BMI 826 / CS 838: Learning Based Methods in Computer vision

Course Information

- Time:
  - Monday / Wednesday 1:05PM - 2:20PM
- Location:
  - 3534 Engineering Hall
- Office Hours:
  - Email me for appointments
- Contact:
  - yin.li@wisc.edu (MSC 6730)
- Website:
  - https://www.biostat.wisc.edu/~yli/bmi826_cs838/

Suggested Course Path

- Prerequisites
  - CS 766 Computer Vision OR
  - BMI/CS 767 Medical Image Analysis

  (We assume that you have already known some of the materials from CS 766 or BMI/CS 767)

- This is a computer vision course!

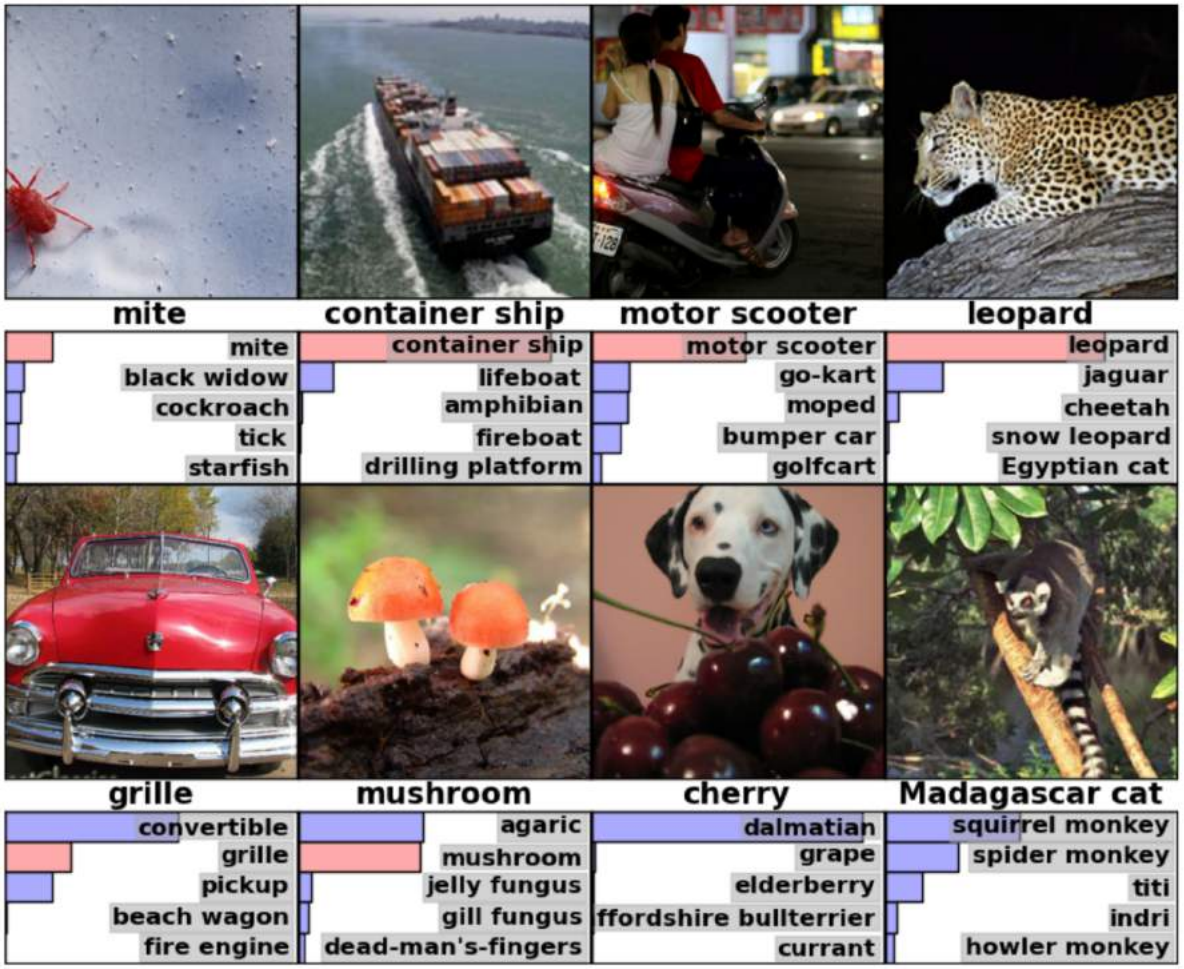
- And this is NOT a machine learning course (although we did cover some aspects of learning)

People

- Instructor: Yin Li (yin.li@wisc.edu)
- TA: Zixuan Huang (zhuang356@wisc.edu)
- **Canvas** for announcements, homework / project submissions, etc.
- **Piazza** for online discussions

What is this course about?

Image Classification



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

Object Detection and Segmentation

“Mask RCNN”, He, Gkioxari, Dollár, Girshick, *ICCV,* 2017

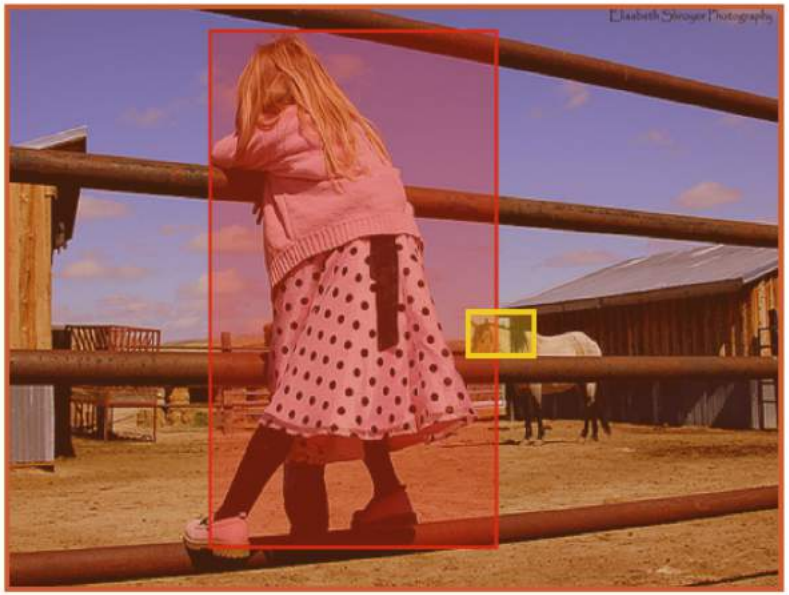
Human Pose Estimation

“OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields”,
Cao, Hidalgo, Simon, Wei and Sheikh, ArXiv 2018

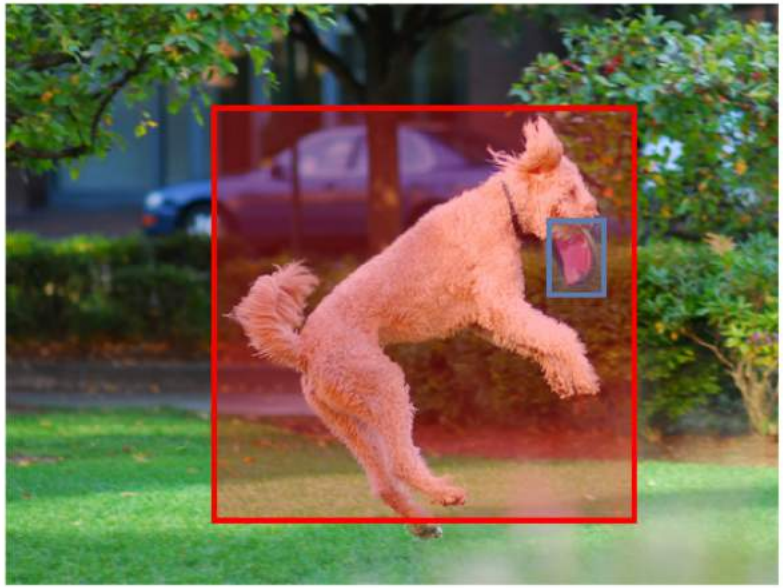
3D Understanding (Pose and Shape)

“3D-RCNN: Instance-level 3D Object Reconstruction via Render-and-Compare”,
Kundu, Li, Rehg. CVPR, 2018

Image-text Resolution



A little girl stands on the fence while peeking through it to look at the horse



A large yellow dog leaps into the air to catch his Frisbee.

“Learning Deep Structure-Preserving Image-Text Embeddings”,
Wang, Li, Lazebnik. CVPR 2016.

Visual Question Answering

Who is wearing glasses?

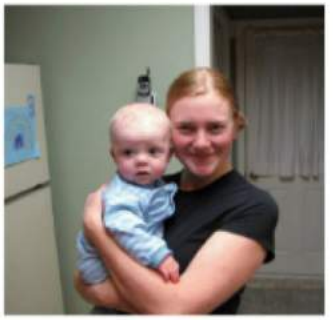
man

woman

Where is the child sitting?

