Object Detection

Jia-Bin Huang

Virginia Tech

ECE 6554 Advanced Computer Vision

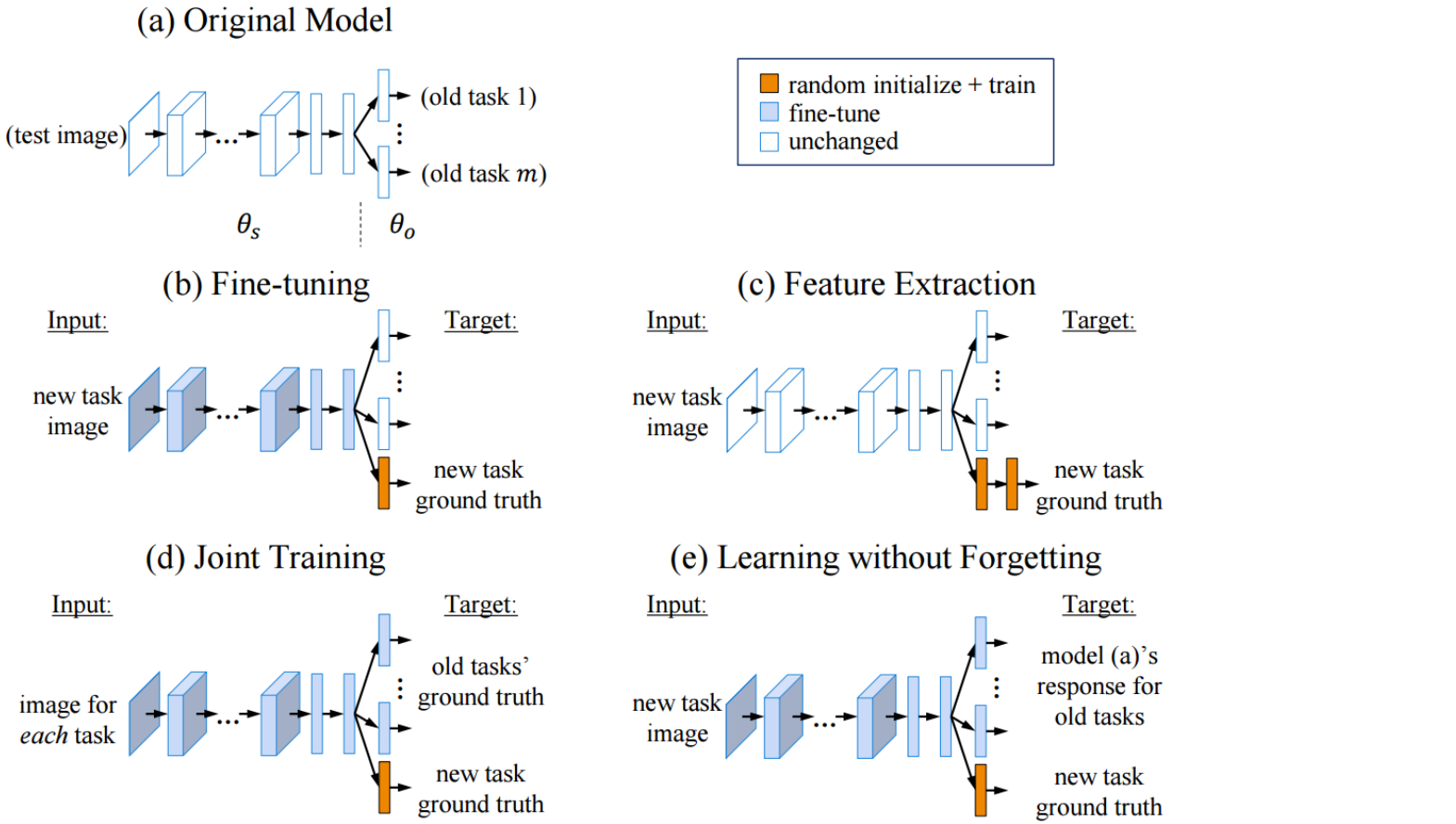
Today's class

- Review on transfer learning
- Discussion of the paper
  - Rich feature hierarchies for accurate object detection and semantic segmentation. R. Girshick, J. Donahue, T. Darrell, J. Malik. CVPR 2014
  - "For" lead: Subhashree
  - "Against" lead: Yousi
- Overview of categorical object detection
- Recent advances

Review: transfer learning

- New dataset is small and similar to original dataset.
  - train a linear classifier on the CNN codes
- New dataset is large and similar to the original dataset
  - fine-tune through the full network
- New dataset is small but very different from the original dataset
  - SVM classifier from activations somewhere earlier in the network
- New dataset is large and very different from the original dataset
  - fine-tune through the entire network

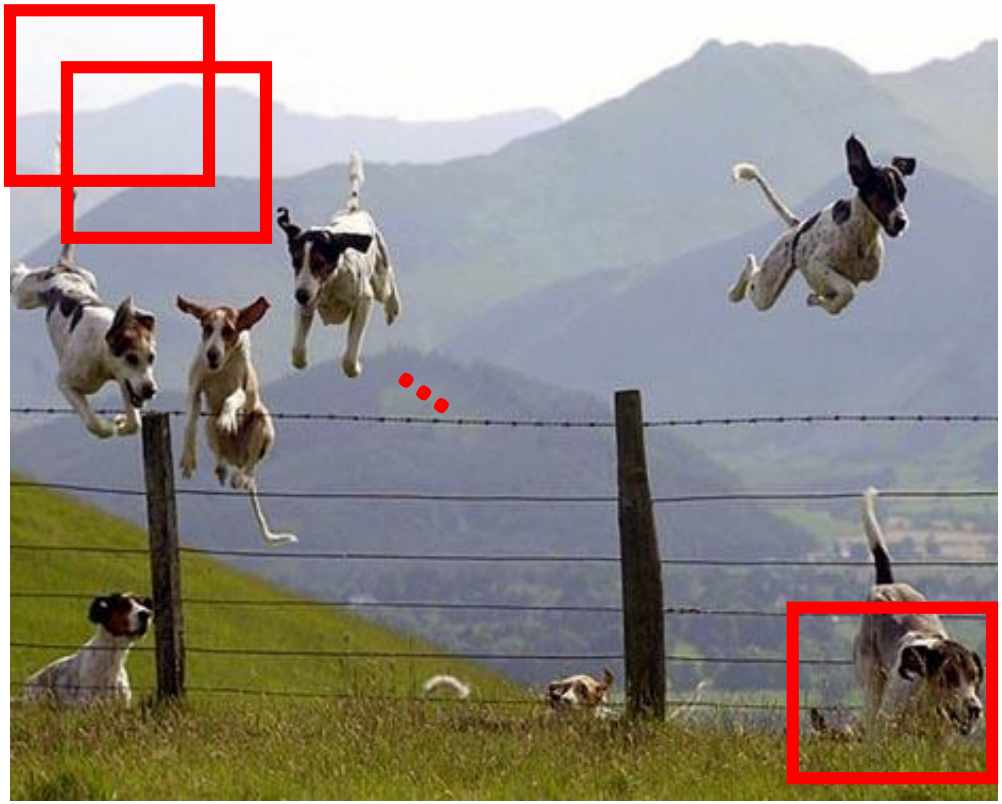
	Fine Tuning	Duplicating and Fine Tuning	Feature Extraction	Joint Training	Learning without Forgetting
new task performance	good	good	X medium	best	✓best
original task performance	X bad	good	good	good	✓good
training efficiency	fast	fast	fast	X slow	✓fast
testing efficiency	fast	X slow	fast	fast	✓fast
storage requirement	medium	X large	medium	X large	✓medium
requires previous task data	no	no	no	X yes	✓no



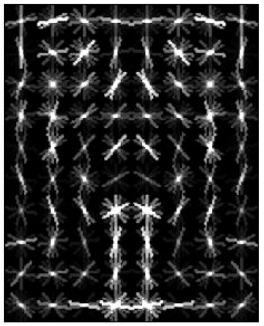
Learning without forgetting, ECCV 2016

Object Category Detection

- Focus on object search: “Where is it?”
- Build templates that quickly differentiate object patch from background patch



Dog Model



Object or
Non-Object?

