estimation and Semantic Part Segmentation” CVPR 2017

Cao et al. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” CVPR 2017

etc.

Mask R-CNN pose model

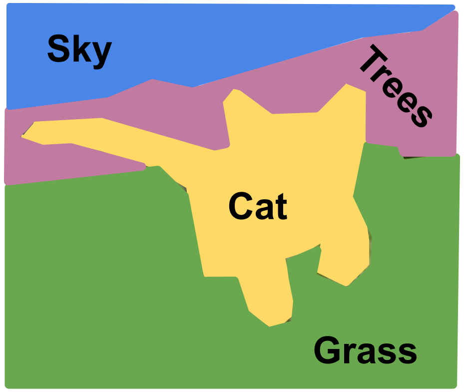


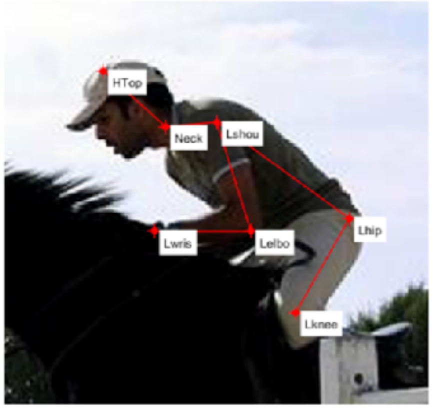
He et al, "Mask R-CNN", ICCV 2017

Mask R-CNN also does pose…

He et al, "Mask R-CNN", ICCV 2017

Bringing it all together





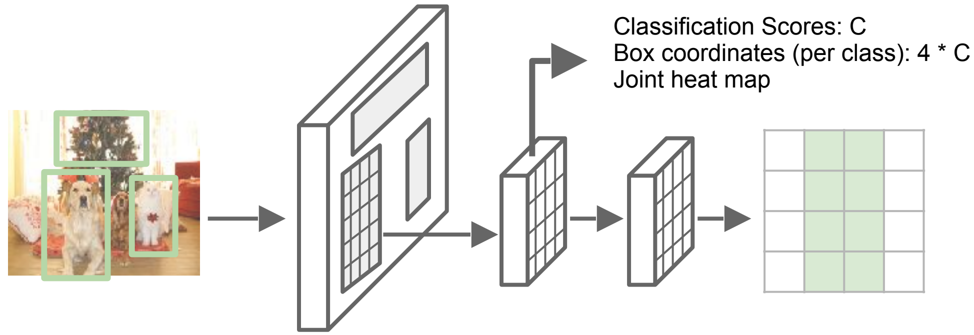
Multi-person pose estimation

Figure 1. Both multi-person pose estimation and instance segmentation are examples of computer vision tasks that require detection of visual elements (joints of the body or pixels belonging to a semantic class) and grouping of these elements (as poses or individual object instances).

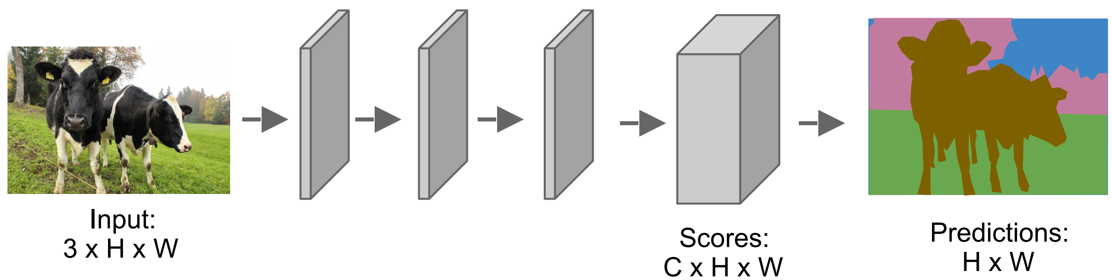
Newell et al. Associative Embedding: End-to-end learning for joint detection and grouping. NIPS 2017