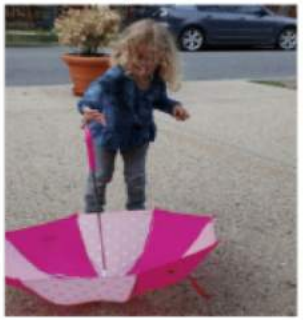
fridge

arms

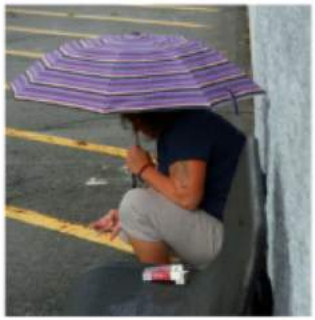


Is the umbrella upside down?

yes



no



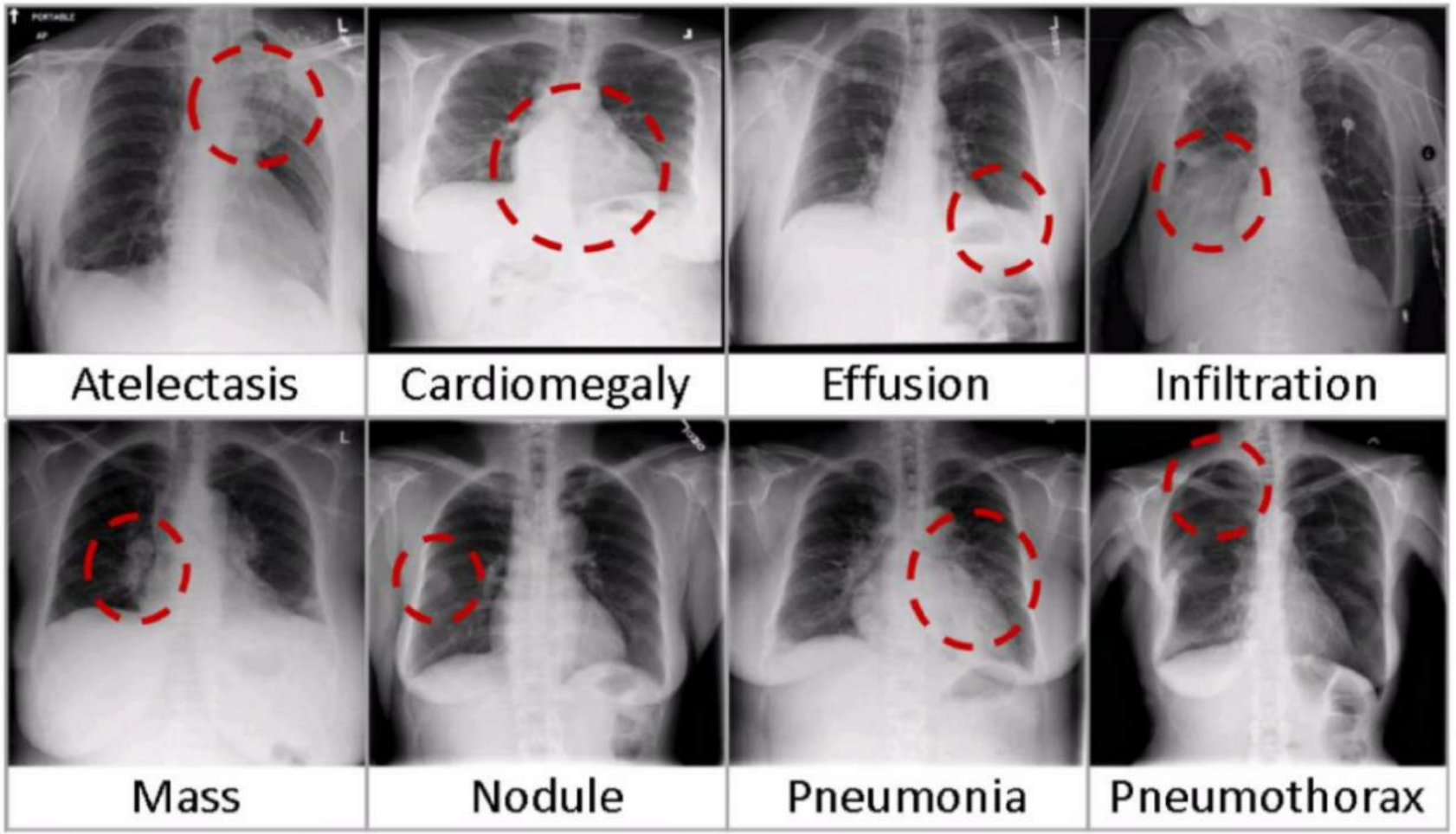
How many children are in the bed?

2

1

"Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering", Yash Goyal, Khot, Summers-Stay, Batra, Parik*, CVPR,* 2017

Medical Image Analysis



"ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases", Wang et al., *CVPR,* 2017

Workload & Rubric

- 3 homework assignments (45% with 15% each)
  - Homework 2 & 3 are team based
- 1 final project (40%)
  - Team based (2-3 people)
  - Multiple milestones (proposal, mid-term update, final report and presentation)
- Single page course write-up (5%)
- 2 in-class quizzes (10% with 5% each)
  - Dates in syllabus

Introduction to Visual Recognition

Yin Li

Many Slides from A. Efros, L. Zitnick, and A. Gupta

Motivation

What is visual recognition?

What did we try?

Where are we now?
(Deep Models)

Goals

15 min: The challenge of visual recognition

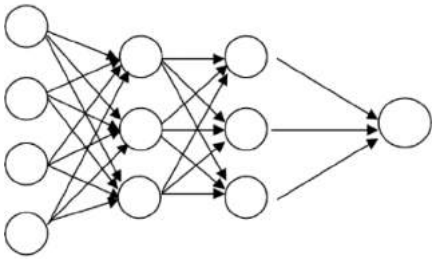
What is visual recognition, why it is hard!

25 min: History of visual recognition

Why it works, why it fails!

20 min: Deep models for recognition

Will deep neural networks solve it all?



The Vision Story Begins

“What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking.”

-- David Marr, *Vision (1982)*

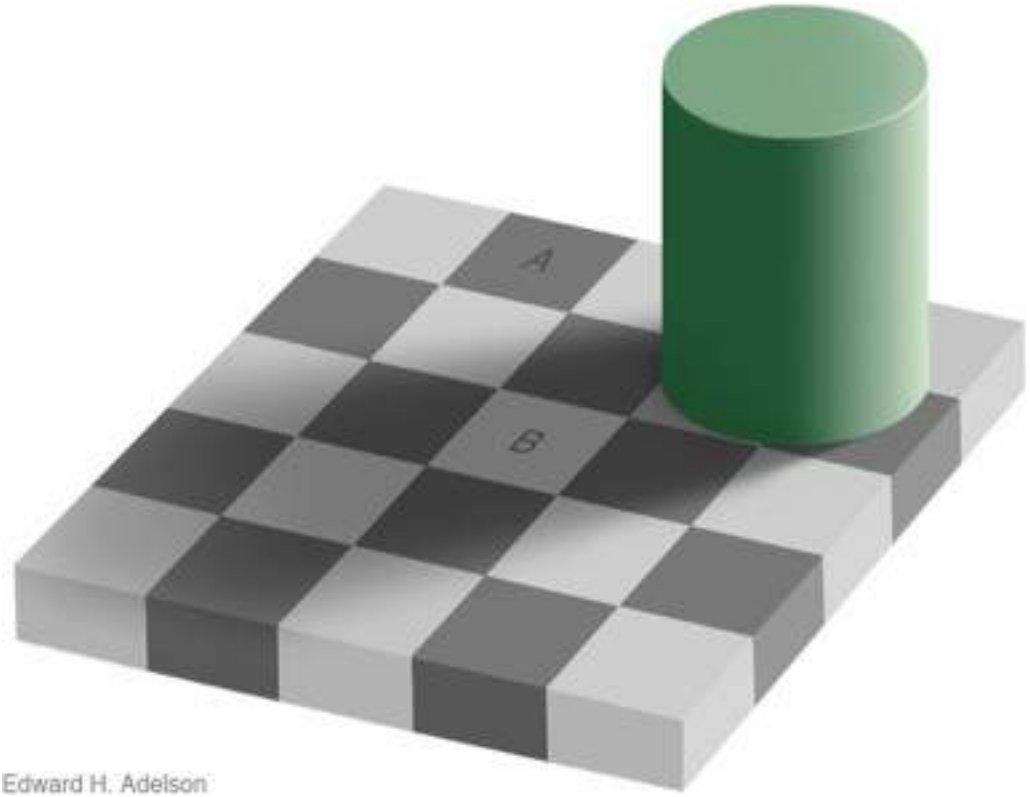
Two Sides of Vision

Which one do you want?

