Detection, Segmentation and Fine-grained Localization

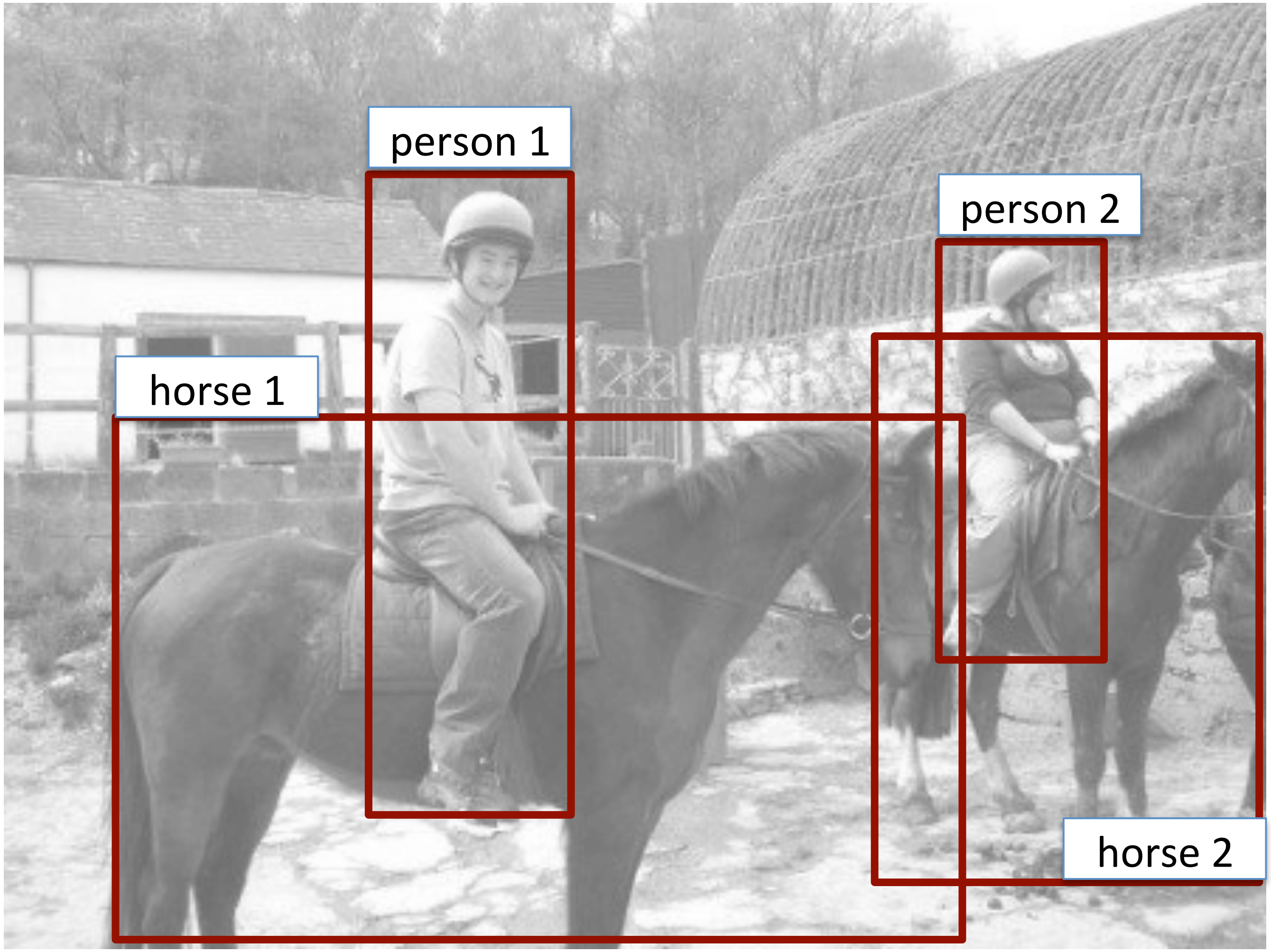
Bharath Hariharan, Pablo Arbeláez, Ross Girshick and Jitendra Malik

UC Berkeley

What is image understanding?

Object Detection

Detect every instance of the category and localize it with a bounding box.



Semantic Segmentation

Label each pixel with a category label

Simultaneous Detection and Segmentation

Detect and segment every instance of the category in the image

Simultaneous Detection, Segmentation and Part Labeling

Detect and segment every instance of the category in the image and label its parts

Goal

A detection system that can describe detected objects in excruciating detail

- Segmentation
- Parts
- Attributes
- 3D models

…

Outline

- Define Simultaneous Detection and Segmentation (SDS) task and benchmark
- SDS by classifying object proposals
- SDS by predicting figure-ground masks
- Part labeling and pose estimation
- Future work and conclusion

Papers

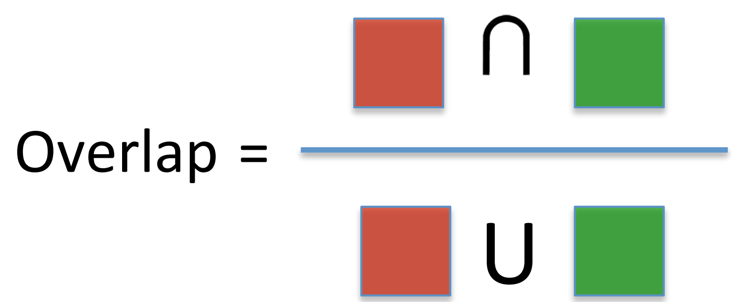
– B. Hariharan, P. Arbeláez, R. Girshick and J. Malik. Simultaneous Detection and Segmentation. ECCV 2014

– B. Hariharan, P. Arbeláez, R. Girshick and J. Malik. Hypercolumns for Object Segmentation and Fine-grained Localization. CVPR 2015

SDS: DEFINING THE TASK AND BENCHMARK

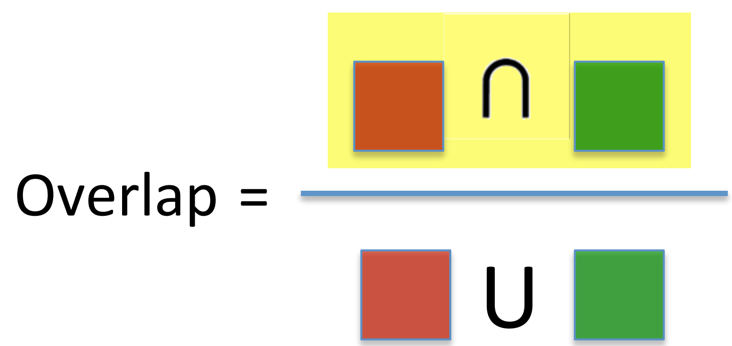
Background:
Evaluating object detectors

- Algorithm outputs ranked list of boxes with category labels
- Compute overlap between detection and ground truth box



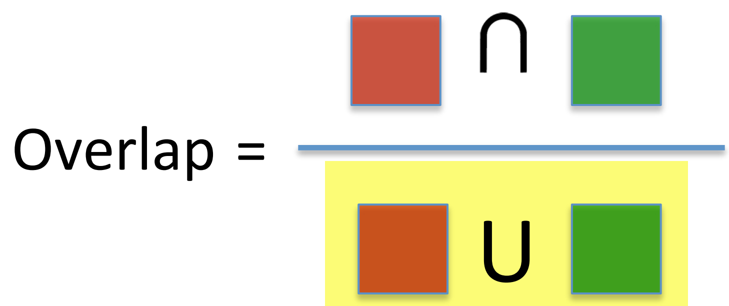
Background: Evaluating object detectors

- Algorithm outputs ranked list of boxes with category labels
- Compute overlap between detection and ground truth box



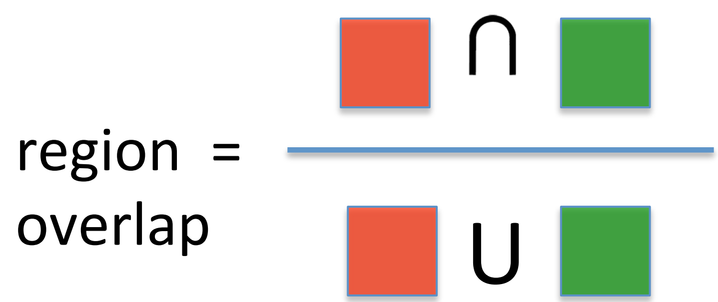
Background:
Evaluating object detectors

- Algorithm outputs ranked list of boxes with category labels
- Compute overlap between detection and ground truth box
- If overlap > thresh, correct
- Compute precision-recall (PR) curve
- Compute area under PR curve : Average Precision (AP)



Evaluating segments

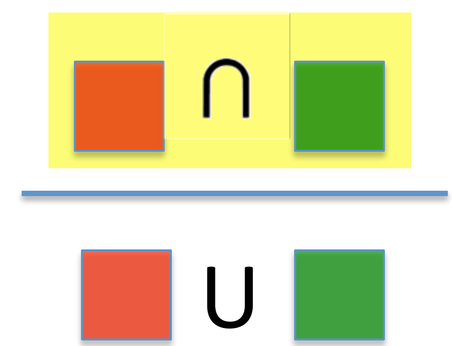
- Algorithm outputs ranked list of segments with category labels
- Compute region overlap of each detection with ground truth instances



Evaluation metric

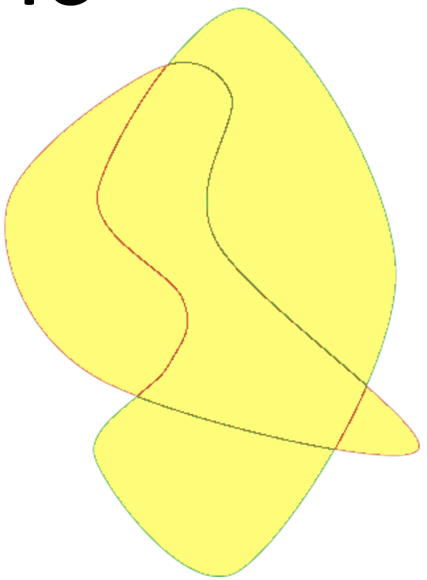
- Algorithm outputs ranked list of segments with category labels
- Compute region overlap of each detection with ground truth instances

region overlap =

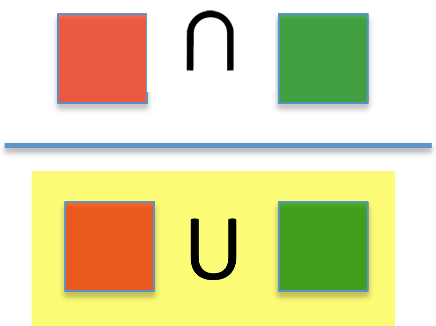


Evaluation metric

- Algorithm outputs ranked list of segments with category labels
- Compute region overlap of each detection with ground truth instances



