CS4670/5670: Computer Vision

Kavita Bala

Lecture 41: Recognition Beyond Classification

Thanks to Andrej Karpathy

Announcements

- Office hours till next Friday
- Drop me a note if you want to meet

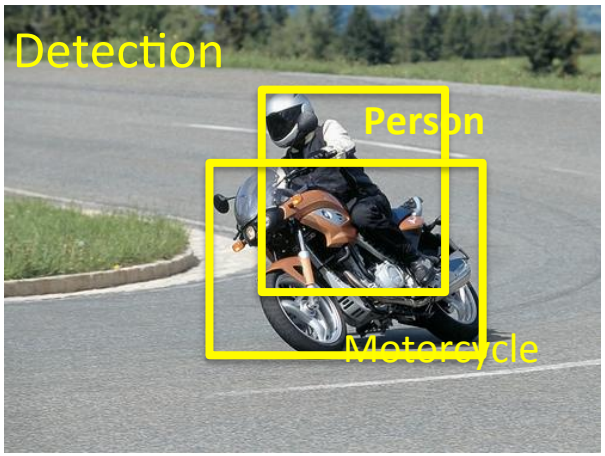
Announcements

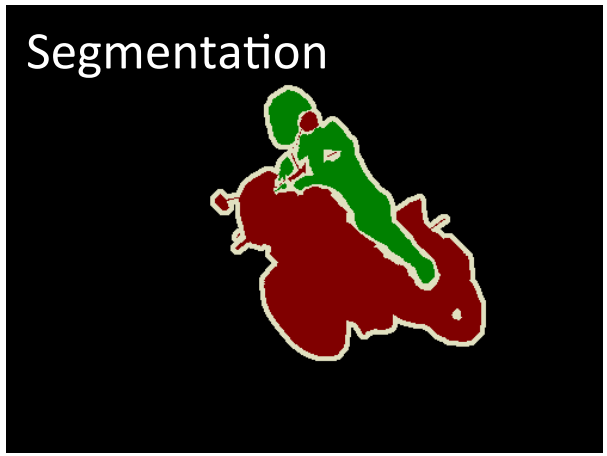
- Exam review will be announced
- Exam topics
  - Lec 18 (Cameras, Mar 9)-lec 39 (ConvNets, May 4)
- Grading: any unresolved issues drop me a note
- **Final: Sunday 2pm, May 22**
  - **BTN100WEST,** Barton Hall 100 West-Main Floor

The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

Three (+2) challenges:

- Classification challenge (is there an X in this image?)
- Detection challenge (draw a box around every X)
- Segmentation challenge (which class is each pixel?)





Slides from Noah Snavely

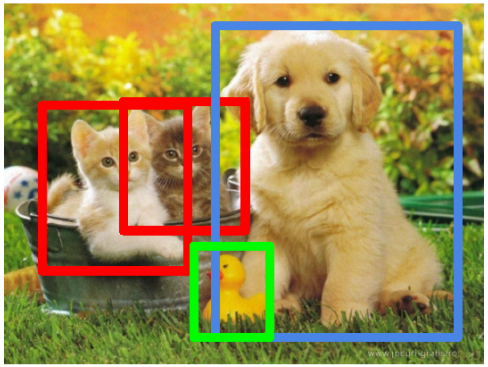
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Instance Segmentation

CAT, DOG, DUCK

Single object

Multiple objects

Classification: C classes

Input: Image

Output: Class label

Evaluation metric: Accuracy

CAT

Localization:

Input: Image

Output: Box in the image (x, y, w, h)

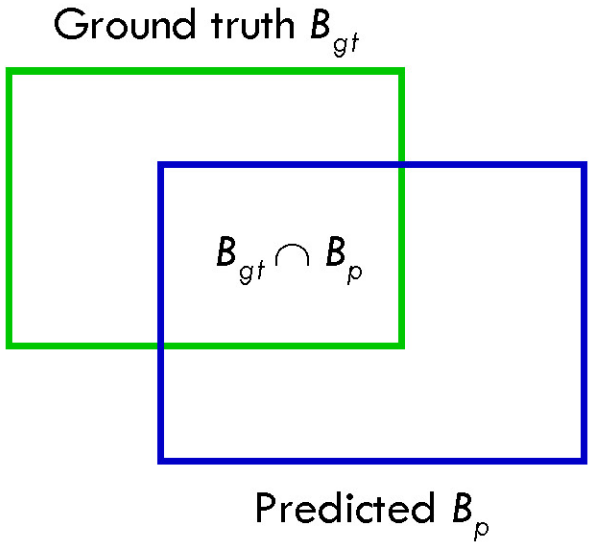
Evaluation metric: Intersection over Union

(x, y, w, h)

Classification + Localization: Do both

Remember from Lec32

- Area of Overlap (AO) Measure



$$AO(B_{gt}, B_p) = \frac{|B_{gt} \bigcap B_p|}{|B_{gt} \bigcup B_p|}$$

  - Need to define a threshold *t* such that $AO(B_{gt}, B_p)$ implies a correct detection: 50%

Classification+Localization: ImageNet

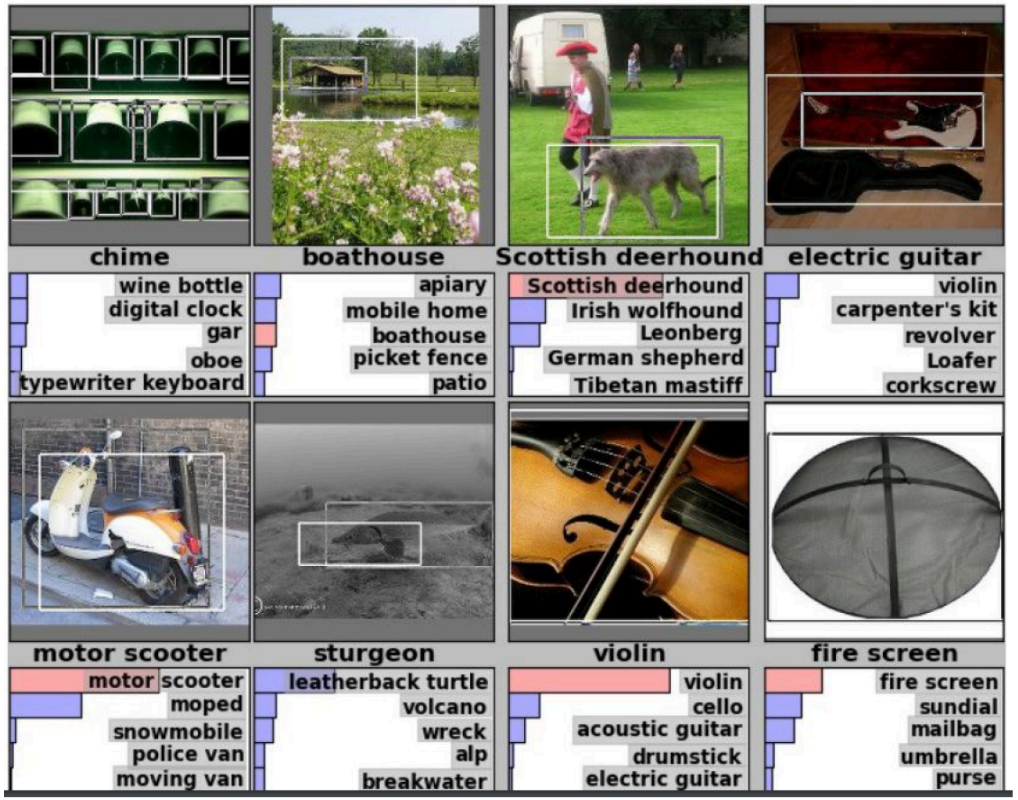
1000 classes (same as classification)

Each image has 1 class, at least one bounding box

~800 training images per class

Algorithm produces 5 (class, box) guesses

Example is correct if at least one one guess has correct class AND bounding box at least 0.5 intersection over union (IoU)



Krizhevsky et. al. 2012

Localization for one object

Idea #1: Localization as Regression

Input: image

Only one object,
simpler than detection

Neural Net

Output:
Box coordinates
(4 numbers)

