CS4670/5670: Computer Vision

Kavita Bala

Lecture 34: Datasets

Visual Object Classes Challenge 2009 (VOC2009)

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

[click on an image to see the annotation]

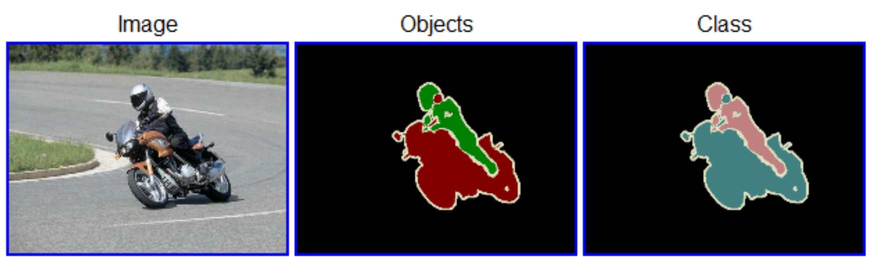
IMAGENET

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

The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- Twenty object categories (aeroplane to TV/monitor)
- Three 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?)



Dataset: Collection

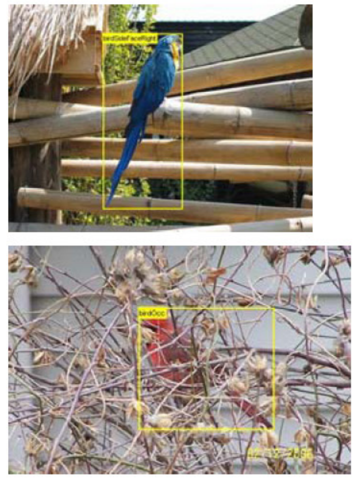
- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation

Examples

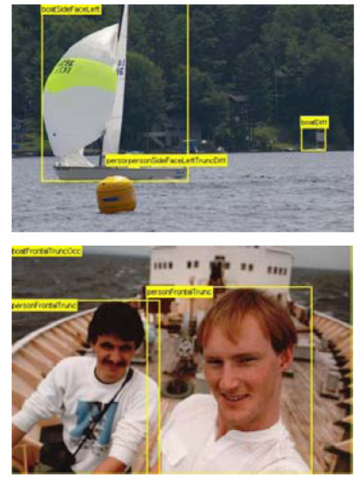
Aeroplane

Bicycle

Bird



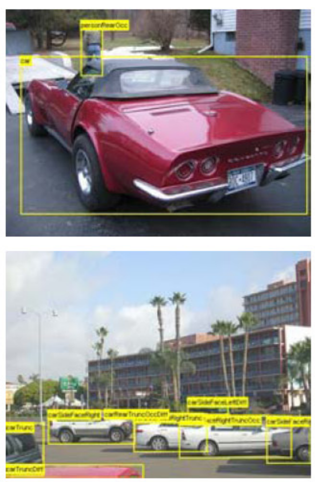
Boat



Bottle

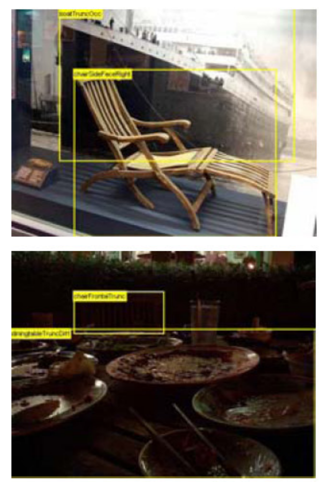
Bus

Car

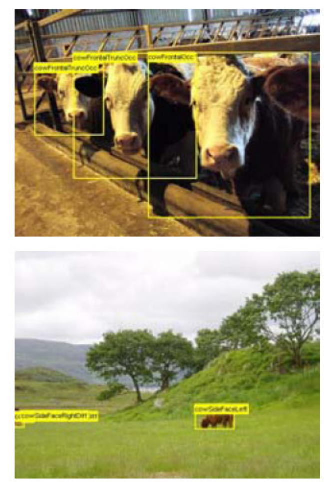


Cat

Chair

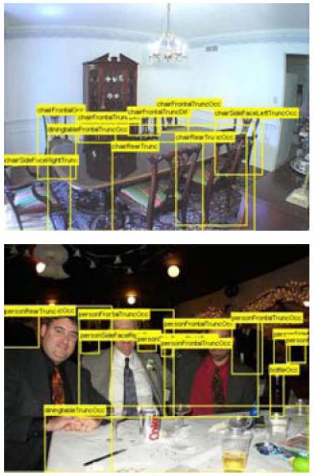


Cow



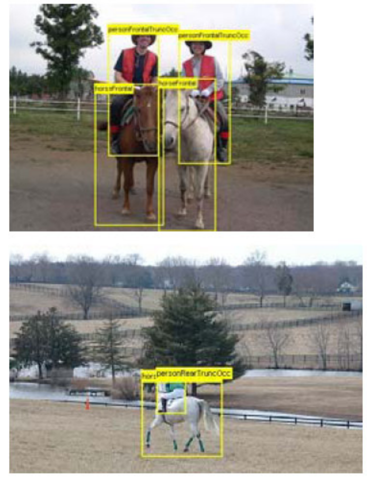
Examples

Dining Table

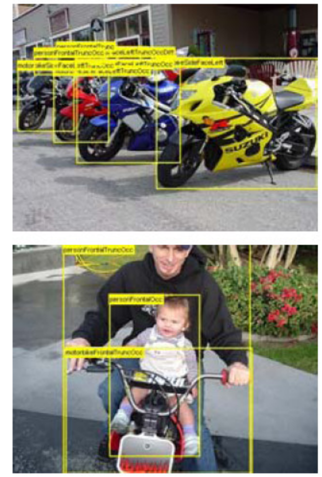


Dog

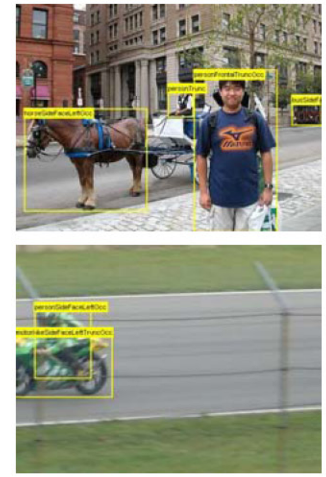
Horse



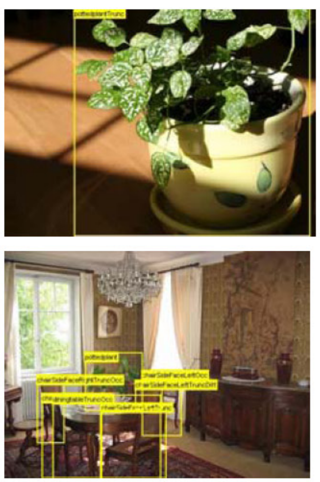
Motorbike



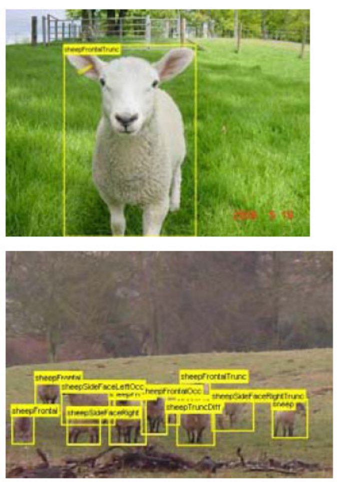
Person



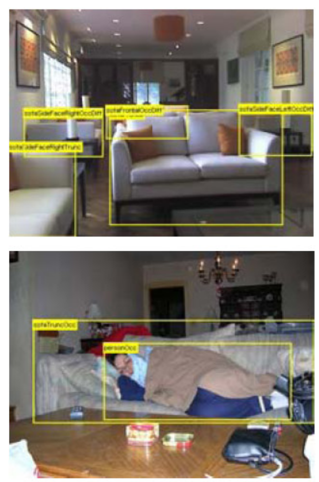
Potted Plant



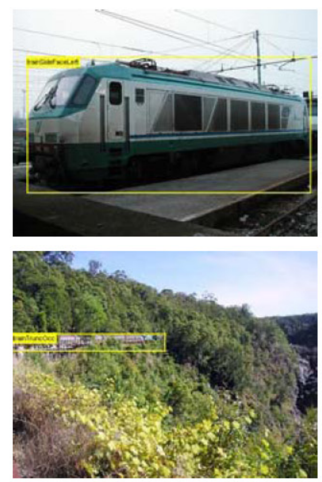
Sheep



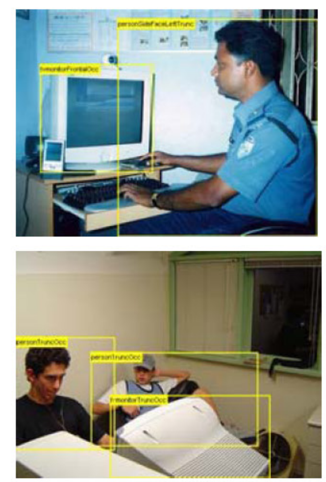
Sofa



Train



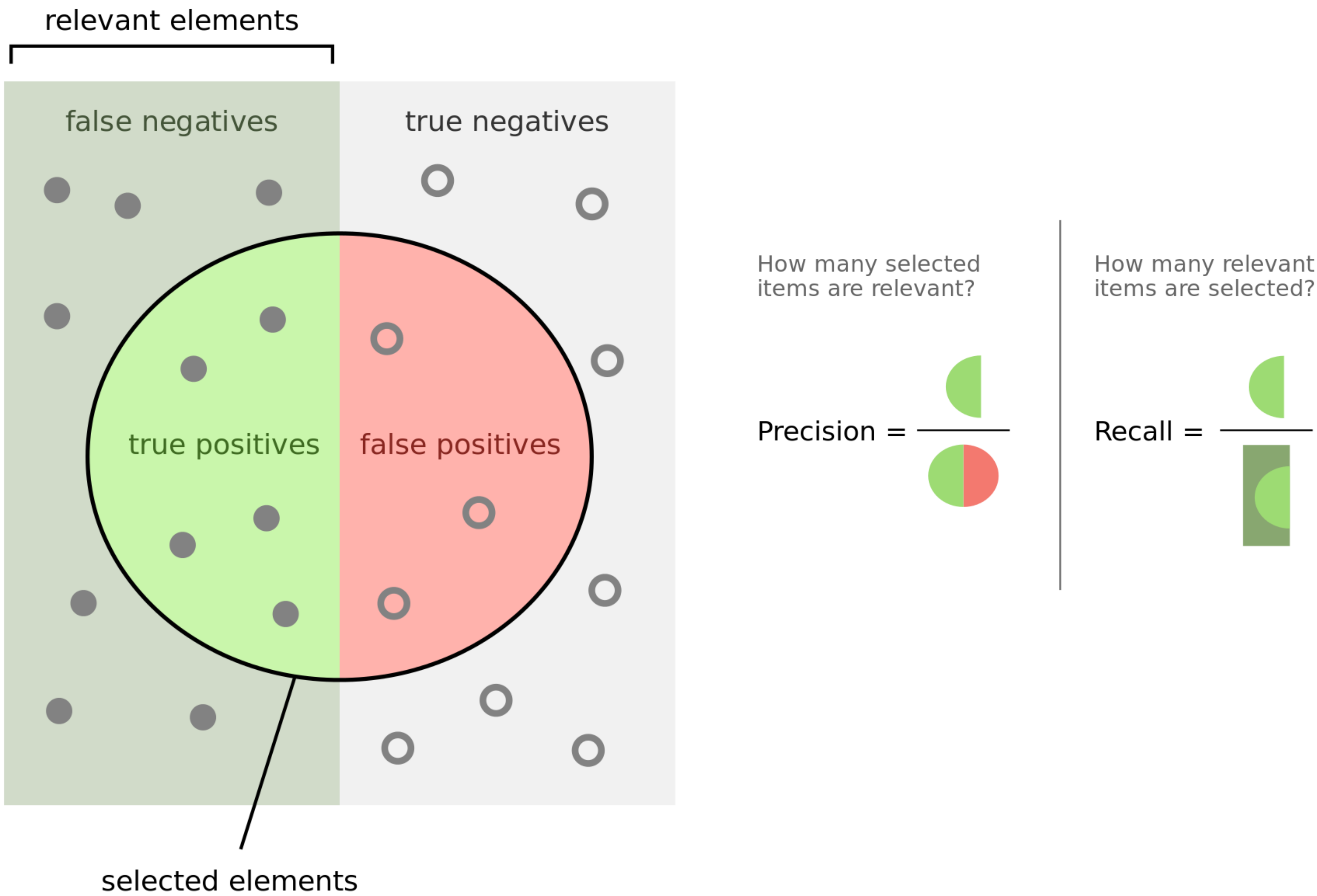
TV/Monitor

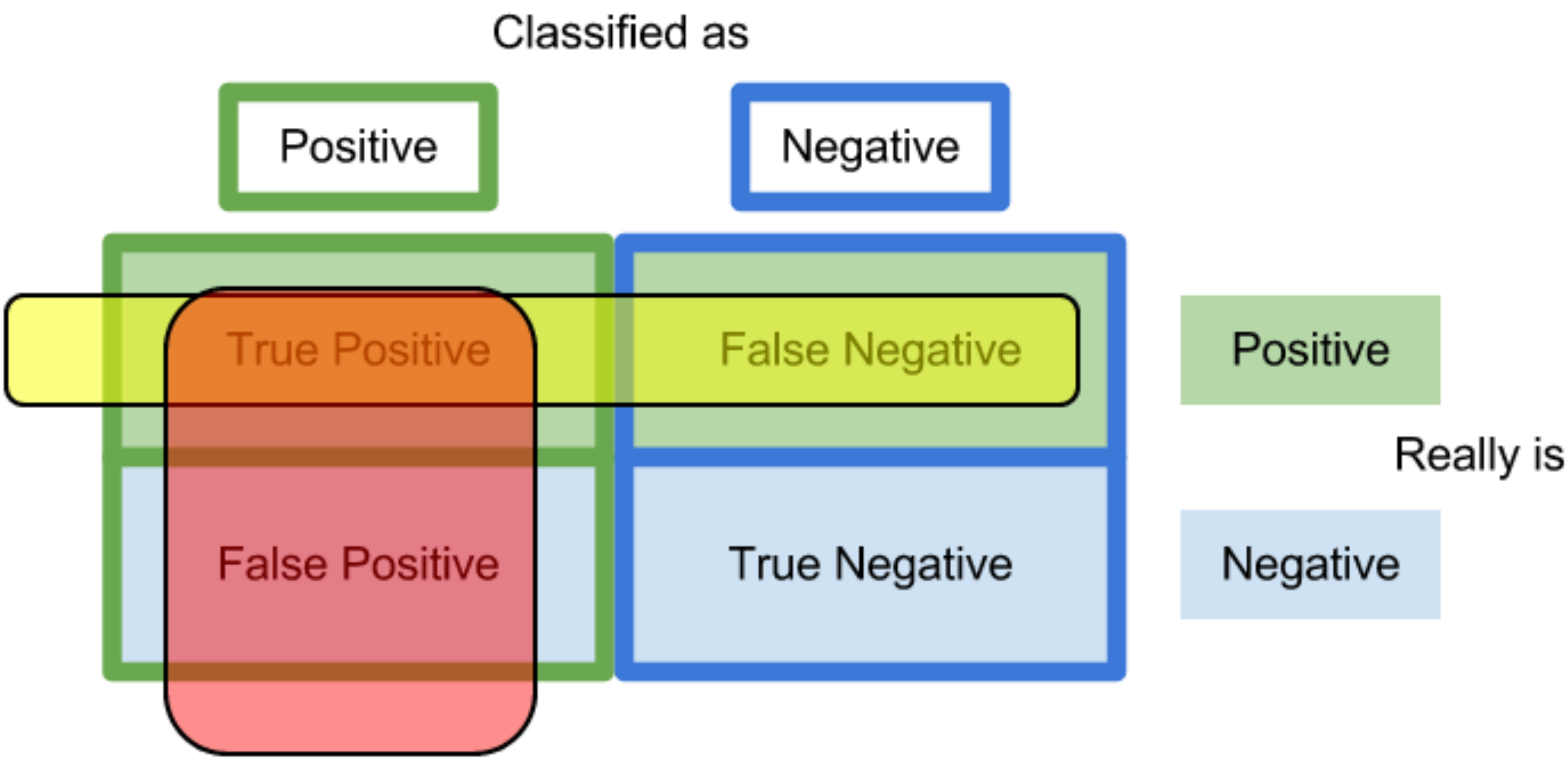


Classification Challenge

- Predict whether at least one object of a given class is present in an image

is there a cat?