Multi-person pose estimation

Mask-RCNN does this automatically but requires going through region proposals



But what about an image-level heatmap



Multi-person pose estimation

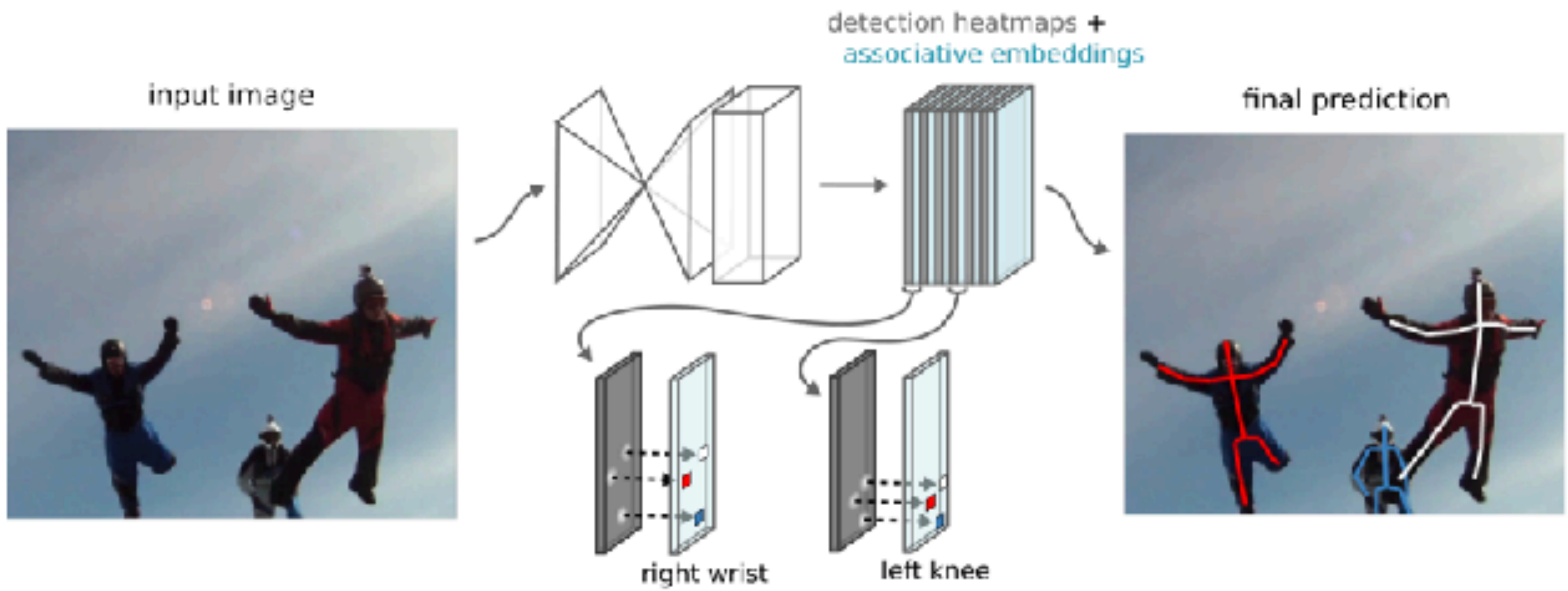


Figure 3. An overview of our approach for producing multi-person pose estimates. For each joint of the body, the network simultaneously produces detection heatmaps and predicts associative embedding tags. We take the top detections for each joint and match them to other detections that share the same embedding tag to produce a final set of individual pose predictions.

Newell et al. Associative Embedding: End-to-end learning for joint detection and grouping. NIPS 2017

Course overview

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.