Answer #2: looks like flat sittable surface of the couch

Answer #1: pixel of brightness 243 at position (124,54) … and depth .7 meters

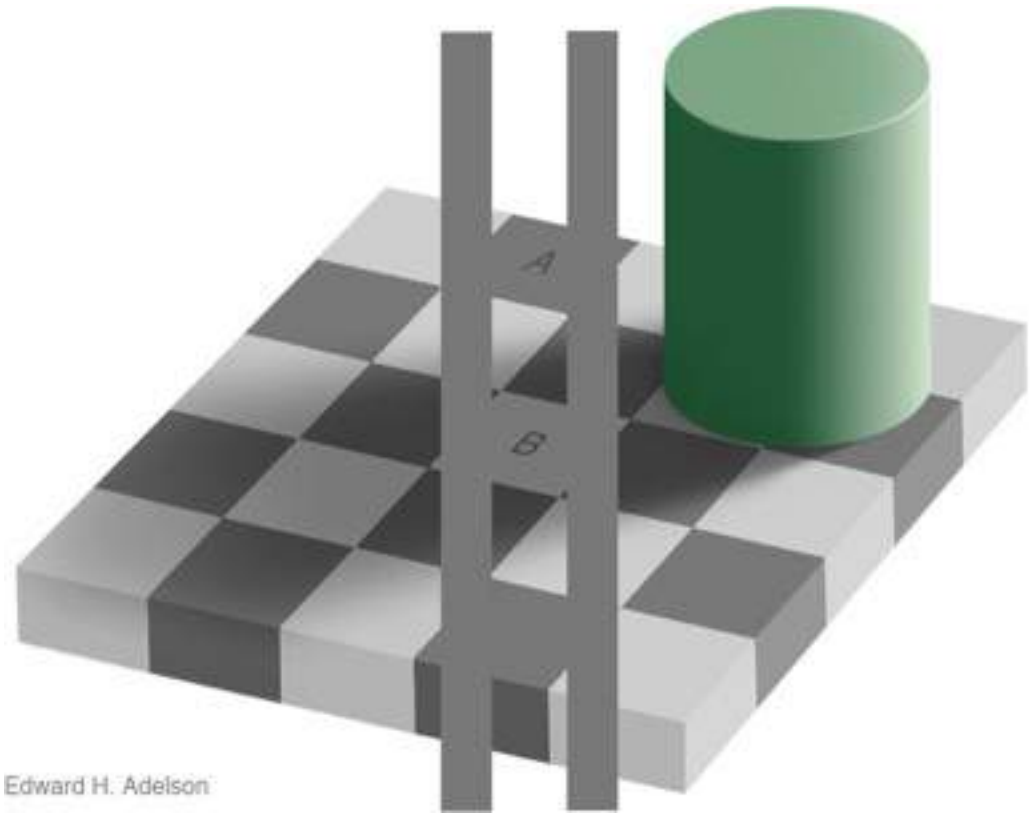
Measurement vs. Perception



Edward H. Adelson

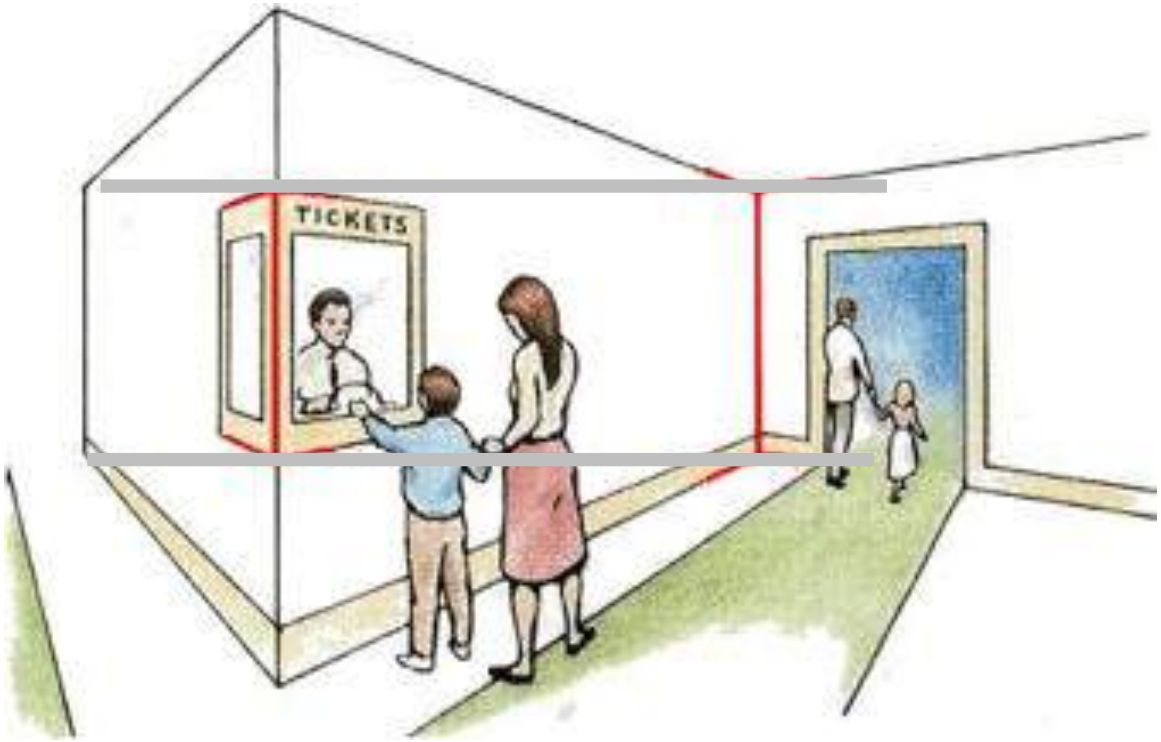
Slide Credit: Alyosha

Measurement vs. Perception



Slide Credit: Alyosha

Measurement vs. Perception



Müller-Lyer Illusion

http://www.michaelbach.de/ot/sze_muelue/index.html

Slide Credit: Alyosha

Measurement

Capturing physical quantities like pixel brightness, depth, etc.

Perception

- a high-level representation that captures the semantic structure of the scene and objects
- can be subjective – Depends on Task and Agent
- intersection of what you see and what you believe

Visual Recognition = Perception

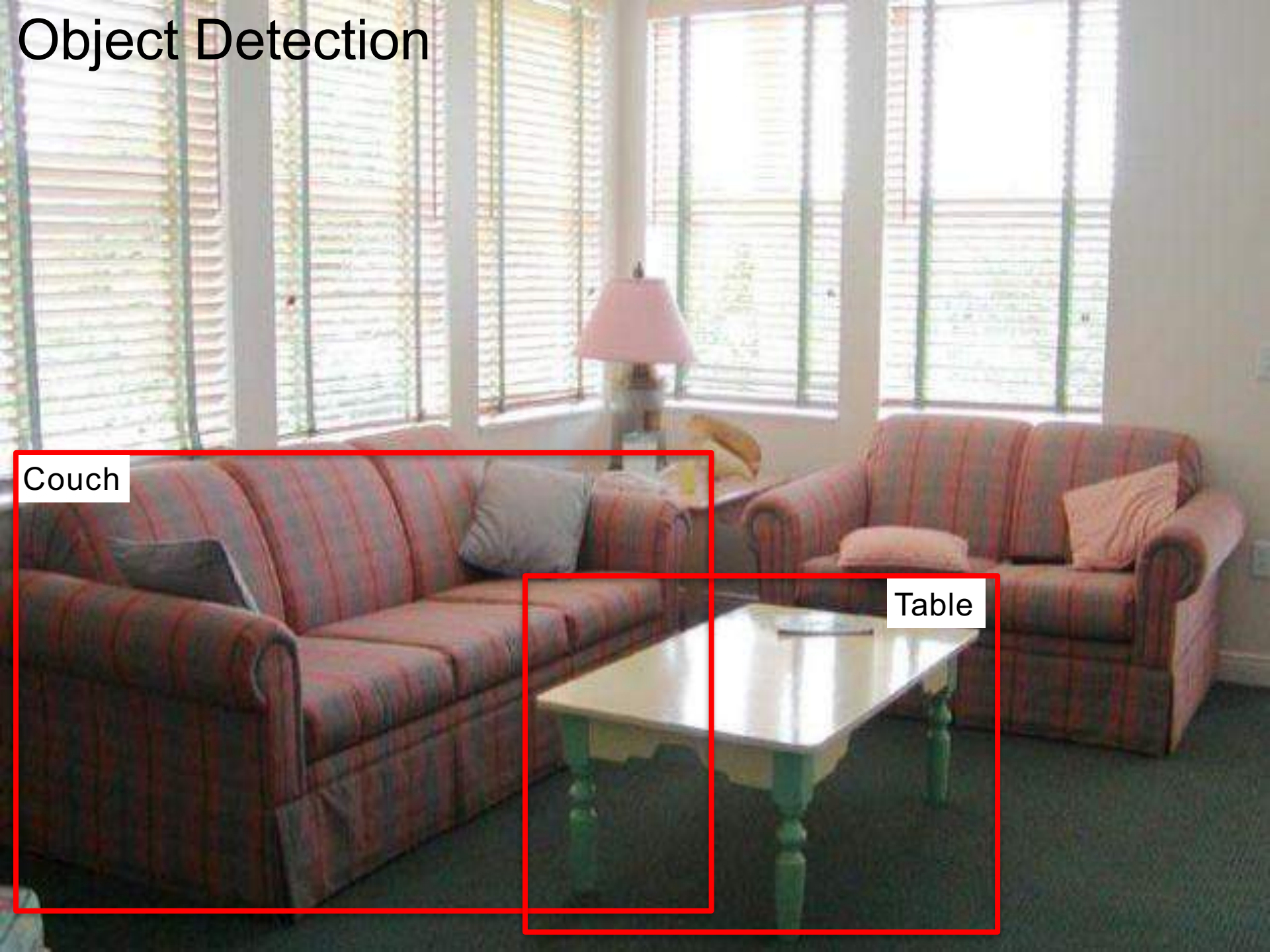
To infer what humans perceive (what + where)

So what do humans care about?