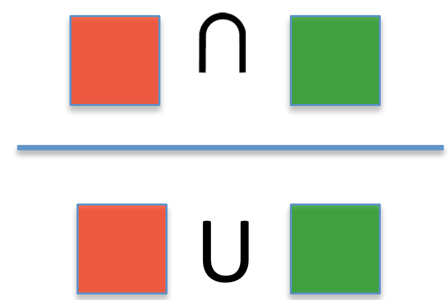
region overlap =



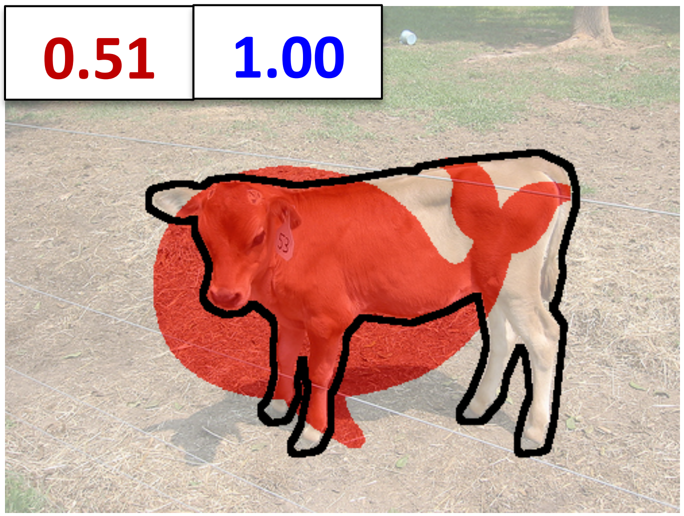
Evaluating segments

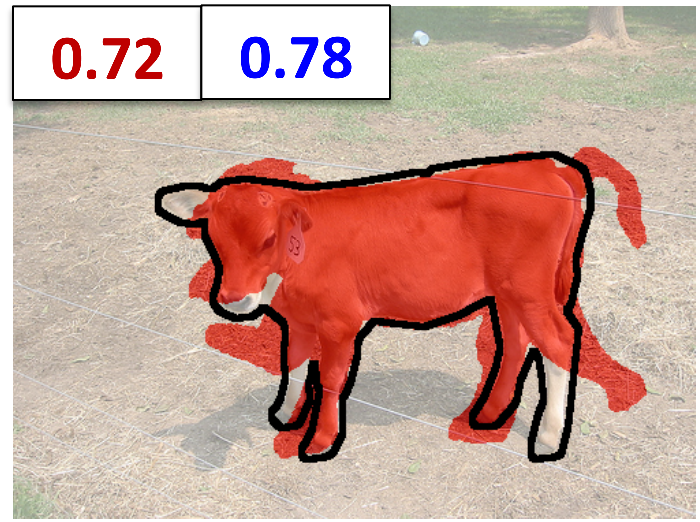
- Algorithm outputs ranked list of segments with category labels
- Compute region overlap of each detection with ground truth instances
- If overlap > thresh, correct
- Compute precision-recall (PR) curve
- Compute area under PR curve : Average Precision (AP^r)

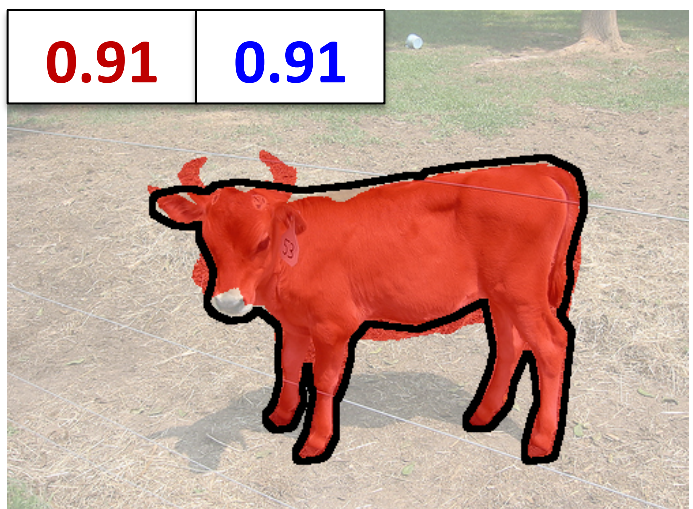
region overlap =



Region overlap vs Box overlap





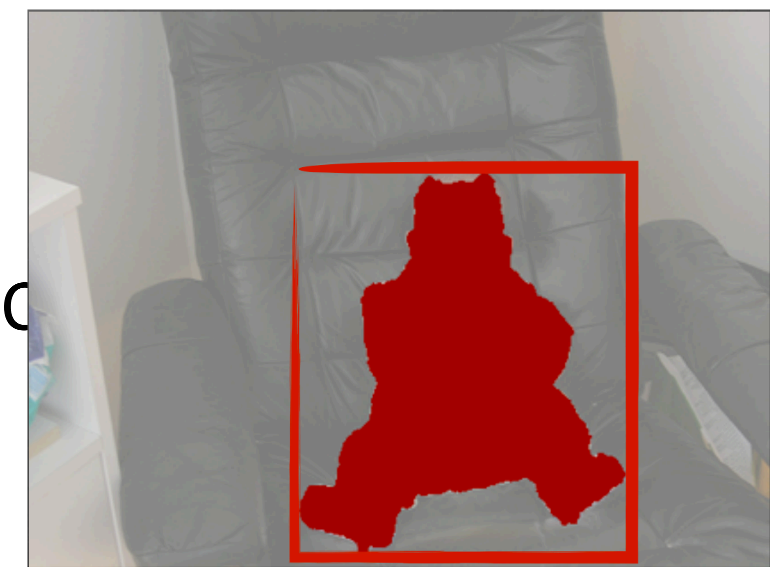


Slide adapted from Philipp Krähenbühl

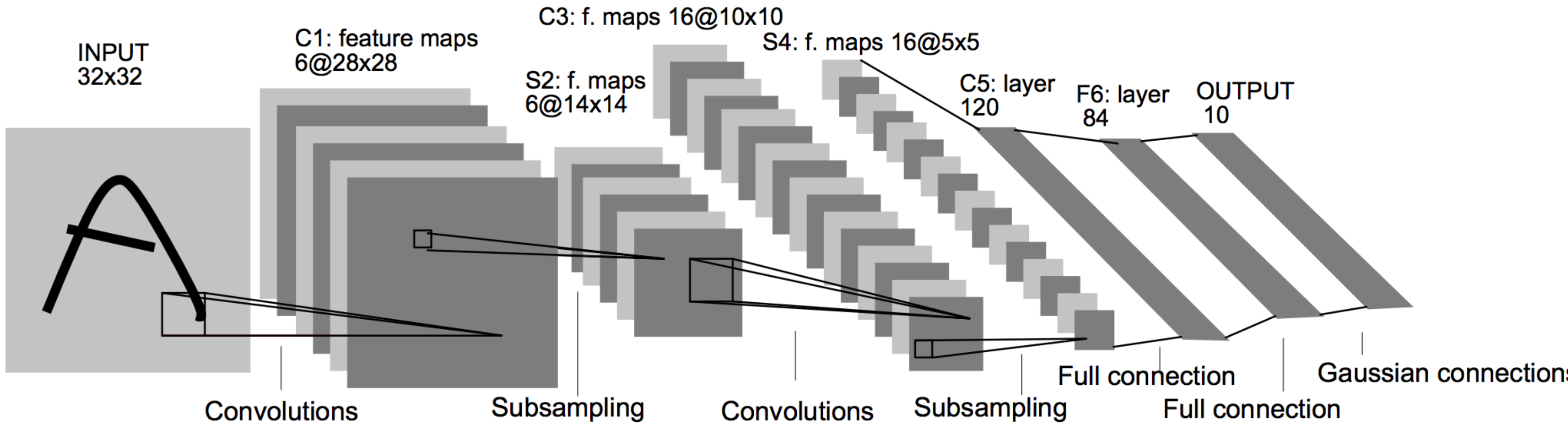
SDS BY CLASSIFYING BOTTOM-UP CANDIDATES

Background : Bottom-up Object Proposals

- Motivation: Reduce search space
- Aim for recall
- Many methods
  - Multiple segmentations (Selective Search)
  - Combinatorial grouping (MCG)
  - Seed/Graph-cut based (CPMC GOP)
  - Contour based (Edge Boxes)



Background : CNN



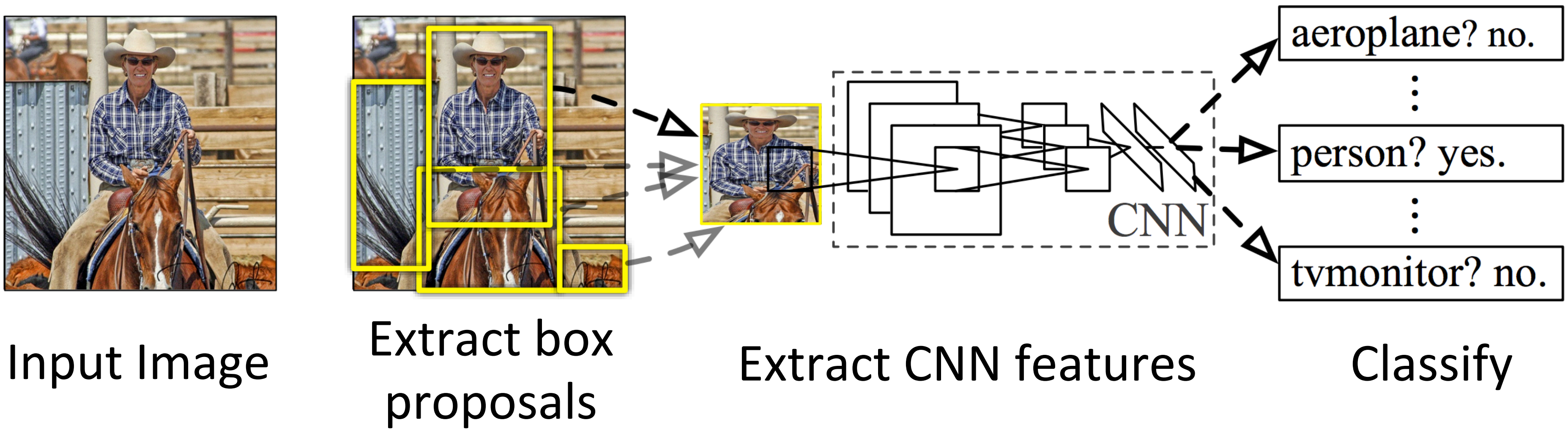
- Neocognitron
  Fukushima, 1980
- Learning Internal Representations by Error Propagation
  Rumelhart, Hinton and Williams, 1986
- Backpropagation applied to handwritten zip code recognition
  Le Cun et al. , 1989

….