THE SUMMER VISION PROJECT

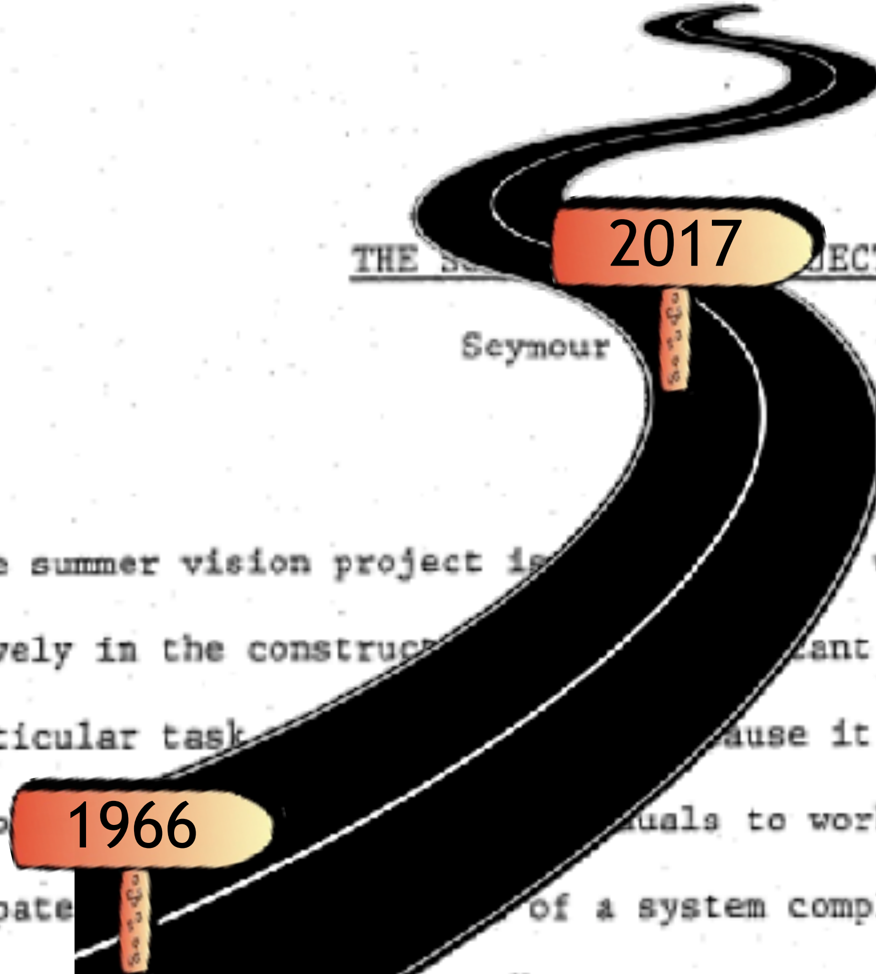
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.



THE S... ...JECT

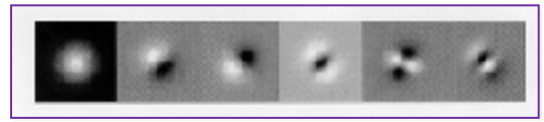
Seymour ...

The summer vision project is ... use our summer workers effectively in the construc... ...ant part of a visual system. The particular task ... ...ause it can be segmented into sub-prob... ...uals to work independently and yet participate ... of a system complex enough to be a real landmark in the development of "pattern recognition".