Challenges in modeling the object class

Illumination

Object pose

Clutter

Occlusions

Intra-class appearance

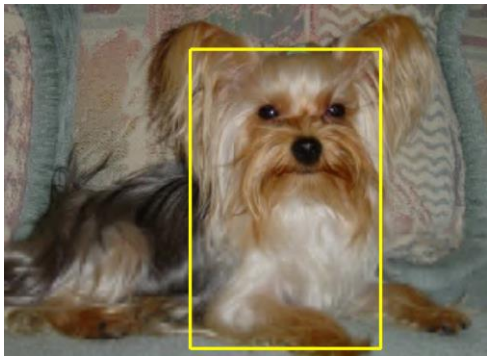
Viewpoint

Slide from K. Grauman, B. Leibe

Challenges in modeling the non-object class

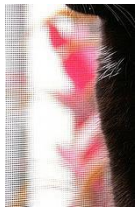
True Detections

Bad Localization



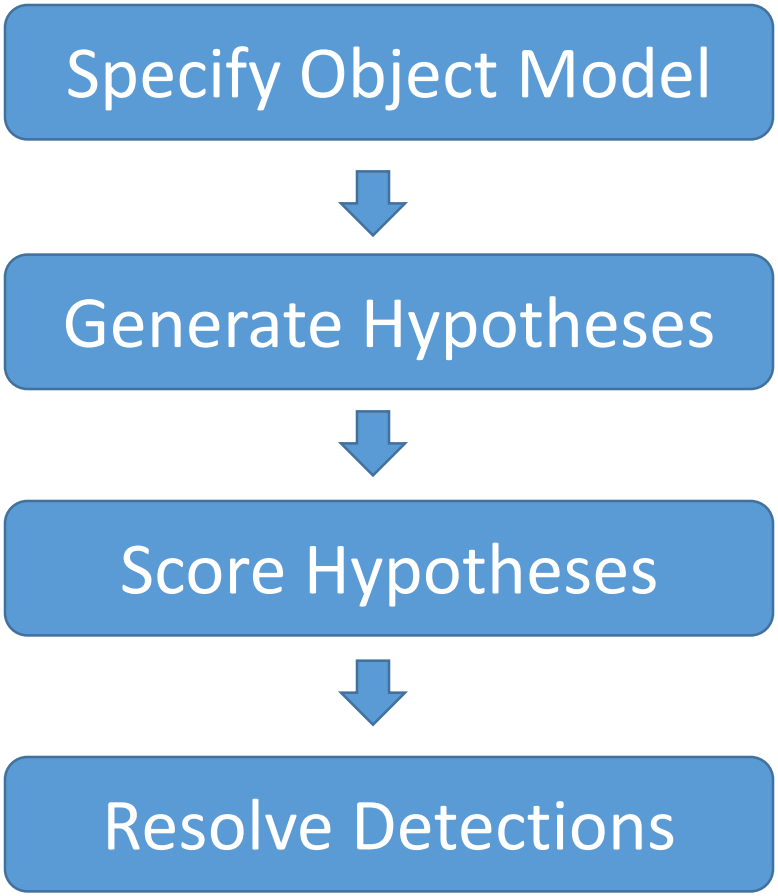
Confused with Similar Object

Misc. Background



Confused with Dissimilar Objects

General Process of Object Recognition

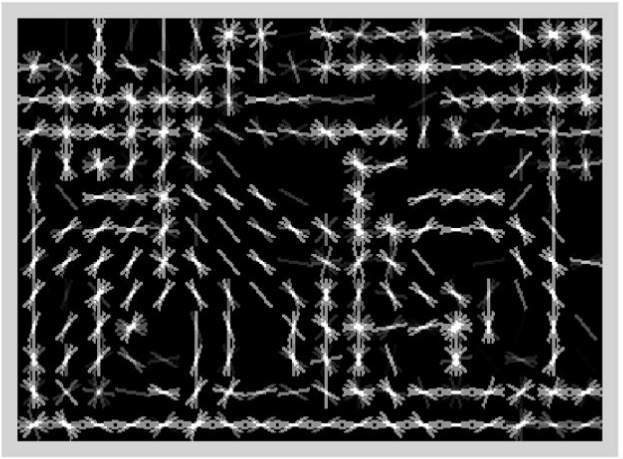


What are the object parameters?

Specifying an object model

1. Statistical Template in Bounding Box
   - Object is some (x,y,w,h) in image
   - Features defined wrt bounding box coordinates

Image

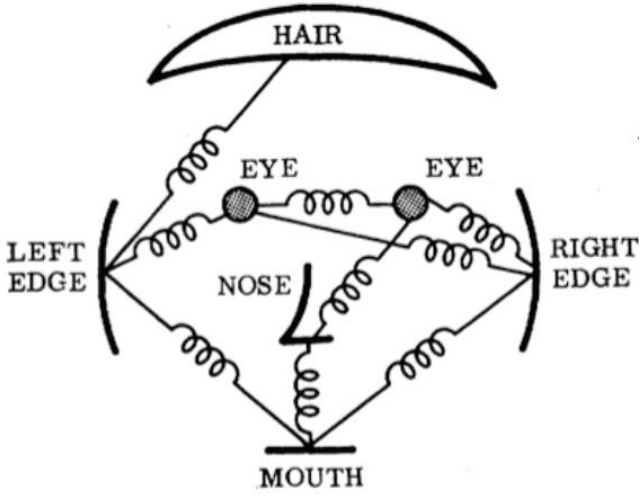


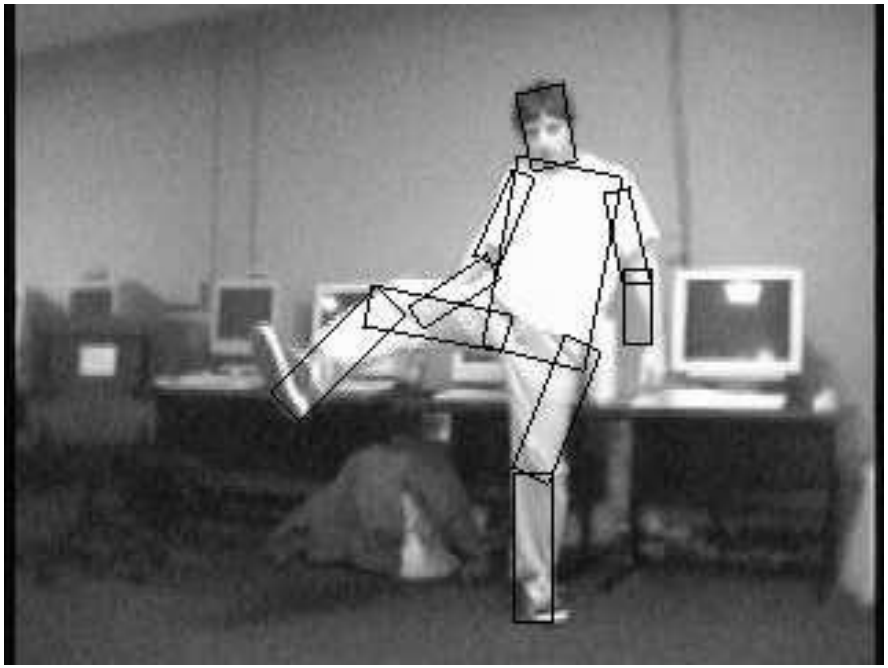
Template Visualization

Images from Felzenszwalb

Specifying an object model

2. Articulated parts model
   - Object is configuration of parts
   - Each part is detectable



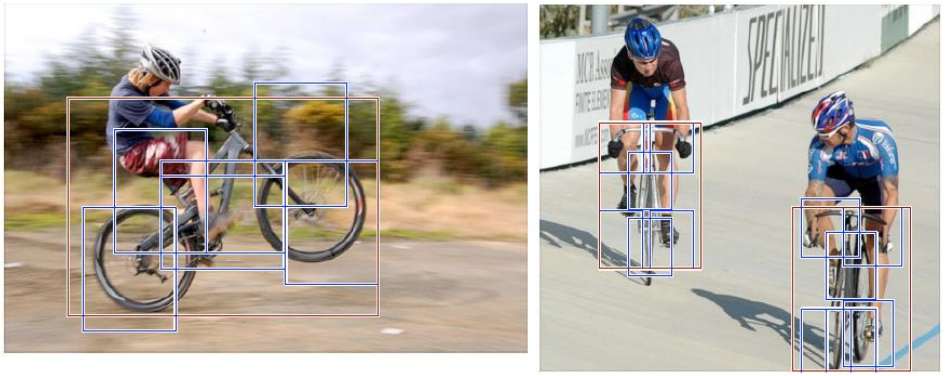


Images from Felzenszwalb

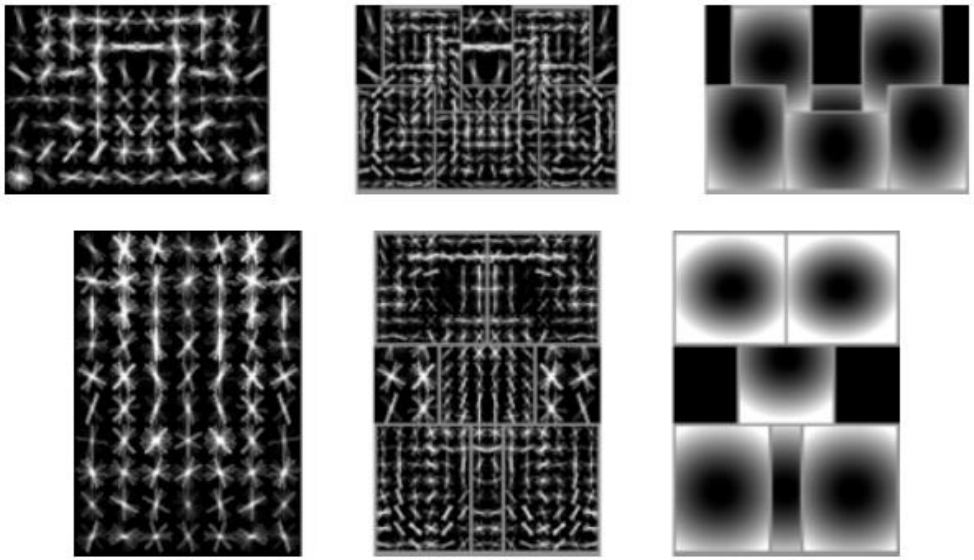
Specifying an object model

3. Hybrid template/parts model

Detections



Template Visualization



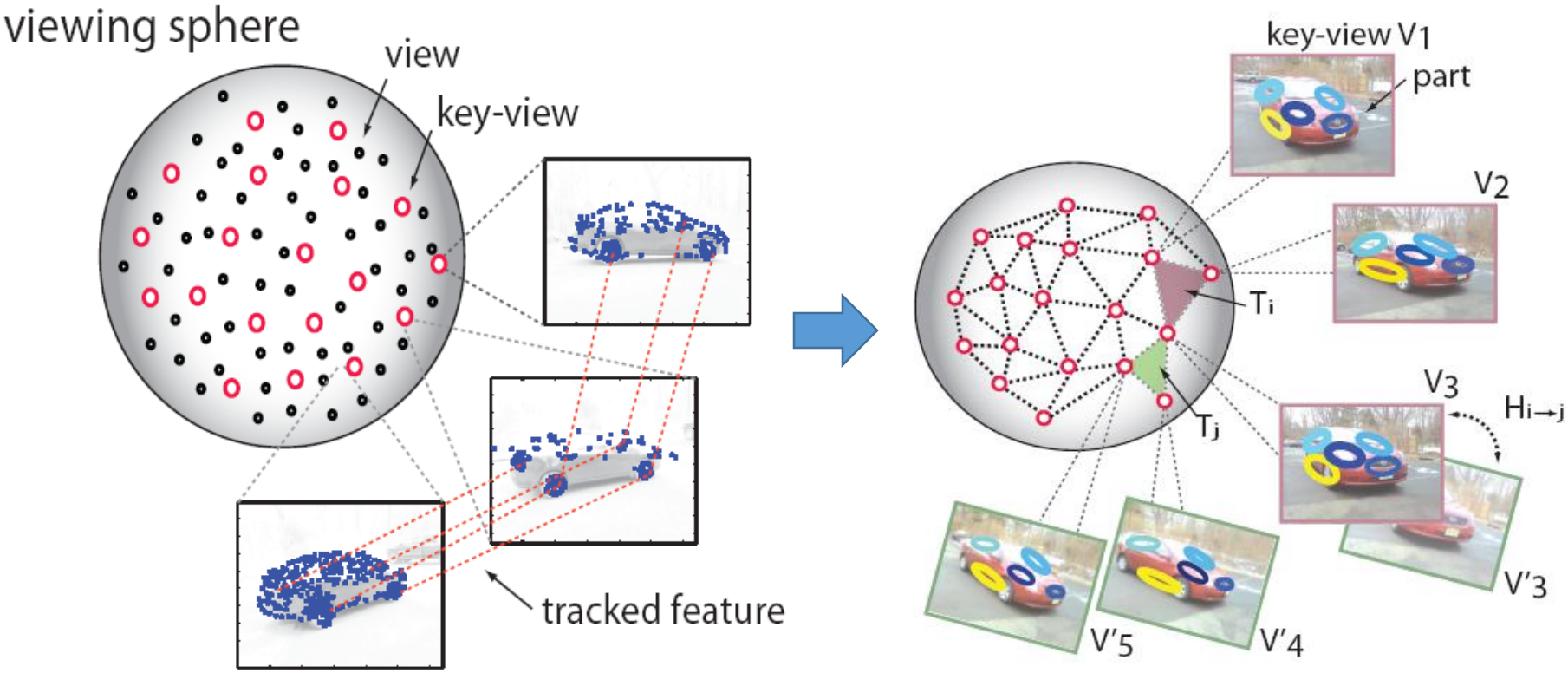
root filters coarse resolution | part filters finer resolution | deformation models