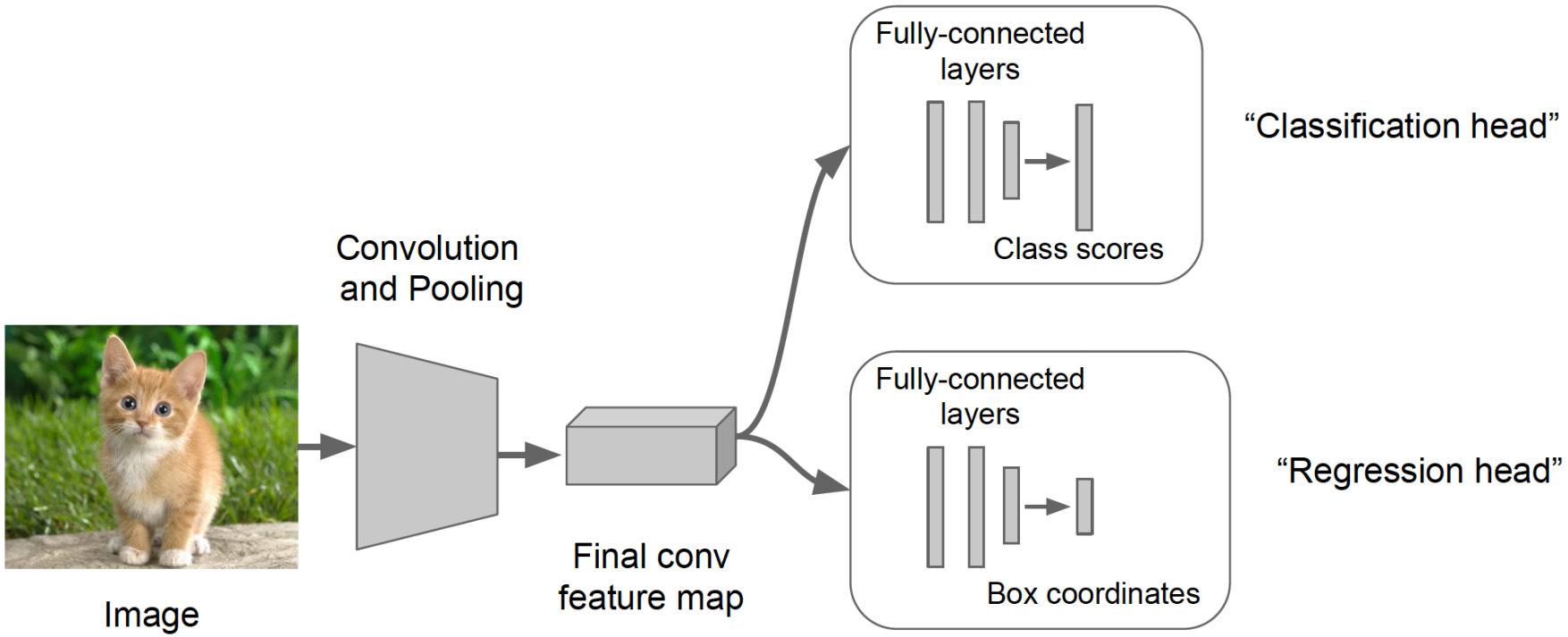
Correct output:
box coordinates
(4 numbers)

Loss:
L2 distance

Classification+Localization for one object

Step 1: Train (or download) a classification model (AlexNet, VGG, GoogLeNet)

Step 2: Attach new fully-connected "regression head" to the network



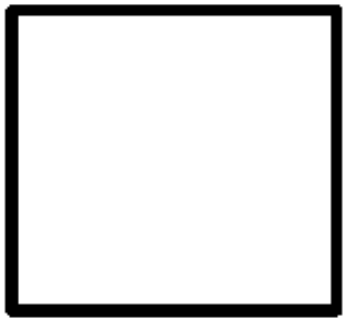
Step 3: Train the regression head only with SGD and L2 loss

Step 4: At test time use both heads

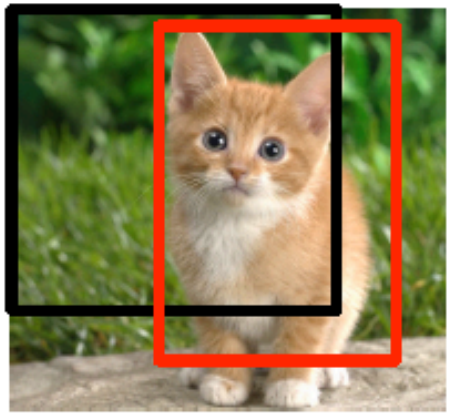
Localization as regression: simple but powerful

Idea #2: Sliding Window

- Run classification + regression network at multiple locations on a high-resolution image
- Convert fully-connected layers into convolutional layers for efficient computation
- Combine classifier and regressor predictions across all scales for final prediction



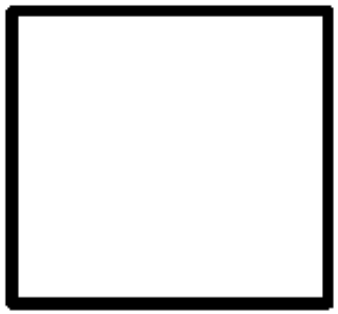
Network input:
3 x 221 x 221



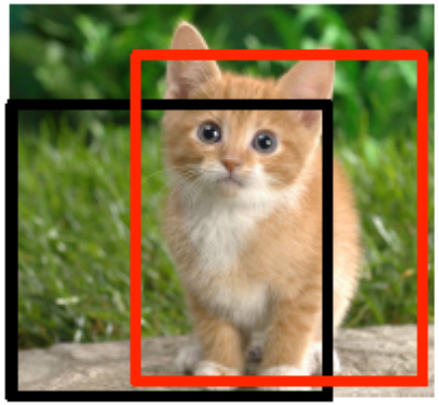
Larger image:
3 x 257 x 257

0.5	

Classification scores:
P(cat)



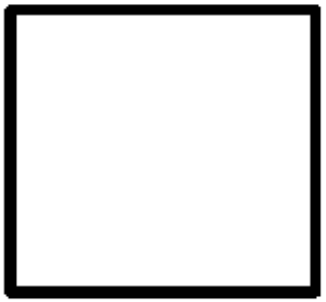
Network input:
3 x 221 x 221



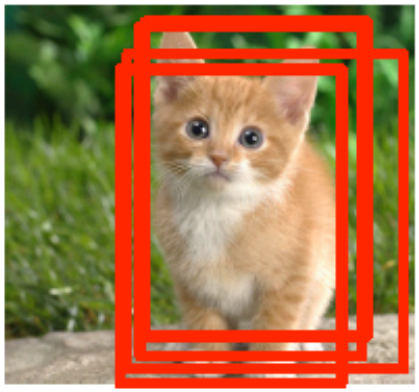
Larger image:
3 x 257 x 257

0.5	0.75
0.6	

Classification scores:
P(cat)



Network input:
3 x 221 x 221

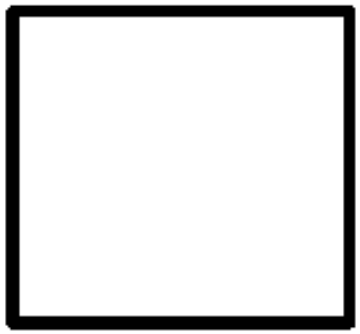


Larger image:
3 x 257 x 257

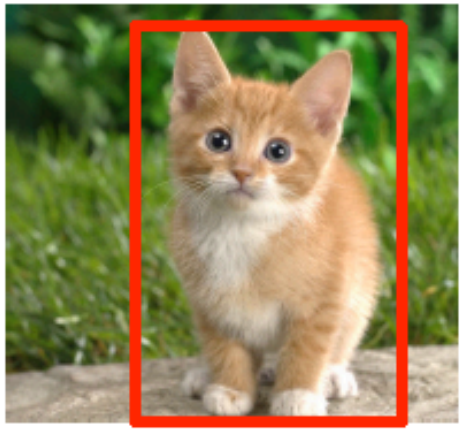
0.5	0.75
0.6	0.8

Classification scores:
P(cat)

Greedily merge boxes and
scores (details in paper)



Network input:
3 x 221 x 221



Larger image:

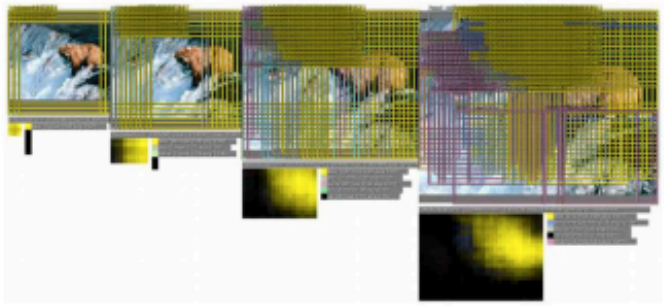
0.8

Classification score: P

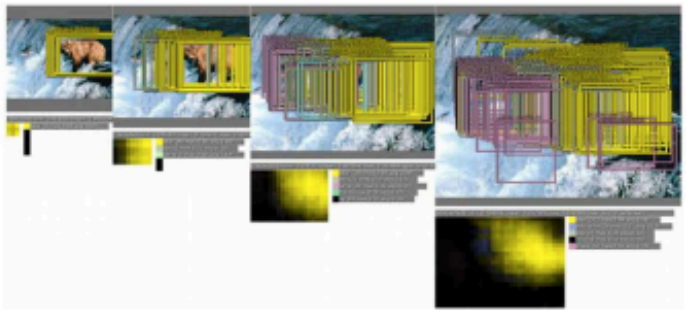
Overfeat

In practice use many sliding window locations and multiple scales

Window positions + score maps

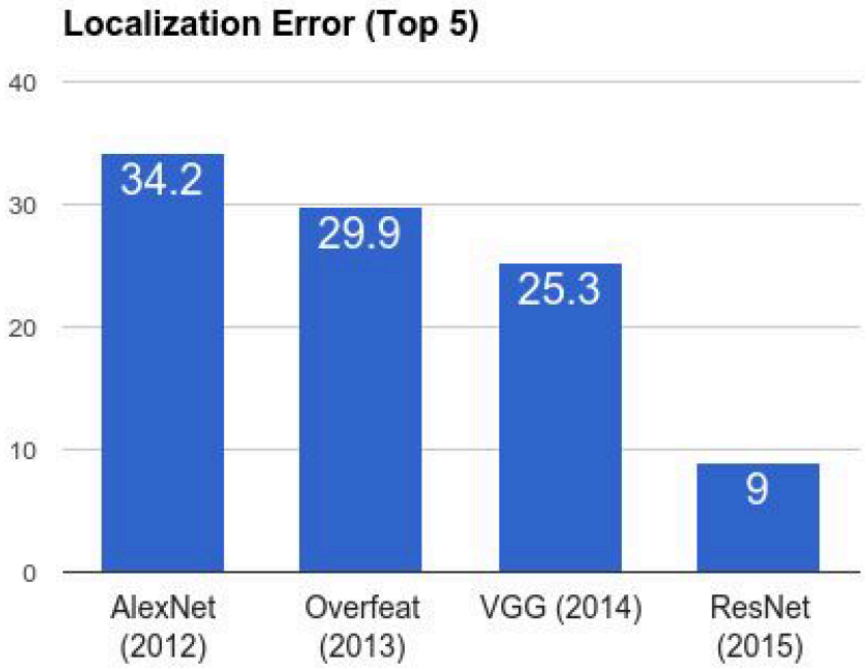


Box regression outputs



Final Predictions

Localization Error



AlexNet: Localization method not published

Overfeat: Multiscale convolutional regression with box merging

VGG: Same as Overfeat, but fewer scales and locations; simpler method, gains all due to deeper features

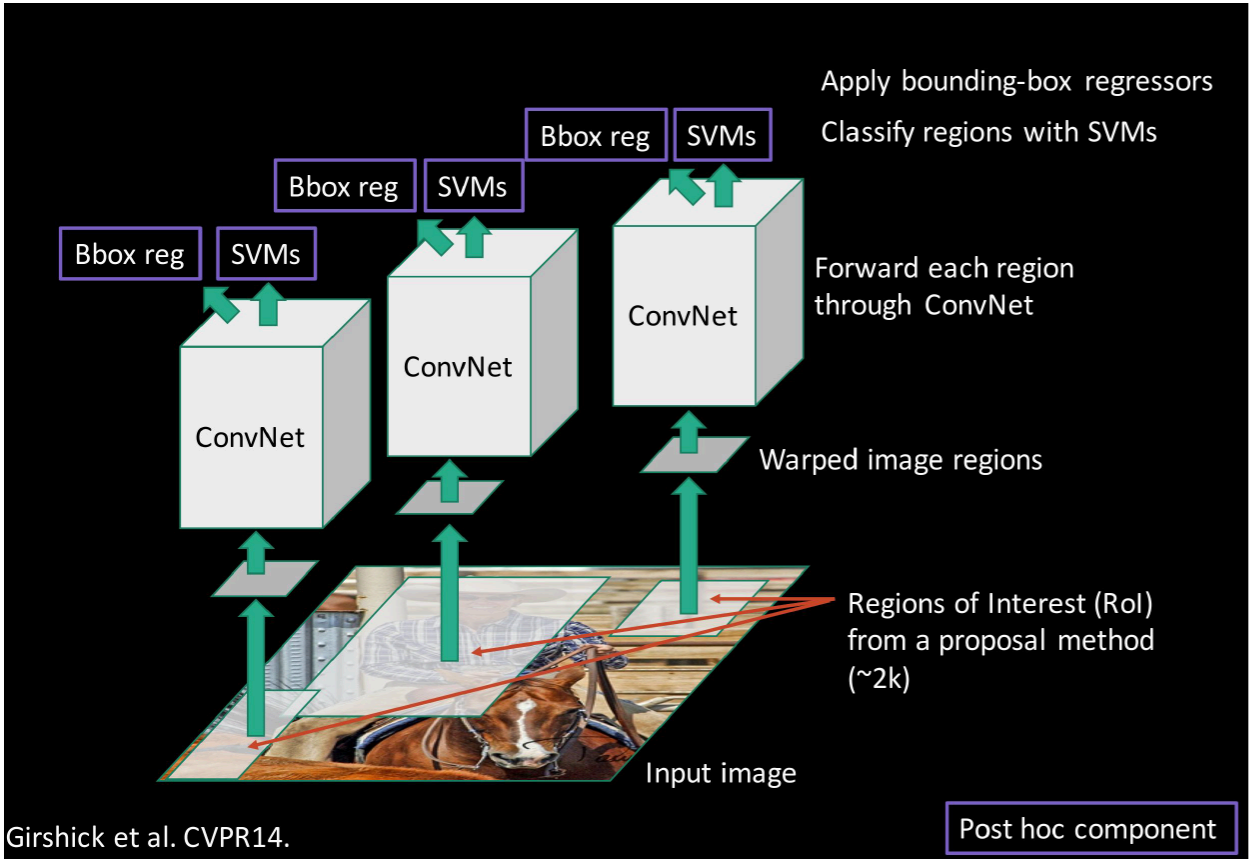
ResNet: Different localization method (RPN) and much deeper features

Object Detection

- Need to test many scales and positions
- Solution: only look at a small set of possible positions