- ImageNet Classification with Deep Convolutional Neural Networks
  Krizhevsky, Sutskever and Hinton, 2012

Slide adapted from Ross Girshick

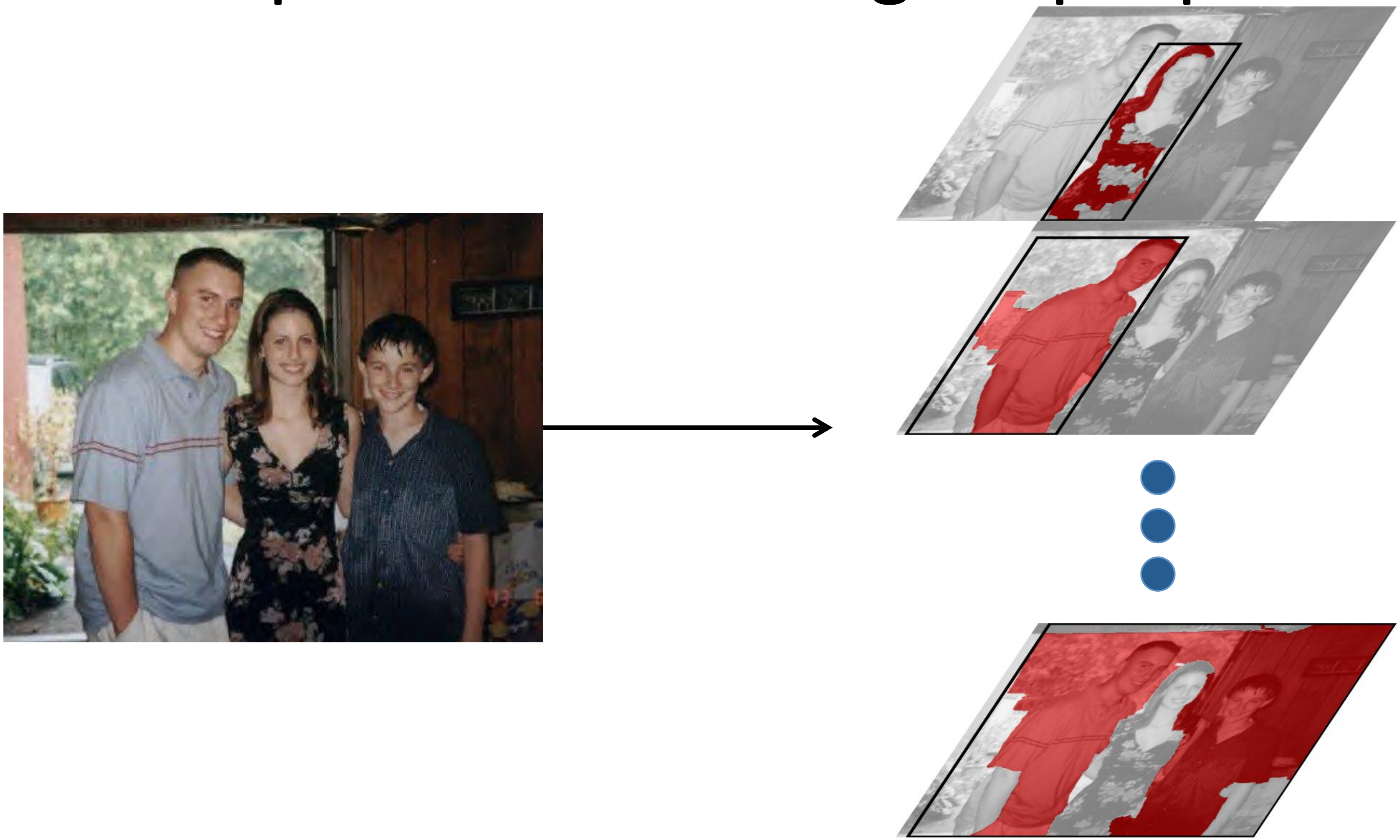
Background : R-CNN



R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In CVPR 2014.