Felzenszwalb et al. 2008

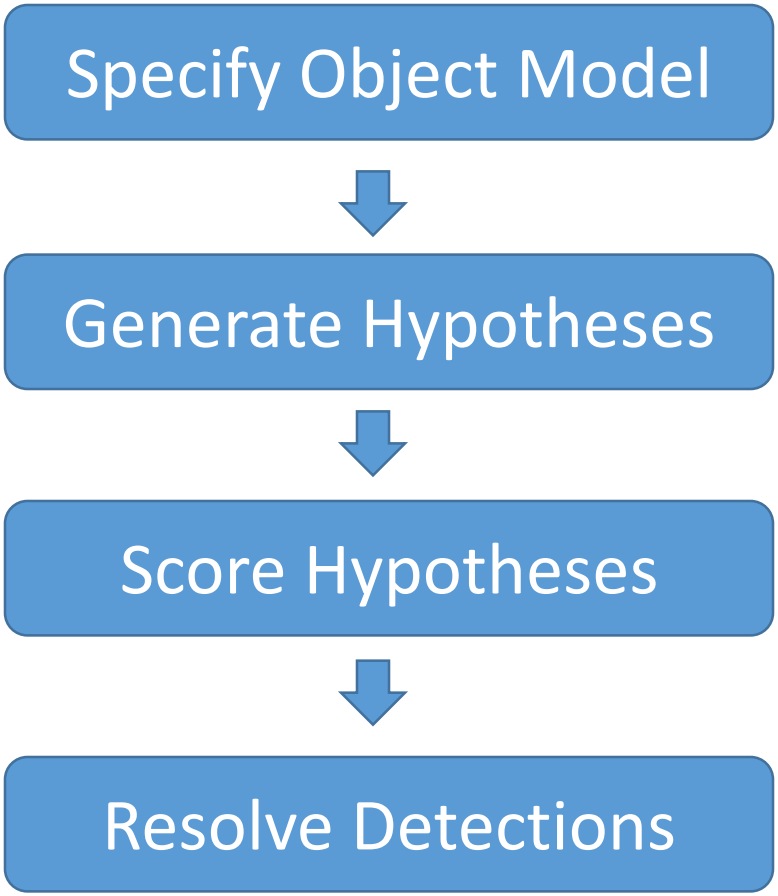
Specifying an object model

4. 3D-ish model

- Object is collection of 3D planar patches under affine transformation



General Process of Object Recognition



Propose an alignment of the model to the image

Generating hypotheses

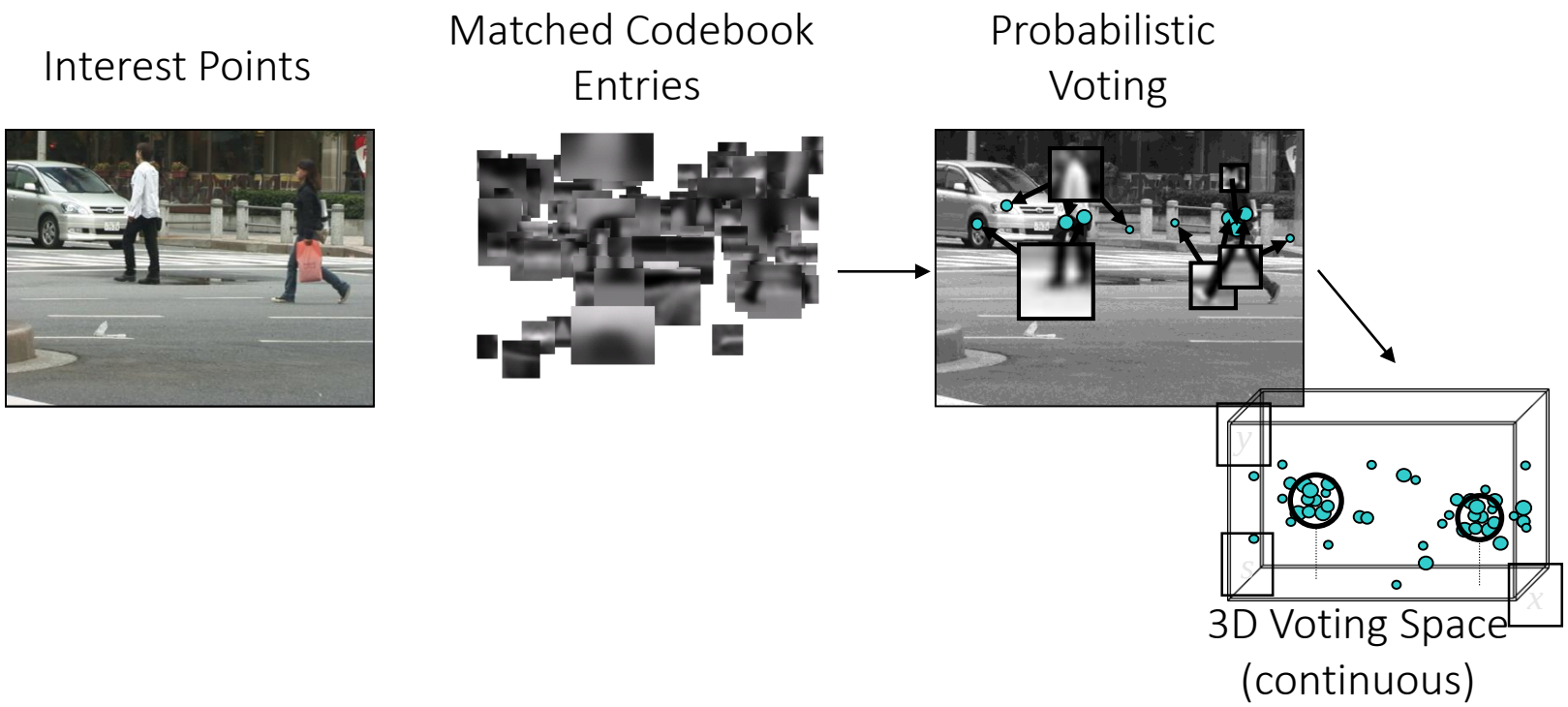
1. Sliding window
   - Test patch at each location and scale

Generating hypotheses

1. Sliding window
   - Test patch at each location and scale

Generating hypotheses

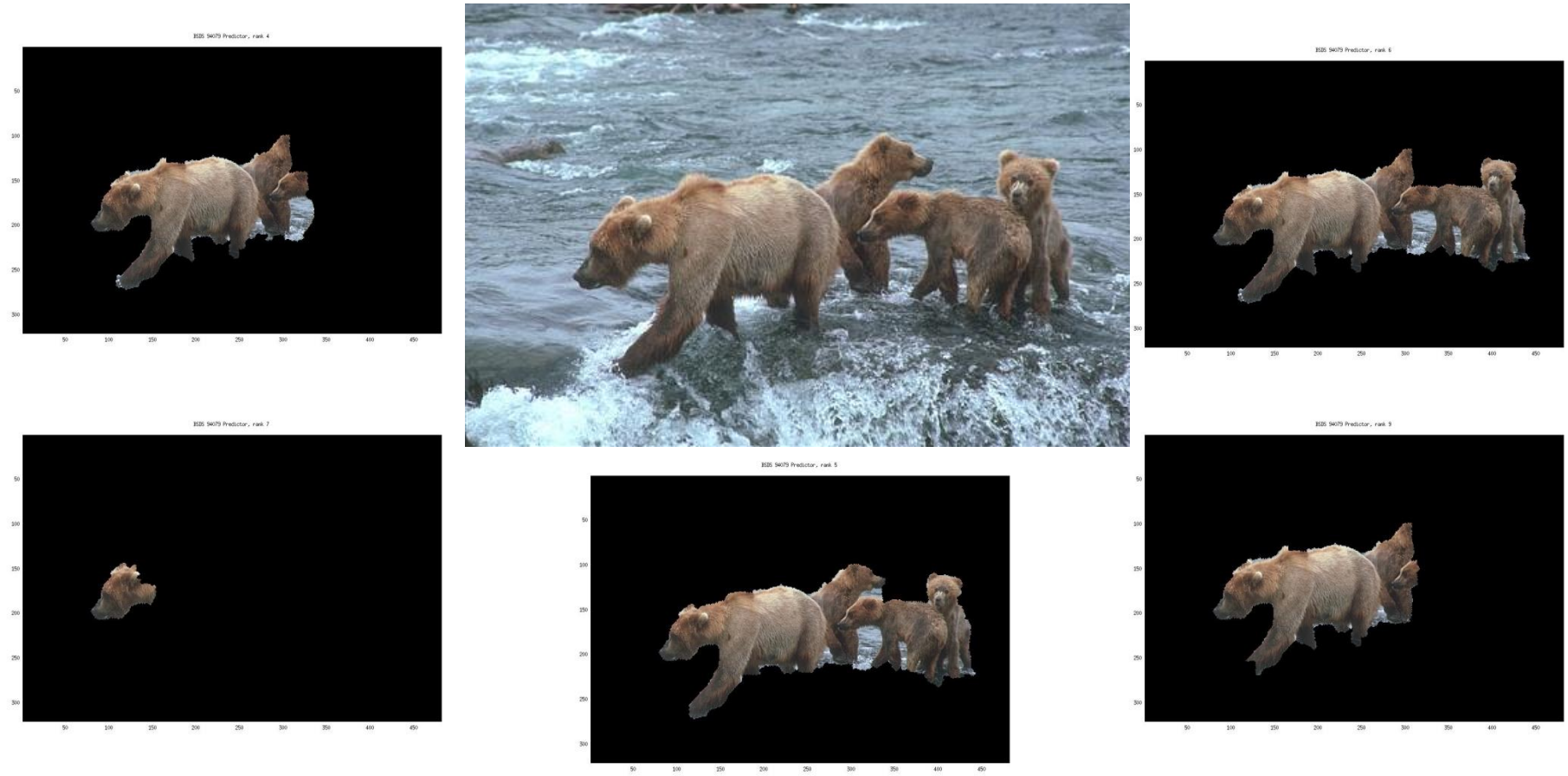
2. Voting from patches/keypoints



ISM model by Leibe et al.