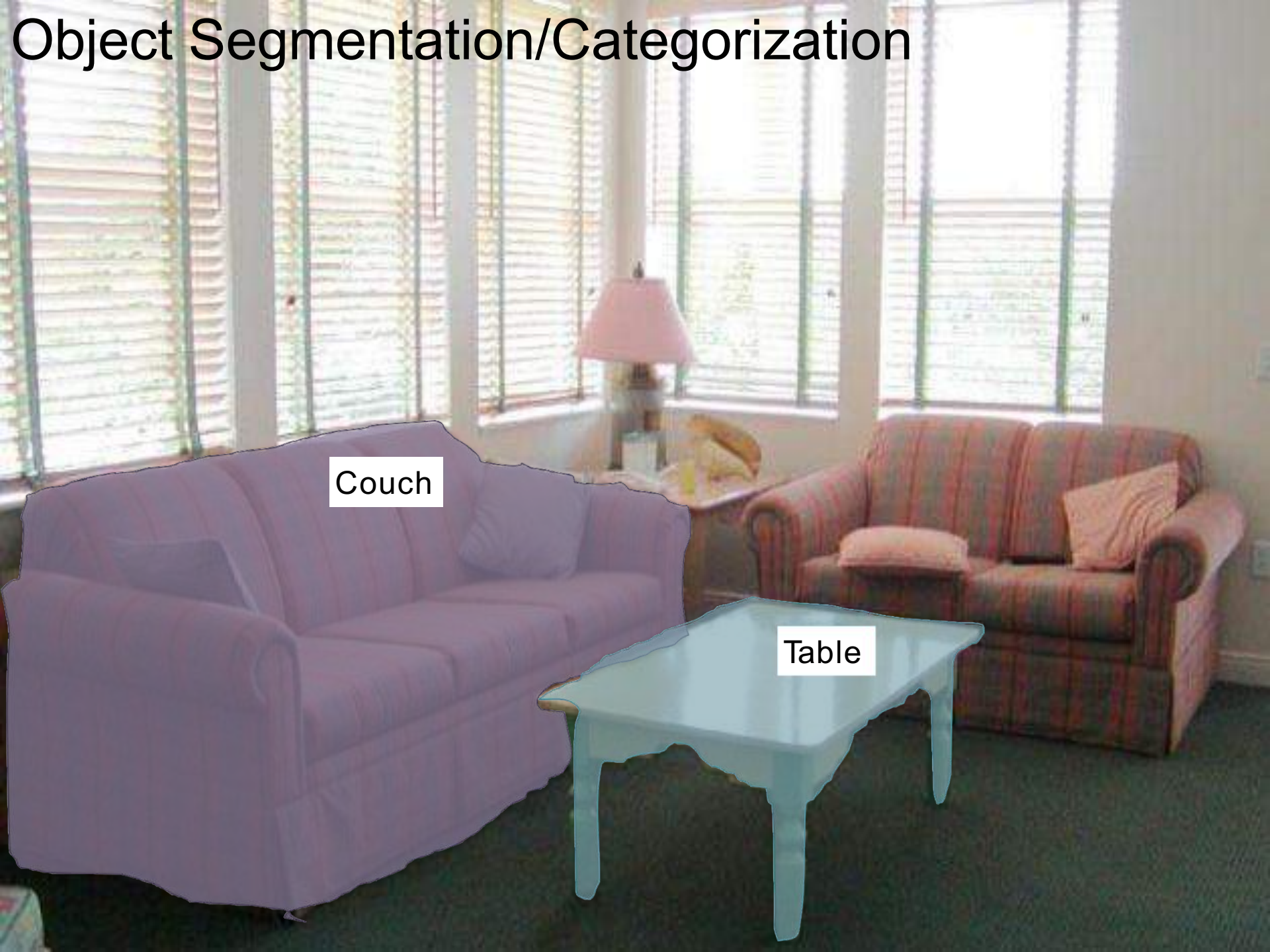
Image Classification/Scene Recognition

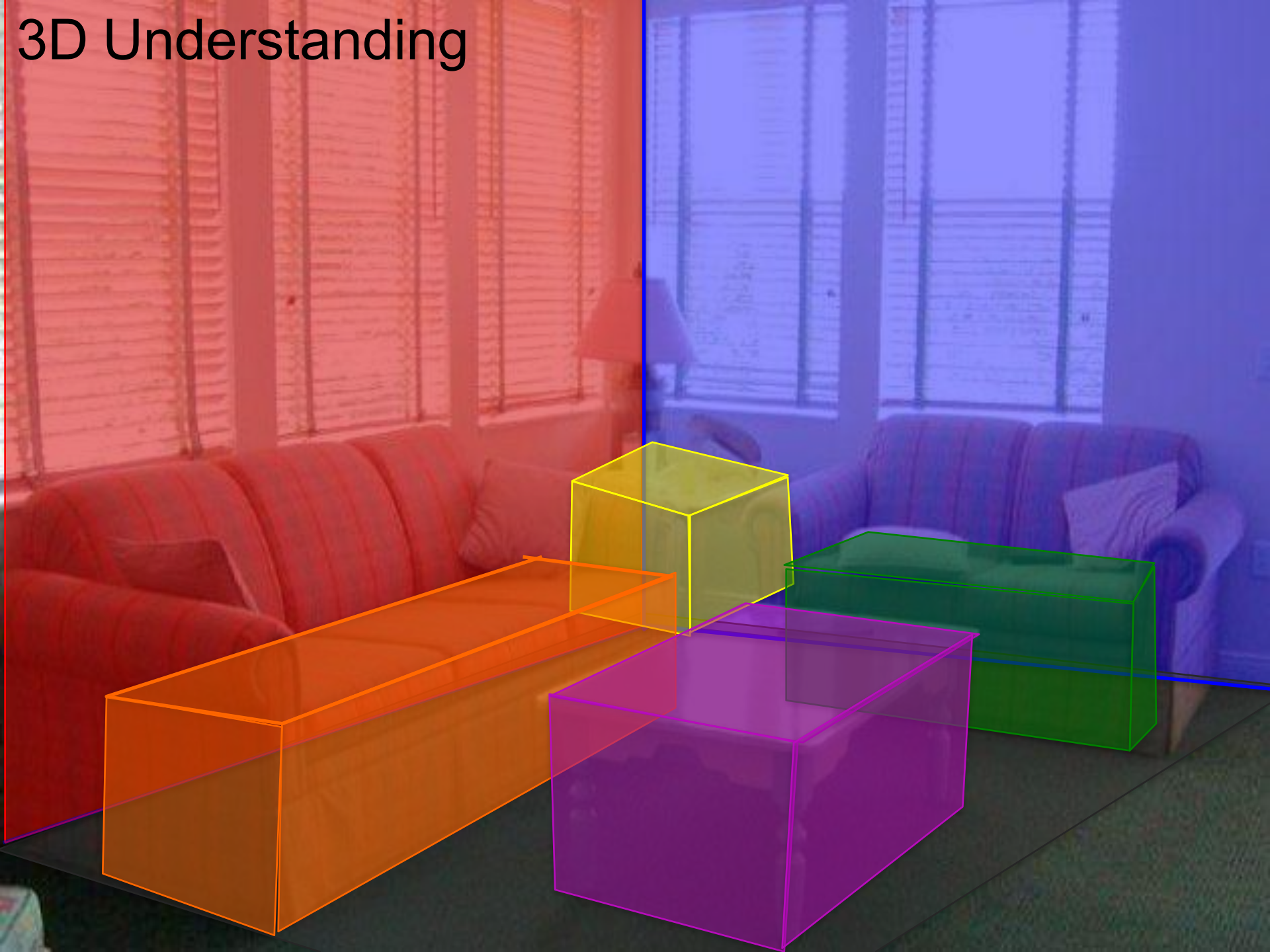
Object Recognition

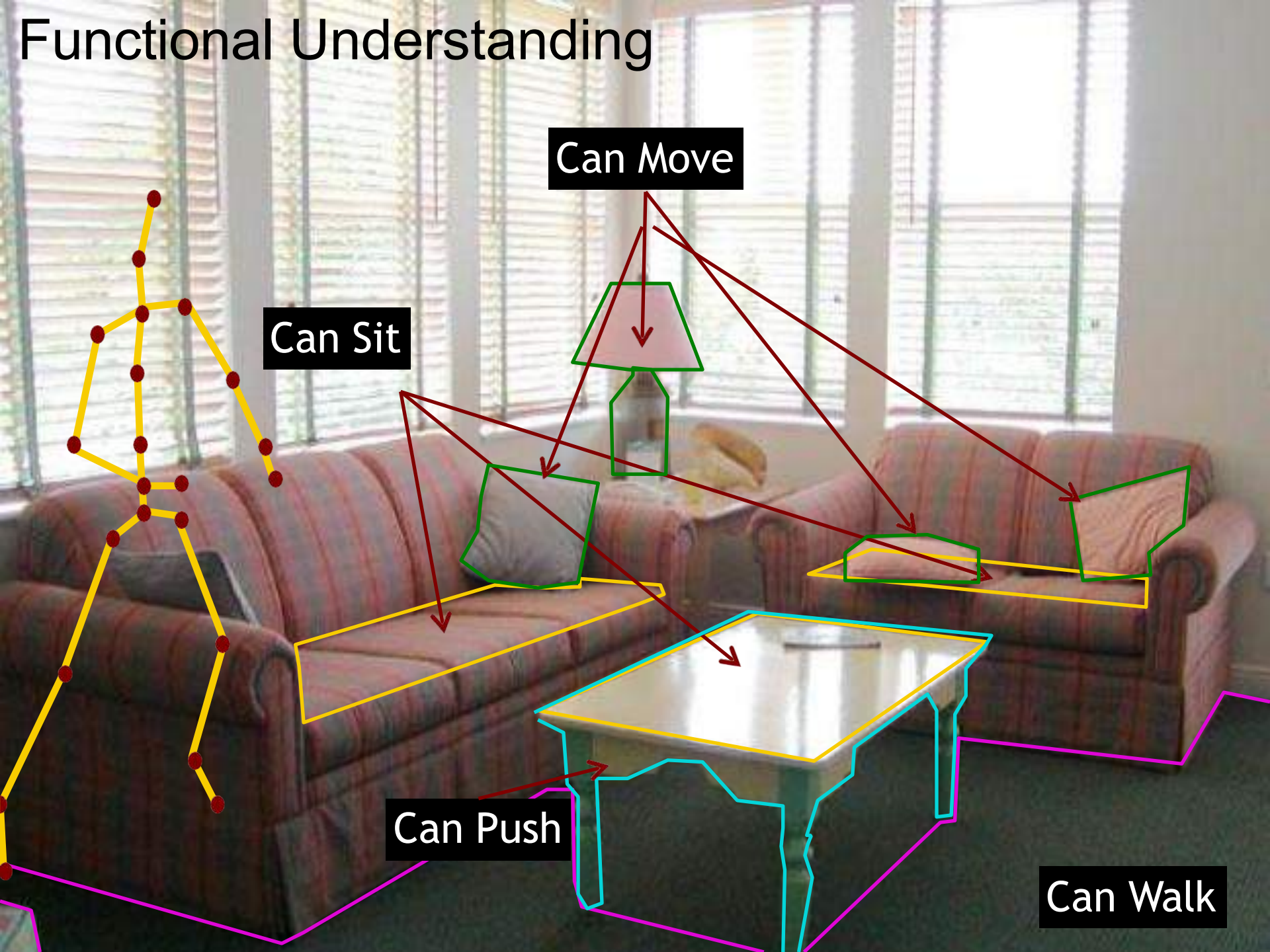
Object Detection