- Approach: propose regions, then regressors

RCNN

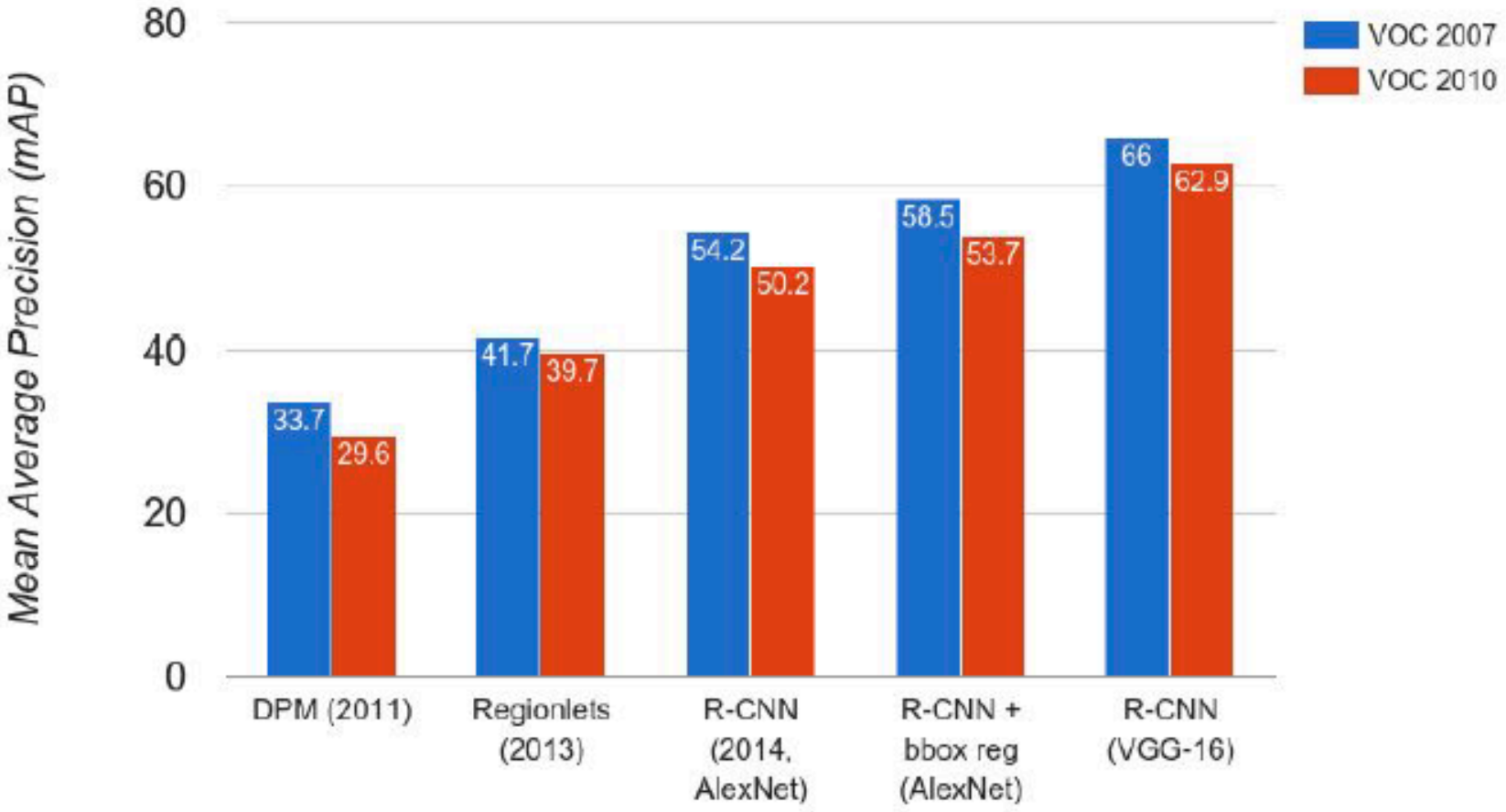


Girshick et al. CVPR14.

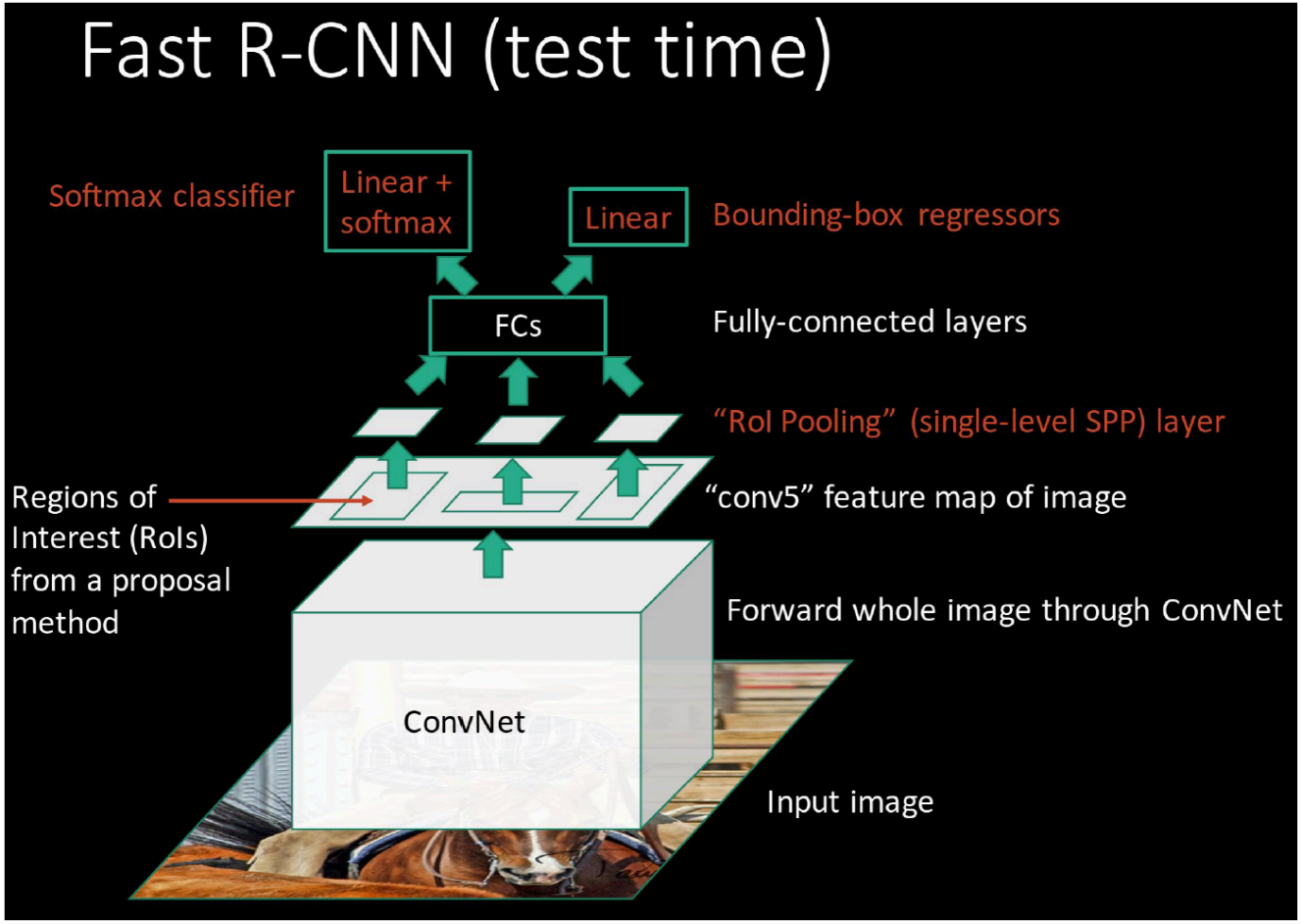
Girschick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014

Slide credit: Ross Girschick

RCNN


Mean Average Precision (mAP)
80
60
40
20
0
VOC 2007
VOC 2010
33.7
29.6
41.7
39.7
54.2
50.2
58.5
53.7
66
62.9
DPM (2011)
Regionlets (2013)
R-CNN (2014, AlexNet)
R-CNN + bbox reg (AlexNet)
R-CNN (VGG-16)