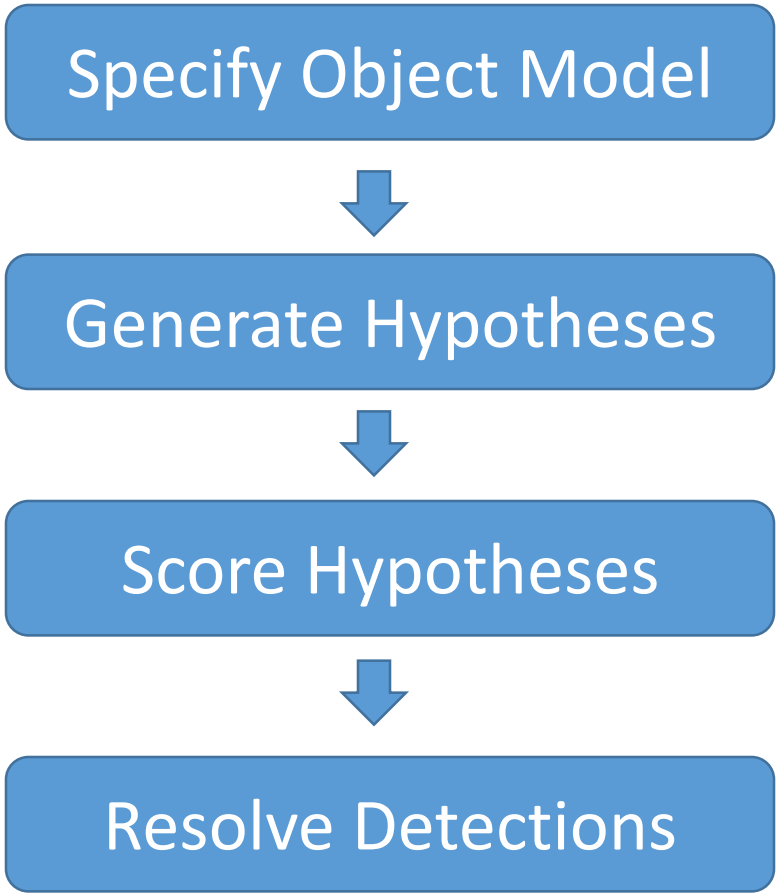
Generating hypotheses

3. Region-based proposal



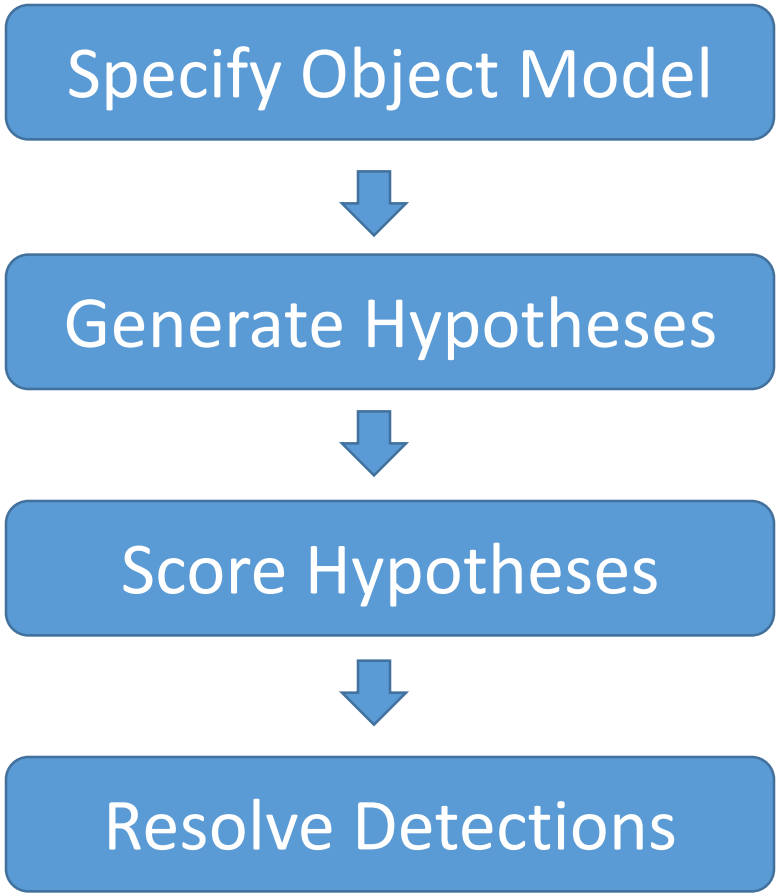
Endres Hoiem 2010

General Process of Object Recognition



Mainly-gradient based or CNN features, usually based on summary representation, many classifiers