Object Segmentation/Categorization

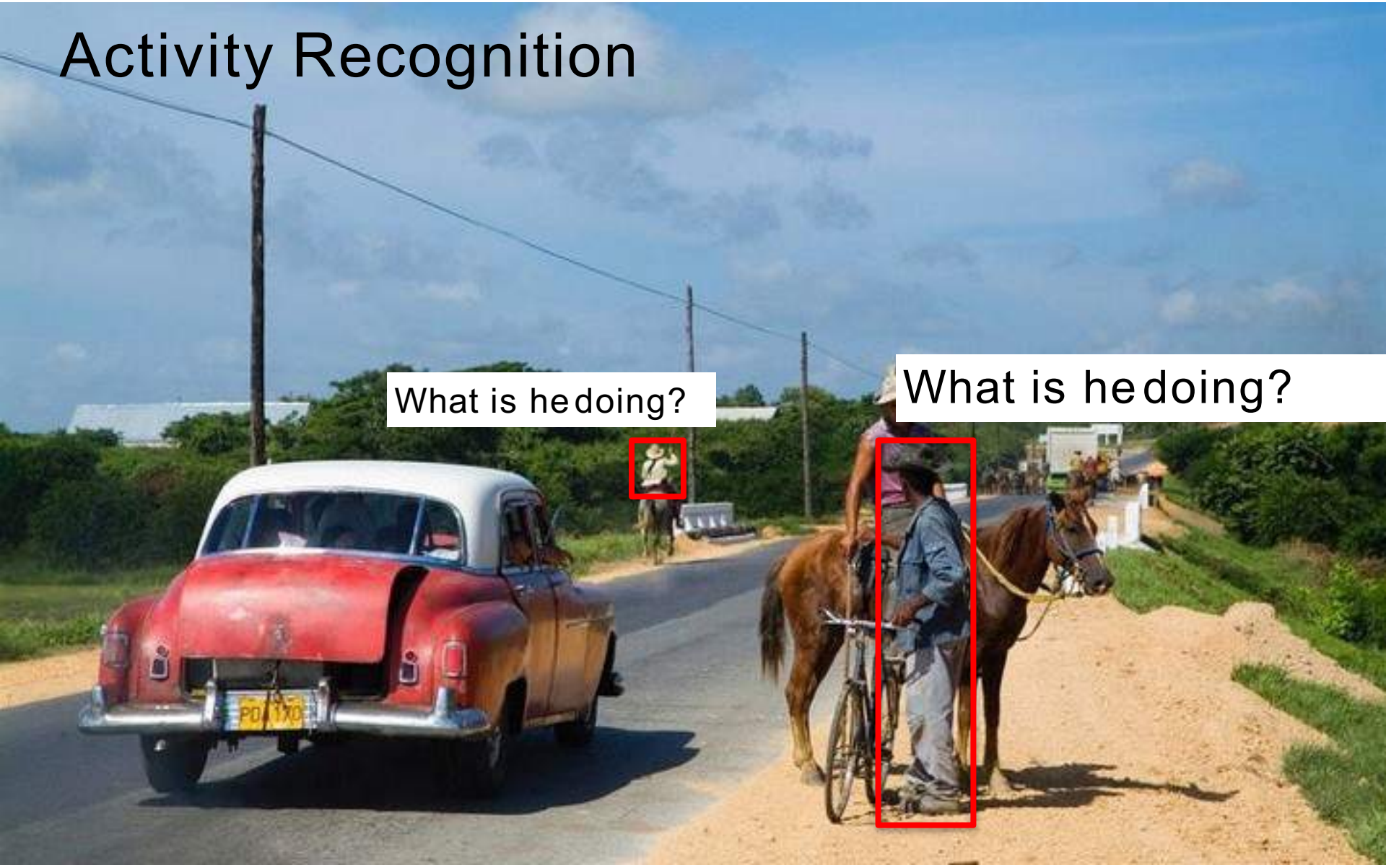



3D Understanding


Functional Understanding
Can Move
Can Sit
Can Push
Can Walk

Pose Estimation

Activity Recognition



Why are the problems hard?