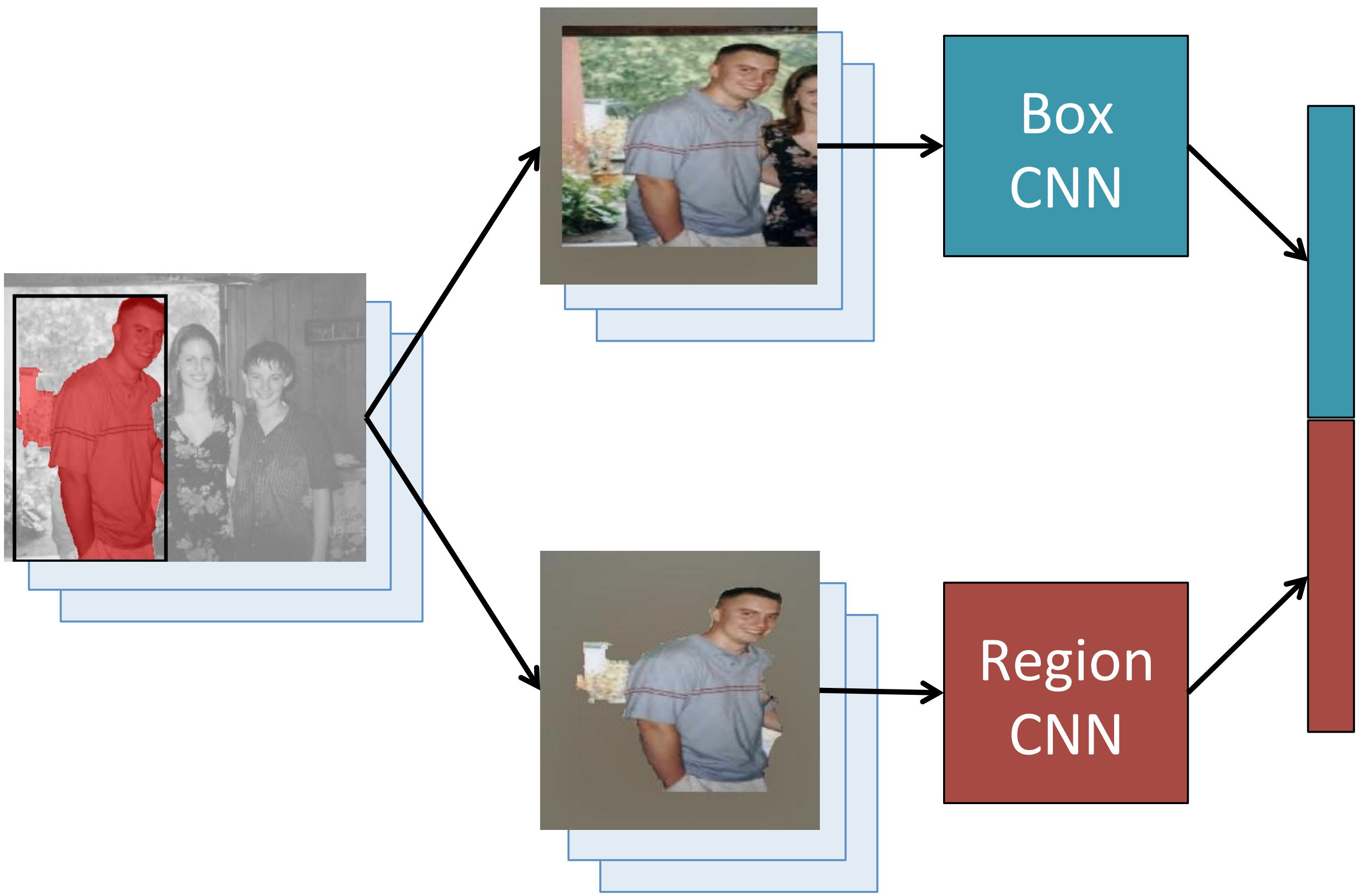
Slide adapted from Ross Girshick

From boxes to segments
Step 1: Generate region proposals

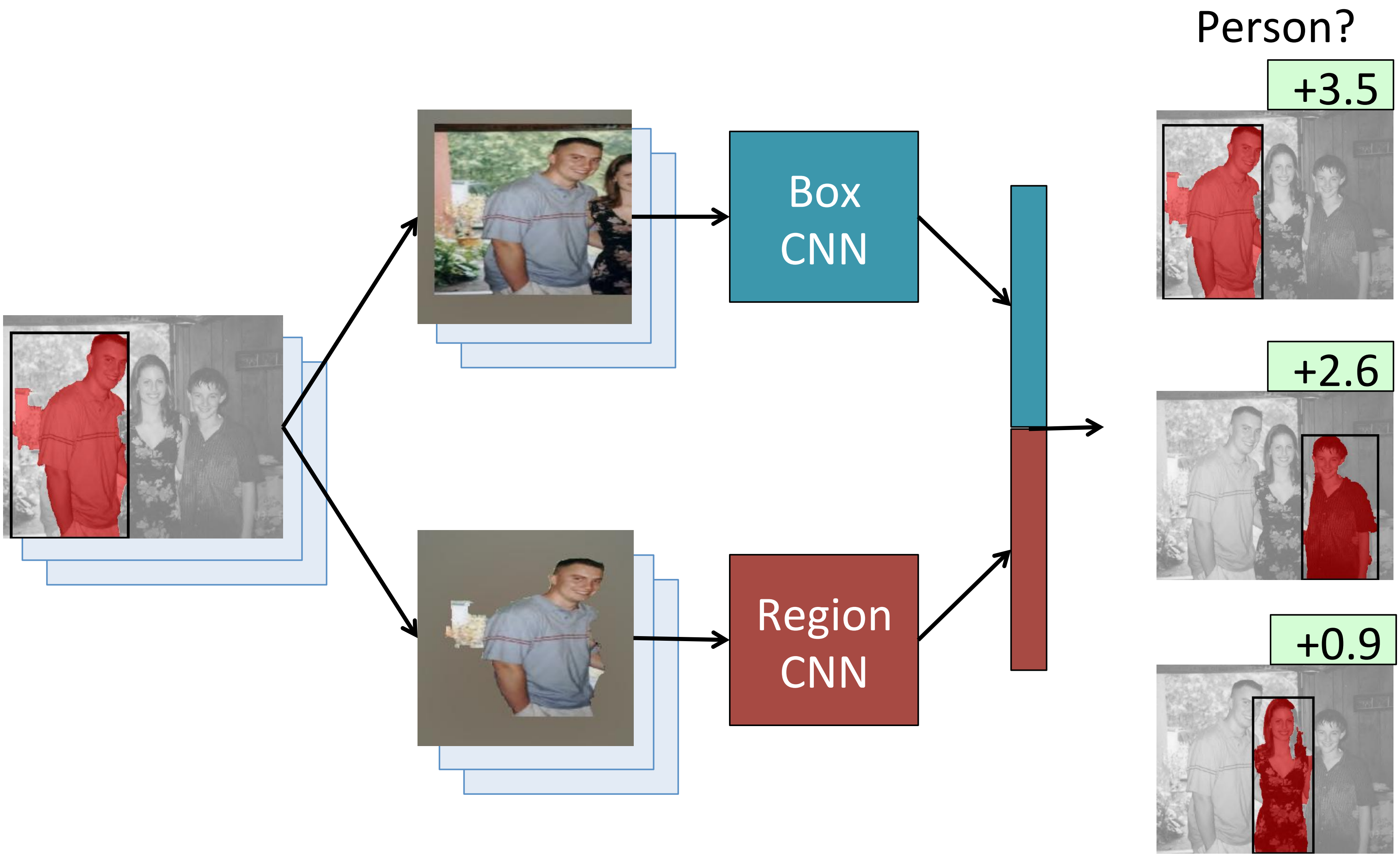


P. Arbeláez*, J. Pont-Tuset*, J. Barron, F. Marques and J. Malik. Multiscale Combinatorial Grouping. In CVPR 2014

From boxes to segments
Step 2: Score proposals

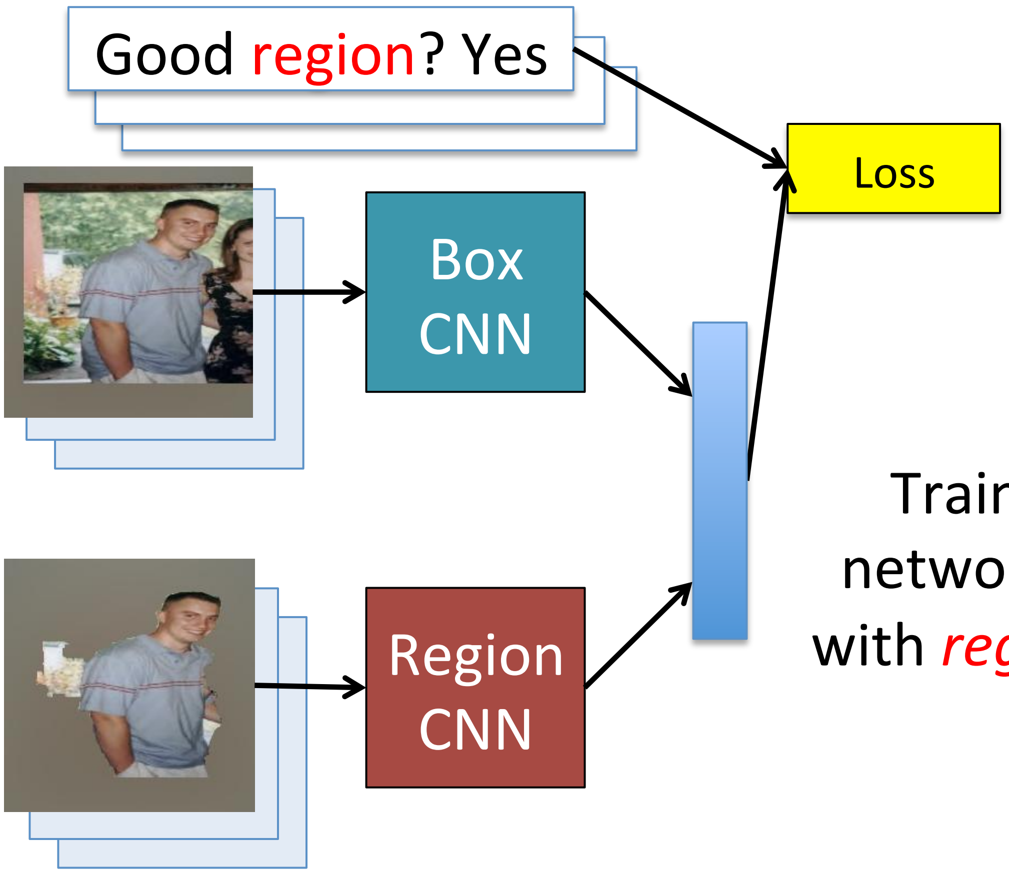


From boxes to segments
Step 2: Score proposals



Network training

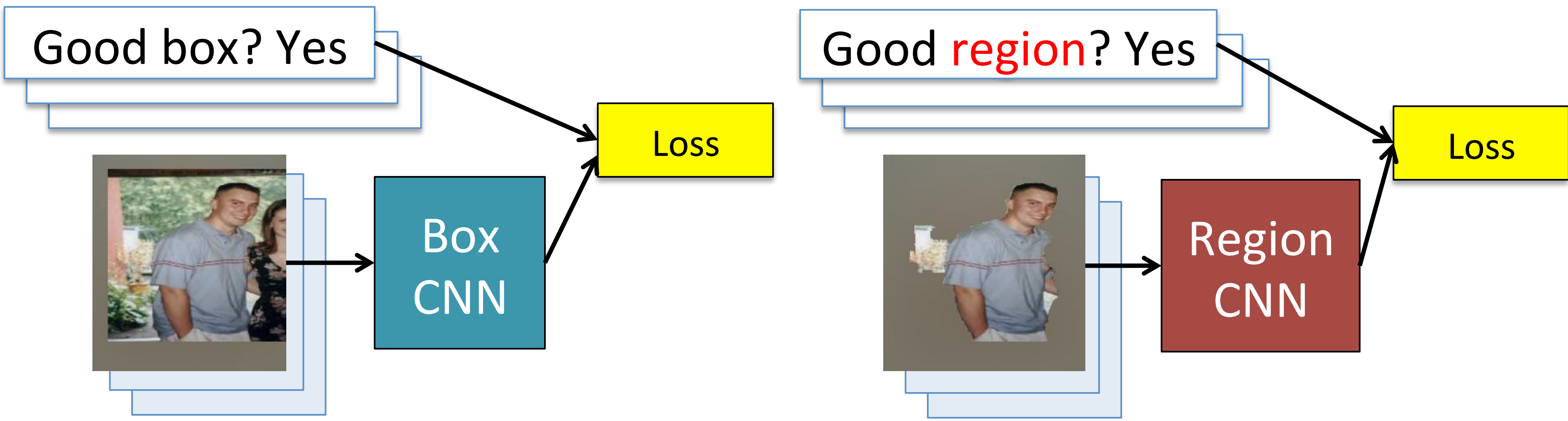
Joint task-specific training



Train entire network as one with *region* labels

Network training

Baseline 1: *Separate task specific training*

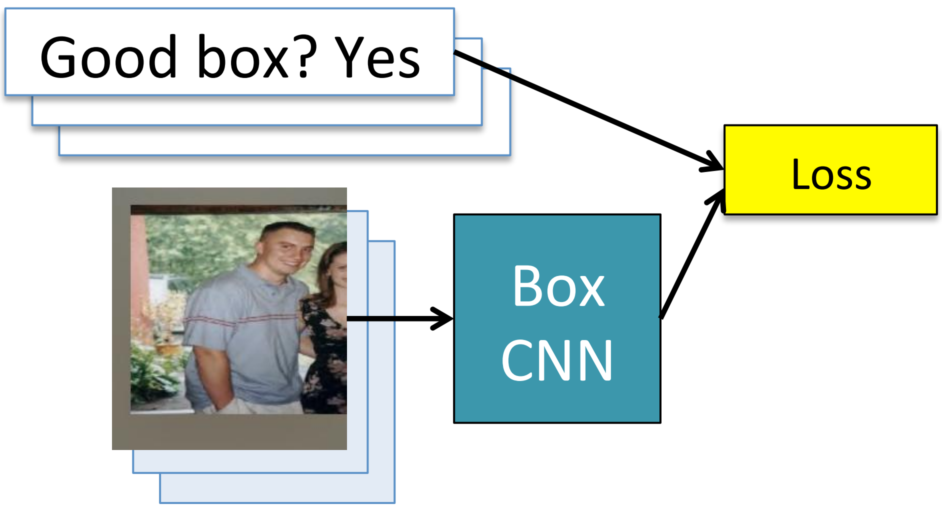


Train Box CNN using bounding box labels

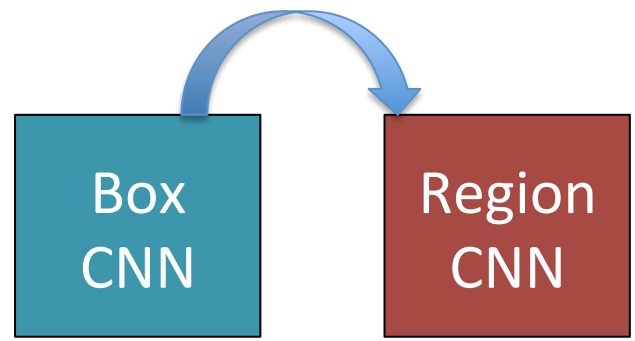
Train Region CNN using *region* labels

Network training

Baseline 2: *Copies of single CNN trained on bounding boxes*



Train Box CNN using bounding box labels



Copy the weights into Region CNN

Experiments

- Dataset : PASCAL VOC 2012 / SBD [1]
- Network architecture : [2]

	AP^r at 0.5	AP^r at 0.7
Joint	**47.7**	**22.9**
Baseline 1	47.0	21.9
Baseline 2	42.9	18.0

- Joint, task-specific training works!

1. B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji and J. Malik. Semantic contours from inverse detectors. ICCV (2011)
2. A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional networks. NIPS(2012)

Results

Error modes

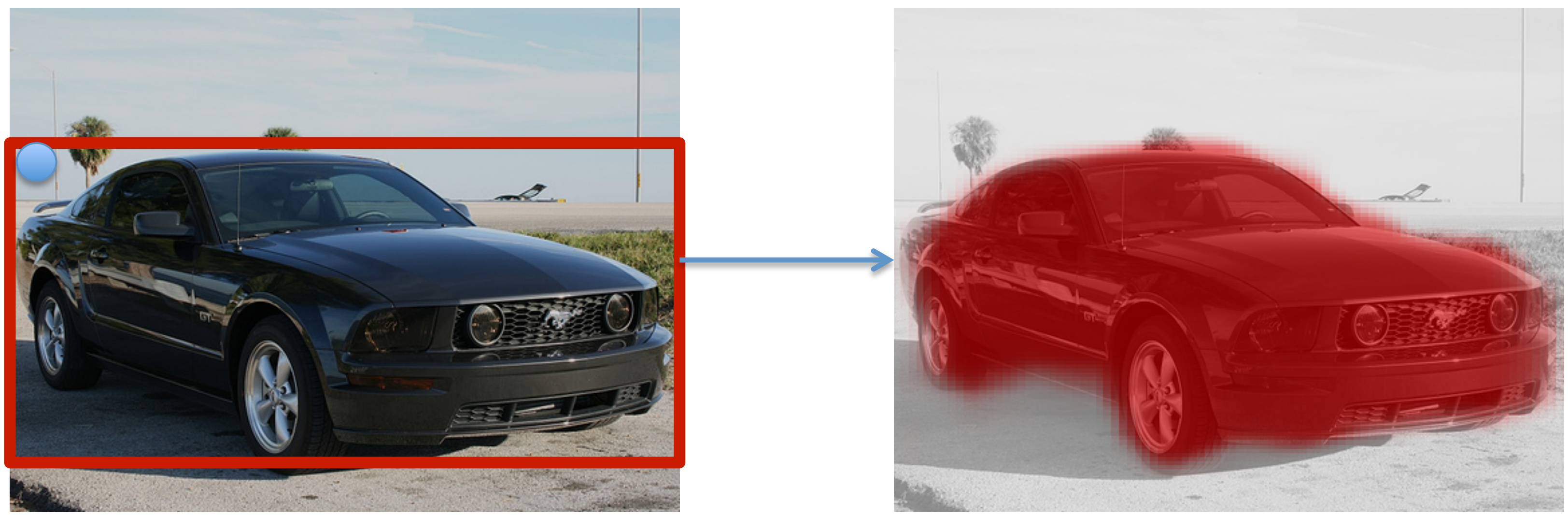
SDS BY TOP-DOWN FIGURE-GROUND PREDICTION

The need for top-down predictions

- Bottom-up processes make mistakes.
- Some categories have distinctive shapes.

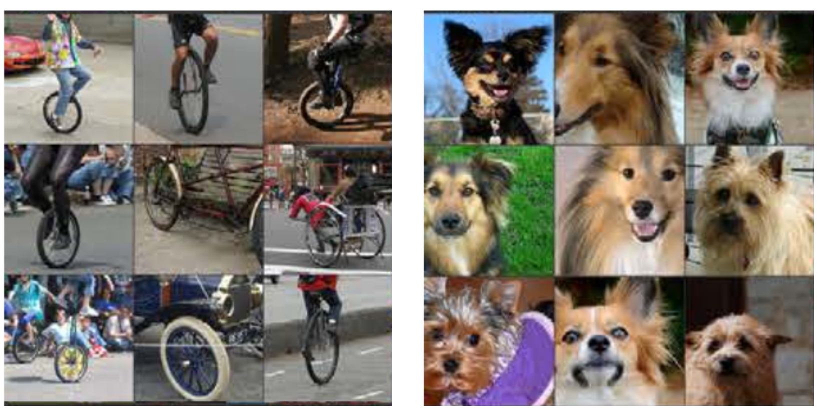
Top-down figure-ground prediction

- Pixel classification
  - For each p in window, does it belong to object?
- Idea: Use features from CNN

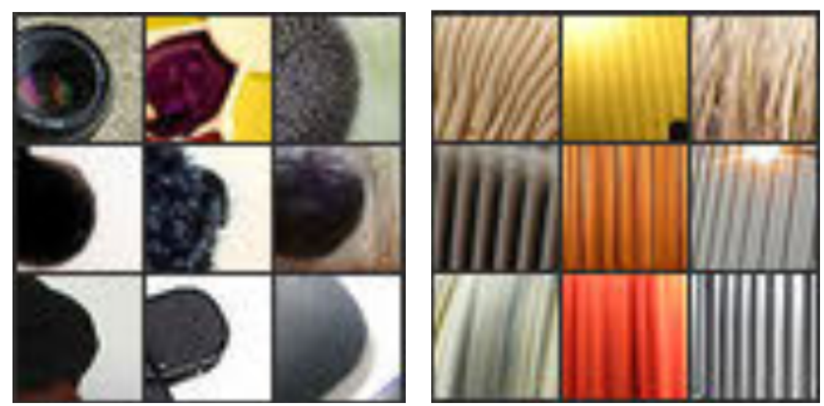


CNNs for figure-ground

- Idea: Use features from CNN
- But which layer?
  - Top layers lose localization information
  - Bottom layers are not semantic enough
- Our solution: use all layers!