Course Outline

- Image formation and capture

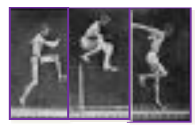
- Filtering and feature detection



- Segmentation and clustering

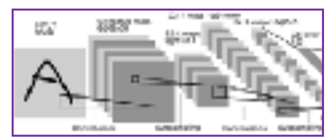
- Recognition and classification

- Motion estimation and tracking

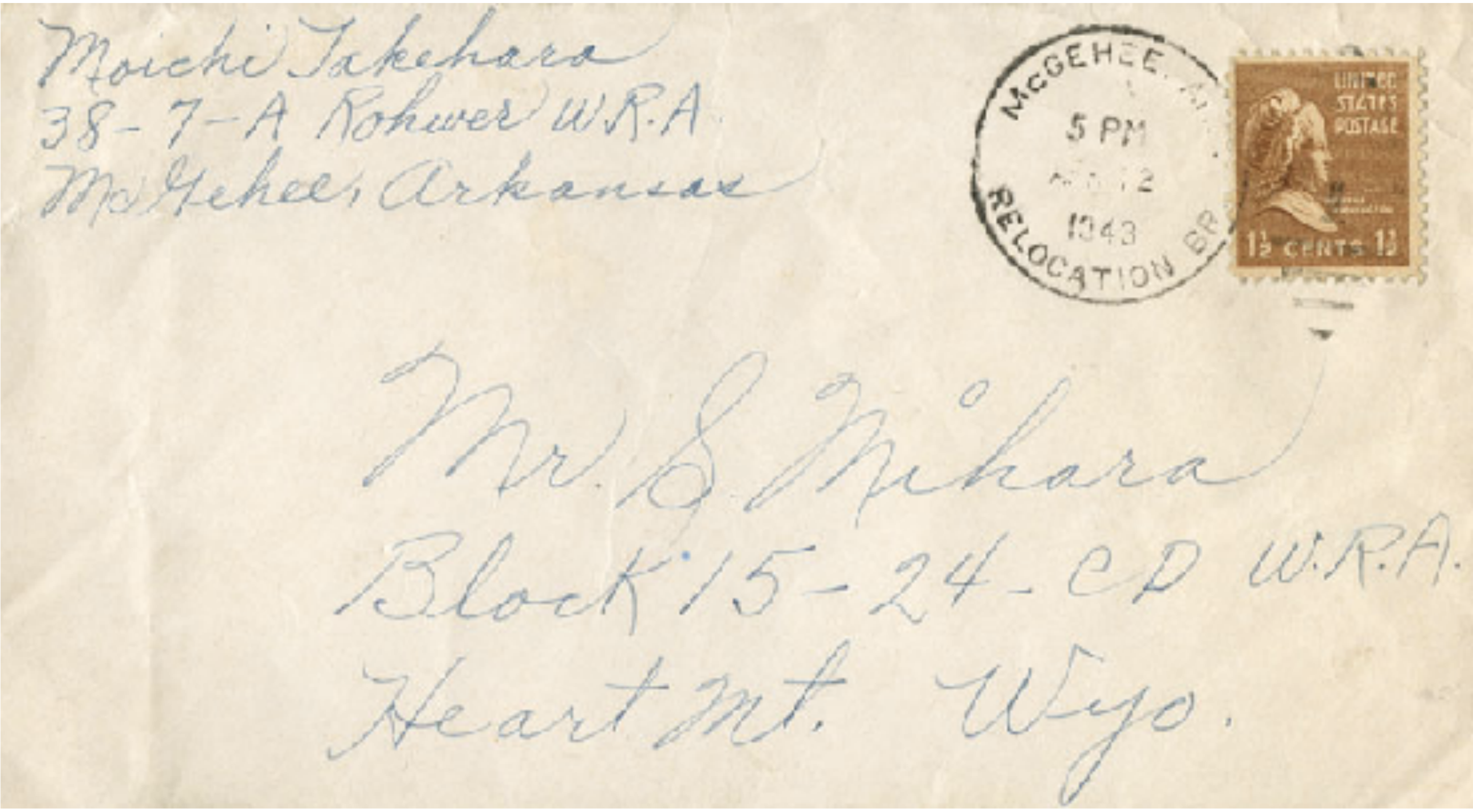


- 3D shape reconstruction

- Convolutional neural nets / deep learning

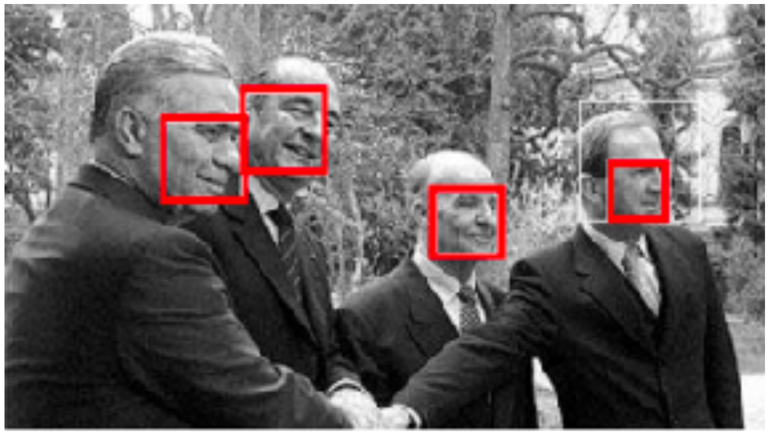


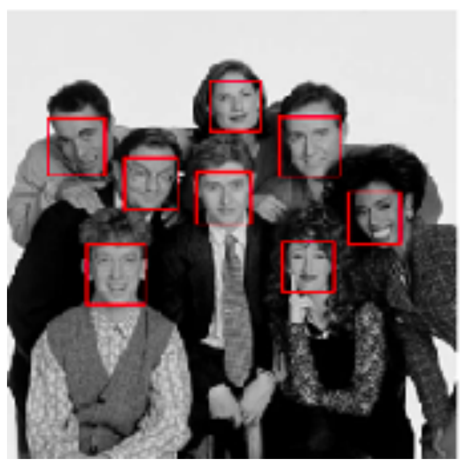
Sorting our mail

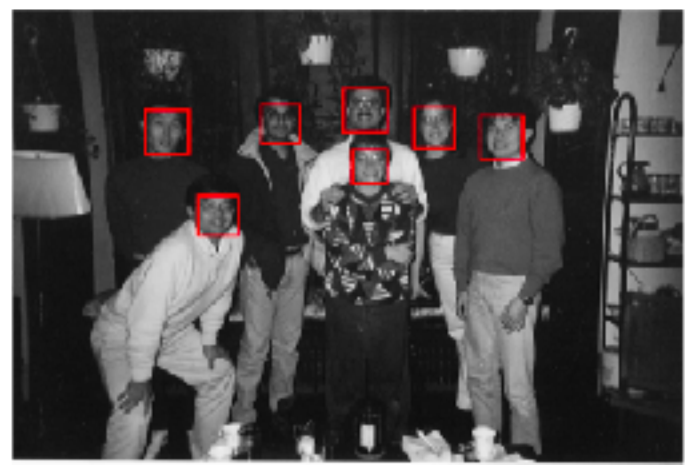