To infer what humans perceive from 2D pixels

Challenges 1: view point variation

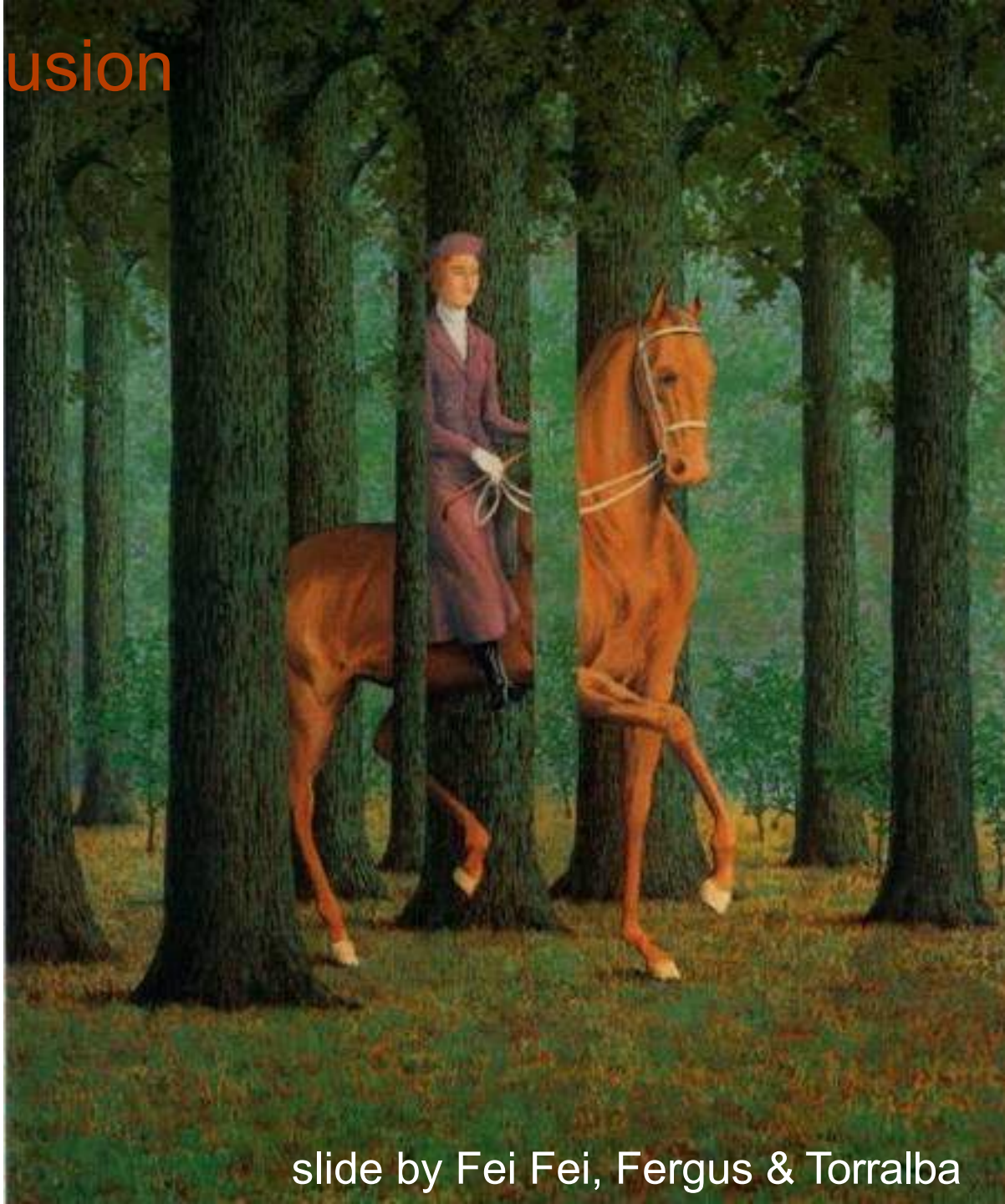
Michelangelo 1475-1564

slide by Fei Fei, Fergus & Torralba

Challenges 2: illumination

slide credit: S. Ullman

Challenges 3: occlusion



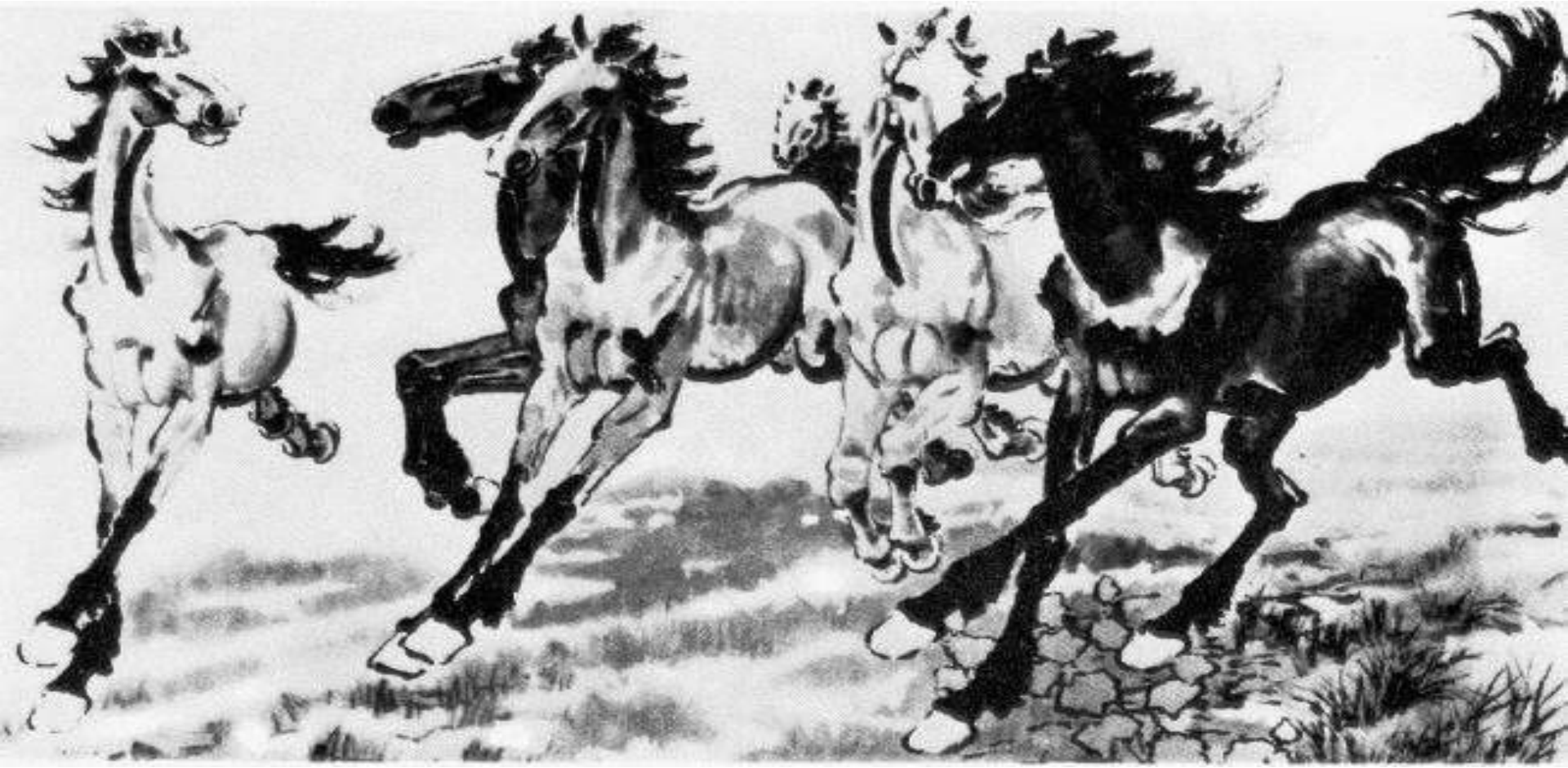
Magritte, 1957

slide by Fei Fei, Fergus & Torralba

Challenges 4: scale

slide by Fei Fei, Fergus & Torralba

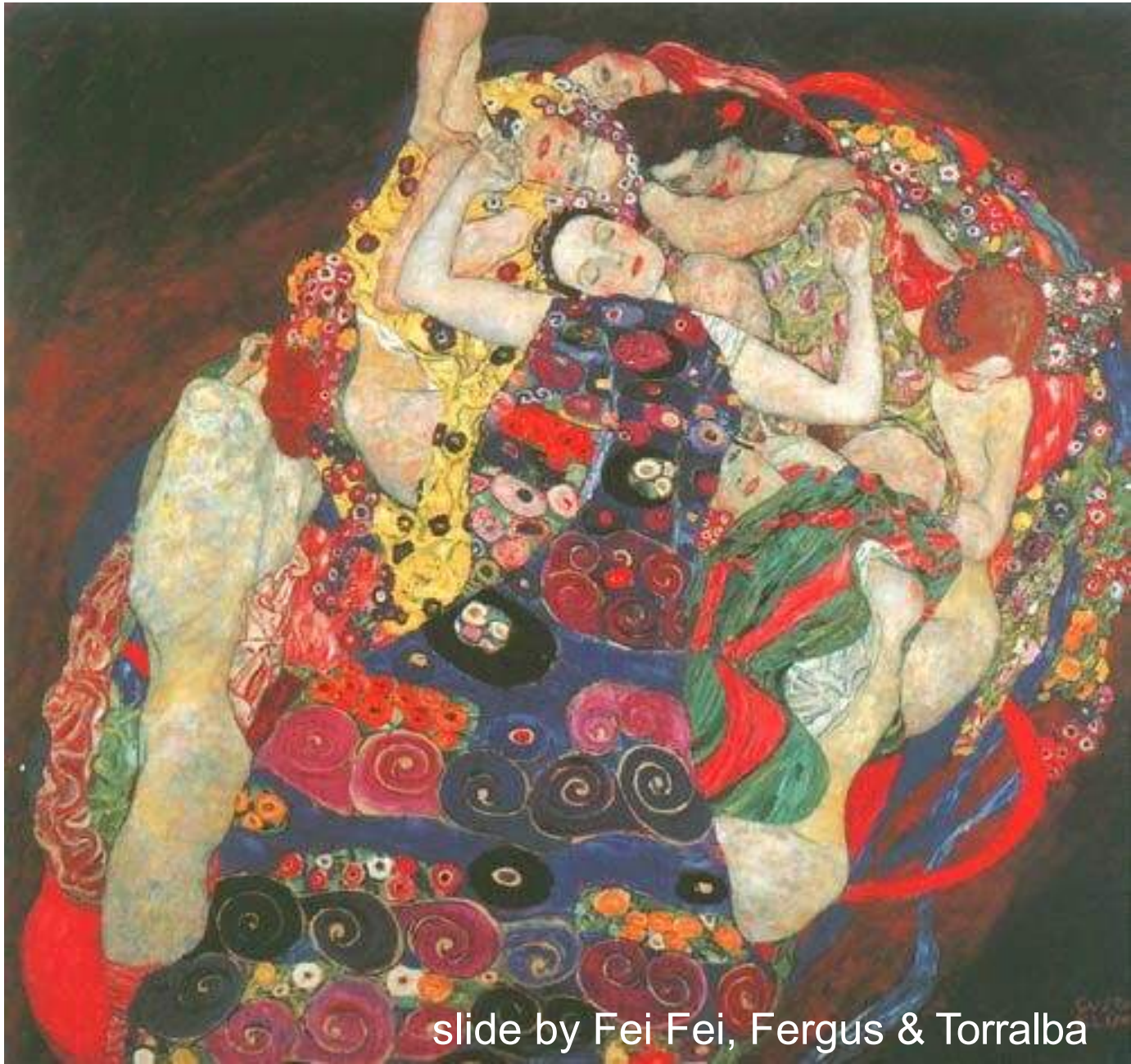
Challenges 5: deformation



slide by Fei Fei, Fergus & Torralba

Xu, Beihong 1943

Challenges 6: background clutter



Klimt, 1913

slide by Fei Fei, Fergus & Torralba

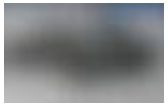
Challenges 7: object intra-class variation