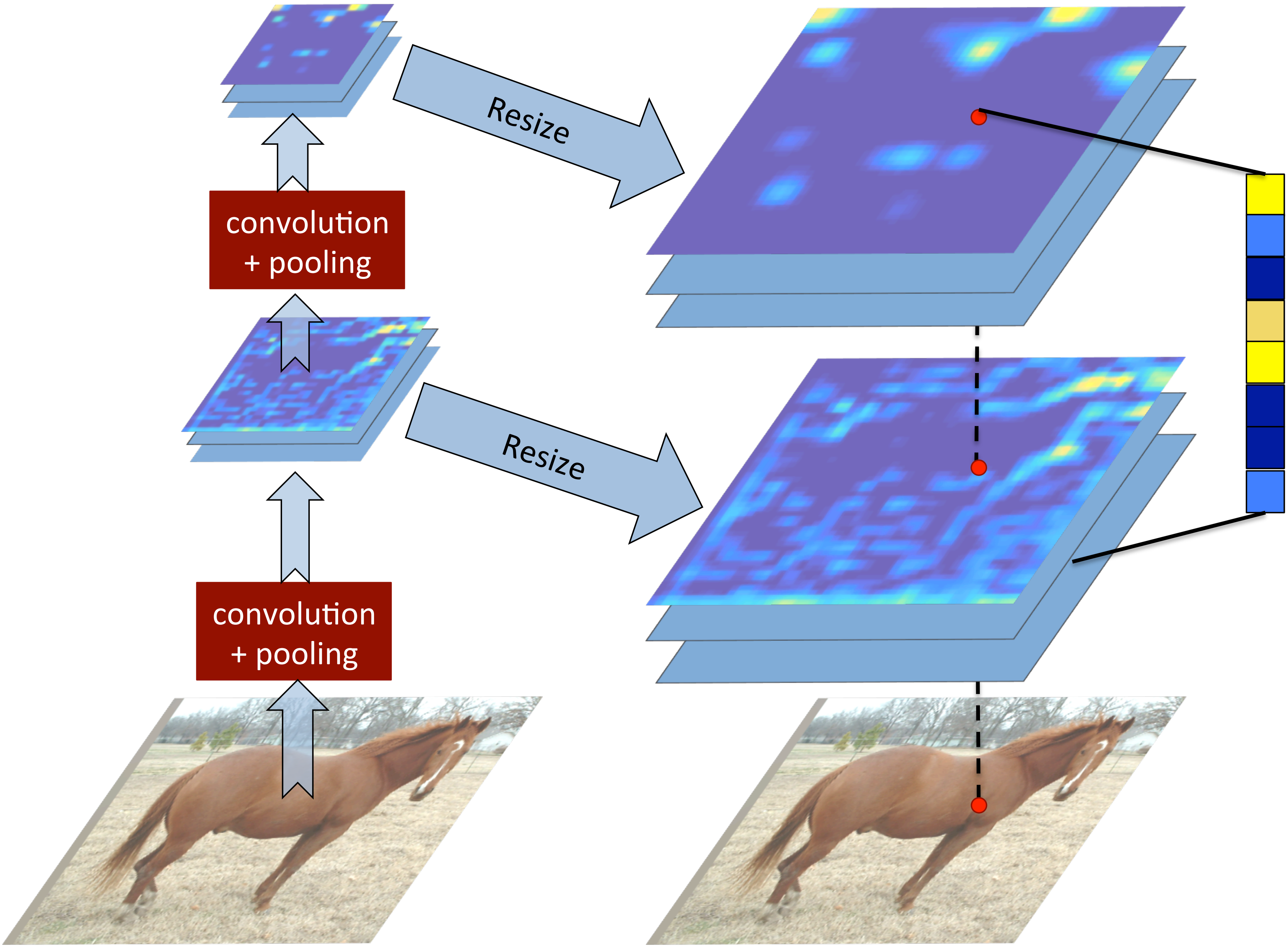


Layer 5



Layer 2

Figure from : M. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In ECCV 2014.

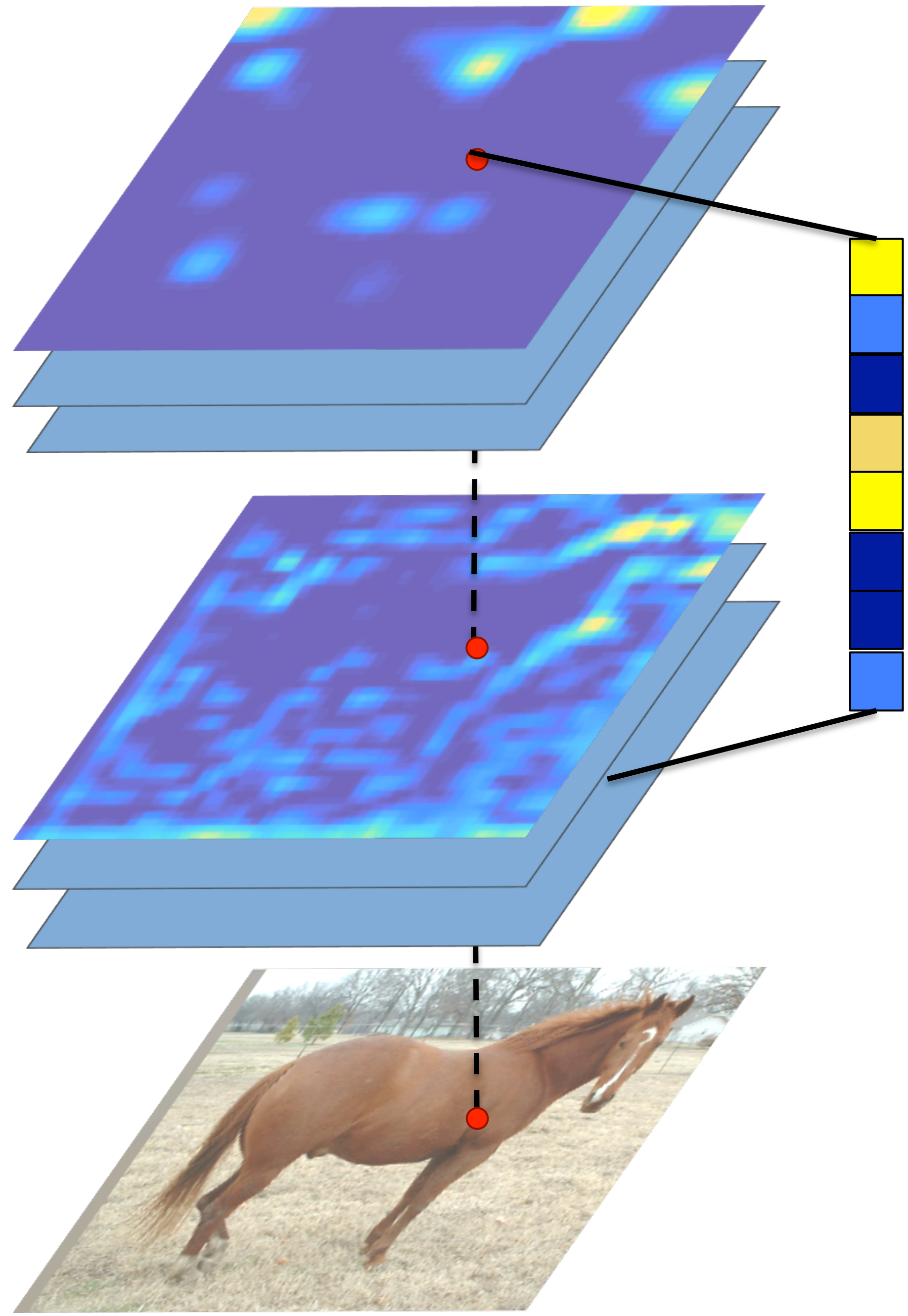

convolution
+ pooling
convolution
+ pooling
Resize
Resize

Hypercolumns*

*D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 1962.

Also called jets: J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. Biological cybernetics, 55(6), 1987.

Also called skip-connections: J. Long, E. Schelhamer and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. arXiv preprint. arXiv:1411.4038

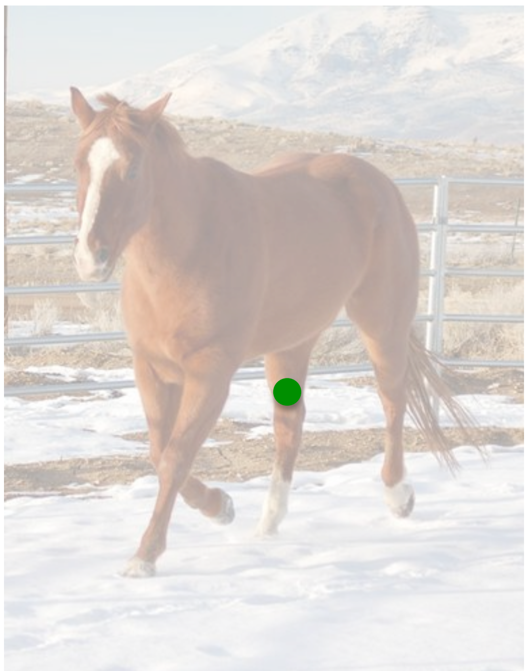


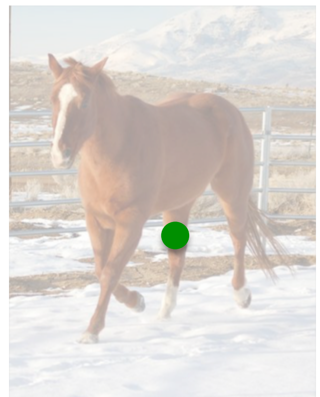
Analogy with image pyramids

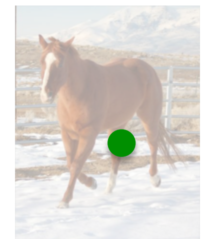
Hard : large coarse displacements
Easy : small fine deformations