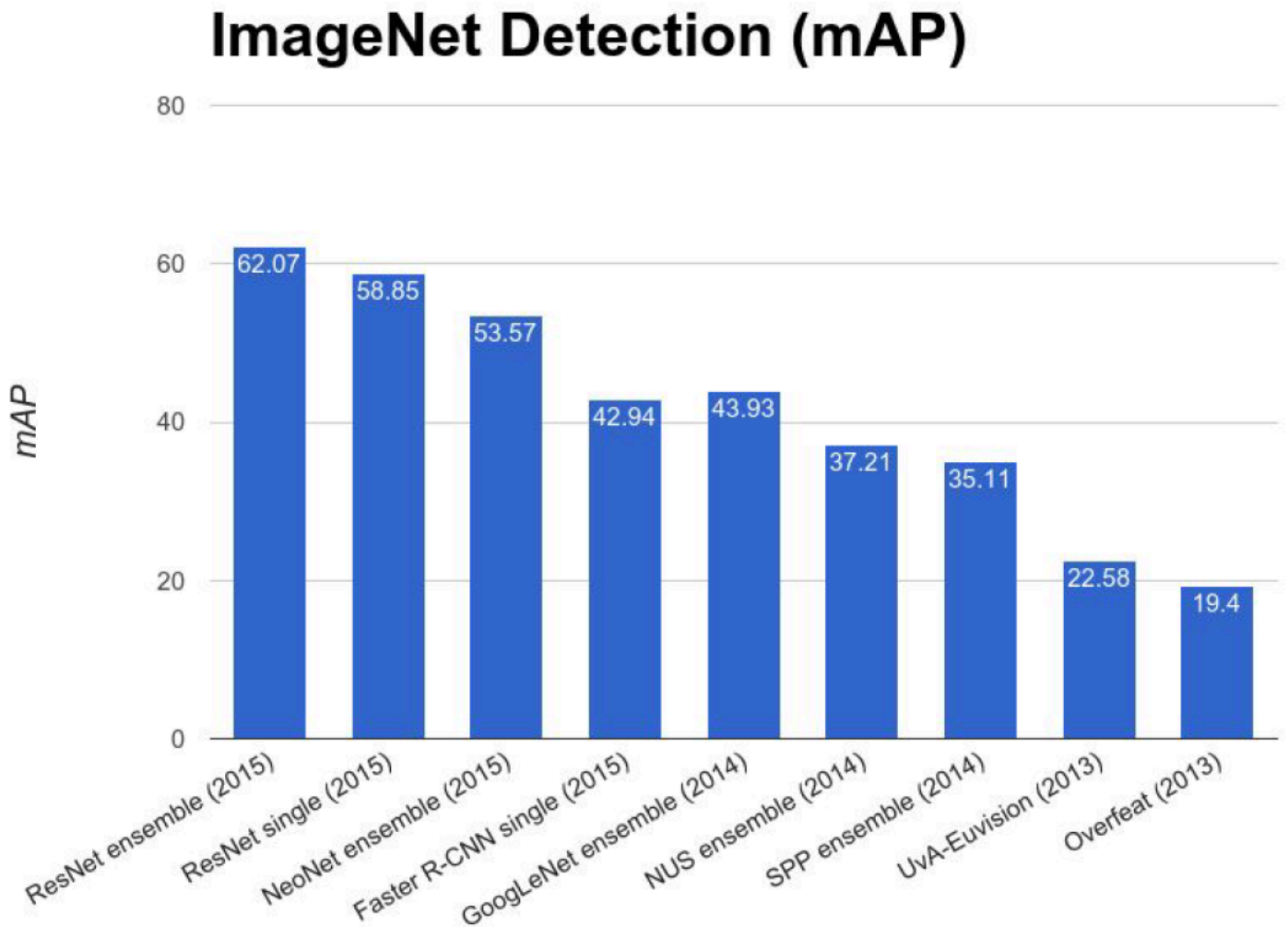
Fast R-CNN



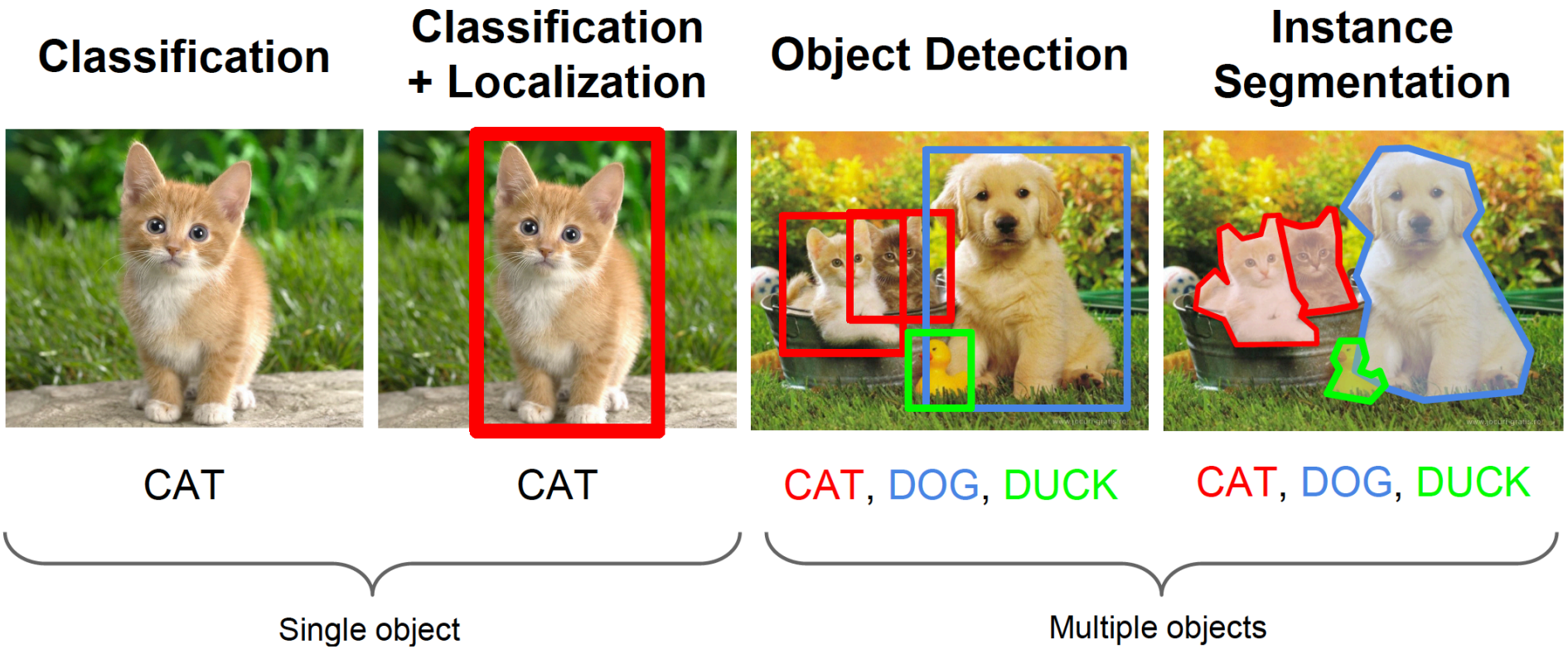
Girschick, "Fast R-CNN", ICCV 2015

Slide credit: Ross Girschick

	R-CNN	**Fast R-CNN**	**Faster R-CNN**
Test time per image (with proposals)	50 seconds	2 seconds	**0.2 seconds**
(Speedup)	1x	25x	**250x**
mAP (VOC 2007)	66.0	**66.9**	**66.9**


ImageNet Detection (mAP)
mAP
80
60
40
20
0
62.07
58.85
53.57
42.94
43.93
37.21
35.11
22.58
19.4
ResNet ensemble (2015)
ResNet single (2015)
NeoNet ensemble (2015)
Faster R-CNN single (2015)
GoogLeNet ensemble (2014)
NUS ensemble (2014)
SPP ensemble (2014)
UvA-Euvision (2013)
Overfeat (2013)