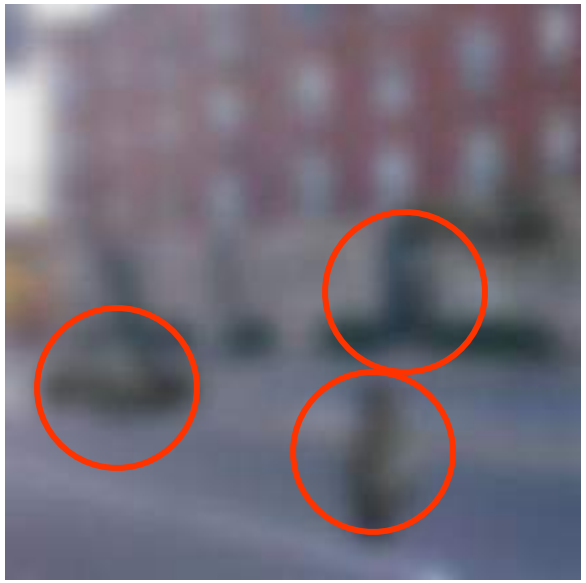





slide by Fei-Fei, Fergus & Torralba

Challenges 8: local ambiguity





slide by Fei-Fei, Fergus & Torralba

How do we solve it?

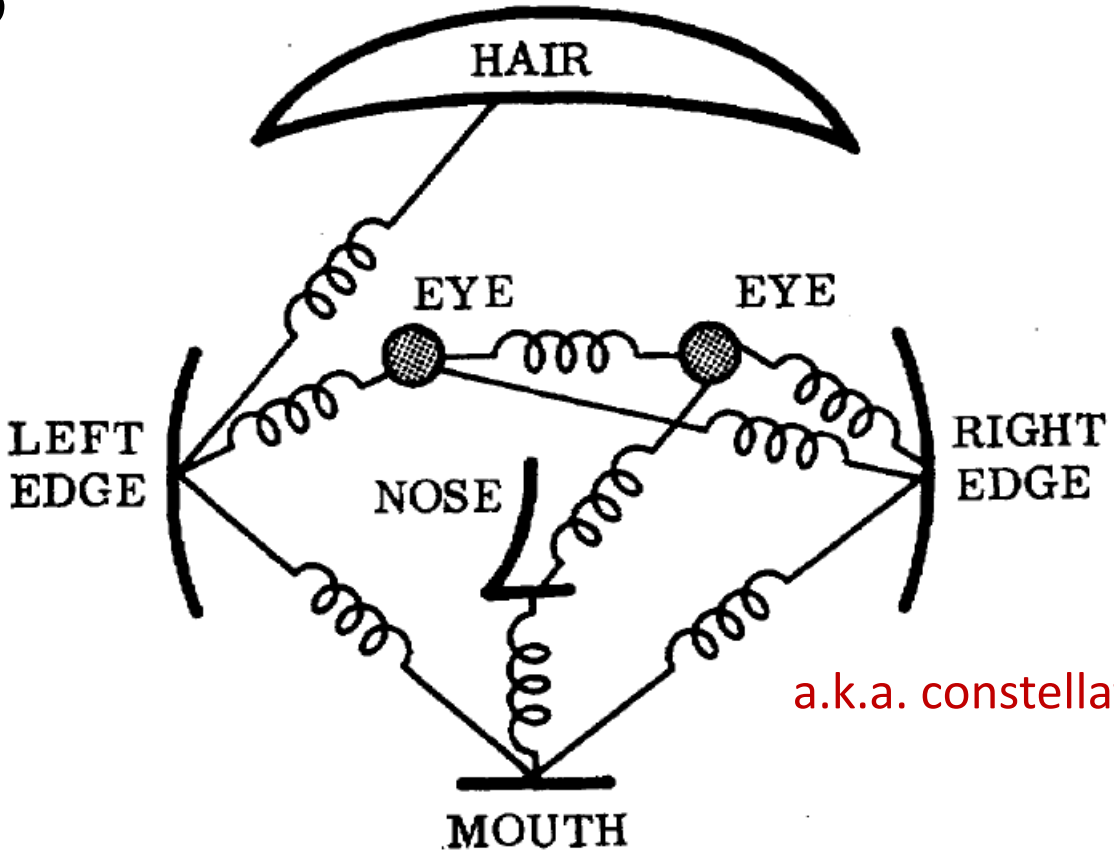
A brief history of visual recognition

1966

"Connect a television camera to a computer and get the machine to describe what it sees."

Marvin Minsky

1973

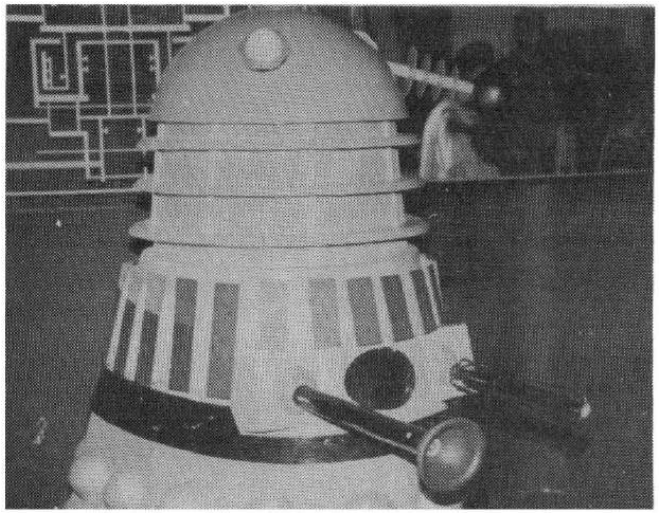


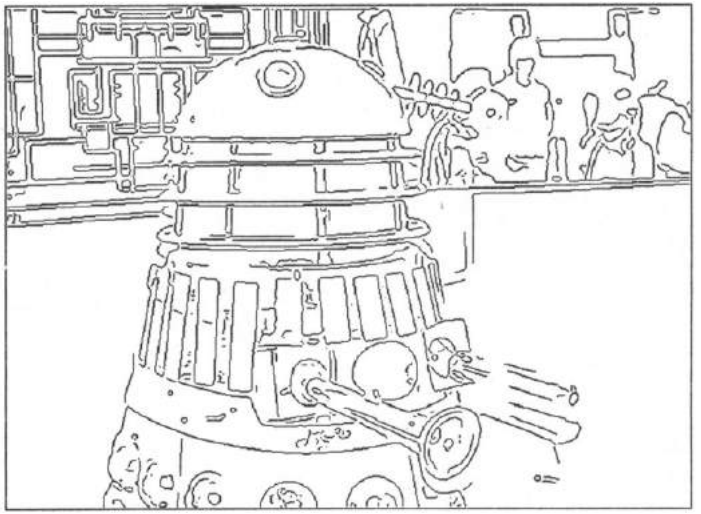
a.k.a. constellation model

The representation and matching of pictorial structures, Fischler and Elschlager, 1973