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =

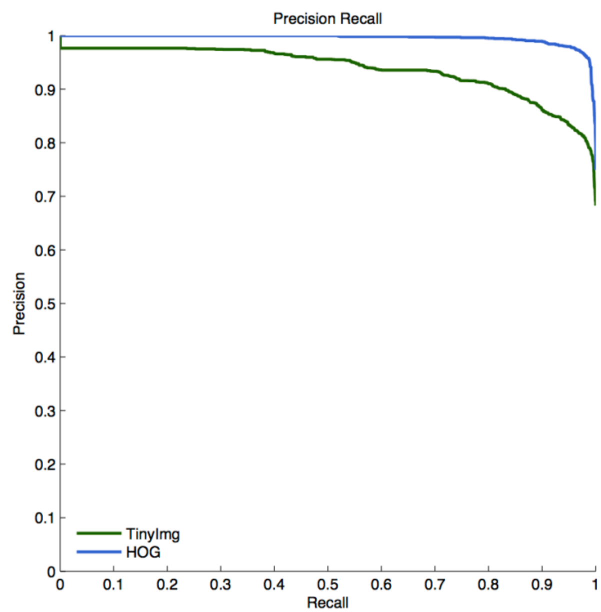


— Precision in red, recall in yellow

https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/

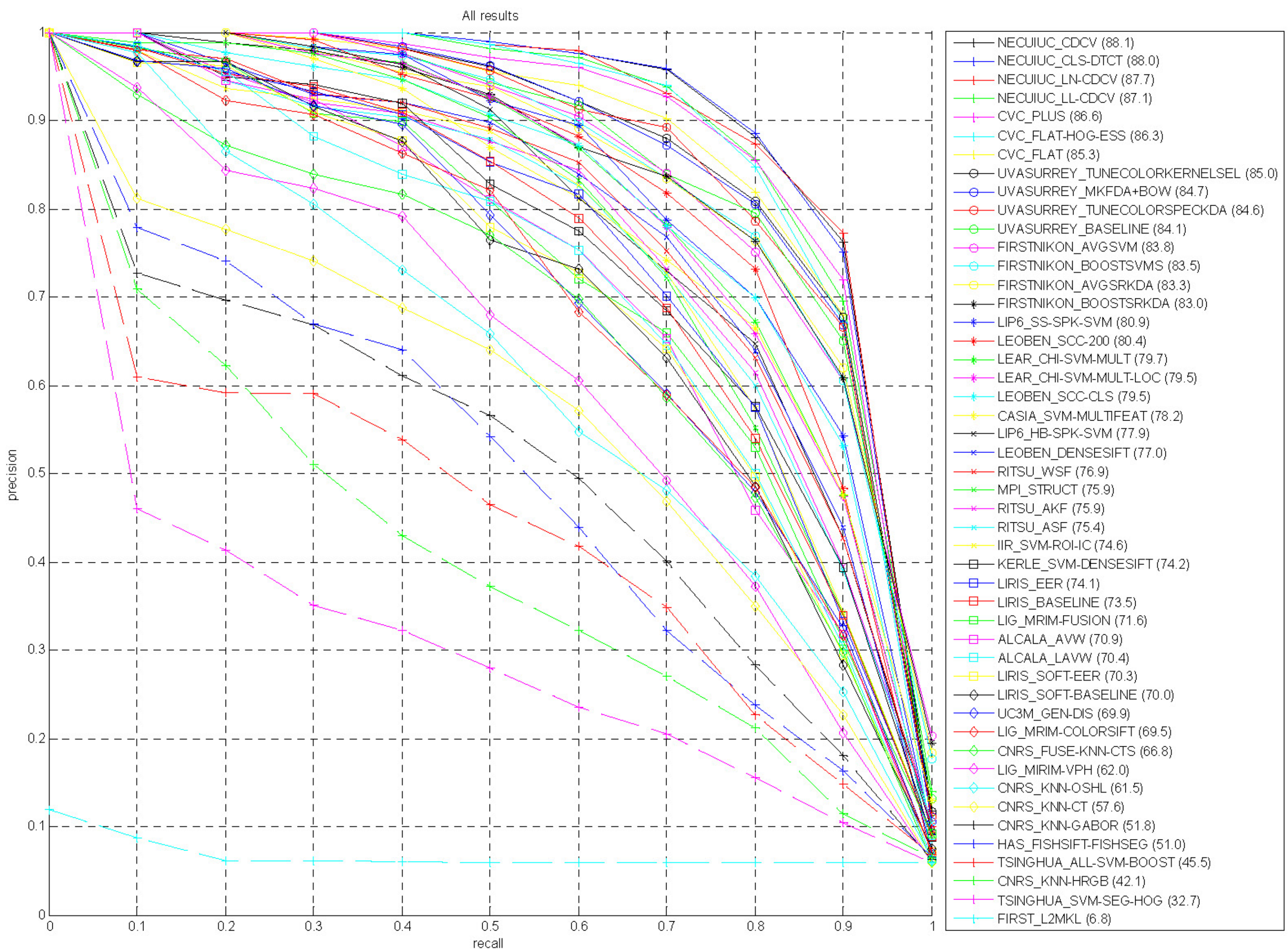
Precision Recall curves

- Related to but different from ROC curves
- Start at (0, 1), higher curves are better

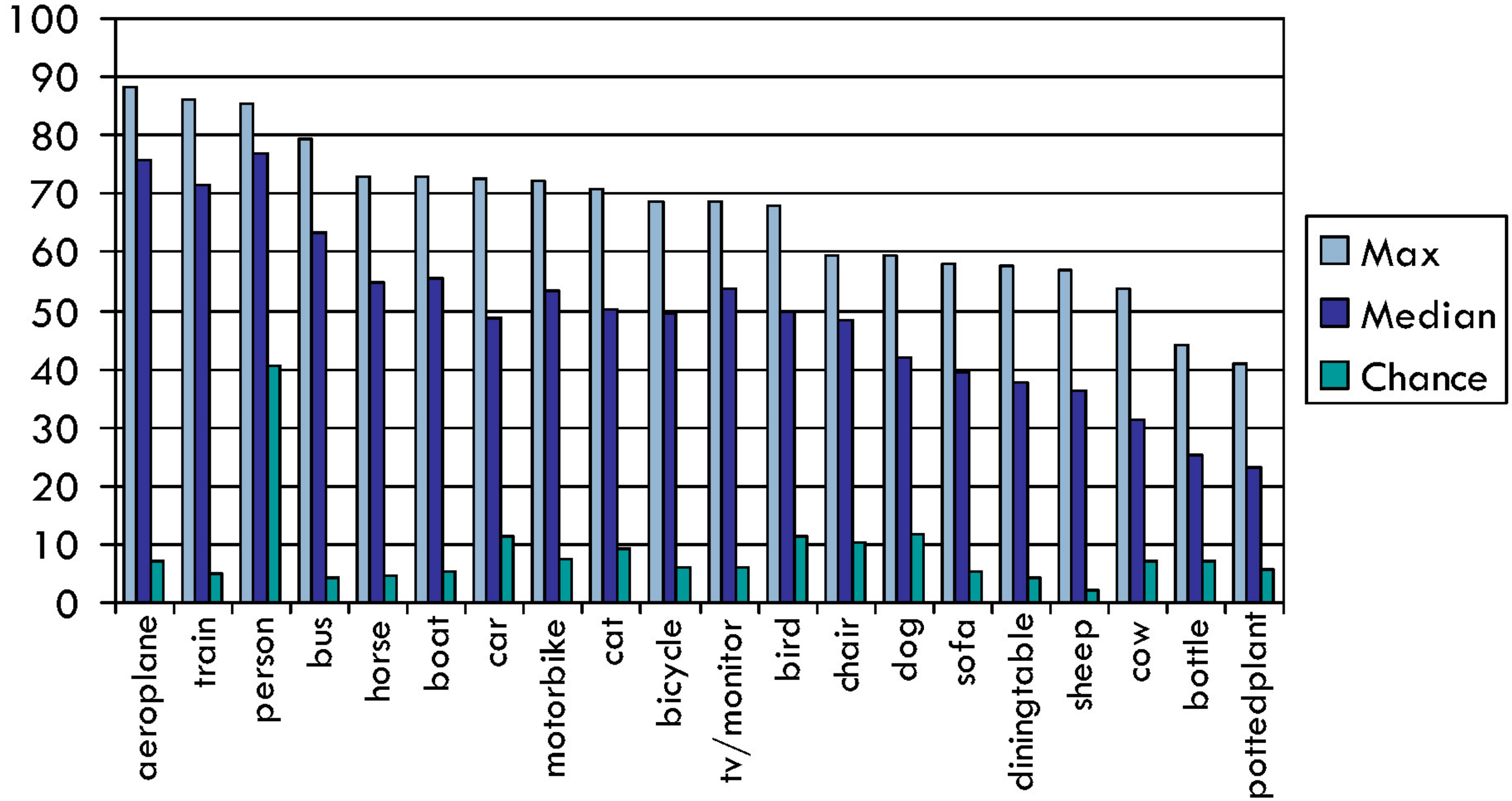


- Average Precision (AP) = area under the curve

Precision/Recall: Aeroplane (All)

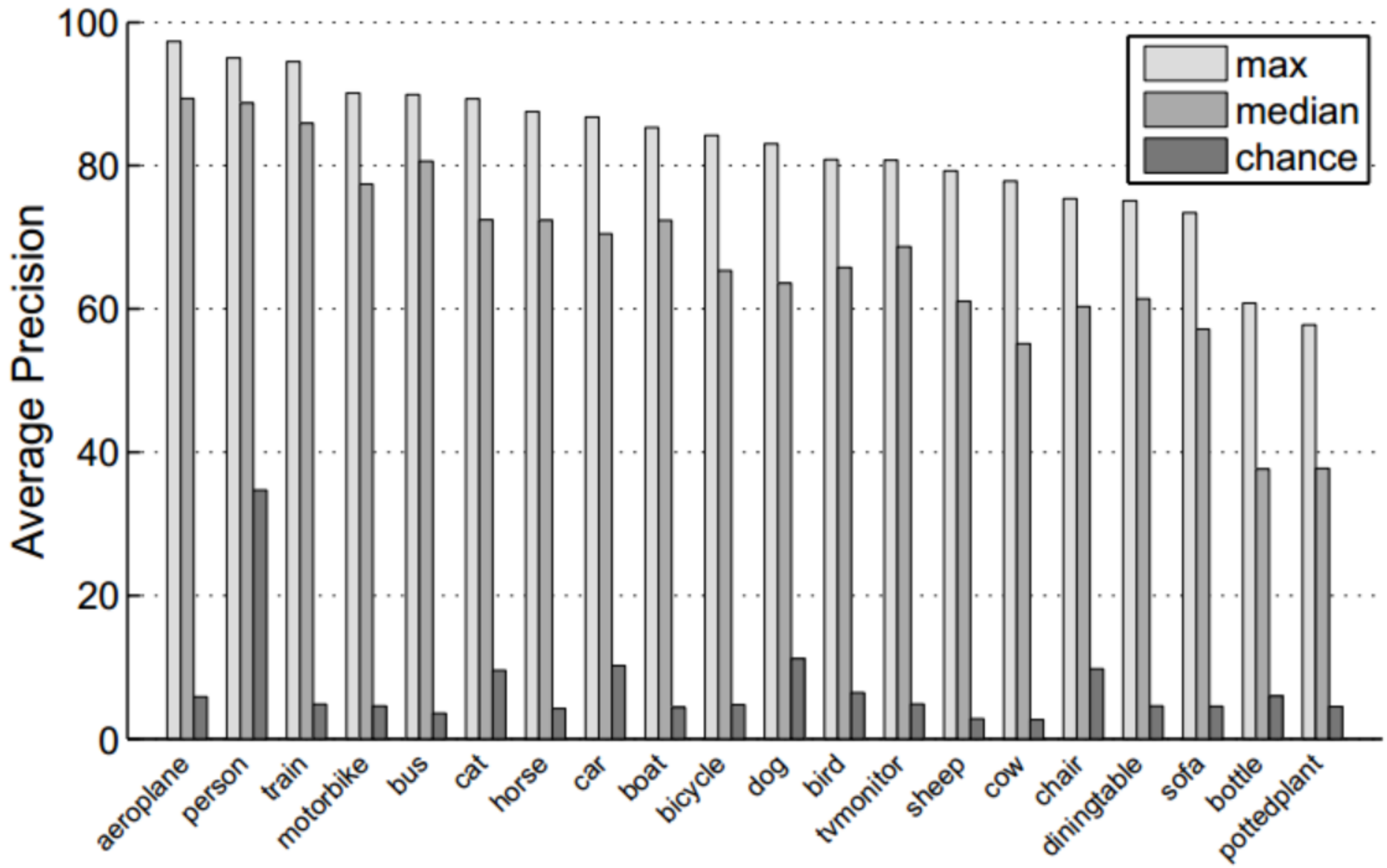


AP by Class



- Max AP: 88.1% (aeroplane) ... 40.8% (potted plant)