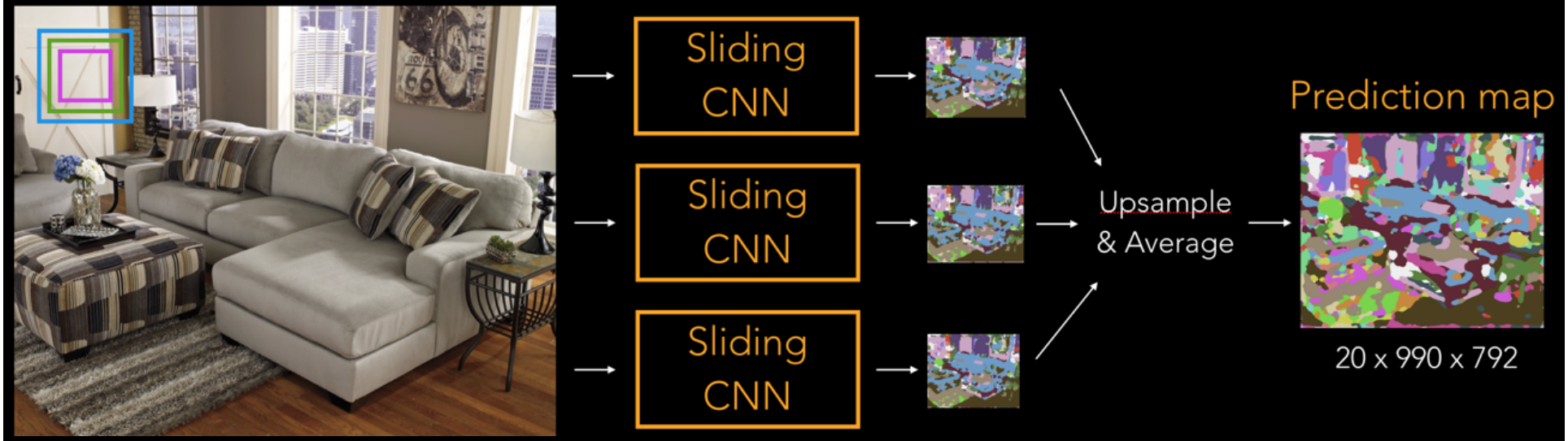
Semantic Segmentation



Deep learning for materials

- Train at different scales

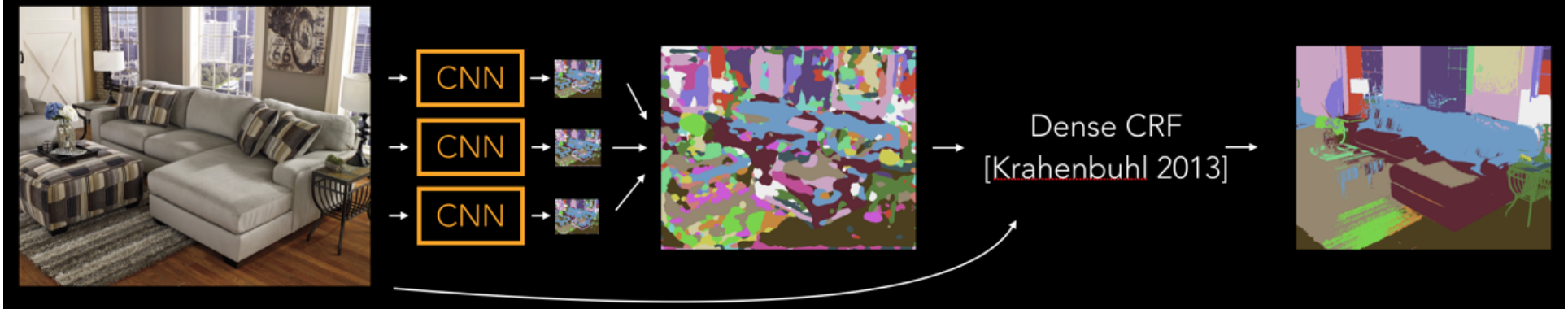
Deep learning to predict materials



Material Predictions

00: brick
01: carpet
02: ceramic
03: fabric
04: foliage
05: food
06: glass
07: hair
08: leather
09: metal
10: other
11: painted
12: paper
13: plastic
14: polishedstone
15: skin
16: tile
17: stone
18: water
19: wood

Semantic segmentation



CRF Runtime: ~1s for 640x480 image

$$E(\mathbf{x}|\mathbf{I},\boldsymbol{\theta}) = \sum_i \psi_i(x_i|\boldsymbol{\theta}) + \sum_{i<j} \psi_{ij}(x_i, x_j|\boldsymbol{\theta})$$

Semantic segmentation

00: brick
01: carpet
02: ceramic
03: fabric
04: foliage
05: food
06: glass
07: hair
08: leather
09: metal
10: other
11: painted
12: paper
13: plastic
14: polishedstone
15: skin
16: tile
17: stone
18: water
19: wood

Mean class accuracy: 84.95% out of 20 categories
Prior work: 41% out of 10 categories

Results

00: brick
01: carpet
02: ceramic
03: fabric
04: foliage
05: food
06: glass
07: hair
08: leather
09: metal
10: other
11: painted
12: paper
13: plastic
14: polishedstone
15: skin
16: tile
17: stone
18: water
19: wood

Summary

- Localization
  - Find fixed number of objects
  - L2 regression from CNN features to box coordinates
- Detection
  - Find variable number of objects
  - Sliding window, too dense
  - Use region proposals: R-CNN and variants
- Segmentation
  - Couple with dense CRF formulations for boundaries

Other Innovations

- Recurrent Neural Nets (RNN)
  - Memory
- Residual Nets (ResNet)
  - Deep
- …

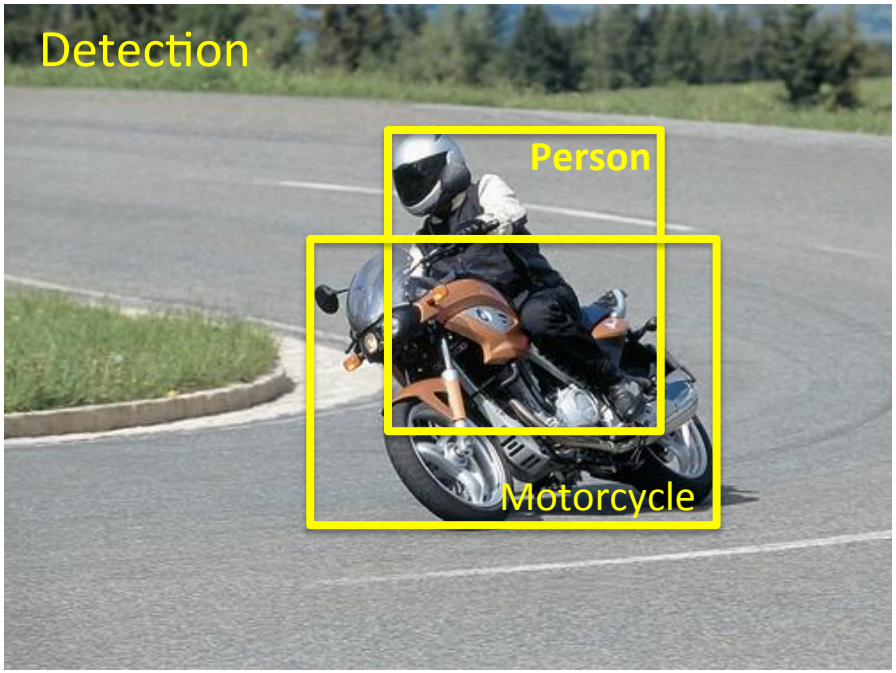
Data Sets

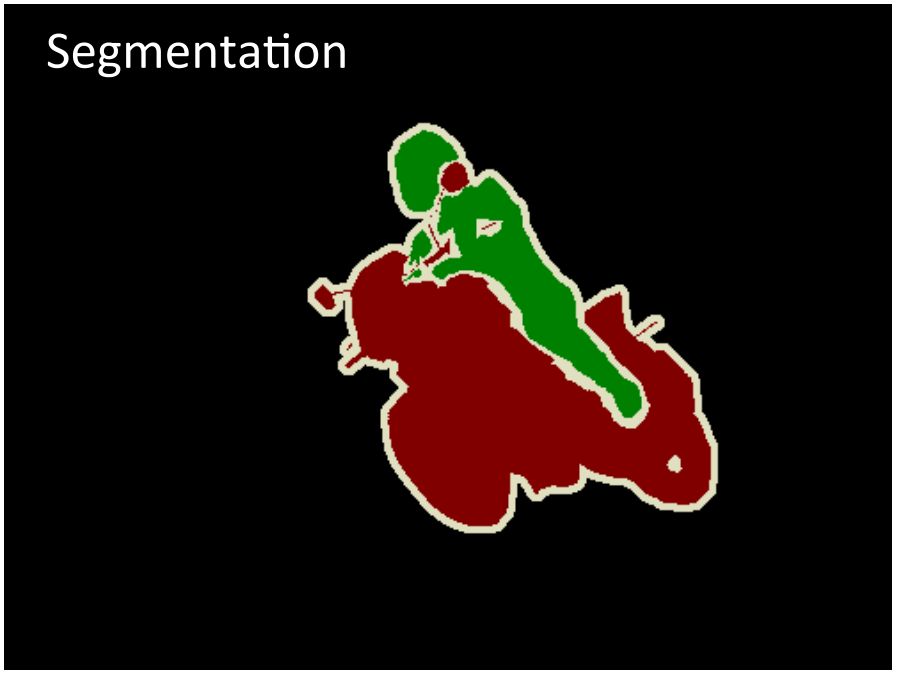
- Critical to the success of deep learning
  - Object classification and segmentation
  - Scene classification
  - Materials
- Examples
  - PASCAL VOC
    - *Not* Crowdsourced, bounding boxes, 20 categories
  - ImageNet
    - Huge, Crowdsourced, Hierarchical, *Iconic* objects
  - SUN Scene Database
    - *Not* Crowdsourced, 397 (or 720) scene categories
  - Microsoft COCO
    - Crowdsourced, large
  - Material Database: OpenSurfaces

PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





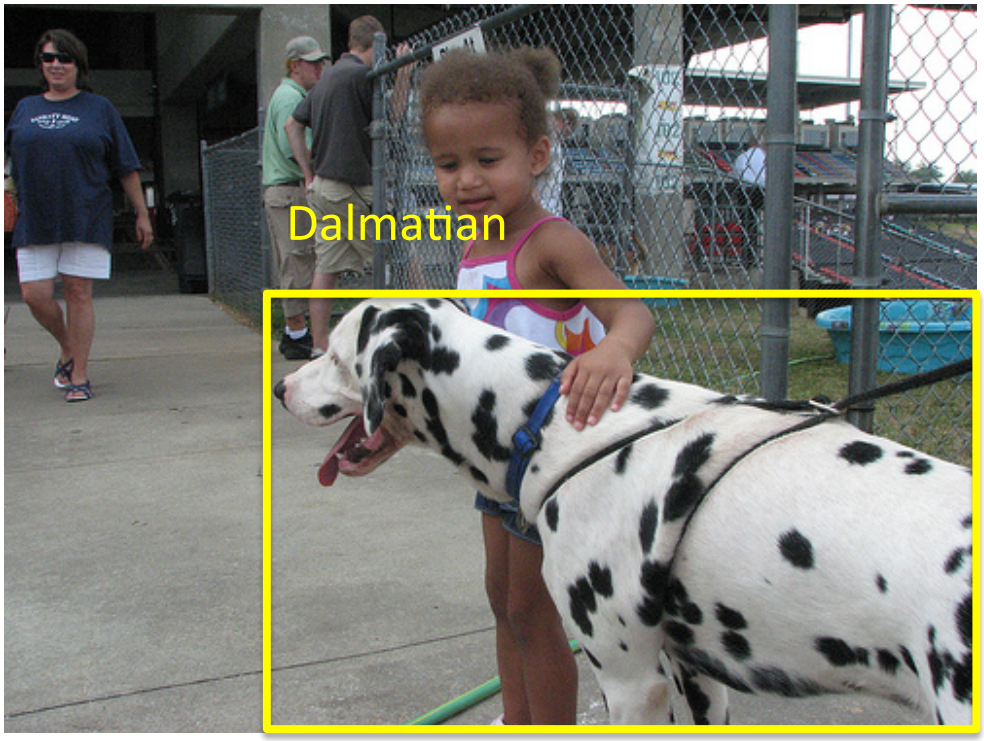
Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