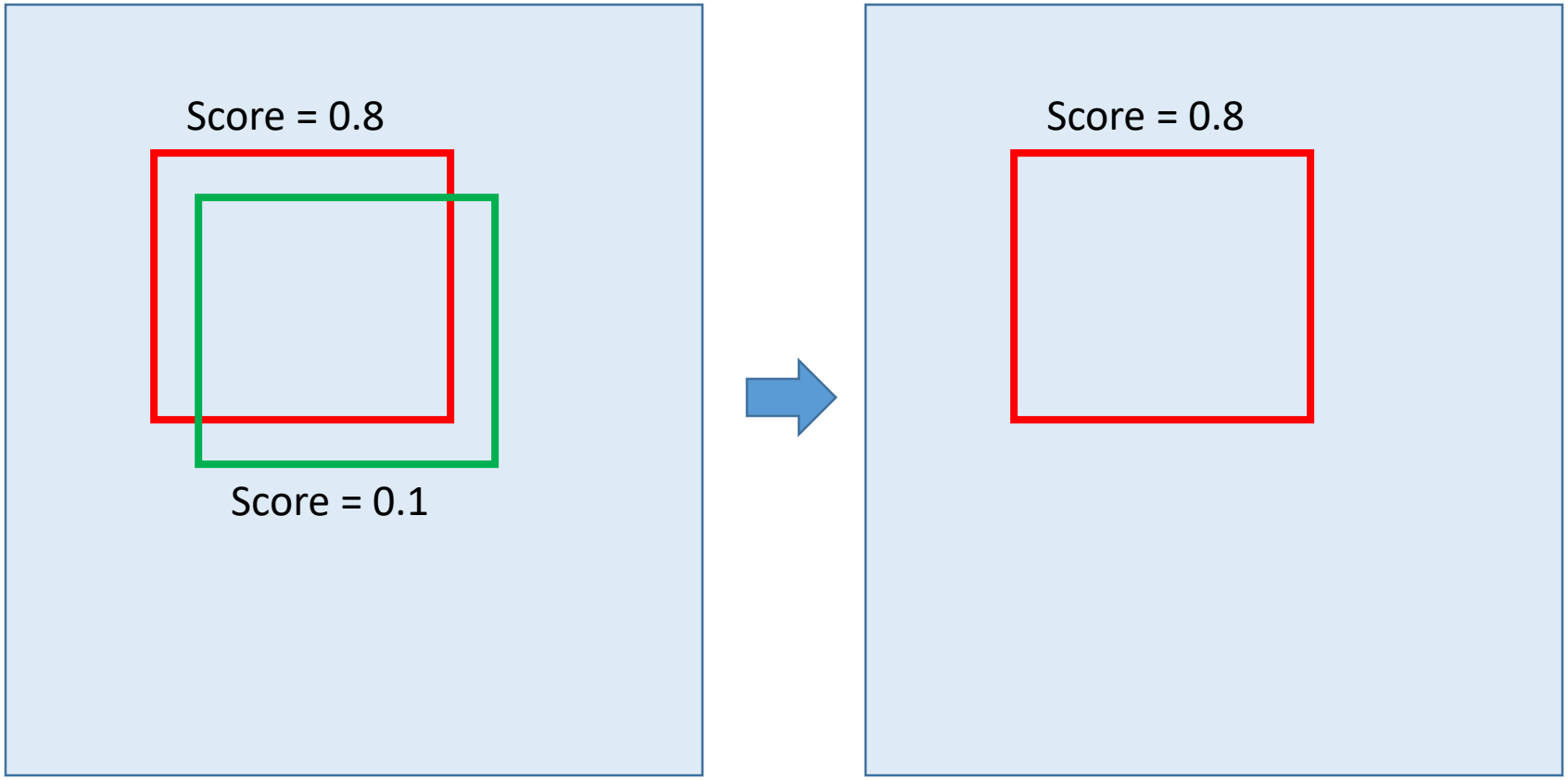
General Process of Object Recognition



Rescore each proposed object based on whole set

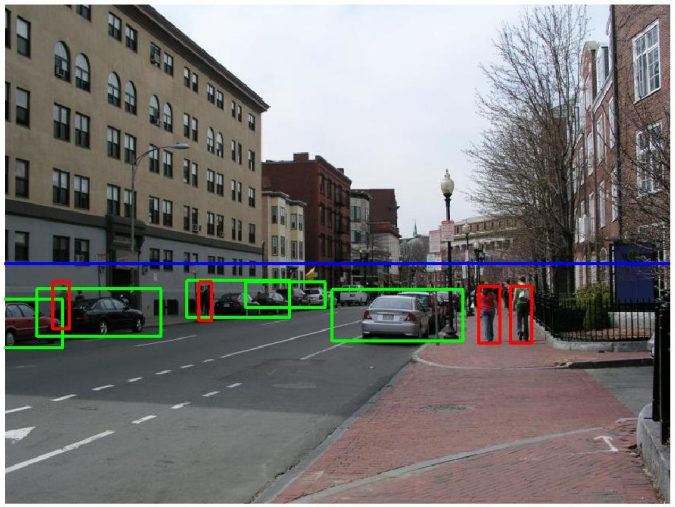
Resolving detection scores

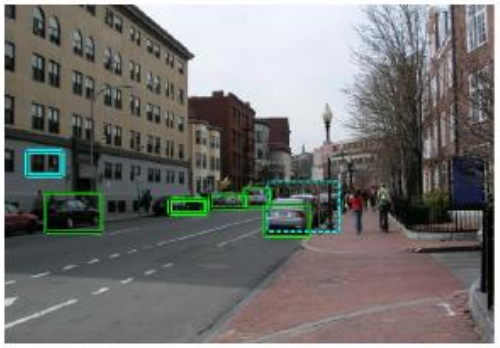
1. Non-max suppression



Resolving detection scores

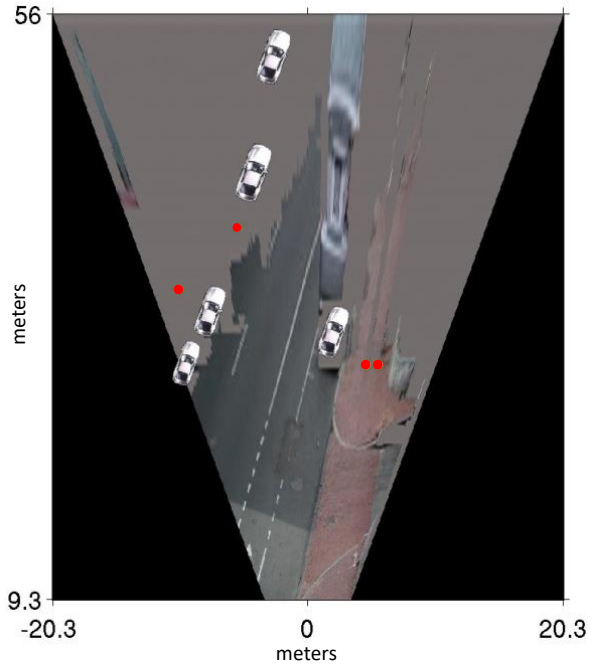
2. Context/reasoning





(g) Car Detections: Local

(h) Ped Detections: Local



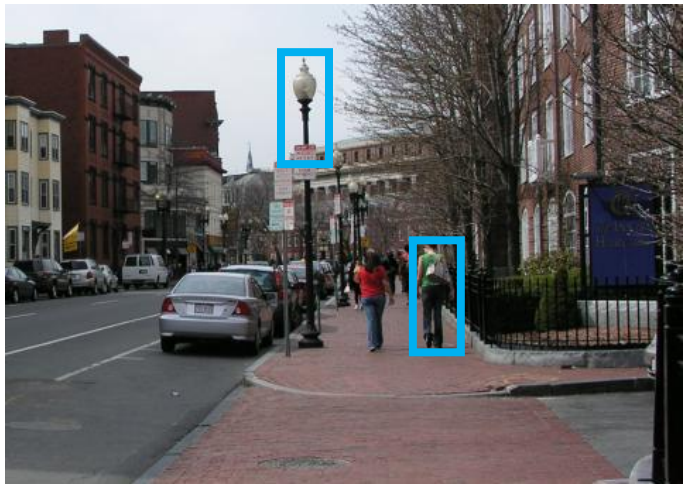
Hoiem et al. 2006

Object category detection in computer vision

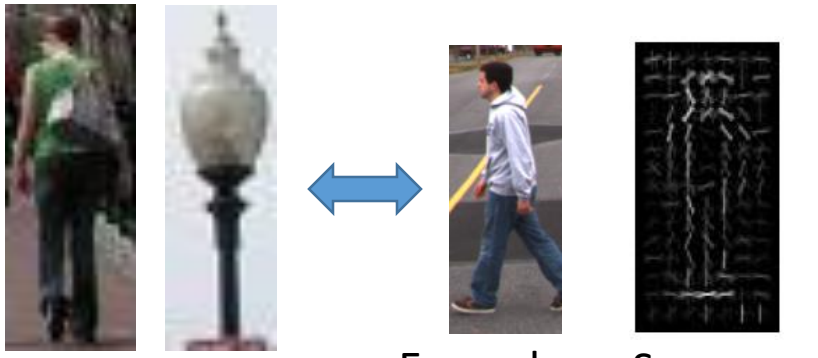
Goal: detect all pedestrians, cars, monkeys, etc in image

Basic Steps of Category Detection

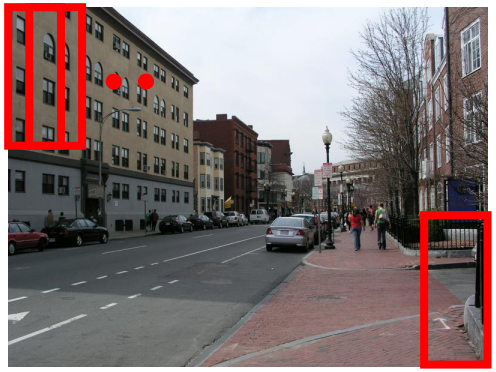
1. Align
   - E.g., choose position, scale orientation
   - How to make this tractable?



2. Compare
   - Compute similarity to an example object or to a summary representation
   - Which differences in appearance are important?



Sliding window: a simple alignment solution



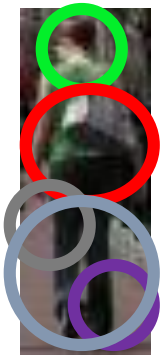
Each window is separately classified

Statistical Template

- Object model = sum of scores of features at fixed positions

+3 +2 -2 -1 -2.5 = -0.5 ? > 7.5

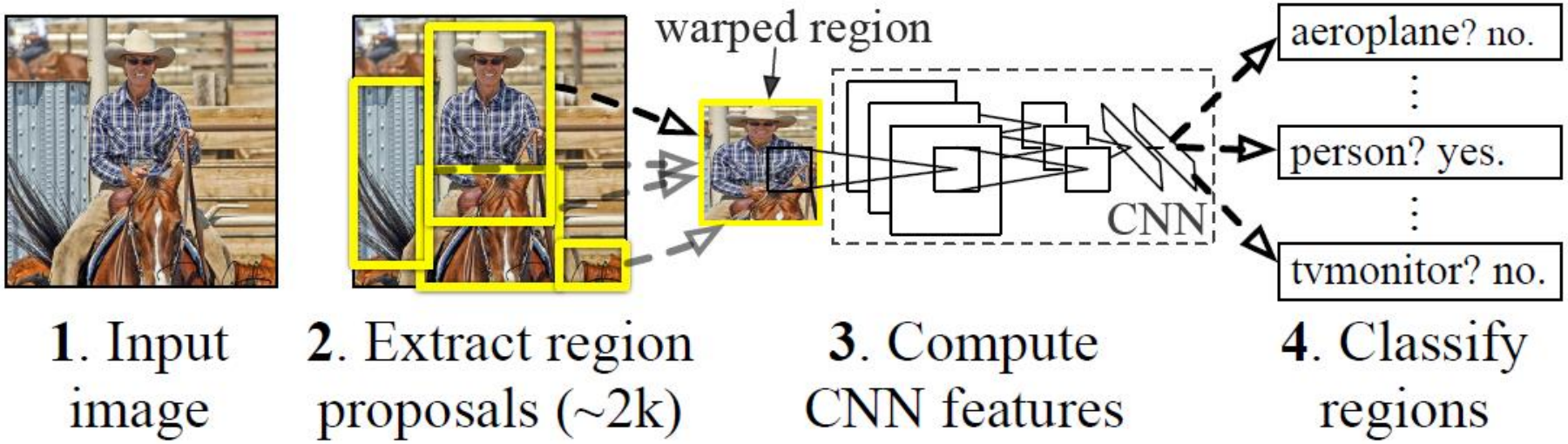
Non-object



+4 +1 +0.5 +3 +0.5 = 10.5 ? > 7.5

Object

R-CNN (Girshick et al. CVPR 2014)



- Replace sliding windows with “selective search” region proposals (Uijilings et al. IJCV 2013)
- Extract rectangles around regions and resize to 227x227
- Extract features with fine-tuned CNN (that was initialized with network trained on ImageNet before training)
- Classify last layer of network features with SVM