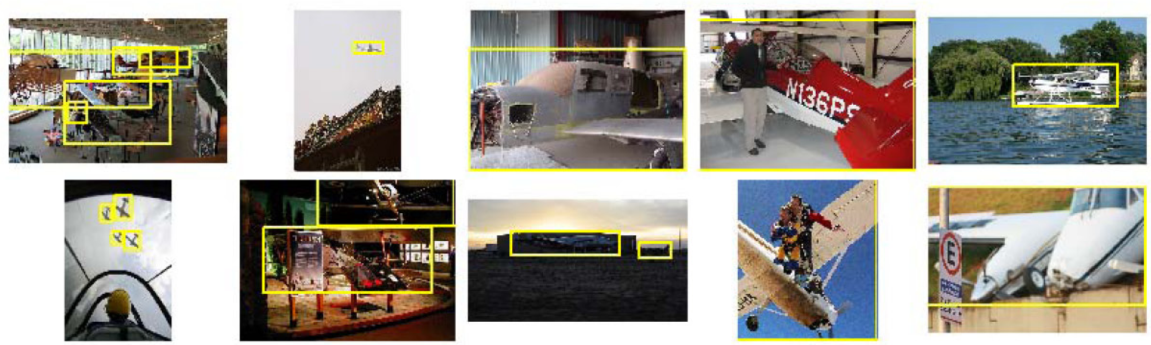
Pascal VOC 2012 Average Precision



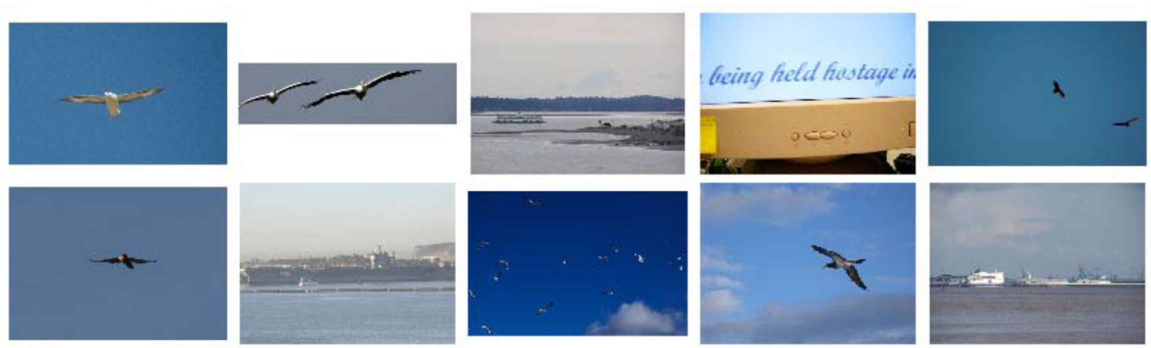
Ranked Images: Aeroplane

- Class images:
  Highest ranked

- Class images:
  Lowest ranked



- Non-class images:
  Highest ranked



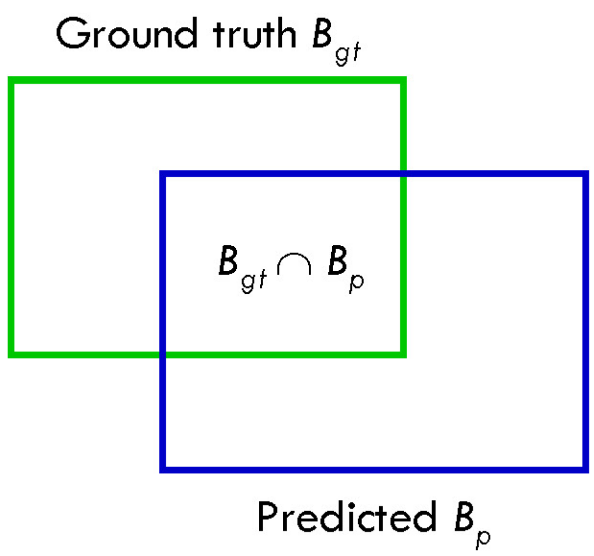
- Context?

Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)

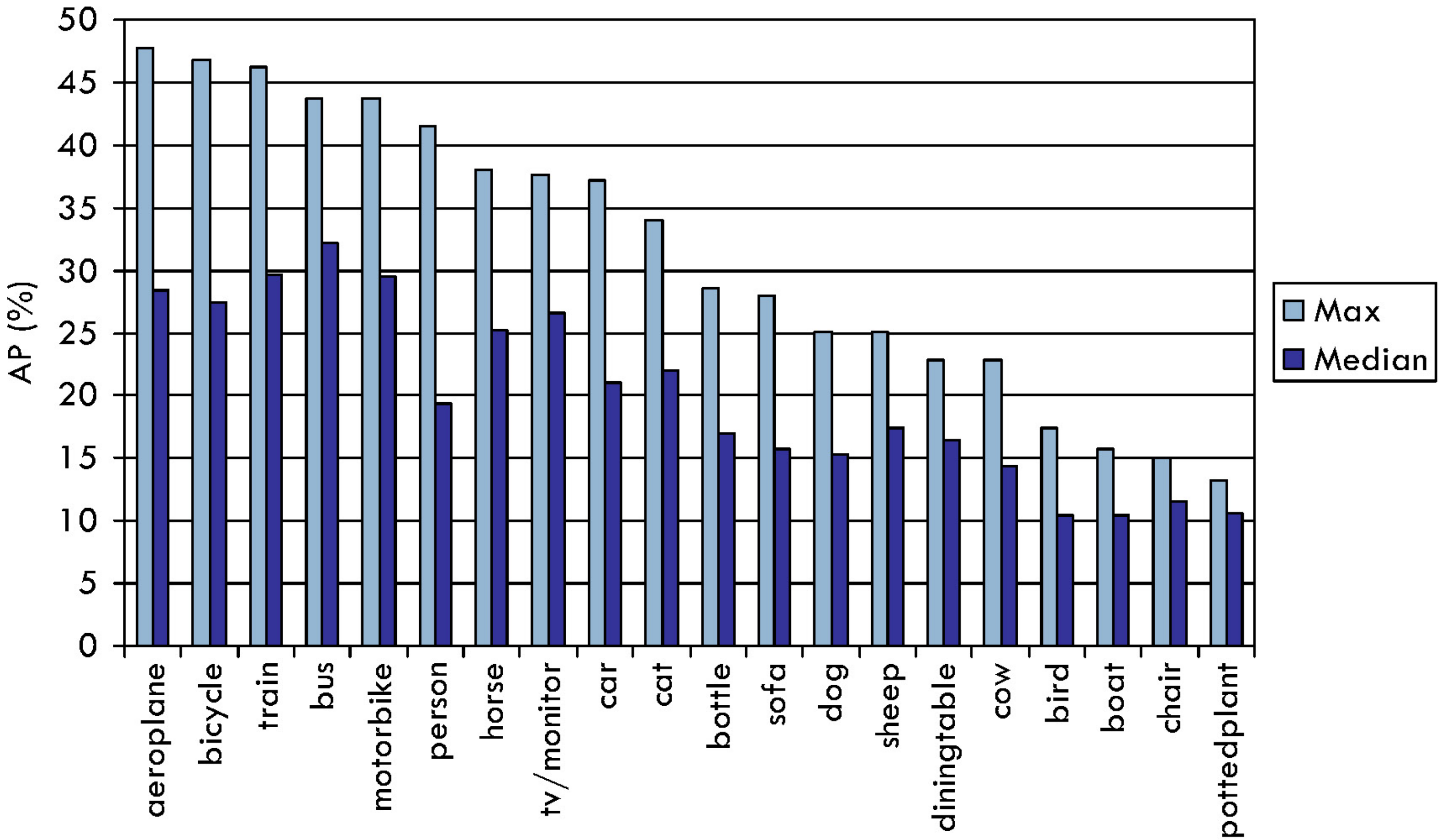
Evaluating Bounding Boxes

- Area of Overlap (AO) Measure



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

AP by Class

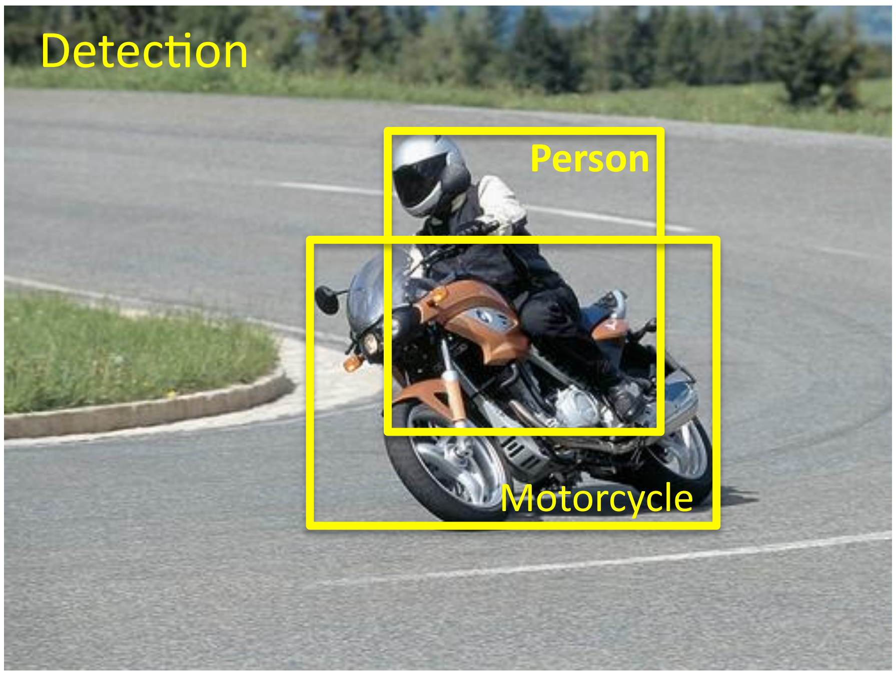


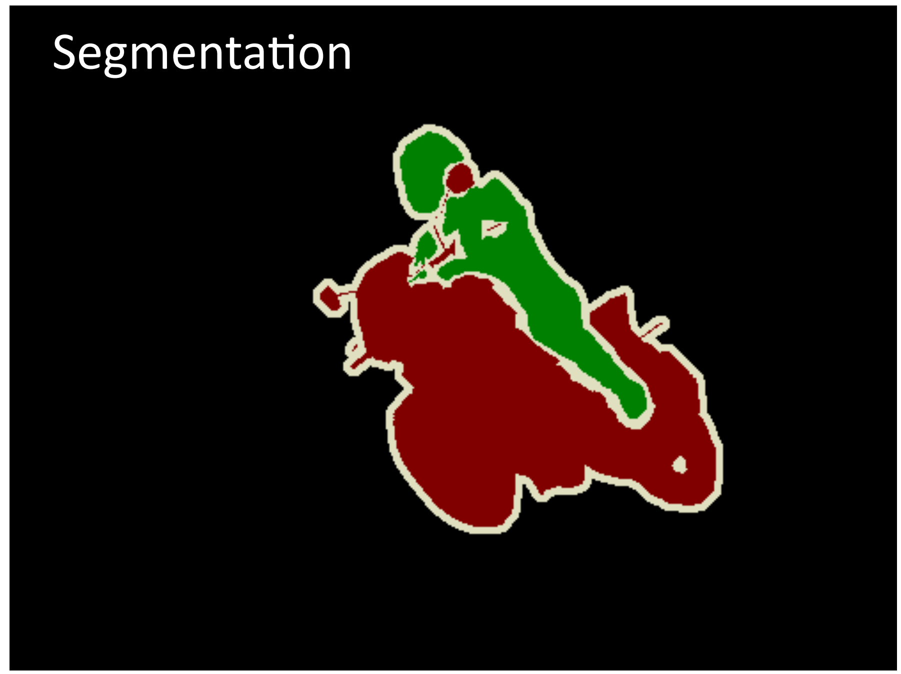
Chance essentially 0

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~~

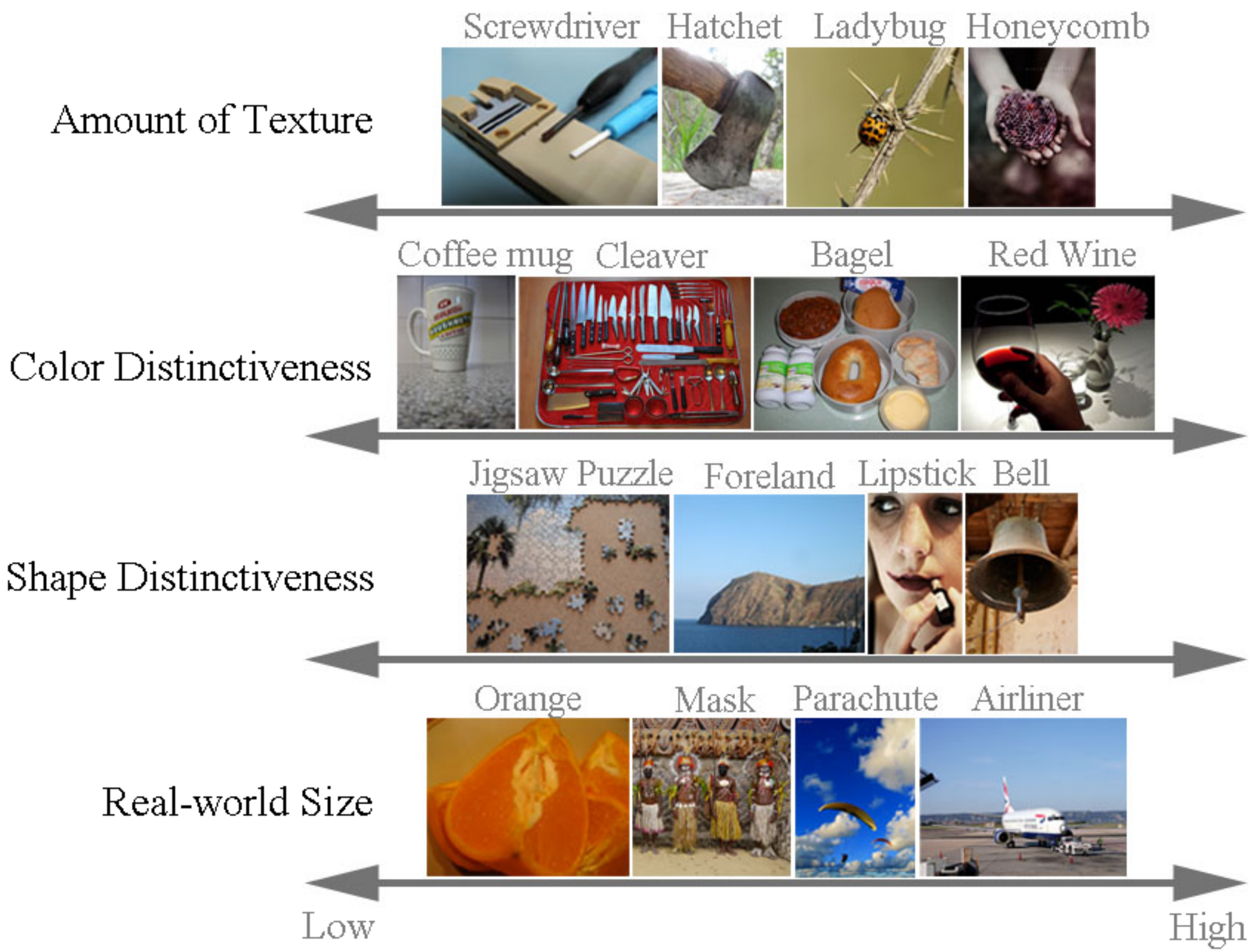
1000 object classes **1,431,167 images**

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

Variety of object classes in ILSVRC


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
. . .