~~20 object classes~~ ~~22,591 images~~

1000 object classes **1,431,167 images**

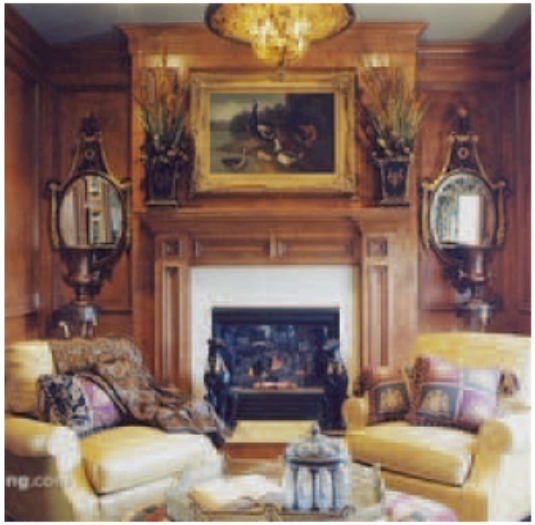


http://image-net.org/challenges/LSVRC/{2010,2011,2012}


PASCAL
ILSVRC
birds
cats
dogs
bird
flamingo
cock
ruffed grouse
quail
partridge
...
cat
Egyptian cat
Persian cat
Siamese cat
tabby
lynx
...
dog
dalmatian
keeshond
miniature schnauzer
standard schnauzer
giant schnauzer
...

How do we classify scenes?

Ceiling
Light
Door
Door
Door
Wall
Door
Wall
Door
Floor



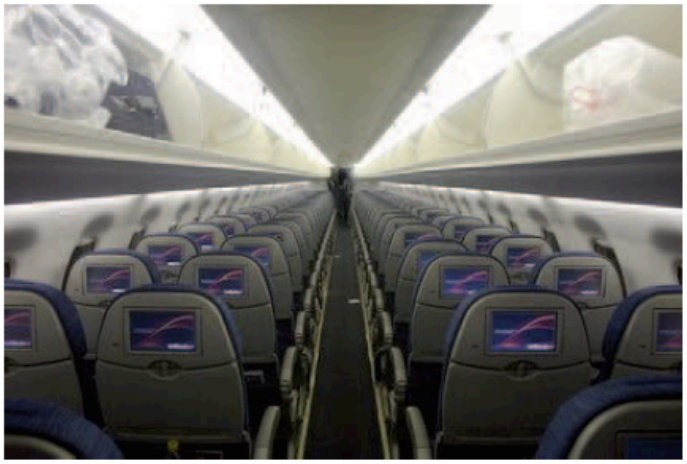
Ceiling
Lamp
Painting
mirror
mirror
wall
Fireplace
armchair
armchair
Coffee table

wall
painting
wall
Lamp
phone
alarm
Bed
Side-table
carpet

Different objects, different spatial layout

Which are the important elements?

cabinets ceiling cabinets window window window seat seat seat seat seat seat seat seat