http://arxiv.org/pdf/1311.2524.pdf

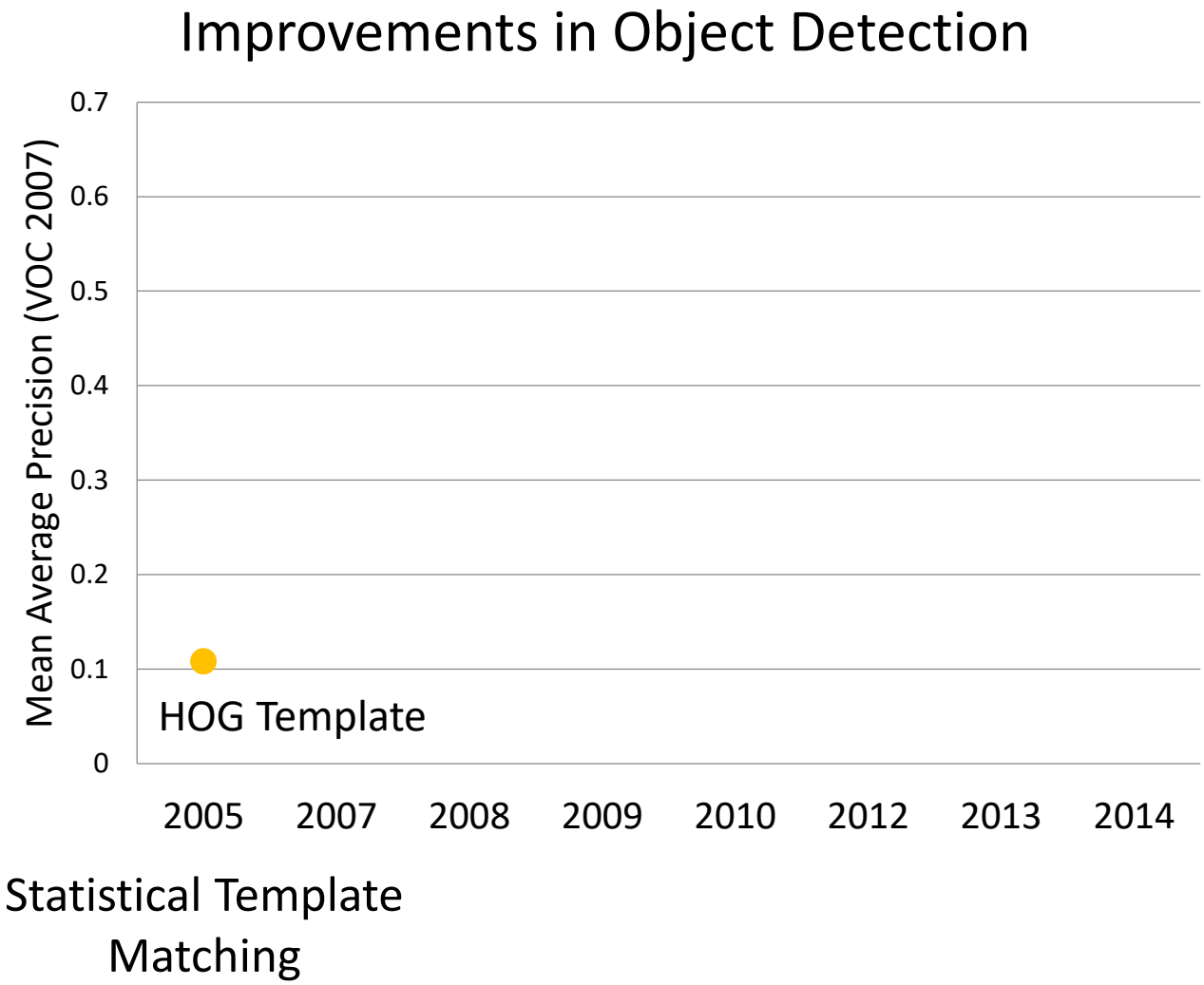
Sliding window vs. region proposals

Sliding window

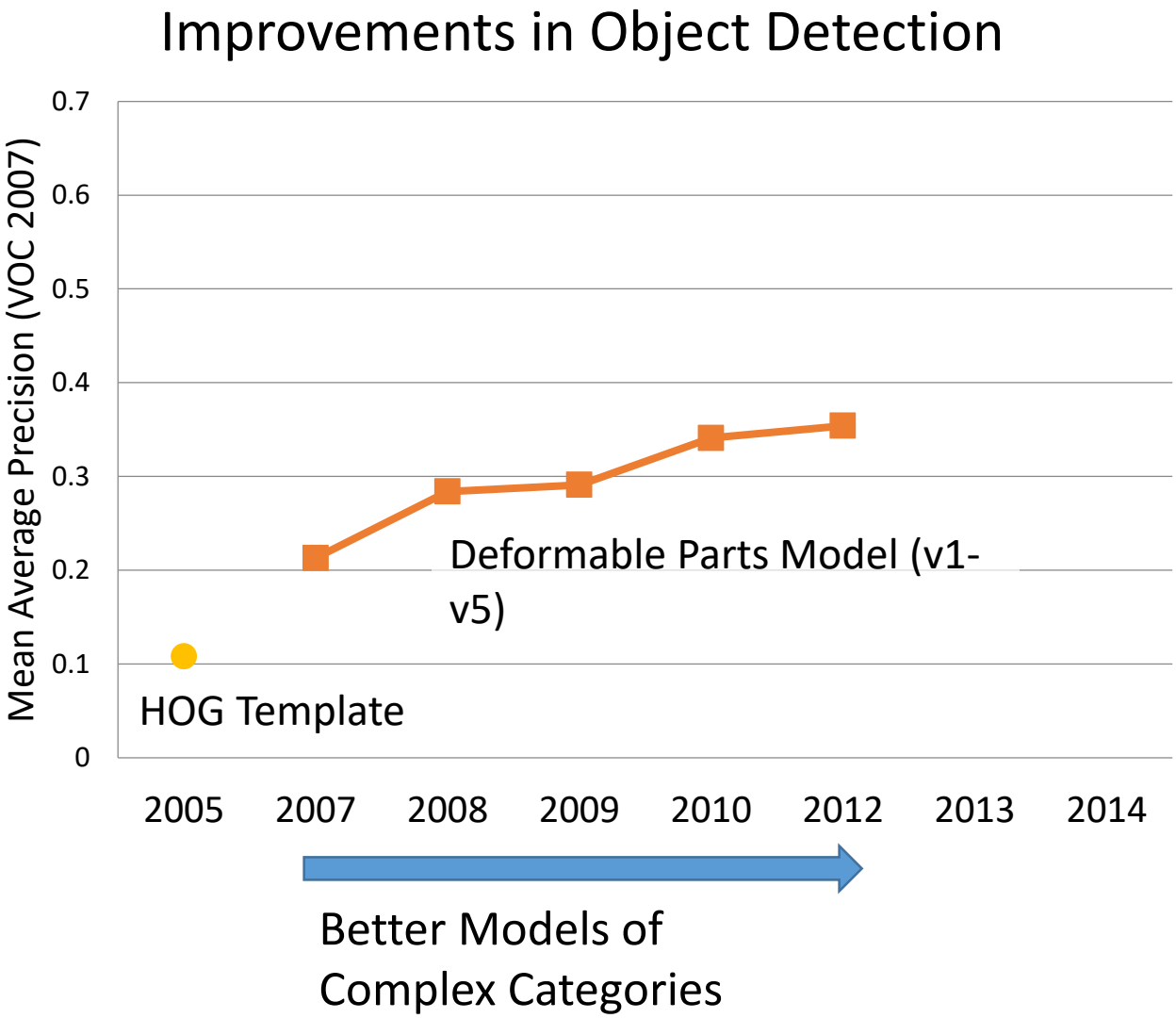
- Comprehensive search over position, scale (sometimes aspect, though expensive)
- Typically 100K candidates
- Simple
- Speed boost through convolution often possible
- Repeatable
- Even with many candidates, may not be a good fit to object

Region proposals

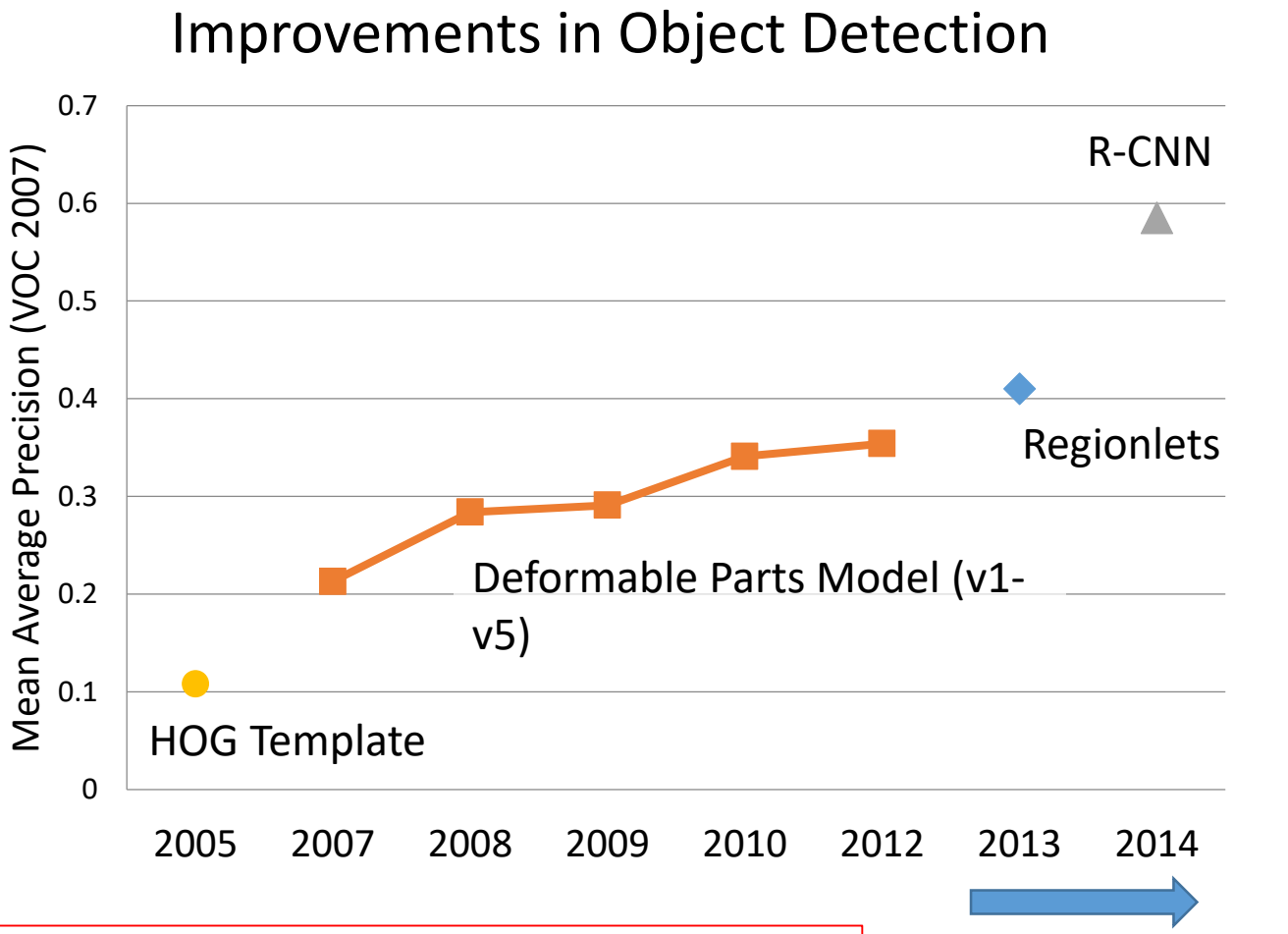
- Search over regions guided by image contours/patterns with varying aspect/size
- Typically 2-10K candidates
- Random (not repeatable)
- Requires a preprocess (currently 1-5s)
- Often requires resizing patch to fit fixed size
- More likely to provide candidates with very good object fit