Depositing checks

Detecting (frontal) faces







FinePix S6000fd, by Fujifilm, 2006

Viola & Jones, 2001

3D Maps

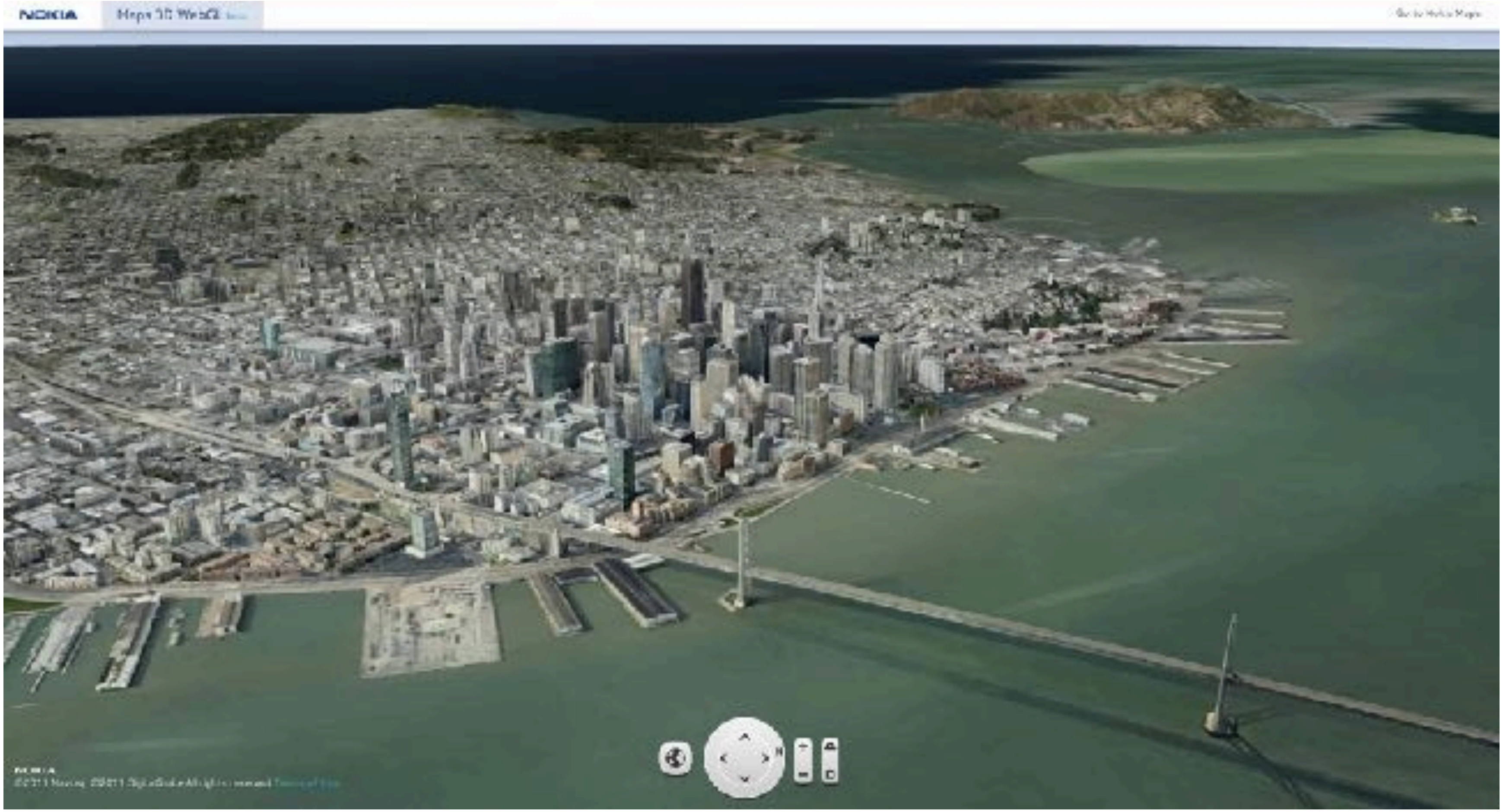


Image from Nokia's Maps 3D WebGL
(see also: Google Maps GL, Google Earth)