Variety of object classes in ILSVRC



How do we classify scenes?

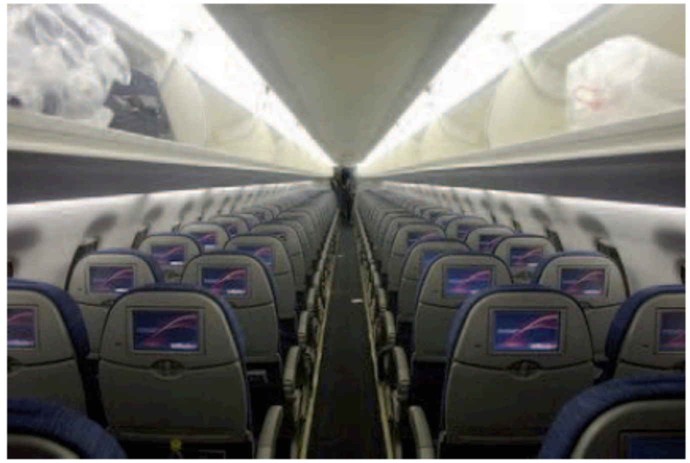
Ceiling
Light
Door Door
Door
Wall Door Wall Door
Floor

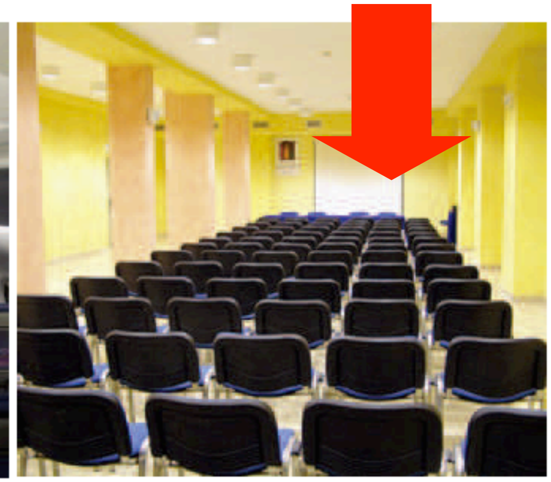
Ceiling
Lamp
Painting
mirror mirror
wall
Fireplace
armchair armchair
Coffee table

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
seat

ceiling
wall column screen
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”

Scene Categorization

Oliva and Torralba, 2001


Coast


Forest


Highway


Inside City


Mountain


Open Country


Street


Tall Building

Fei Fei and Perona, 2005

\+


Bedroom


Kitchen


Living Room


Office


Suburb

Lazebnik, Schmid, and Ponce, 2006

\+


Industrial


Store

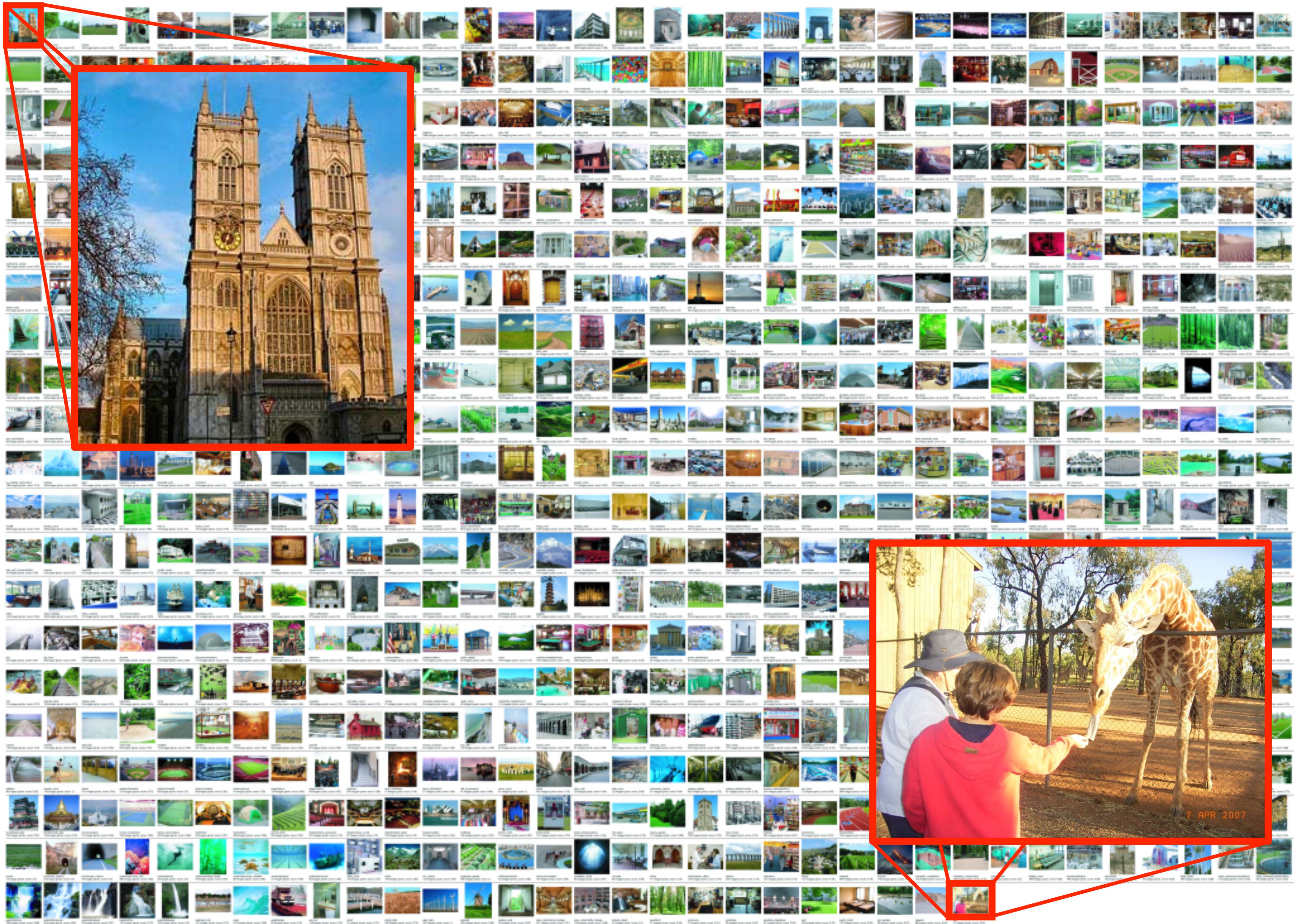
15 Scene Database

SUN Database: Large-scale Scene Categorization and Detection

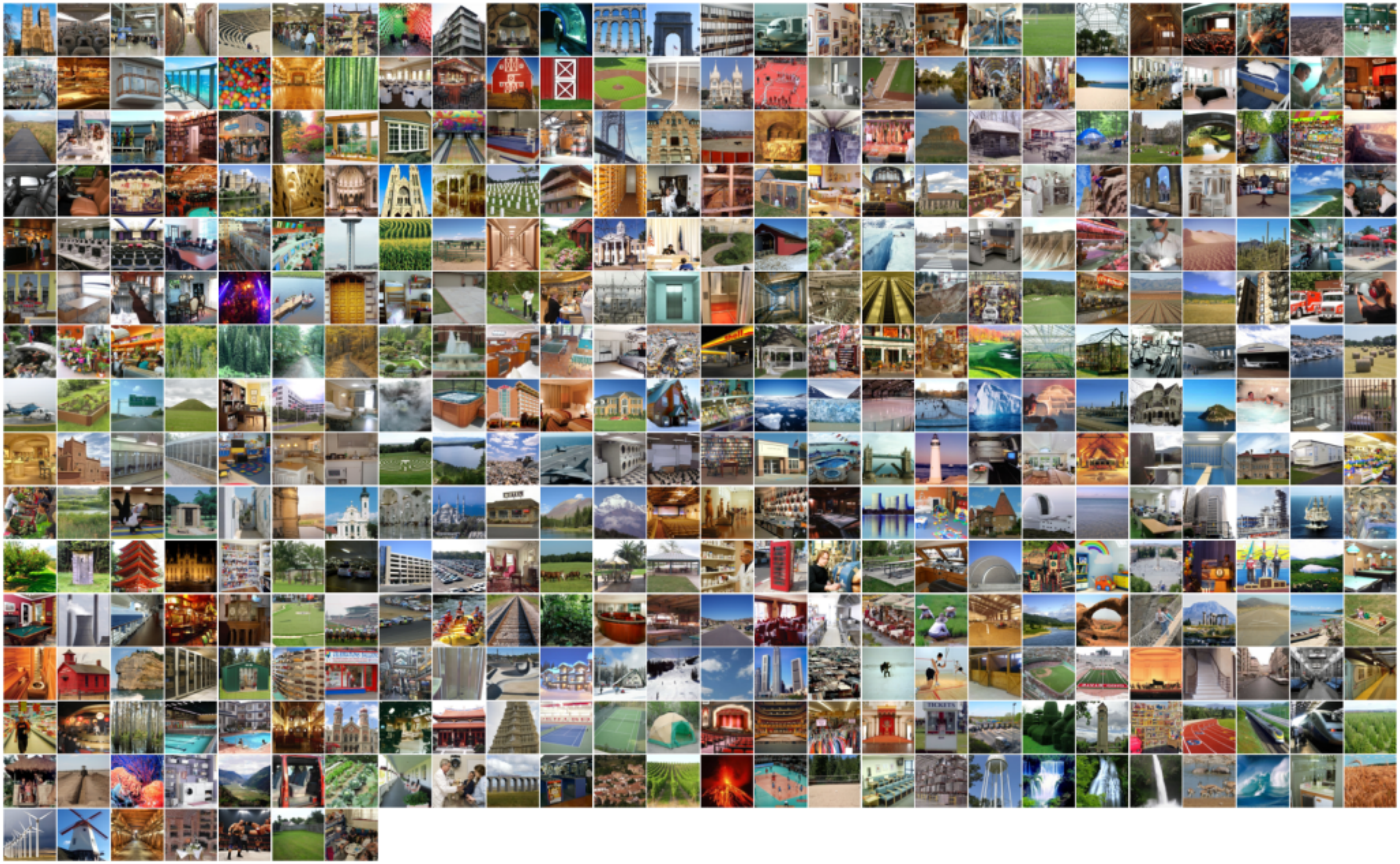
Jianxiong Xiao, James Hays[†], Krista A. Ehinger, Aude Oliva, Antonio Torralba

Massachusetts Institute of Technology

[†] Brown University


7 APR 2007

397 Well-sampled Categories



bathroom(100%)

beauty salon(100%)

bedroom(100%)

bullring(100%)

playground(100%)

phone booth(100%)

greenhouse outdoor(100%)

podium outdoor(100%)

tennis court outdoor(100%)

wind farm(100%)

veterinarians office(100%)

riding arena(100%)

Scene category

Most confusing categories

Inn (0%)

Restaurant patio (44%)

Chalet (19%)

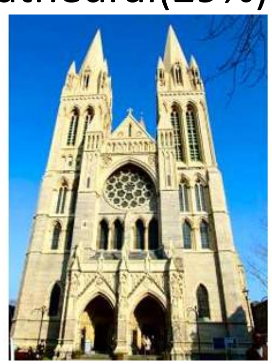
Bayou (0%)

River (67%)

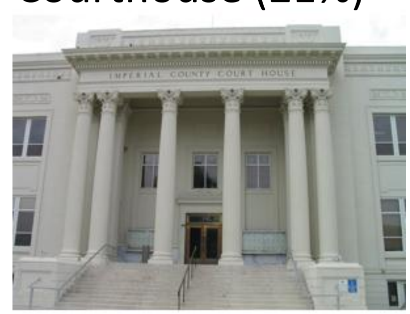
Coast (8%)

Basilica (0%)

Cathedral(29%)

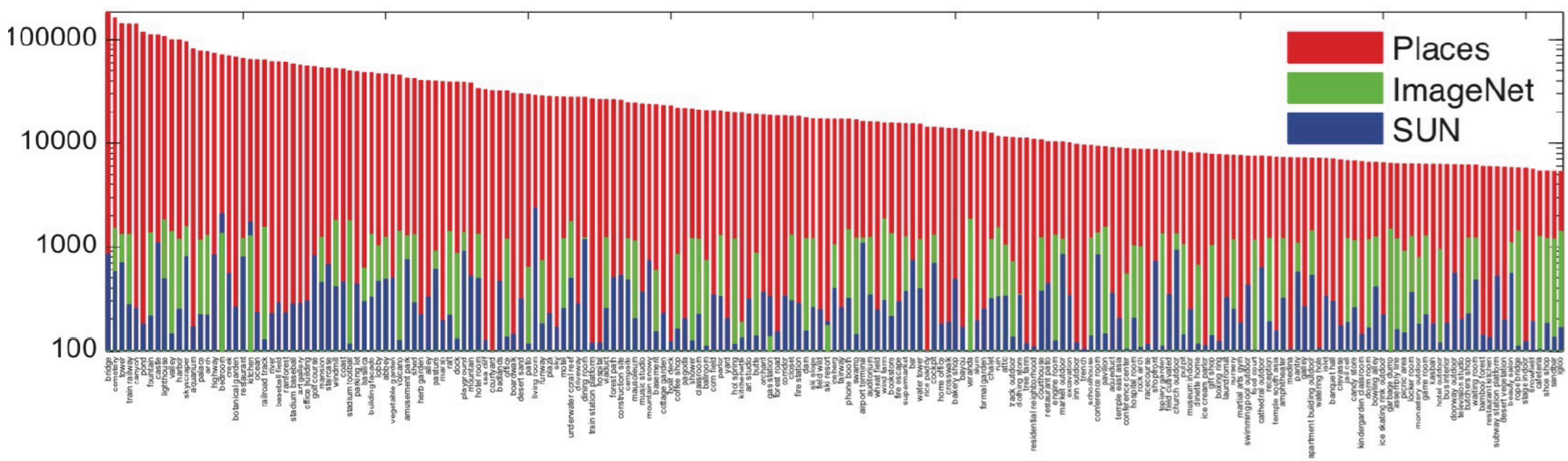


Courthouse (21%)