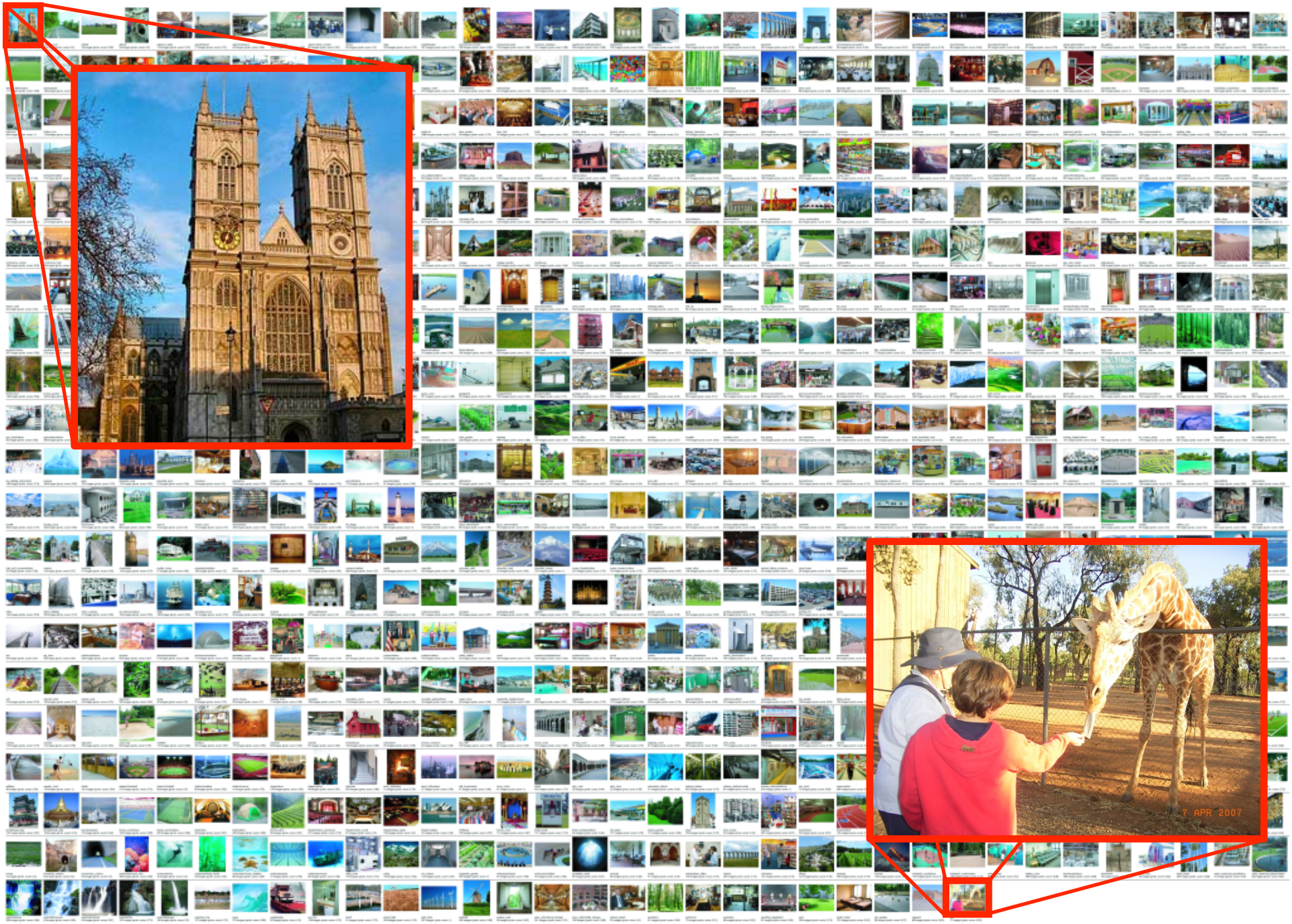


cabinets ceiling cabinets window seat seat window seat seat seat seat seat

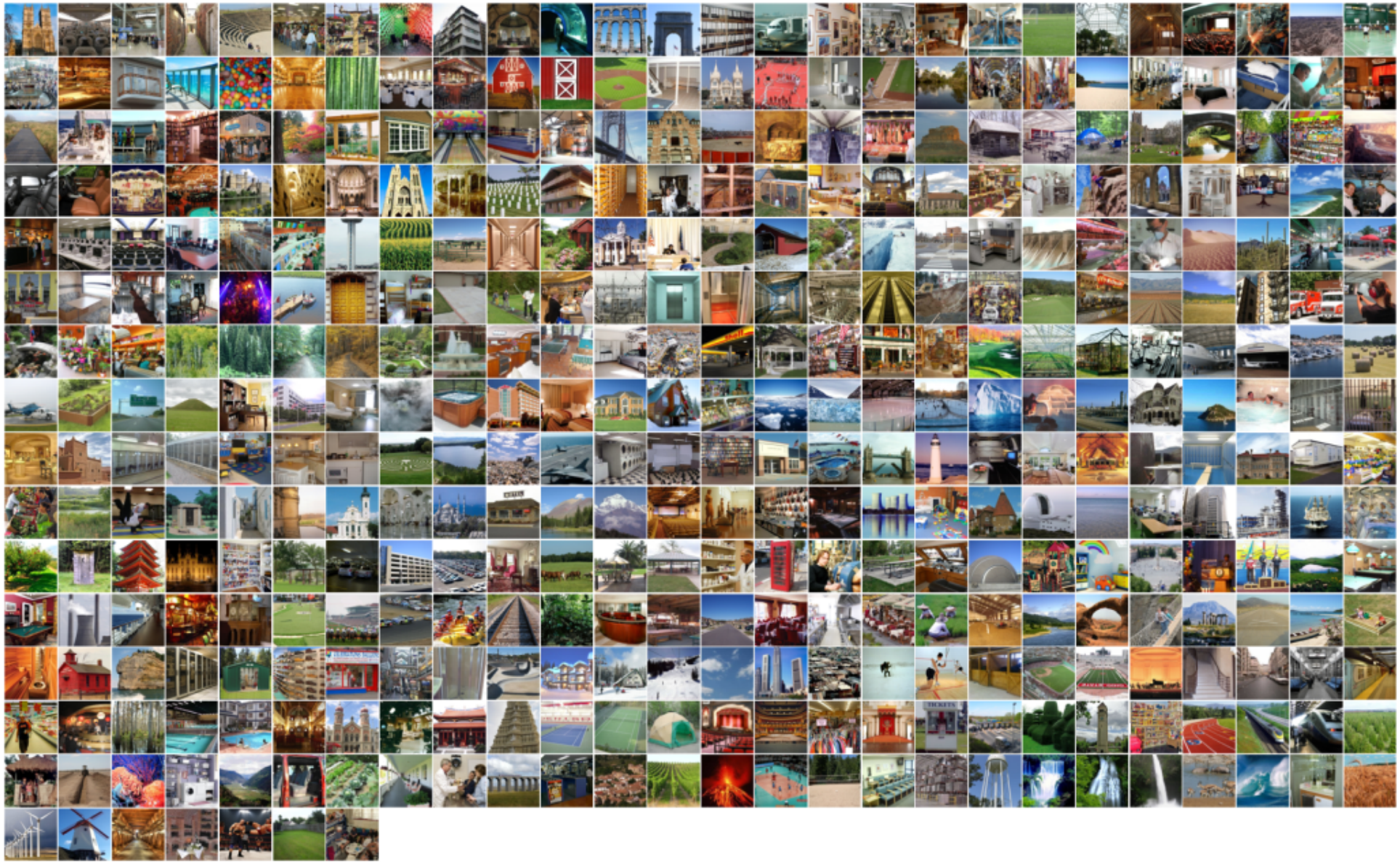
ceiling wall column screen seat seat seat seat seat seat seat seat seat seat seat seat seat seat seat seat seat seat

Similar objects, and similar spatial layout

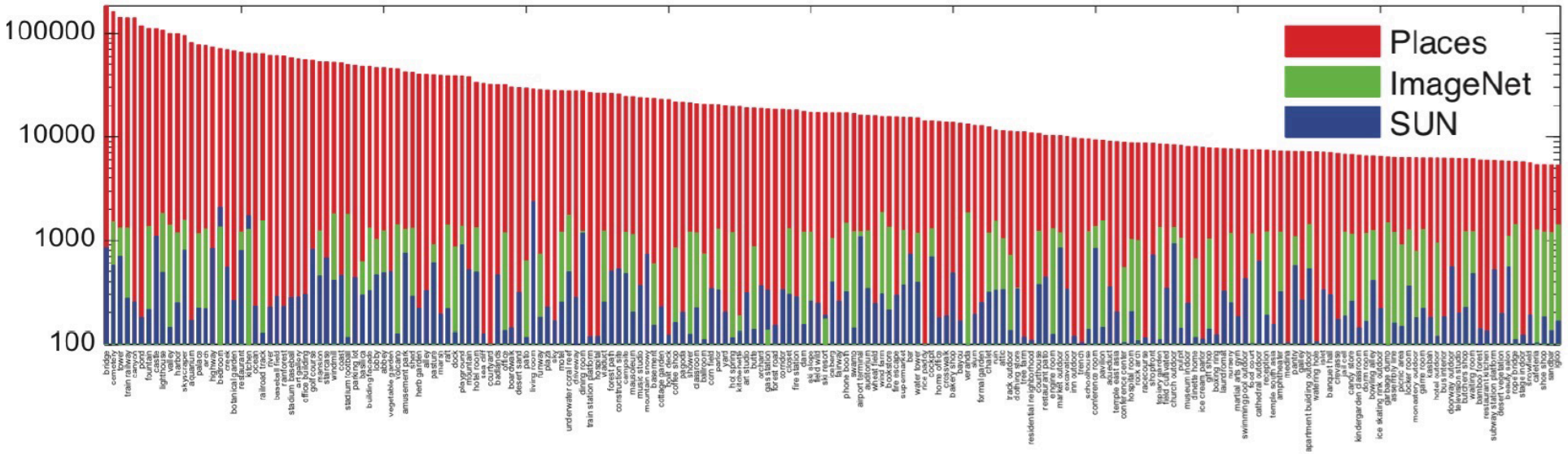
Different lighting, different materials, different “stuff”


7 APR 2007

397 Well-sampled Categories



- **Places Database**
- ~7 million images from 476 scene categories



ImageNet-CNN and Places-CNN

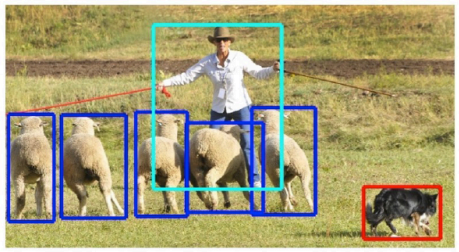
- Same structure as AlexNet, but trained on different databases.

	SUN397	MIT Indoor67	Scene15	SUN Attribute
Places-CNN feature	**54.32±0.14**	**68.24**	**90.19±0.34**	**91.29**
ImageNet-CNN feature	42.61±0.16	56.79	84.23±0.37	89.85
	Caltech101	Caltech256	Action40	Event8
Places-CNN feature	65.18±0.88	45.59±0.31	42.86±0.25	94.12±0.99
ImageNet-CNN feature	**87.22±0.92**	**67.23±0.27**	**54.92±0.33**	**94.42±0.76**

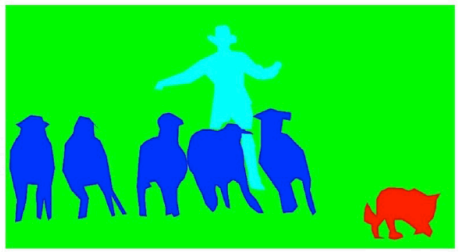
Microsoft COCO

We present a new dataset with the goal of advancing the state-of-the-art in object recognition by placing the question of object recognition in the context of the broader question of scene understanding. This is achieved by gathering images of complex everyday scenes containing common objects in their natural context. Objects are labeled using per-instance segmentations to aid in precise object localization. Our dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. With a total of 2.5 million labeled instances in 328k images, the creation of our dataset drew upon extensive crowd worker involvement via novel user interfaces for category detection, instance spotting and instance segmentation. We present a detailed statistical analysis of the dataset in comparison to PASCAL, ImageNet, and SUN. Finally, we provide baseline performance analysis for bounding box and segmentation detection results using a Deformable Parts Model.

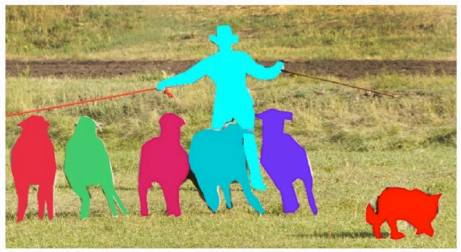
(a) Image classification



(b) Object localization



(c) Semantic segmentation



(d) This work

✓ Instance segmentation
✓ Non-iconic Images