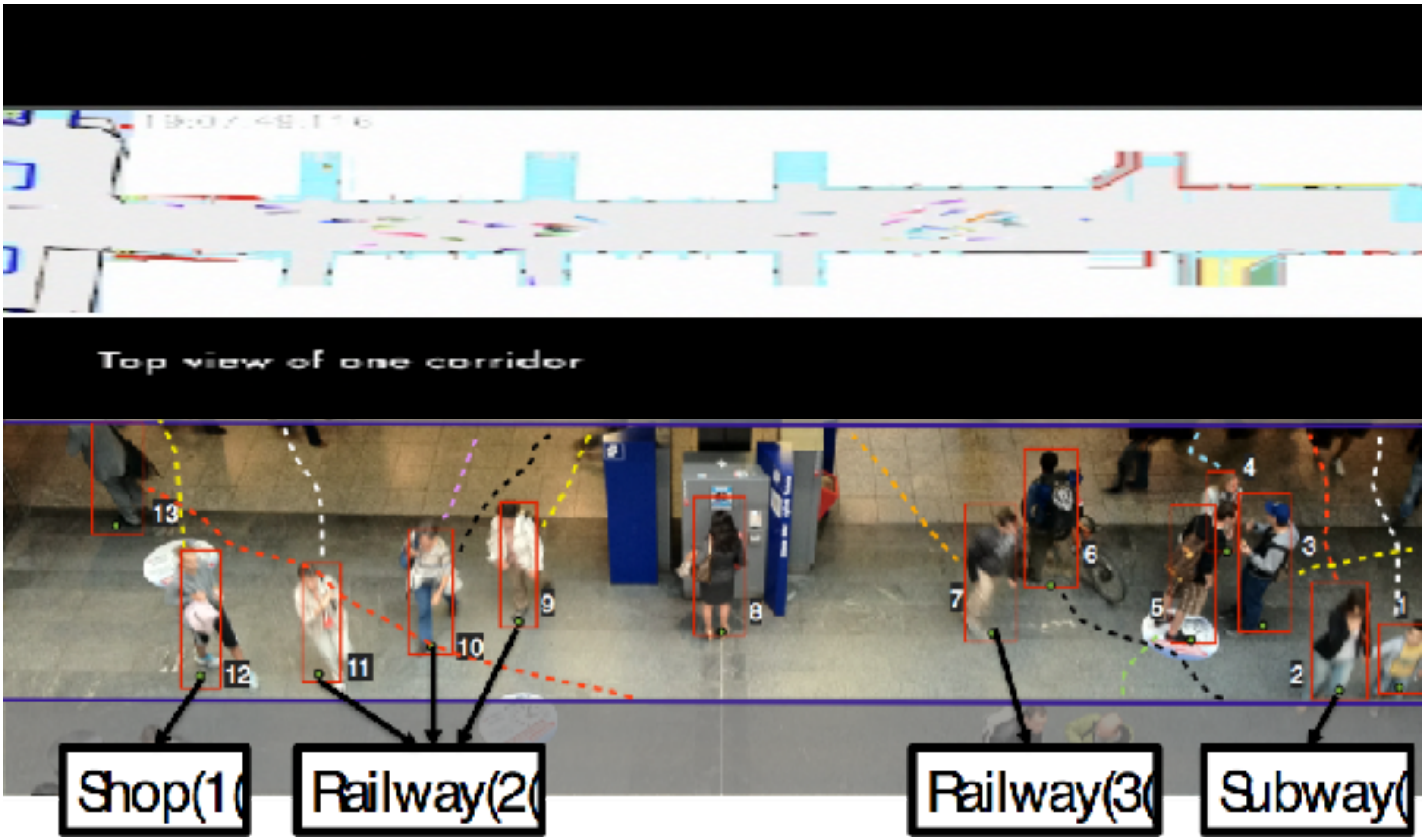
Slide credit: Deva Ramanan

Photo tourism

Reconstructing the 4D world
(UWashington/Microsoft)

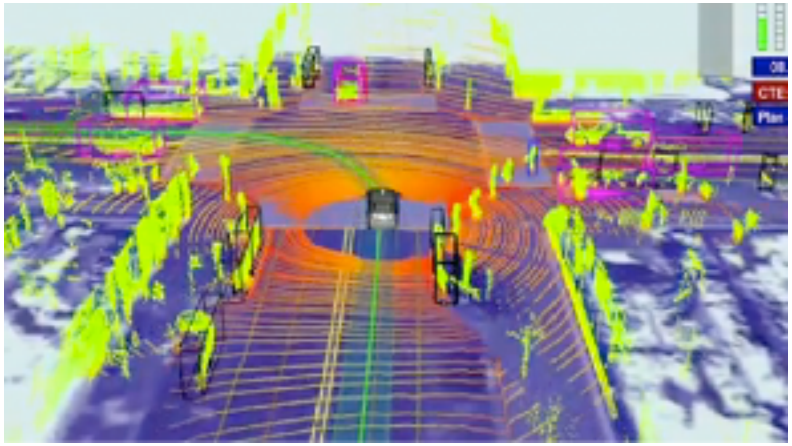
Slide credit: Deva Ramanan

Understanding traffic patterns

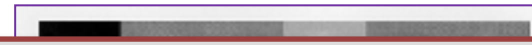
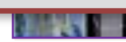


Alahi & Fei-Fei, 2014

Self-Driving Cars



Course Outline

- Image formation and capture 
- Filtering a

- Segmenta
- Recogniti
- Motion es
- 3D shape

- Convolutional neural nets / deep learning 
- *Guest lecture on Thursday: video understanding*
- *Your projects: deep dive into your favorite topic*
- *COS 598B seminar: More advanced deep learning, closer examination of vision data, language + vision (VQA), action recognition in video*

Course evaluations:

- What did you like about the course?
- What were your favorite topics?
- What didn’t work for you?