Now it’s the era of Big Data and Deep Learning

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



ImageNet-CNN and Places-CNN

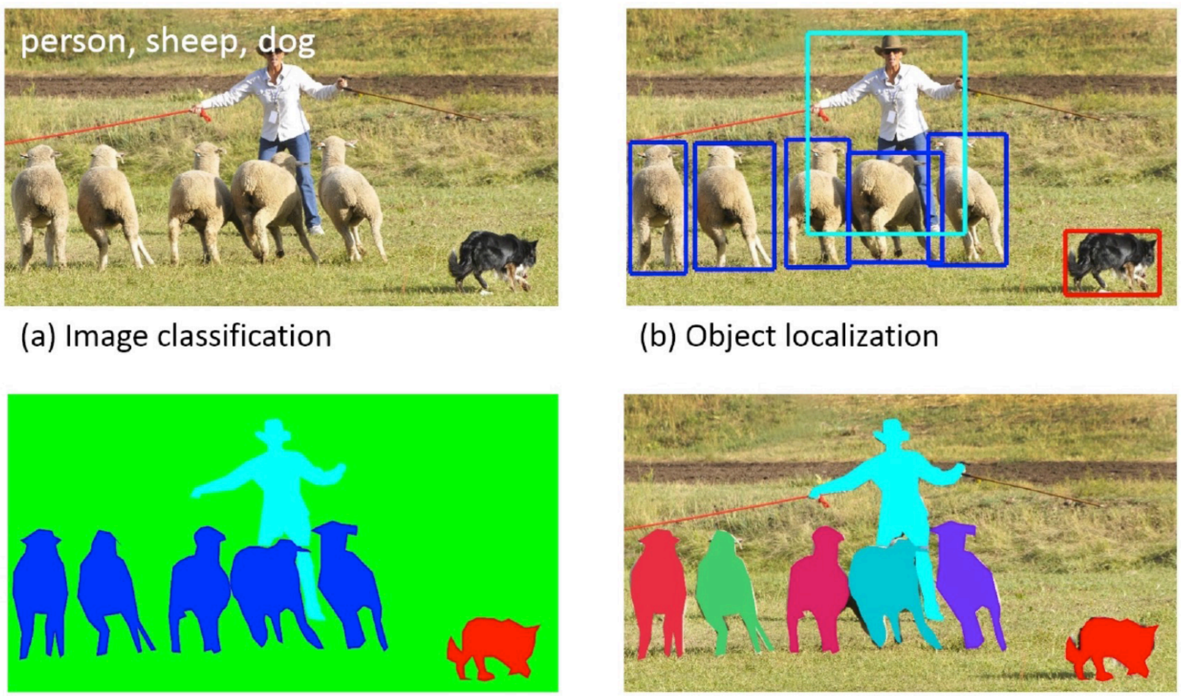
- Same structure as AlexNet, but trained on different databases.

	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
comparison to PA
bounding box and



(a) Image classification

(b) Object localization

(c) Semantic segmentation

(d) This work

- ✓ Instance segmentation
- ✓ Non-iconic Images

(a) Iconic object images (b) Iconic scene images (c) Non-iconic images

Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.

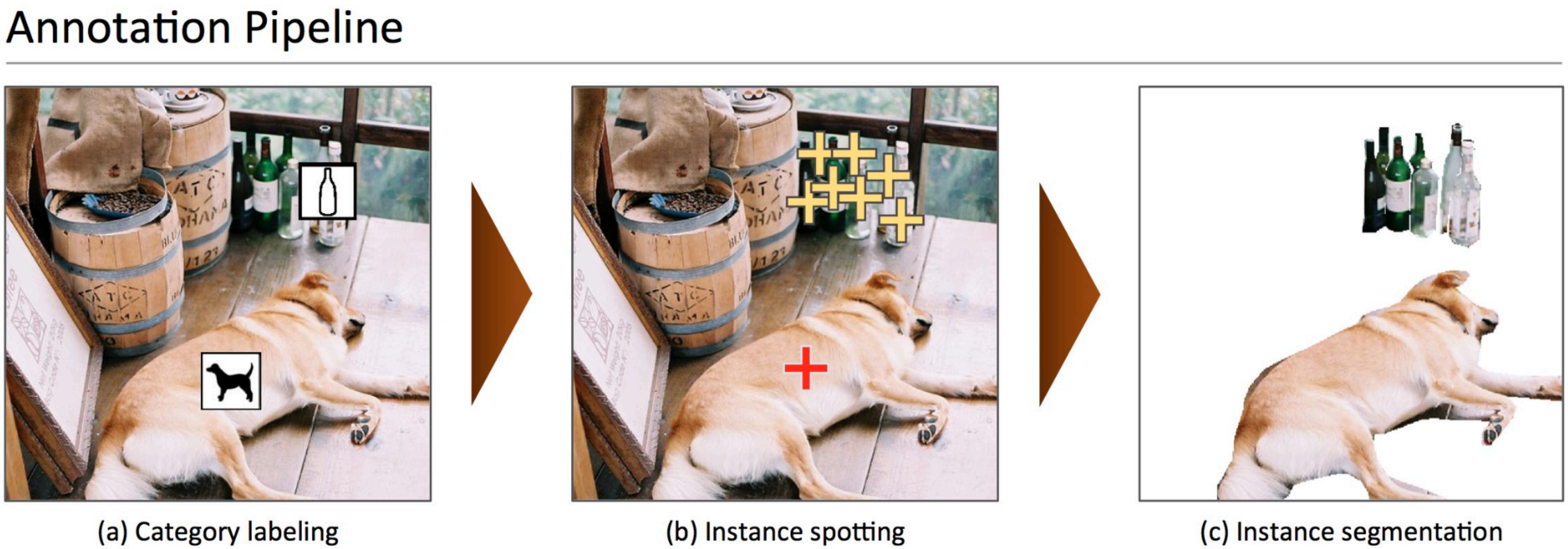


Fig. 3: Our annotation pipeline is split into 3 primary tasks: (a) labeling the categories present in the image (§4.1), (b) locating and marking all instances of the labeled categories (§4.2), and (c) segmenting each object instance (§4.3).

Material Database

- Different domain
  - Most of the focus has been on objects
  - Our focus on materials
- Sean Bell, Paul Upchurch, Noah Snavely, Kavita Bala
  - OpenSurfaces [2013]
  - Segmentation interface used by Microsoft COCO

OpenSurfaces

Get novice workers to accurately describe material appearance in [scalable, verifiable, and economical] way

- Open Surfaces: opensurfaces.cs.cornell.edu

Pipeline Preview

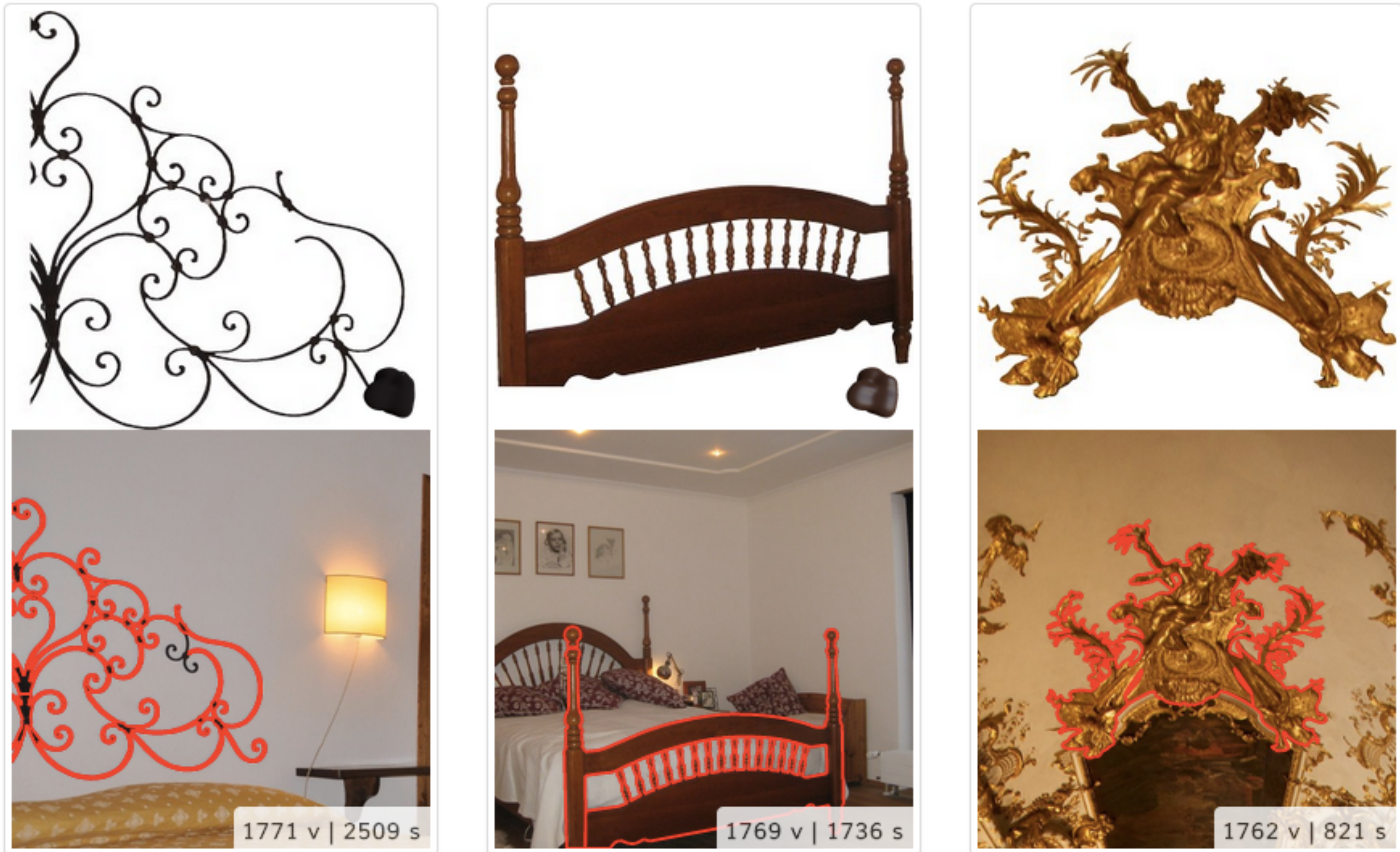
1. Material segmentation: Draw boundaries
2. Name material
3. Reflectance
4. Texture


110,000 Segmentations

[redacted]

Material Segmentation

Context



[redacted]


25,000 Textures


50,000 Reflectances

Data

- More is more….

Classification

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

→ cat

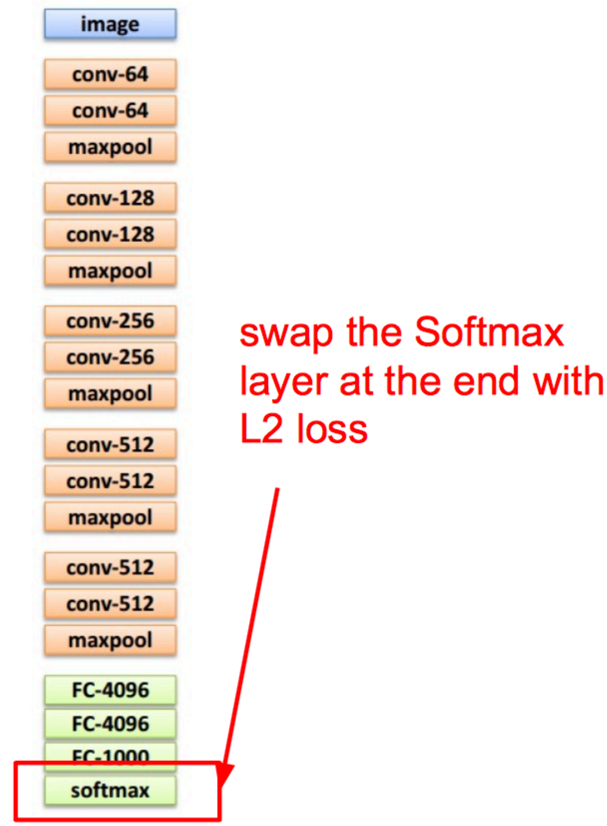
Localization

Model must output:

- class (integer)
- x1,y1,x2,y2 bounding box coordinates

Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

Idea: train a Localization net
Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.



Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

Idea: train a Localization net
Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.

predictions: instead of class scores, now interpreted as the 4 bounding box coords **(also 4D vector from net)**

targets: true bounding box **4D vector of [x1,y1,x2,y2]**

$$L_i = ||f - y_i||_2^2$$

swap the Softmax layer at the end with L2 loss

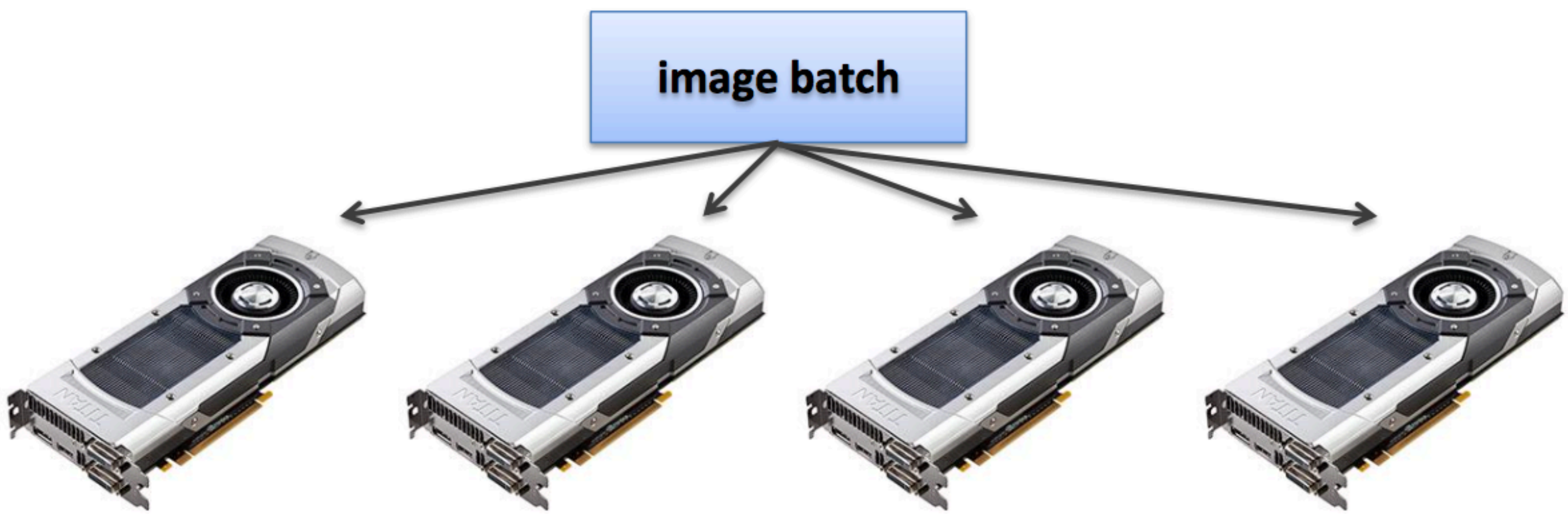
Very Deep Convolutional Networks for Large-Scale Image Recognition,
Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,
Sermanet et al., 2014

In practice:

- It works better to predict a **4D vector for every class** (e.g. 4000D vector for 1000 ImageNet classes). During training only backprop the loss for the correct class
- apply at **multiple locations and scales**

swap the Softmax layer at the end with L2 loss

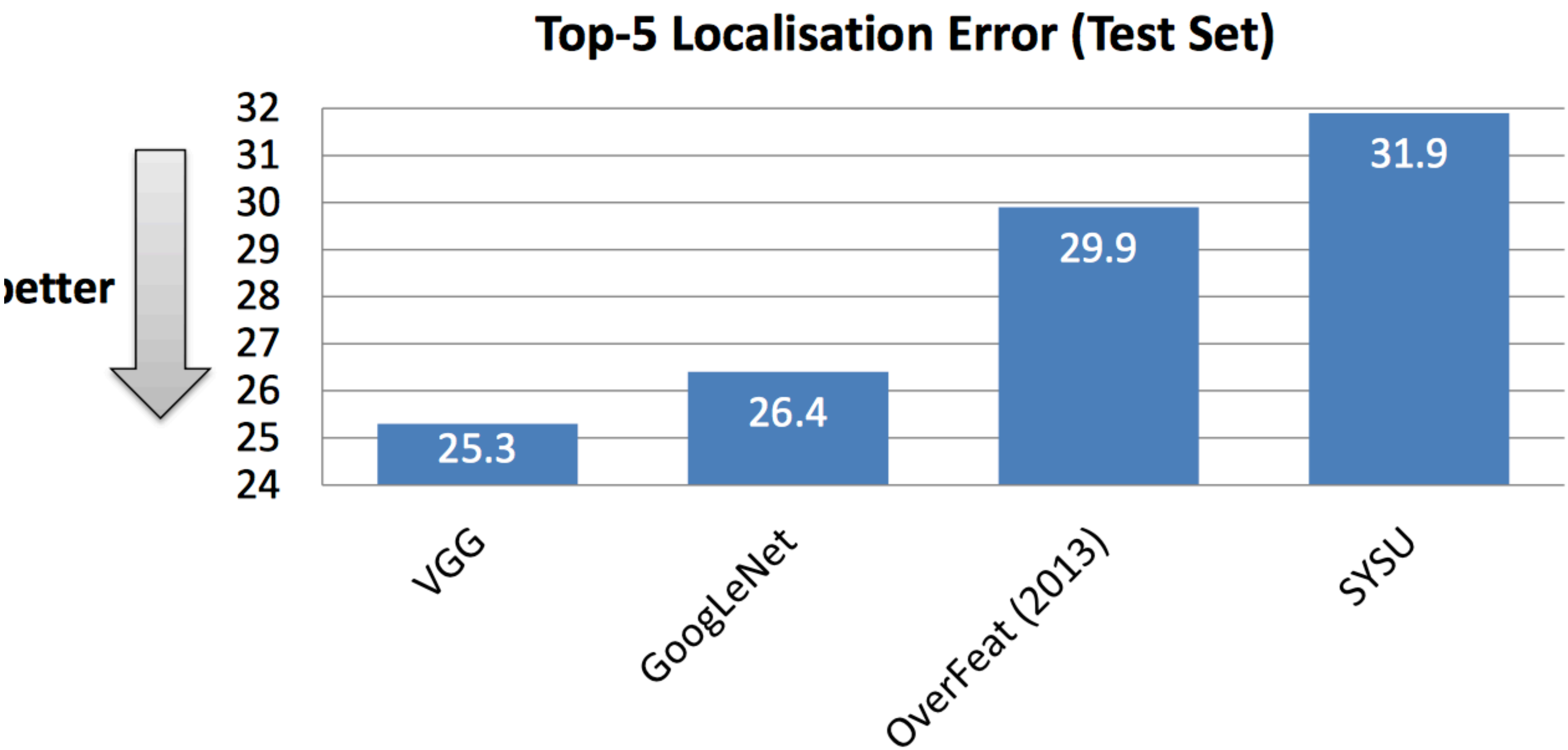
- Heavily-modified Caffe C++ toolbox
- Multiple GPU support
  - 4 x NVIDIA Titan, off-the-shelf workstation
  - data parallelism for training and testing
  - ~3.75 times speed-up, 2-3 weeks for training



Summary of VGG

- Localisation task
  - 1st place, 25.3% error
- Classification task
  - 2nd place, 7.3% error

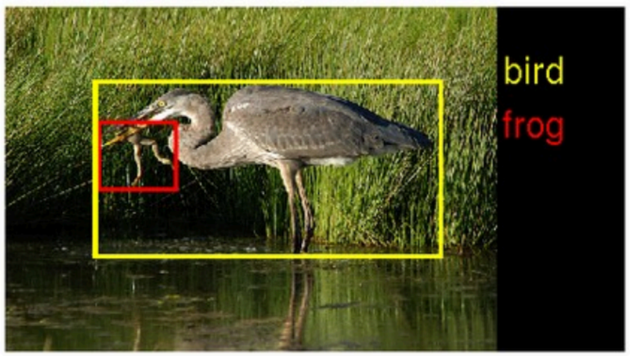
- Deep: 19 weight layers

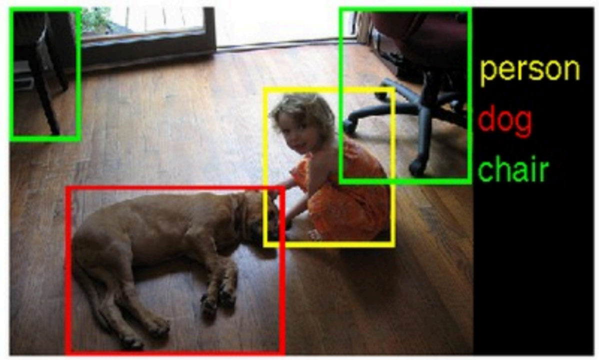


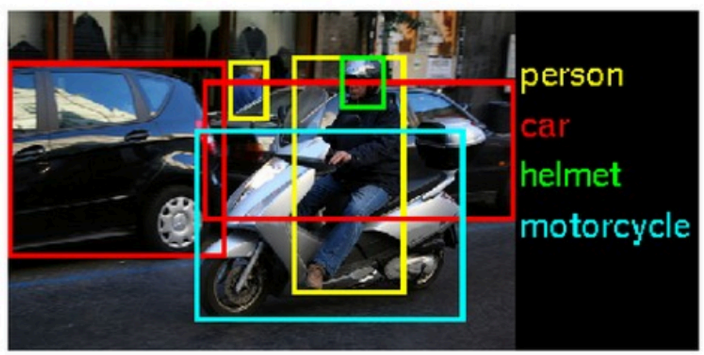
Top-5 Localisation Error (Test Set)
32
31
30
29
28
27
26
25
24
etter
25.3
26.4
29.9
31.9
VGG
GoogLeNet
OverFeat (2013)
SYSU

Detection

Needs to find all instances of the various classes







Model must output:

A set of detections

Each detection has:

- confidence
- class (integer)
- x1,y1,x2,y2 bounding box coordinates

Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

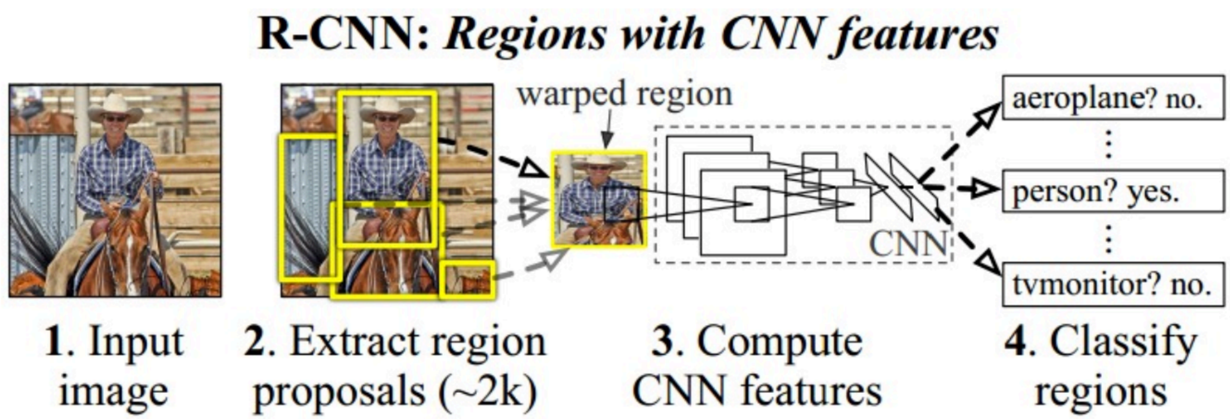
Idea*: Turn a Detection Problem into an Image Classification problem (but over image regions).*

Content of every labeled bounding box for is a positive example for a class.

Every other bounding box in the image is a special **negative class**.

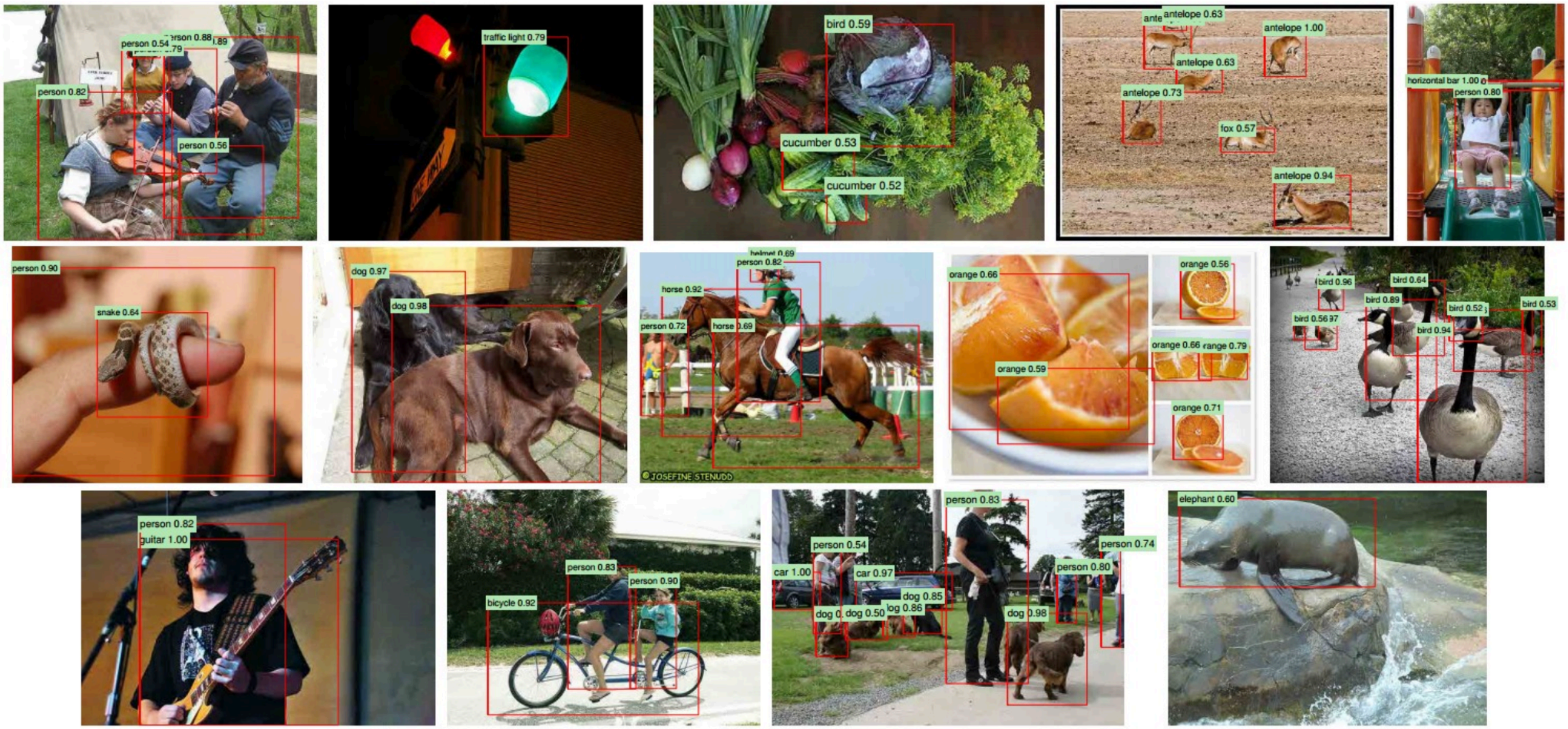
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