Easy : large coarse displacements
Hard : small fine deformations

Analogy with image pyramids



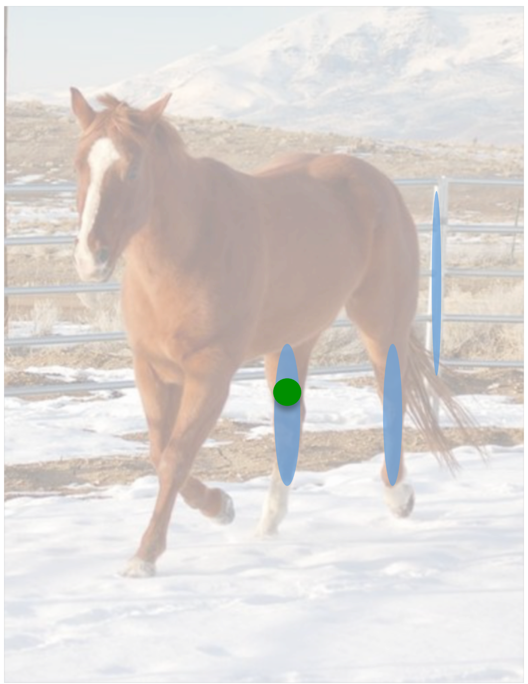




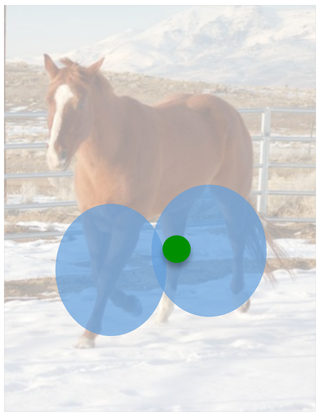
Hard : large coarse displacements
Easy : small fine deformations

Easy : large coarse displacements
Hard : small fine deformations

Analogy with image pyramids



High resolution “vertical bar” detector



Medium resolution “animal leg” detector

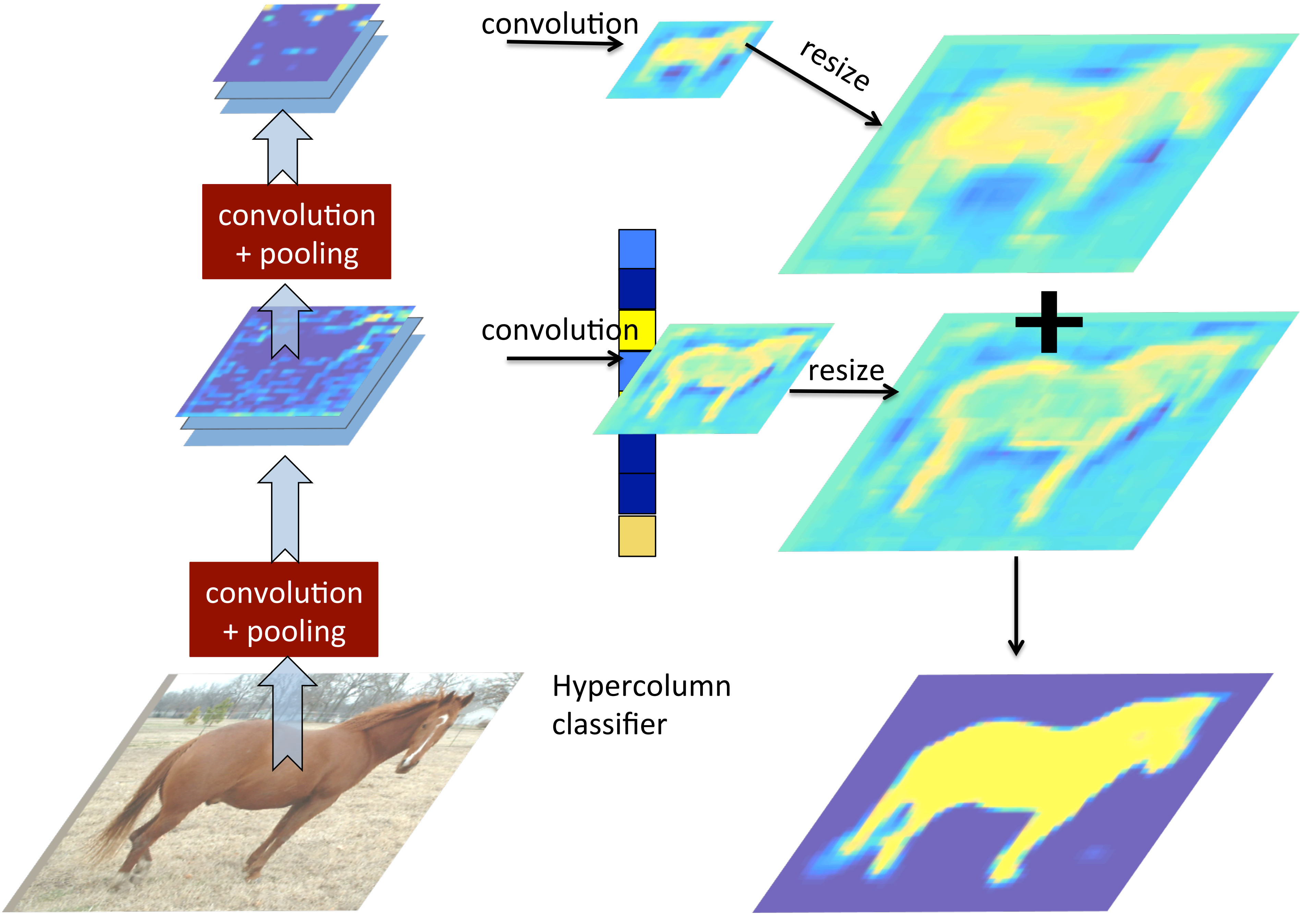
High resolution “horse” detector

Hypercolumns

- Layer outputs are feature maps
- Concatenate to get hypercolumn feature maps
- Feature maps are of coarser resolution
  - Resize (bilinear interpolate) to image resolution

Efficient pixel classification

- Upsampling large feature maps is expensive!
- Linear classification (bilinear interpolation) = bilinear interpolation (linear classification)
- Linear classification = 1x1 convolution
  - extension : use nxn convolution
- *Classification = convolve, upsample, sum, sigmoid*



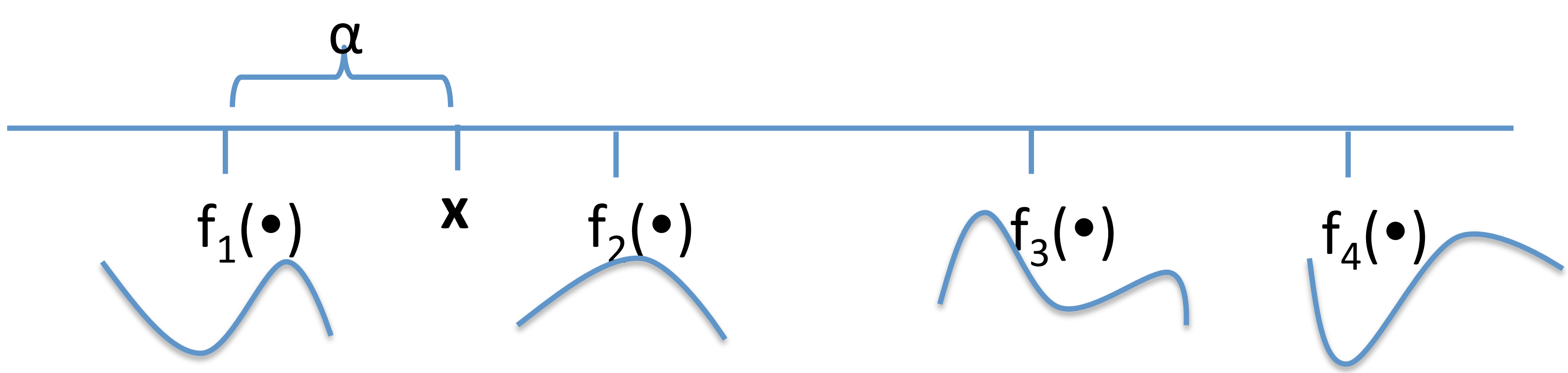
convolution
resize
convolution + pooling
convolution
resize
convolution + pooling
Hypercolumn classifier

Using pixel location

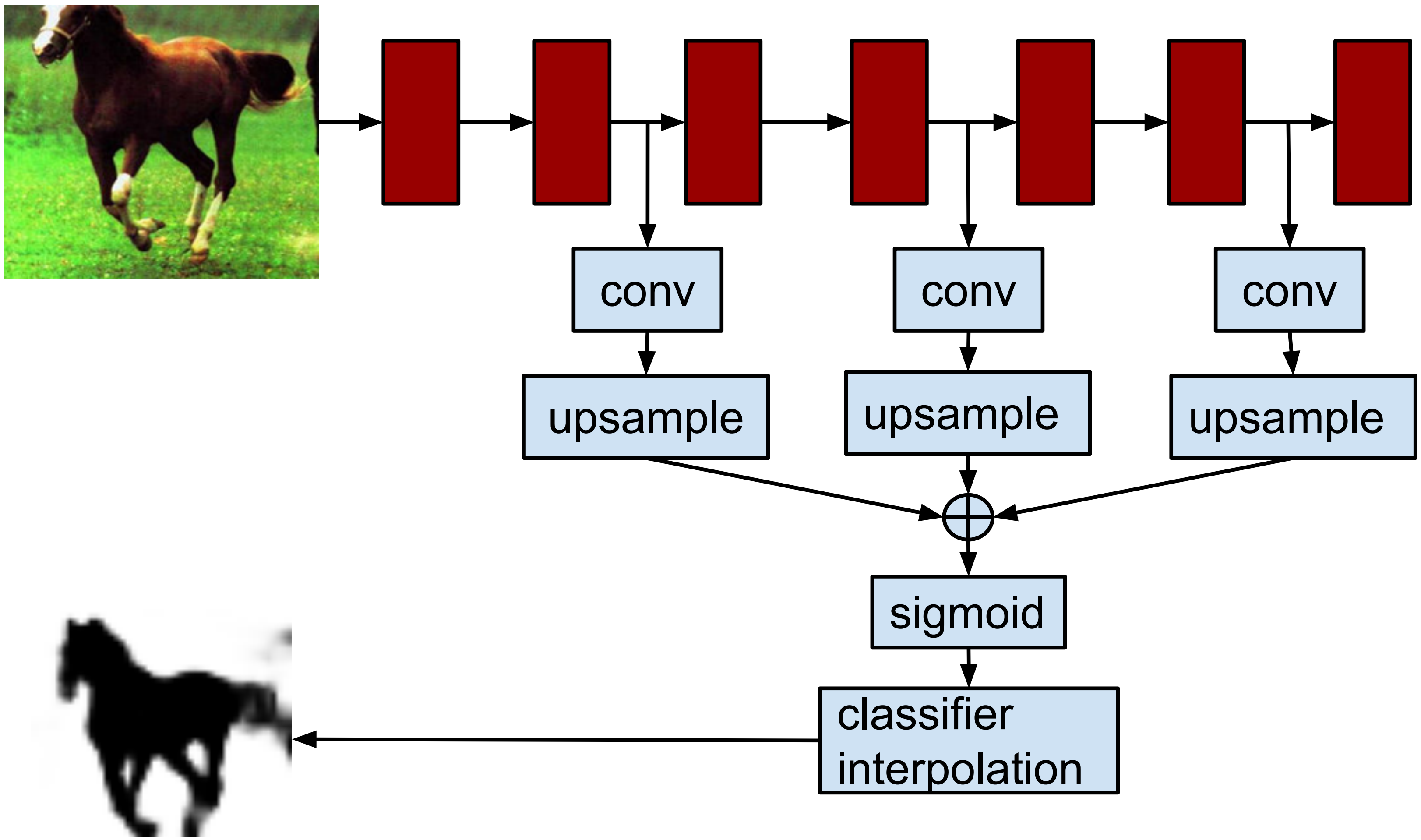
Using pixel location

- Separate classifier for each location?
  - Too expensive
  - Risk of overfitting
- Interpolate into coarse grid of classifiers

$$f(\mathbf{x}) = \alpha f_2(\mathbf{x}) + (1-\alpha) f_1(\mathbf{x})$$



Representation as a neural network



Using top-down predictions

- For refining bottom-up proposals
  - Start from high scoring SDS detections
  - Use hypercolumn features + binary mask to predict figure-ground
- For segmenting bounding box detections