HOG: Dalal-Triggs 2005



HOG: Dalal-Triggs 2005 DPM: Felzenszwalb et al. 2008-2012

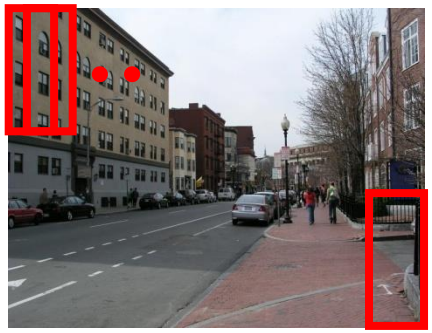


Key Advance: Learn effective features from massive amounts of labeled data *and* adapt to new tasks with less data

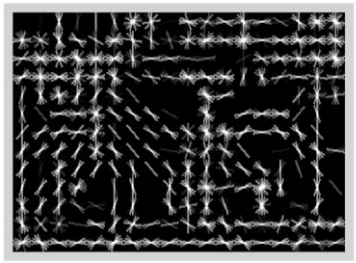
HOG: Dalal-Triggs 2005 DPM: Felzenszwalb et al. 2008-2012 Regionlets: Wang et al. 2013 R-CNN: Girshick et al. 2014

Summary: templates

Propose Window → Extract Features → Classify → Post-process



Sliding window: scan image pyramid



HOG

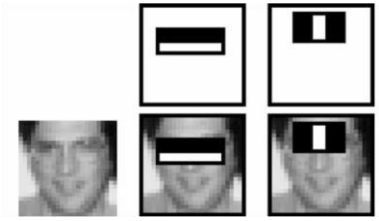
SVM

Boosted stubs

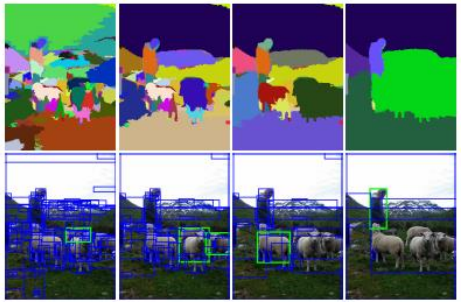
Neural network

Non-max suppression

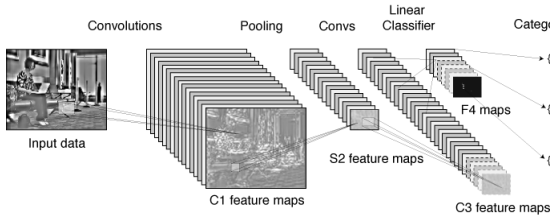
Segment or refine localization



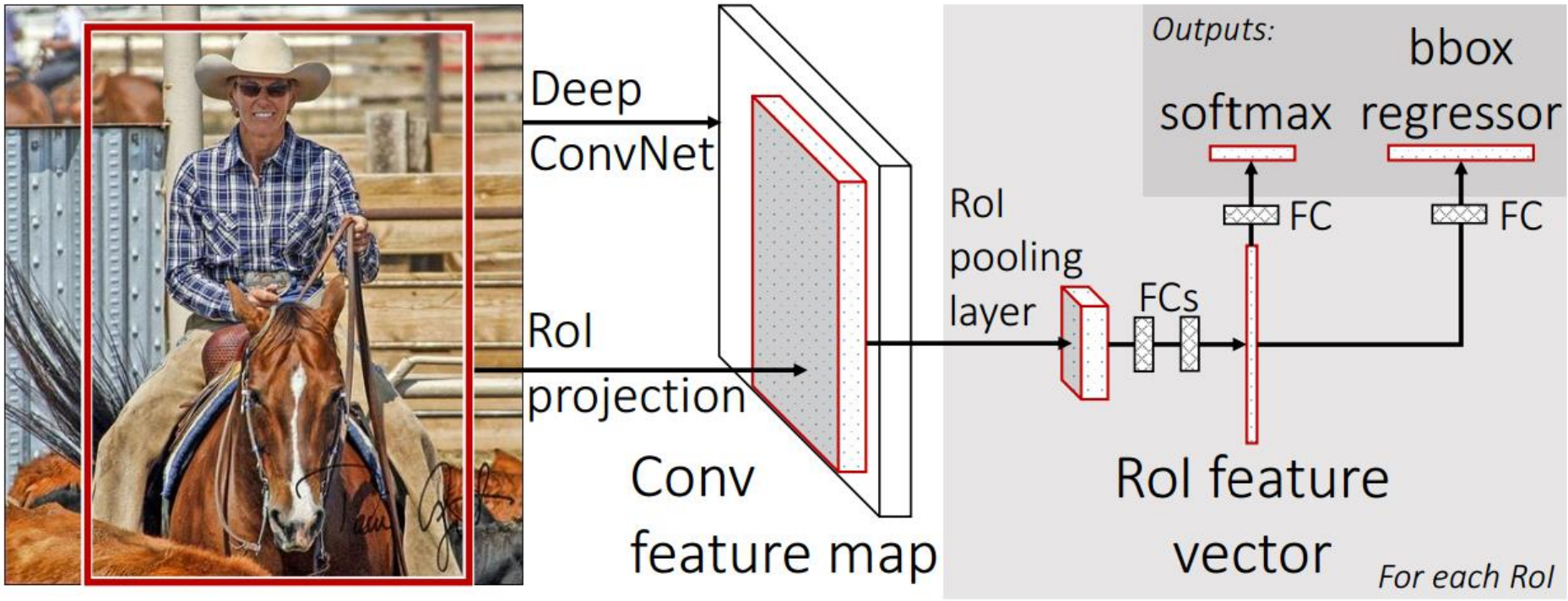
Fast randomized features



Region proposals: edge/region-based, resize to fixed window

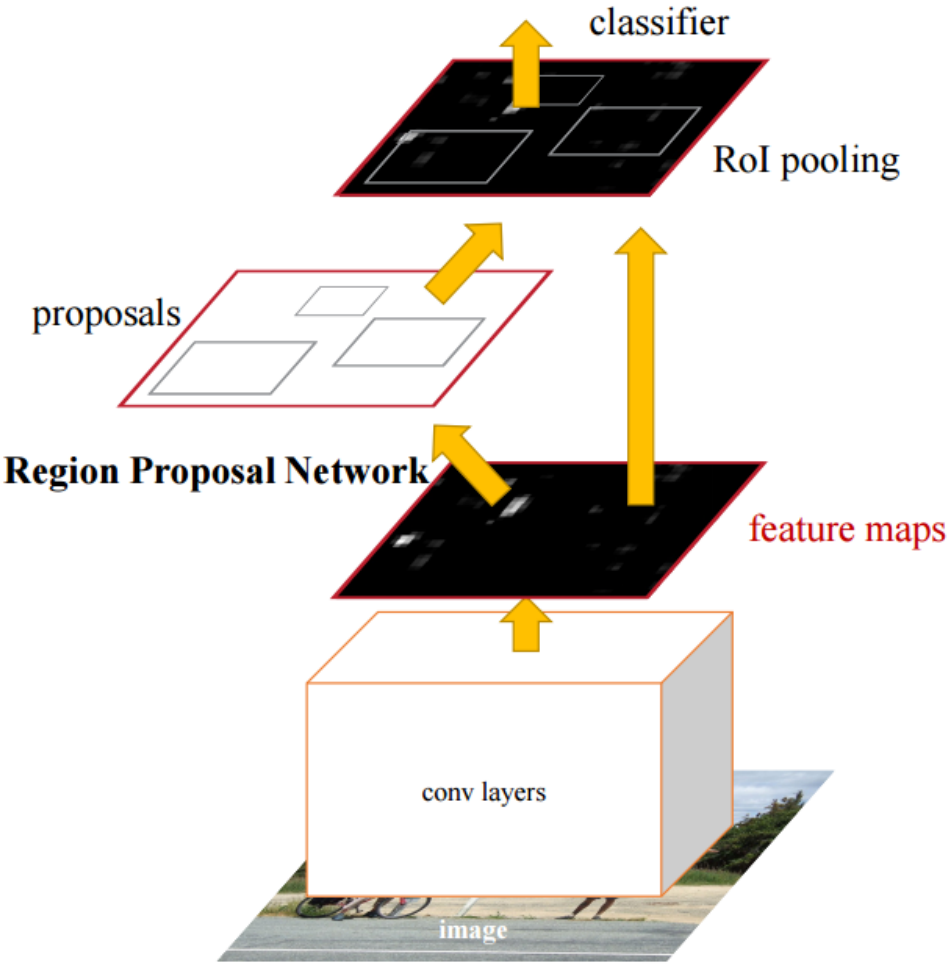


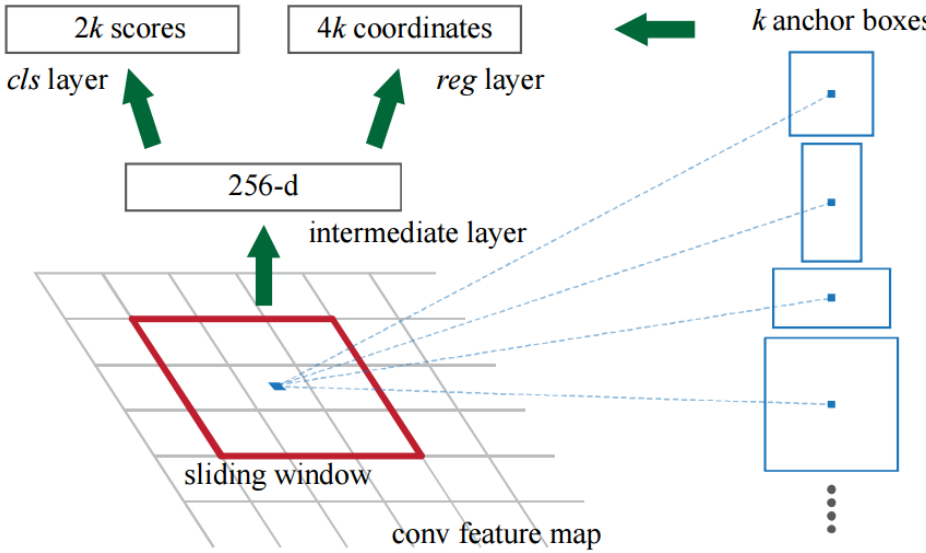
CNN features



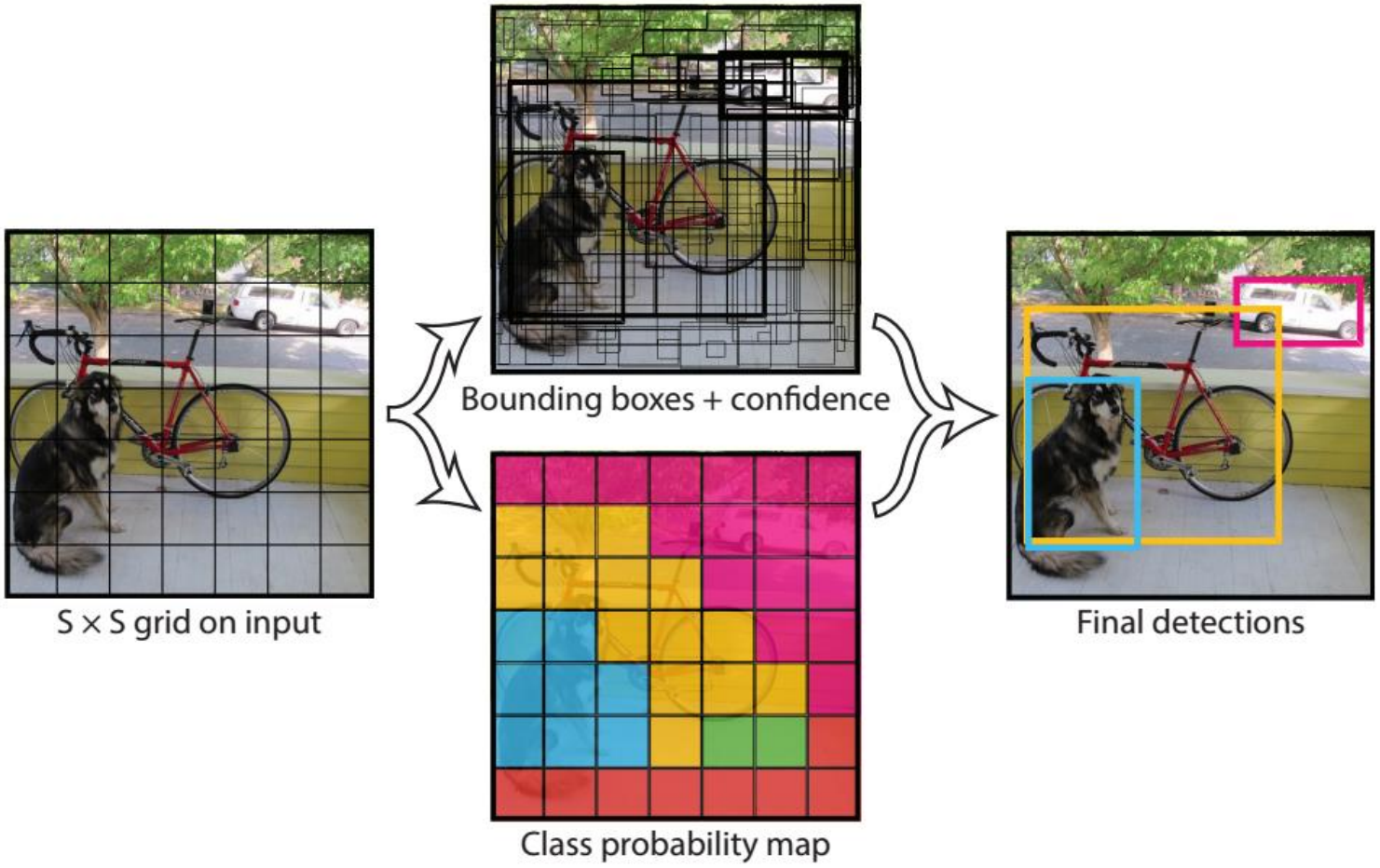
method	train set	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
SPPnet BB [11]†	07 \ diff	73.9	72.3	62.5	51.5	44.4	74.4	73.0	74.4	42.3	73.6	57.7	70.3	74.6	74.3	54.2	34.0	56.4	56.4	67.9	73.5	63.1
R-CNN BB [10]	07	73.4	77.0	63.4	45.4	**44.6**	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	**35.6**	66.8	67.2	70.4	**71.1**	66.0
FRCN [ours]	07	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8	66.9