***Idea**: Turn a Detection Problem into an Image Classification problem (but over image regions).*

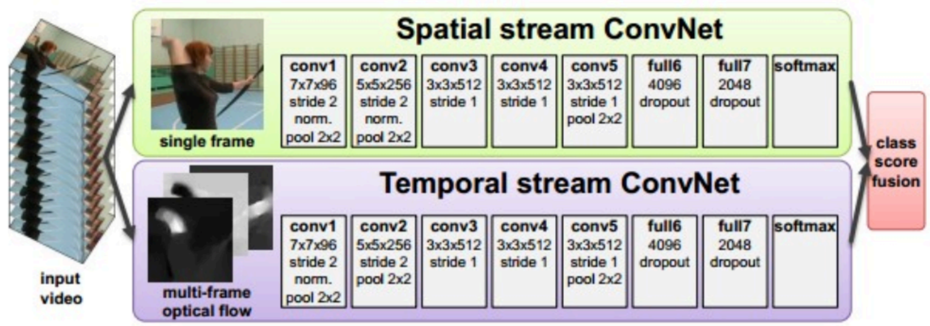


Rich feature hierarchies for accurate object detection and semantic segmentation

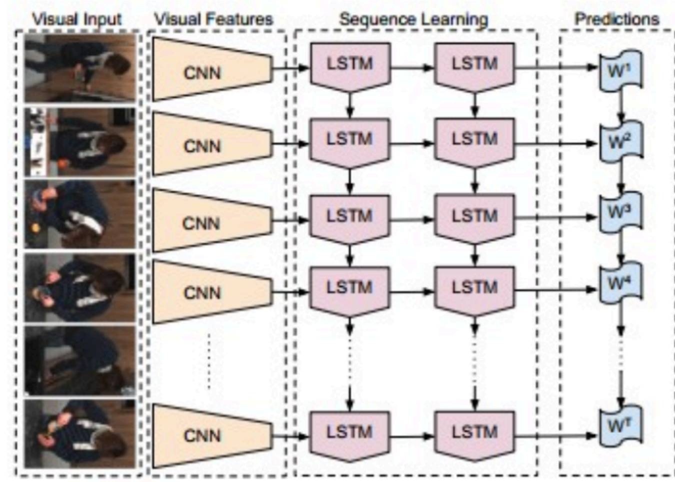
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]



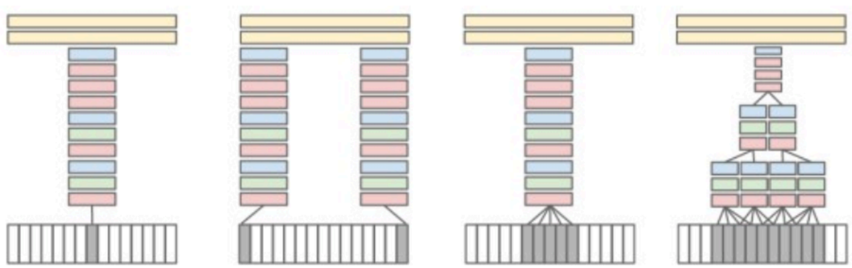
Video Classification



Two-Stream Convolutional Networks for Action Recognition in Videos *[Simonyan et al.], 2014*



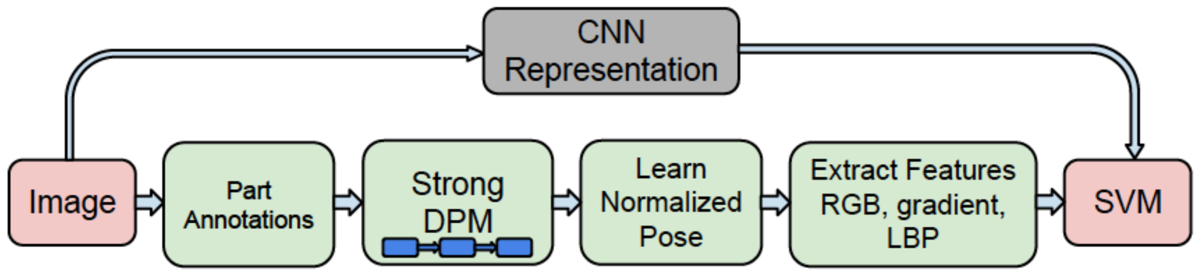
Long-term Recurrent Convolutional Networks for Visual Recognition and Description *[Donahue et al.], 2014*



Large-scale Video Classification with Convolutional Neural Networks *[Karpathy et al.], 2014*

CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden



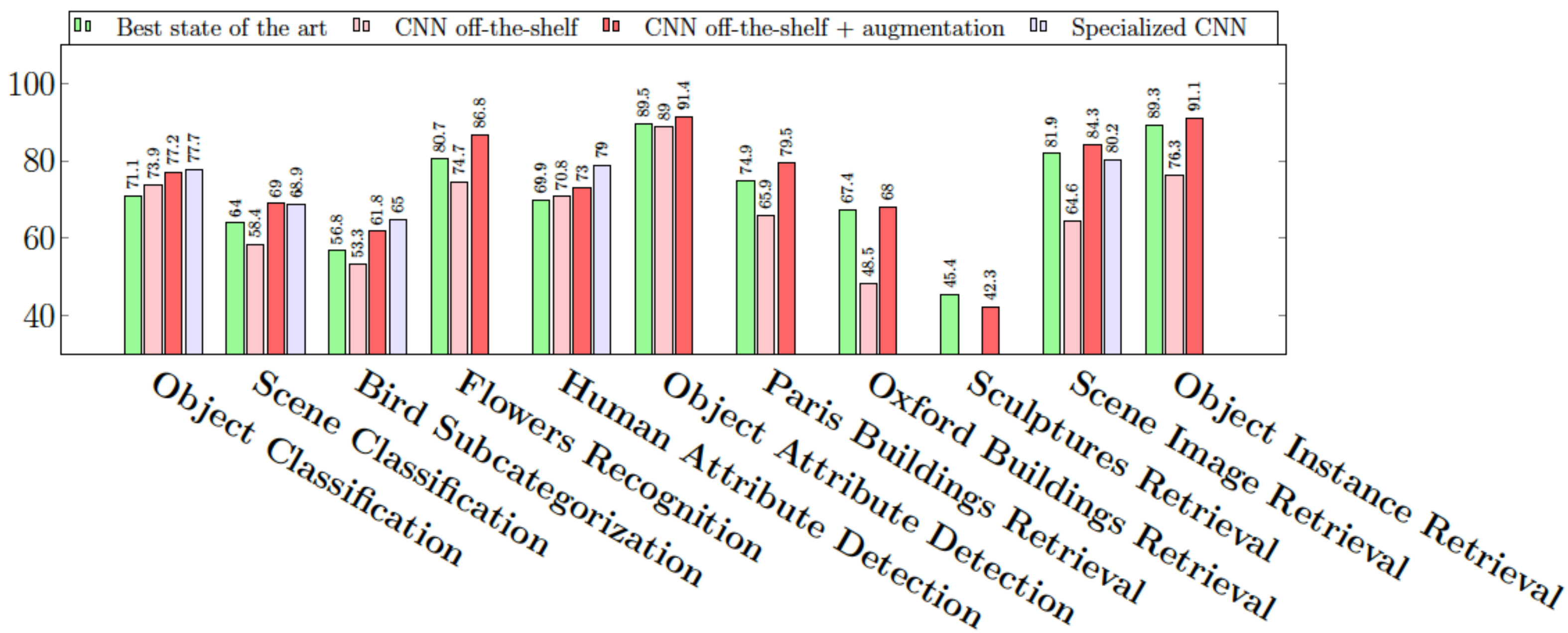


Image Captioning

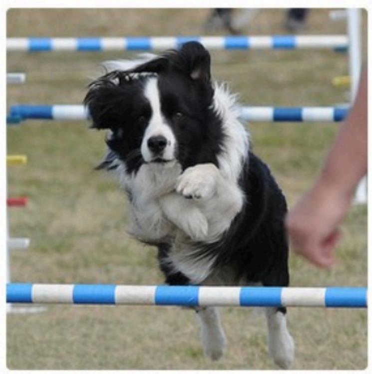
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."



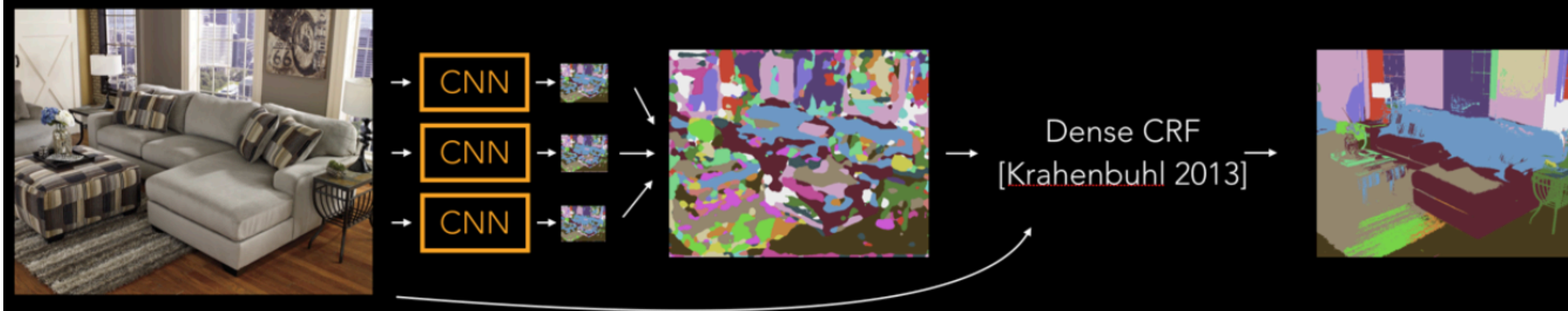
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Material Segmentation[CVPR15]

CNNs + CRFs



CRF Runtime: ~1s for 640x480 image

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

ConvNets breakthroughs for visual tasks

	Dataset	Performance	Score
[Sermanet et al 2014]: OverFeat (fine-tuned features for each task) (tasks are ordered by increasing difficulty)			
• image classification	ImageNet LSVRC 2013	competitive	13.6 % error
	Dogs vs Cats Kaggle challenge 2014	**state of the art**	98.9%
• object localization	ImageNet LSVRC 2013	**state of the art**	29.9% error
• object detection	ImageNet LSVRC 2013	competitive	24.3% mAP
[Razavian et al, 2014]: public OverFeat library (no retraining) + SVM **(simplest approach possible on purpose, no attempt at more complex classifiers)** (tasks are ordered by "distance" from classification task on which OverFeat was trained)			
• image classification	Pascal VOC 2007	competitive	77.2% mAP
• scene recognition	MIT-67	**state of the art**	69% mAP
• fine grained recognition	Caltech-UCSD Birds 200-2011	competitive	61.8% mAP
	Oxford 102 Flowers	**state of the art**	86.8% mAP
• attribute detection	UIUC 64 object attributes	**state of the art**	91.4% mAUC
	H3D Human Attributes	competitive	73% mAP
• image retrieval (search by image similarity)	Oxford 5k buildings	**state of the art**	68% mAP?
	Paris 6k buildings	**state of the art**	79.5% mAP?
	Sculp6k	competitive	42.3% mAP?
	Holidays	**state of the art**	84.3% mAP?
	UKBench	**state of the art**	91.1% mAP?

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun, **OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks**, http://arxiv.org/abs/1312.6229, ICLR 2014

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, **CNN Features off-the-shelf: an Astounding Baseline for Recognition**, http://arxiv.org/abs/1403.6382, DeepVision CVPR 2014 workshop

ConvNets breakthroughs for visual tasks

	Dataset	Performance	Score
[Zeiler et al 2013]			
• image classification	ImageNet LSVRC 2013	**state of the art**	11.2% error
	Caltech-101 (15, 30 samples per class)	competitive	83.8%, 86.5%
	Caltech-256 (15, 60 samples per class)	**state of the art**	65.7%, 74.2%
	Pascal VOC 2012	competitive	79% mAP
[Donahue et al, 2014]: DeCAF+SVM			
• image classification	Caltech-101 (30 classes)	**state of the art**	86.91%
• domain adaptation	Amazon -> Webcam, DSLR -> Webcam	**state of the art**	82.1%, 94.8%
• fine grained recognition	Caltech-UCSD Birds 200-2011	**state of the art**	65.0%
• scene recognition	SUN-397	competitive	40.9%
[Girshick et al, 2013]			
• image detection	Pascal VOC 2007	**state of the art**	48.0% mAP
	Pascal VOC 2010 (comp4)	**state of the art**	43.5% mAP
	ImageNet LSVRC 2013	**state of the art**	31.4% mAP
• image segmentation	Pascal VOC 2011 (comp6)	**state of the art**	47.9% mAP
[Oquab et al, 2013]			
• image classification	Pascal VOC 2007	**state of the art**	77.7% mAP
	Pascal VOC 2012	**state of the art**	82.8% mAP
	Pascal VOC 2012 (action classification)	**state of the art**	70.2% mAP

M.D. Zeiler, R. Fergus, **Visualizing and Understanding Convolutional Networks,** Arxiv 1311.2901 http://arxiv.org/abs/1311.2901

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. **Decaf: A deep convolutional activation feature for generic visual recognition**. In ICML, 2014, http://arxiv.org/abs/1310.1531

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. **Rich feature hierarchies for accurate object detection and semantic segmentation**. arxiv:1311.2524 [cs.CV], 2013, http://arxiv.org/abs/1311.2524

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. **Learning and transferring mid-level image representations using convolutional neural networks.** Technical Report HAL-00911179, INRIA, 2013. http://hal.inria.fr/hal-00911179