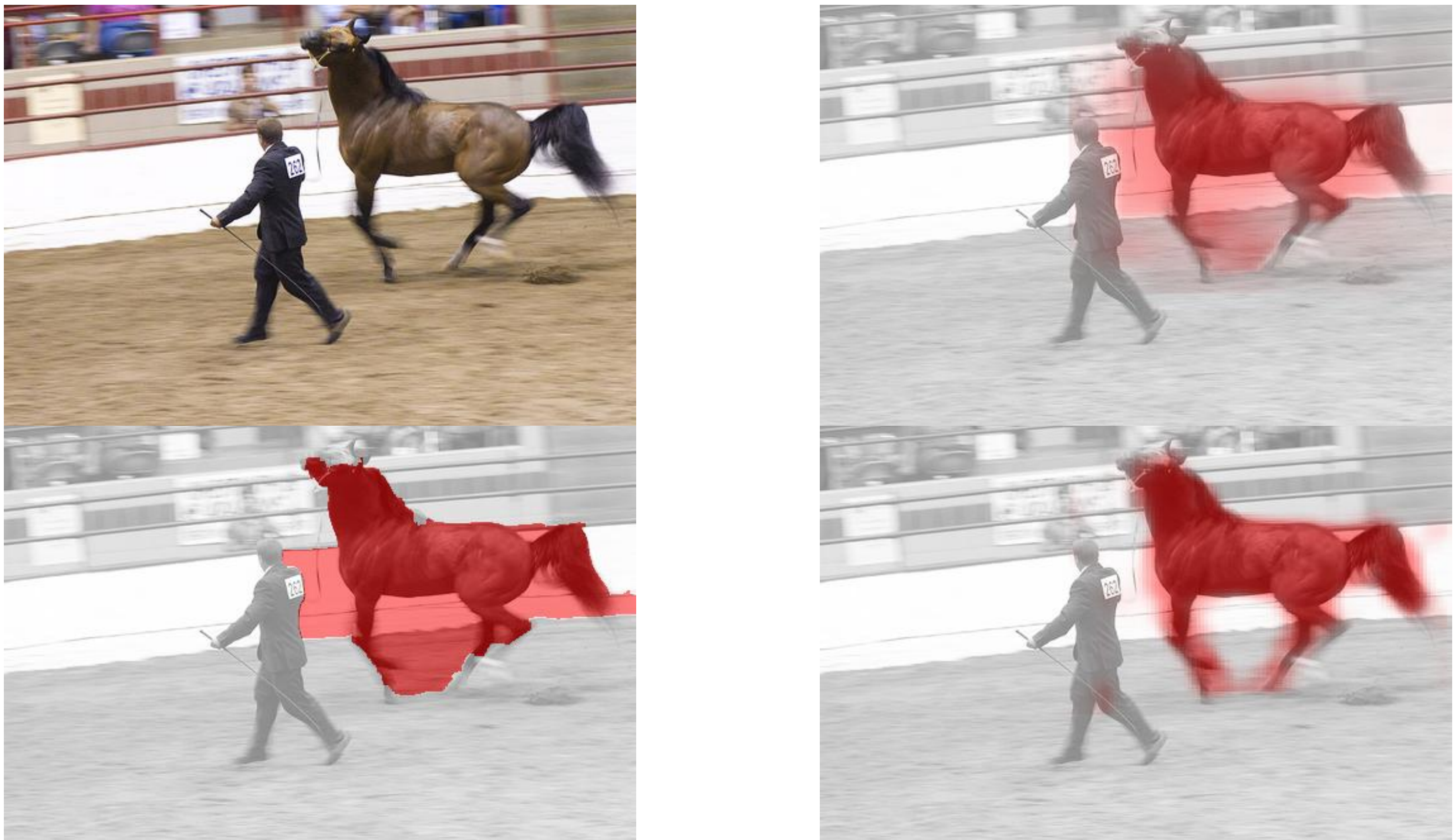
Refining proposals

	AP^r at 0.5	AP^r at 0.7
No refinement	47.7	22.8
Top layer (layer 7)	49.7	25.8

Refining proposals: Using multiple layers

Image

Layer 7



Bottom-up candidate

Layers 7, 4 and 2

Refining proposals: Using multiple layers

Image

Layer 7

Bottom-up candidate

Layers 7, 4 and 2

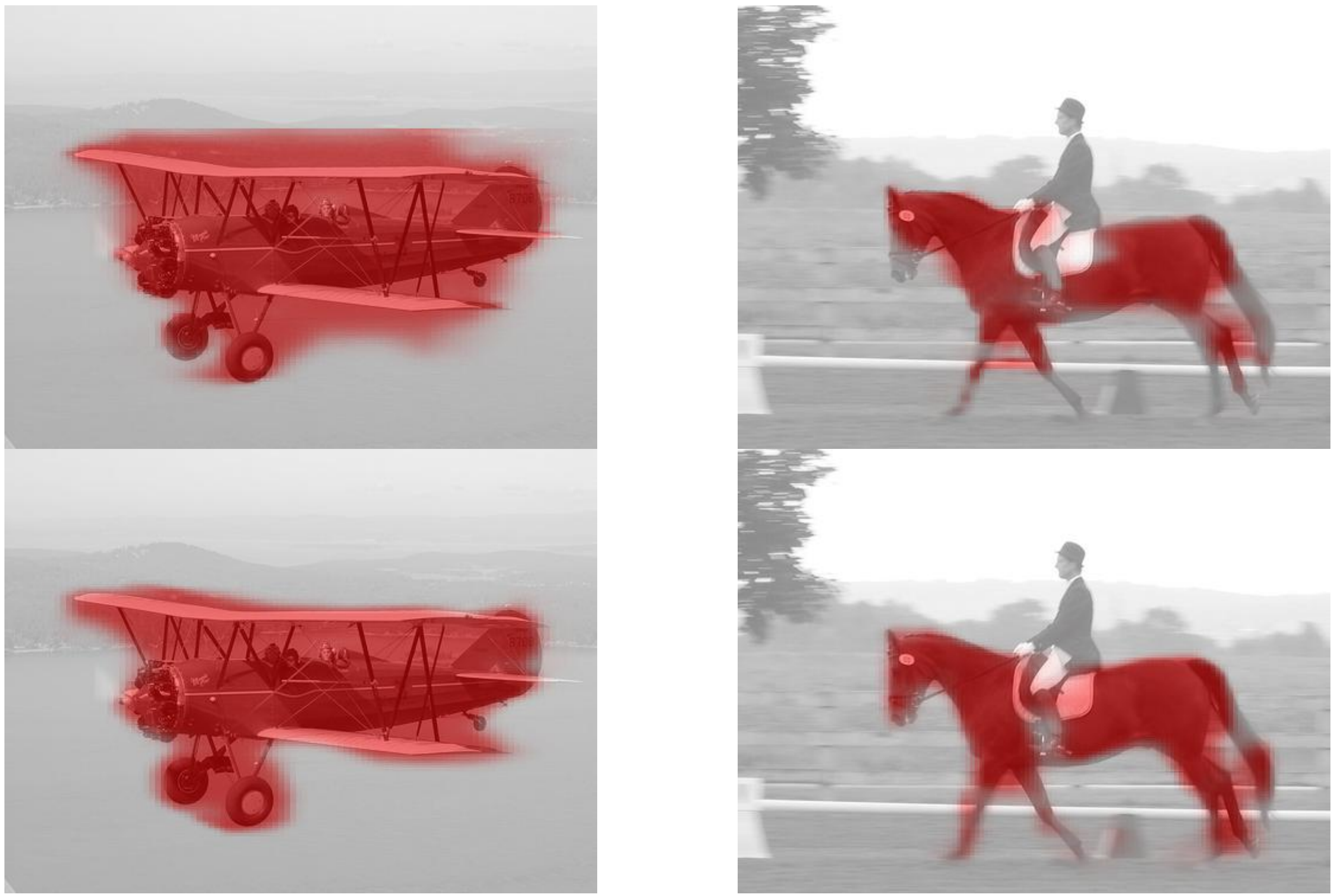
Refining proposals: Using location

Grid size	AP^r at 0.5	AP^r at 0.7
1x1	50.3	28.8
2x2	**51.2**	30.2
5x5	**51.3**	**31.8**
10x10	**51.2**	**31.6**

Refining proposals: Using location

1 x 1

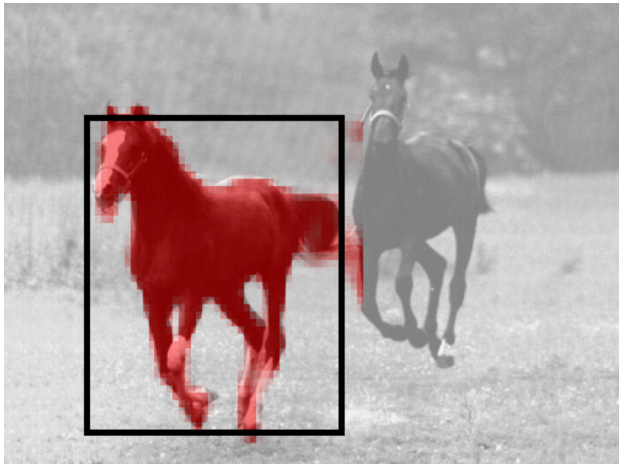
5 x 5

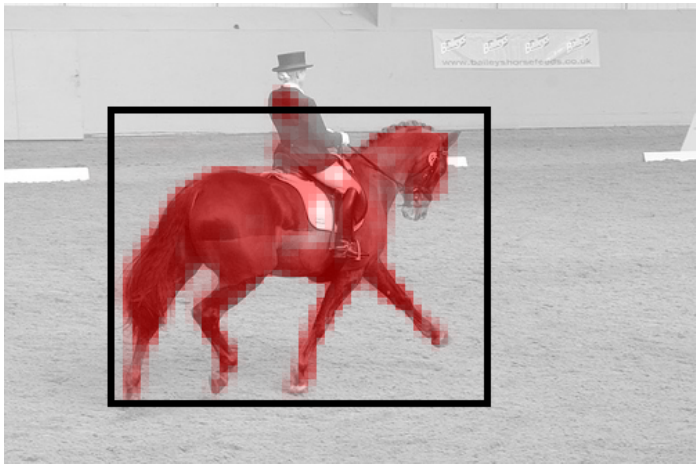


Refining proposals: Finetuning and bbox regression

	AP^r at 0.5	AP^r at 0.7
Hypercolumn	51.2	31.6
+Bbox Regression	51.9	32.4
+Bbox Regression+FT	**52.8**	**33.7**

Segmenting bbox detections





Segmenting bbox detections

	Network	APr at 0.5	APr at 0.7
Classify segments + Refine	T-net[1]	51.9	32.4

1. A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional networks. NIPS(2012)
2. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