FRCN [ours]	07 \ diff	74.6	**79.0**	68.6	57.0	39.3	79.5	**78.6**	81.9	**48.0**	74.0	67.4	80.5	80.7	74.1	69.6	31.8	67.1	68.4	75.3	65.5	68.1
FRCN [ours]	07+12	**77.0**	78.1	**69.3**	**59.4**	38.3	**81.6**	**78.6**	**86.7**	42.8	**78.8**	**68.9**	**84.7**	**82.0**	**76.6**	**69.9**	31.8	**70.1**	**74.8**	**80.4**	70.4	**70.0**

Fast RCNN, ICCV 2015

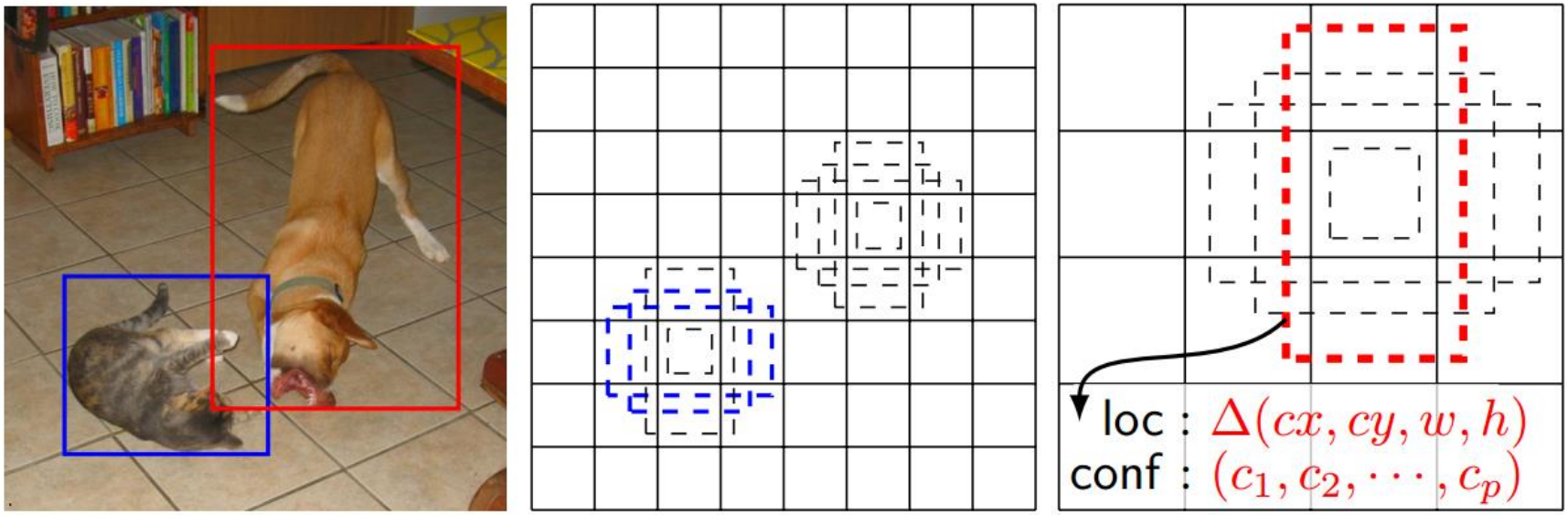




[Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015]



YOLO: Real-Time Object Detection, CVPR 2016

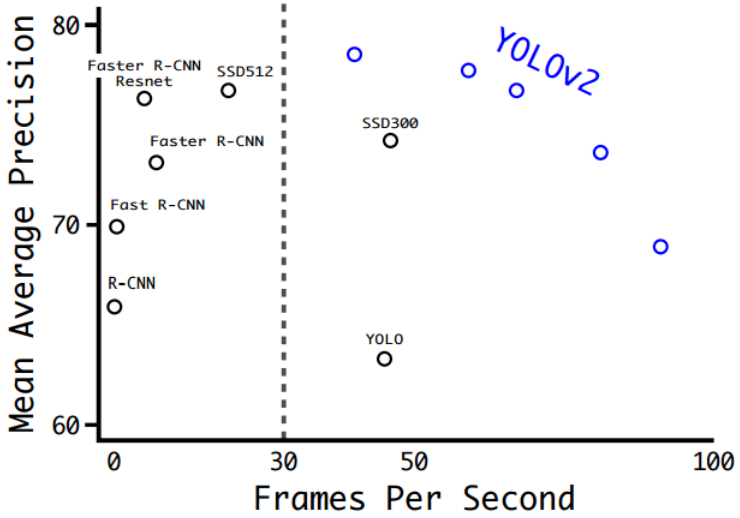


(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	$\sim 1000 \times 600$
Fast YOLO	52.7	155	1	98	448×448
YOLO (VGG16)	66.4	21	1	98	448×448
SSD300	74.3	46	1	8732	300×300
SSD512	76.8	19	1	24564	512×512
SSD300	74.3	59	8	8732	300×300
SSD512	76.8	22	8	24564	512×512

SSD: Single Shot MultiBox Detector, ECCV 2016

	YOLO							YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓
anchor boxes?				✓	✓			
new network?					✓	✓	✓	✓
dimension priors?						✓	✓	✓
location prediction?						✓	✓	✓
passthrough?							✓	✓
multi-scale?								✓
hi-res detector?								✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	**78.6**



YOLO9000: Better, Faster, Stronger, arXiv 2016

YOLO9000: Better, Faster, Stronger, arXiv 2016

Things to remember

- Specify Object Model
  - Statistical template
- Generate Hypotheses
  - Sliding windows
  - Object proposal algorithms
  - Region proposal network
- Score Hypotheses
  - CNN
- Resolve Detections
  - Non-maximum suppression