ConvNets breakthroughs for visual tasks

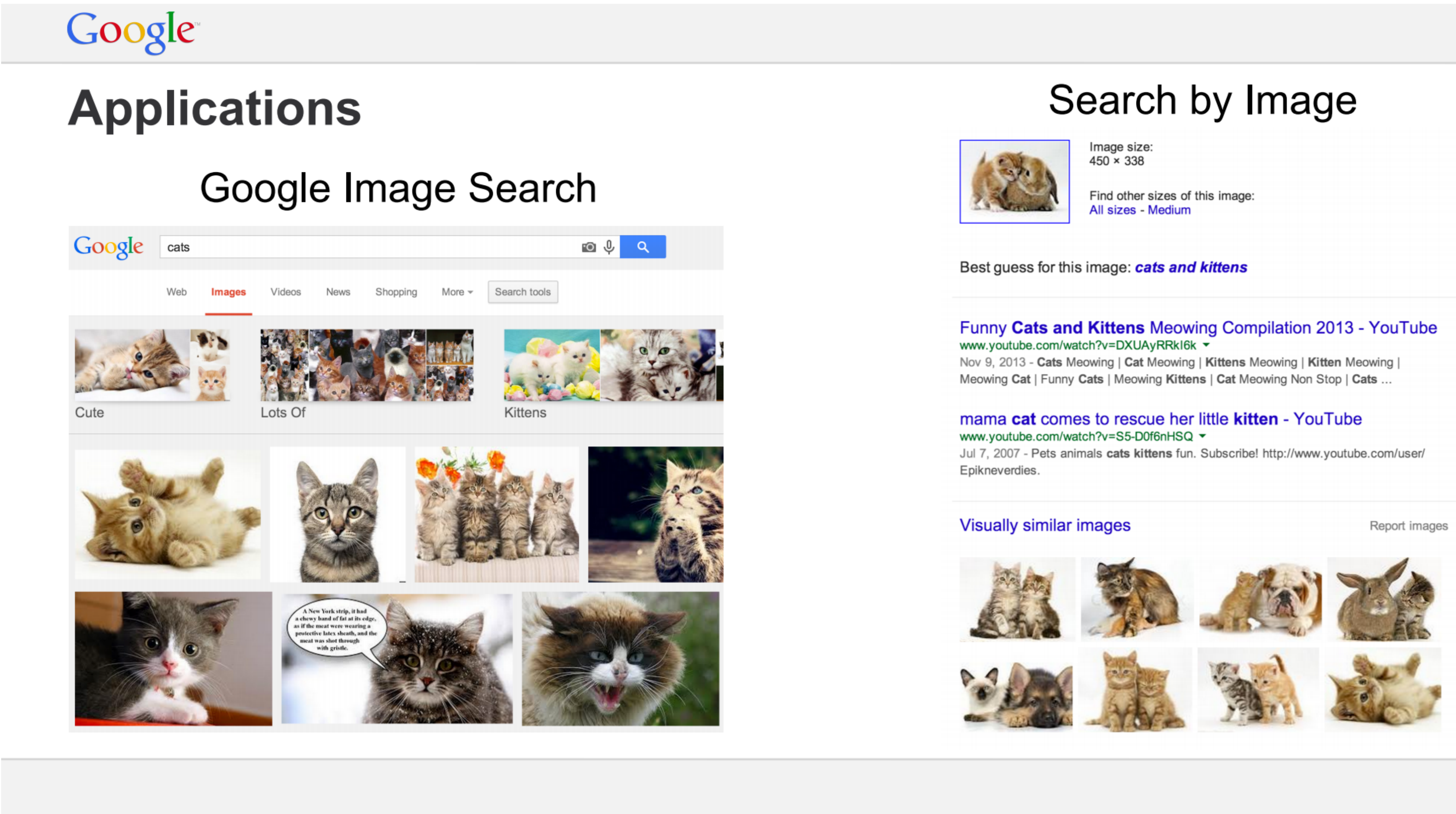
	Dataset	Performance	Score
[Khan et al 2014] • shadow detection	UCF CMU UIUC	**state of the art** **state of the art** **state of the art**	90.56% 88.79% 93.16%
[Sander Dieleman, 2014] • image attributes	Kaggle Galaxy Zoo challenge	**state of the art**	0.07492

S. H. Khan, M. Bennamoun, F. Sohel, R. Togneri. **Automatic Feature Learning for Robust Shadow Detection,** CVPR 2014
Sander Dieleman, Kaggle Galaxy Zoo challenge 2014 http://benanne.github.io/2014/04/05/galaxy-zoo.html

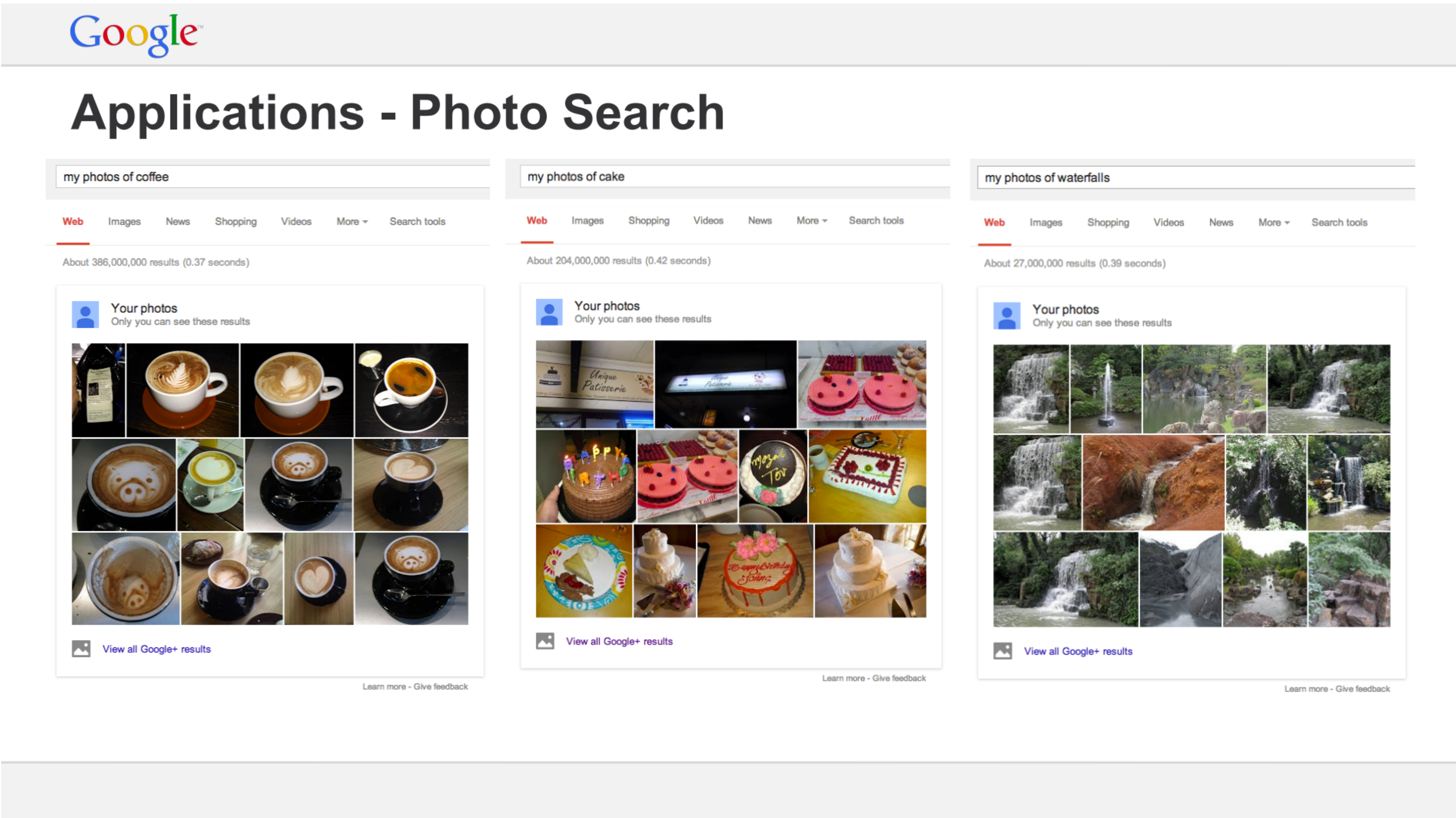
"It can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task."

[Razavian 2014]

CNNs at Google (as of 2014)



CNNs at Google (as of 2014)

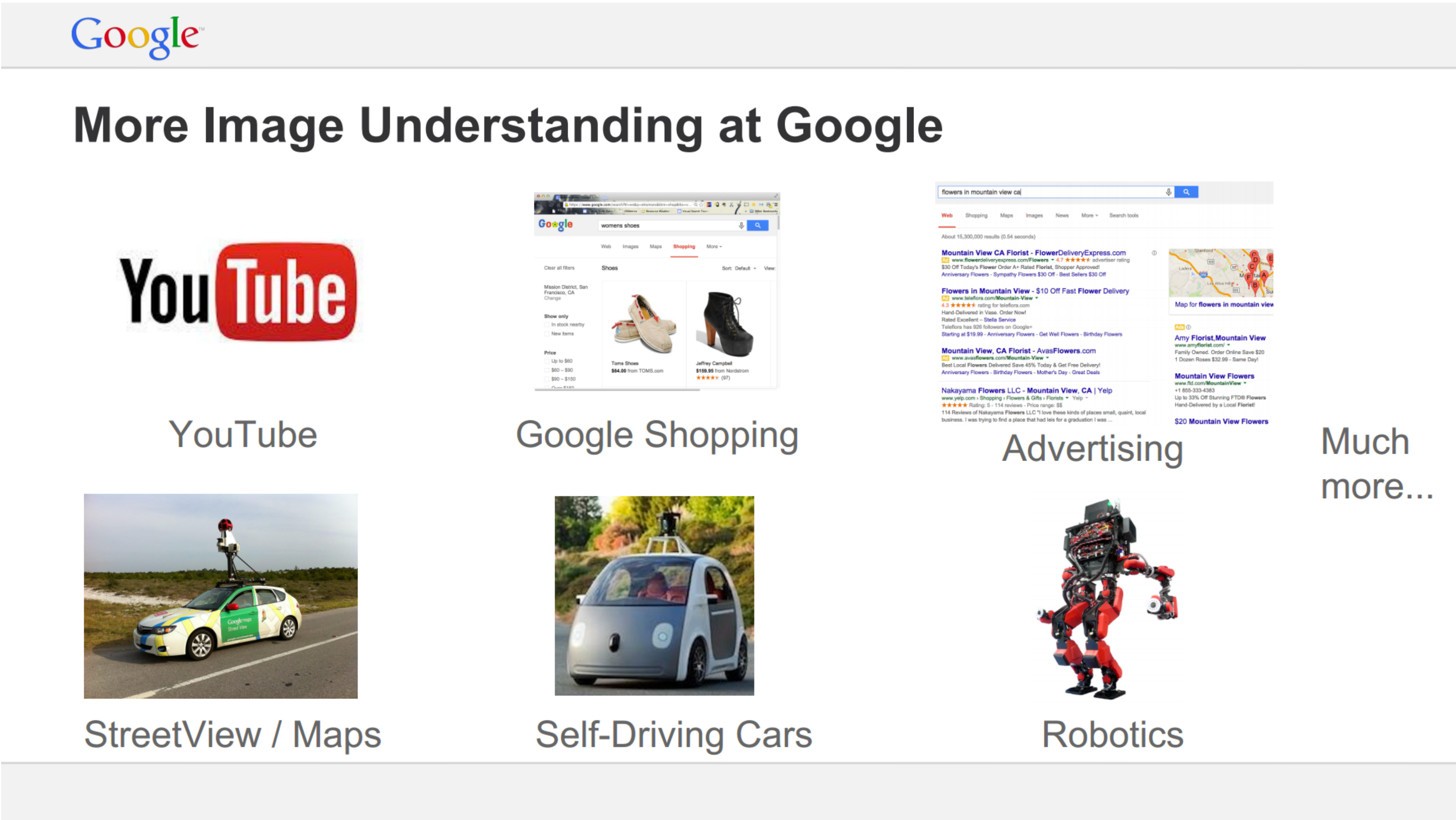


CNNs at Google (as of 2014)

Google

Google Photos - Auto Awesome

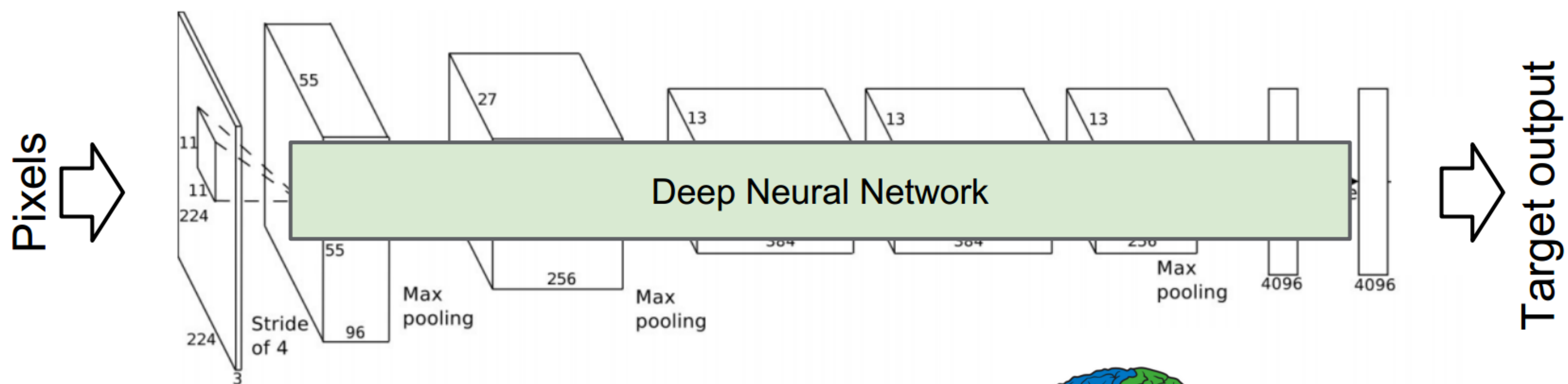
CNNs at Google (as of 2014)



CNNs at Google (as of 2014)

Google

The Deep and now Deeper Hammer



Deep learning infrastructure by the Google Brain team

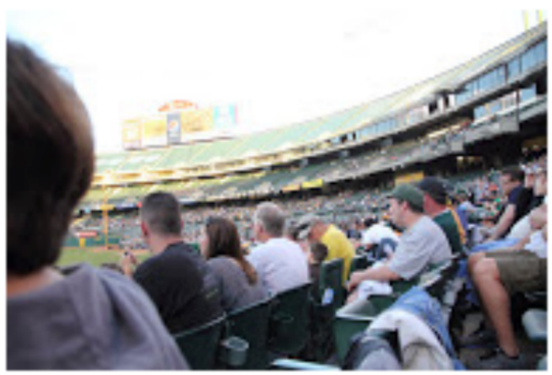
"ImageNet Classification with Deep Convolutional Neural Networks", Krizhevsky, Sutskever, Hinton, NIPS 2012

CNNs at Google (as of 2014)

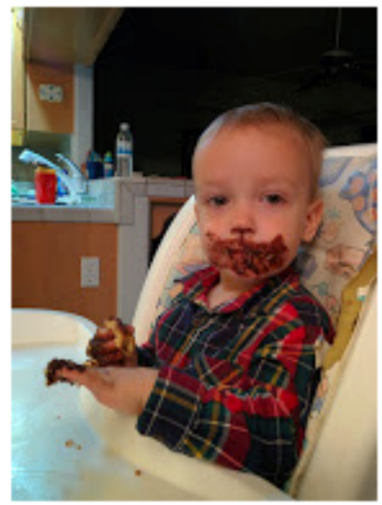
Google

Personal Photos - Example Annotations

Christmas tree
Red
Christmas decoration
Christmas



Crowd
Cheering
People
Stadium



Play
Meal
Cake
Child

Hummingbird
Macro photography
Reflection
Red