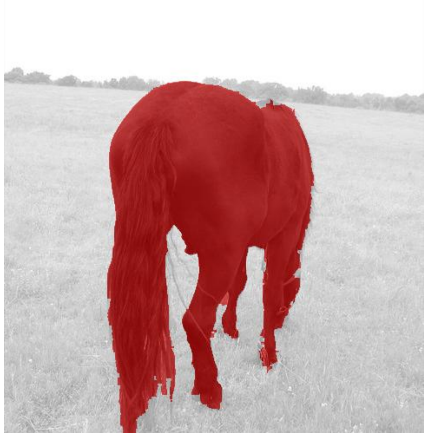
Segment + Rescore

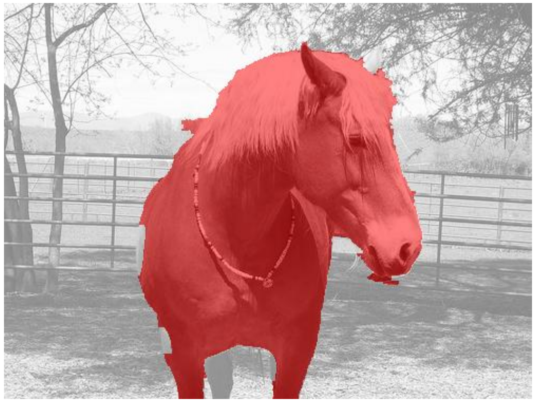
Segmenting bbox detections

	Network	APr at 0.5	APr at 0.7
Classify segments + Refine	T-net[1]	51.9	32.4
Segment bbox detections	T-net	49.1	29.1
Segment bbox detections	O-net[2]	56.5	37.0

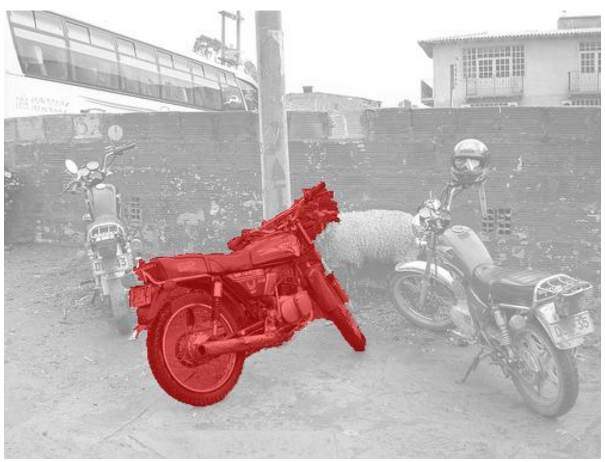
1. A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional networks. NIPS(2012)
2. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

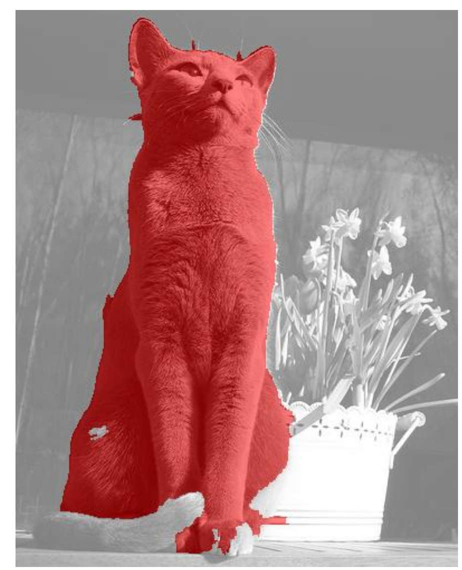
Qualitative results

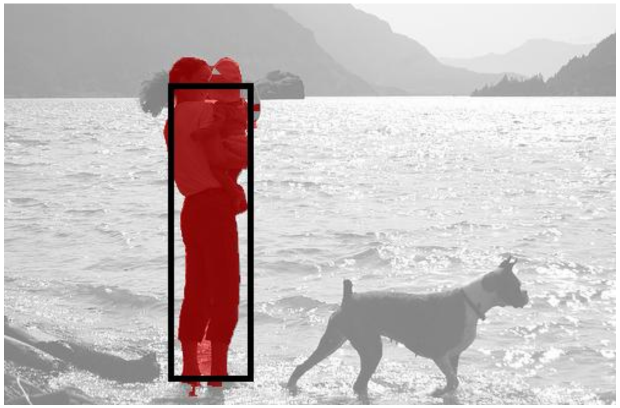


Qualitative results





Error modes

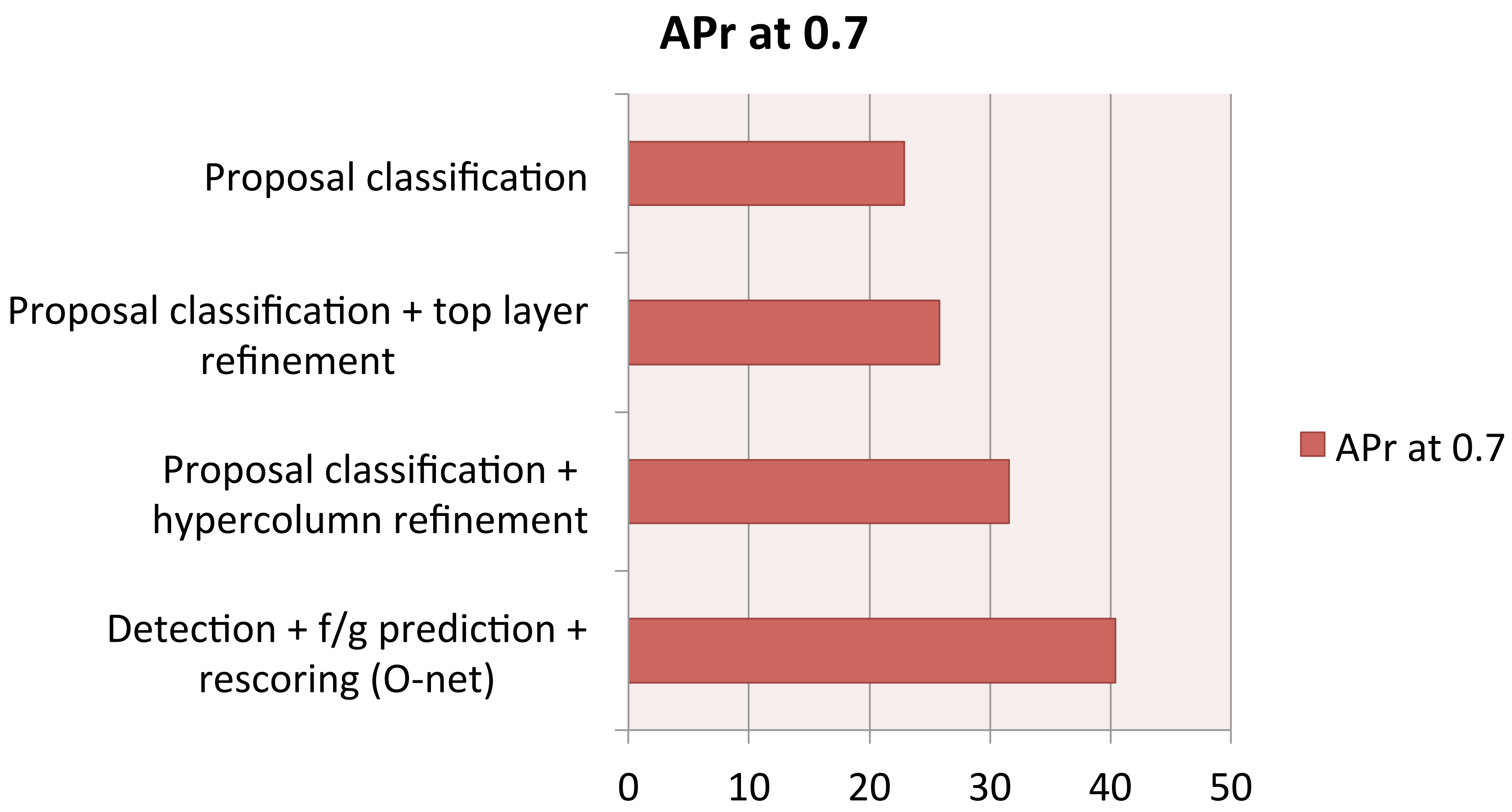


Multiple objects

Non-prototypical poses

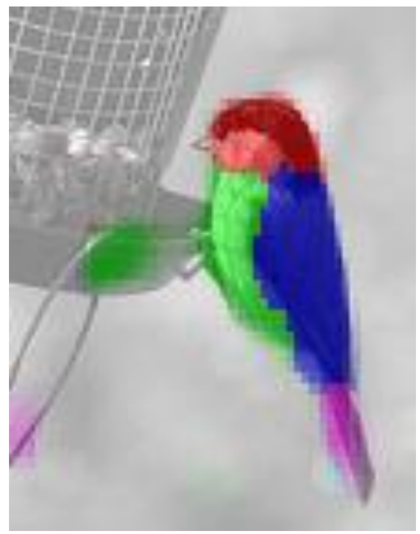
Occlusion

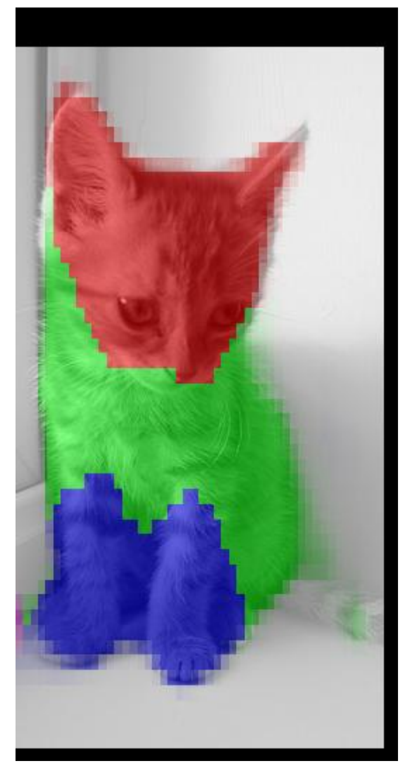
Summary of SDS



Part Labeling

- Same (hypercolumn) features, different labels!





Part Labeling - Experiments

- Dataset: PASCAL Parts [1]
- Evaluation: Detection is correct if #(correctly labeled pixels) / union > threshold

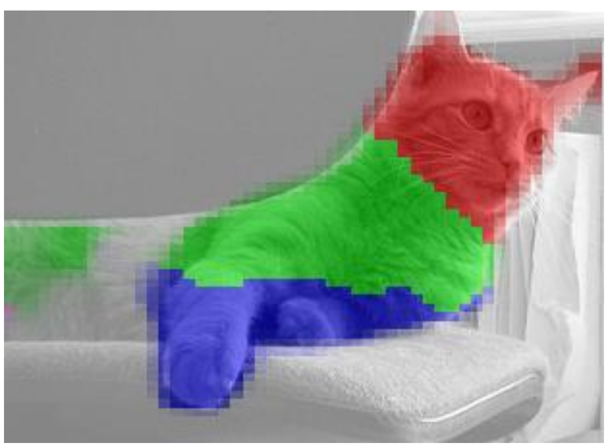
	Bird	Cat	Cow	Dog	Horse	Person	Sheep
Layer 7	**15.4**	19.2	14.5	8.5	16.6	21.9	38.9
Layers 7, 4 and 2	14.2	**30.3**	**21.5**	**14.2**	**27.8**	**28.5**	**44.9**

1. X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun and A. Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts . CVPR 2014

Error modes

Disjointed parts

Misclassification



Wrong figure/ground

Conclusion

- A detection system that can
  - Provide pixel accurate segmentations
  - Provide part labelings and pose estimates
- A general framework for fine-grained localization using CNNs.