1980's

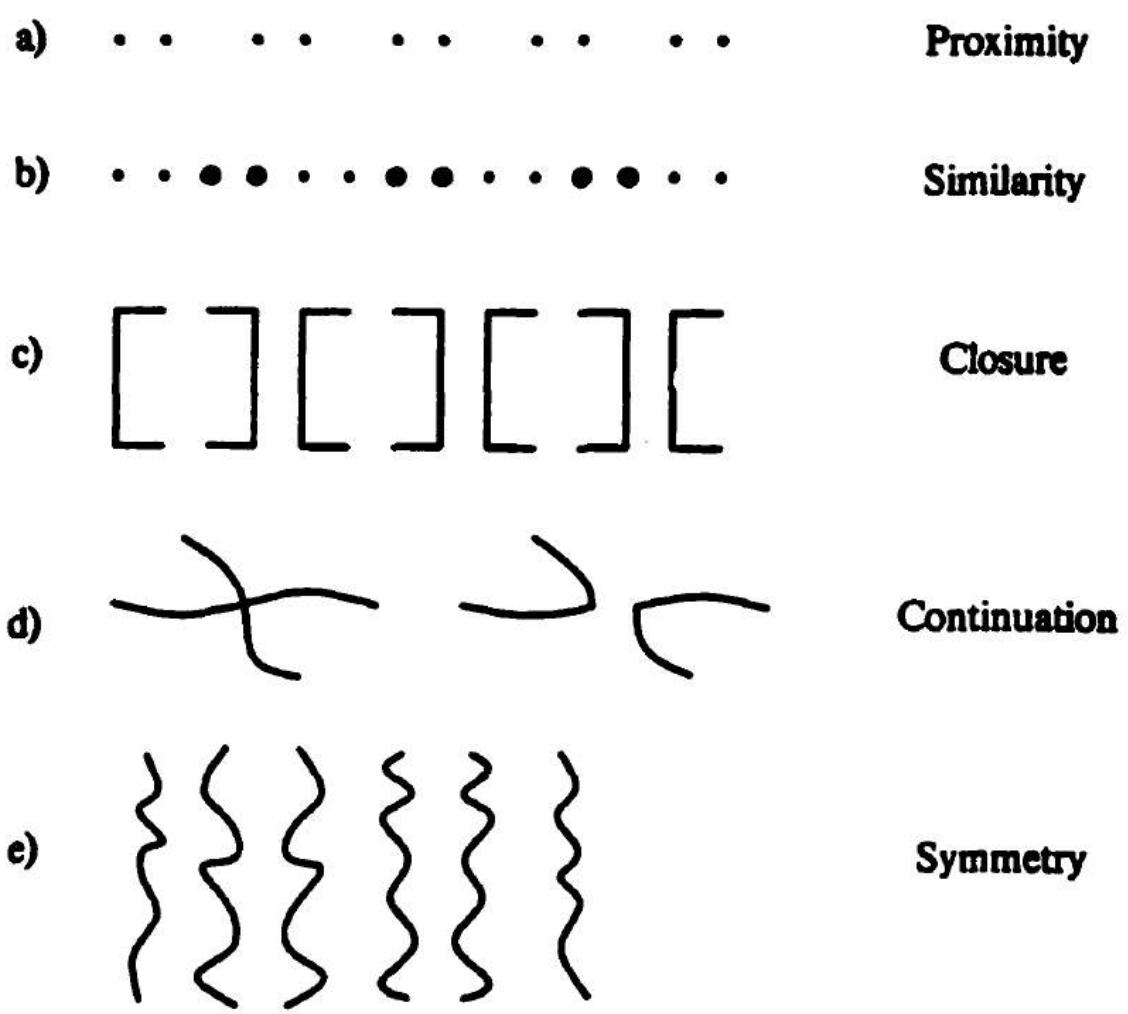
AI winter… …back to basics





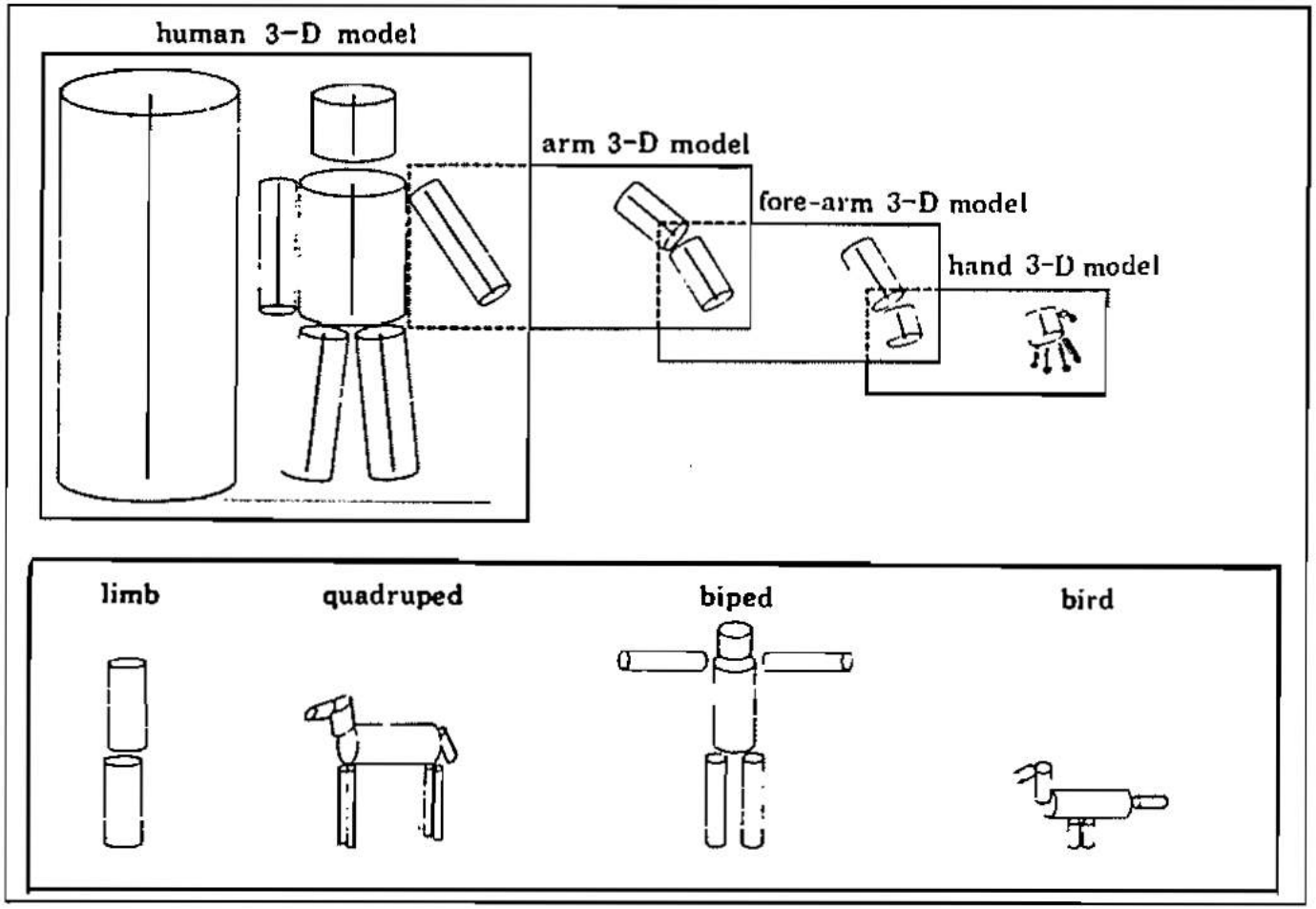
A Computational Approach to Edge Detection, Canny 1986

1984



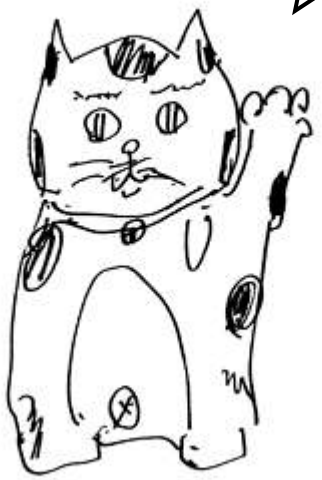
Perceptual Organization and Visual Recognition,
David Lowe, 1984

1986

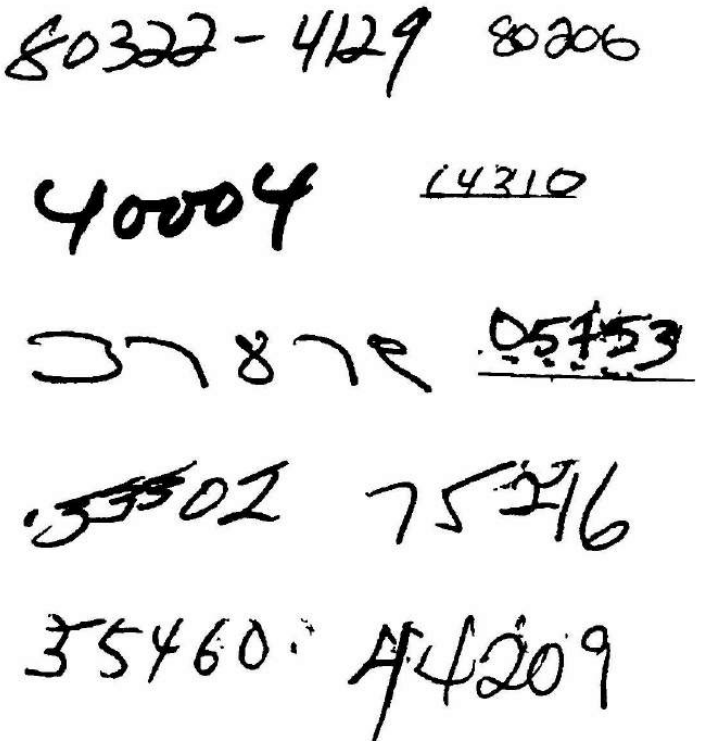


Perceptual organization and the representation of natural form, Alex Pentland, 1986

Hello,
Neural Networks!



1989



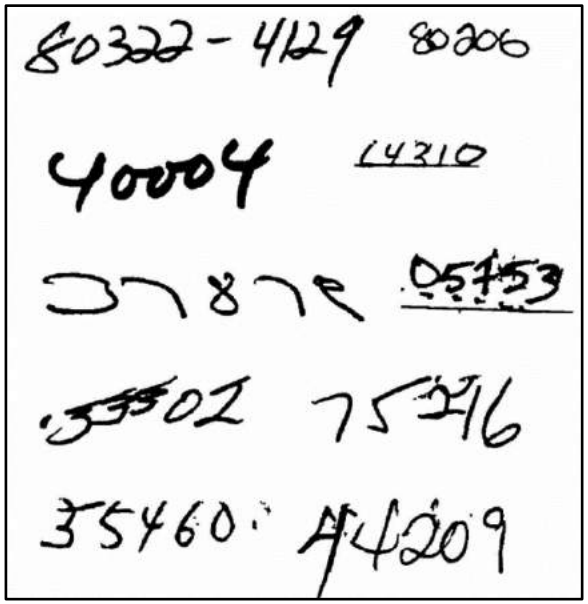
Zip codes

MNIST

Backpropagation applied to handwritten zip code recognition, Lecun et al., 1989

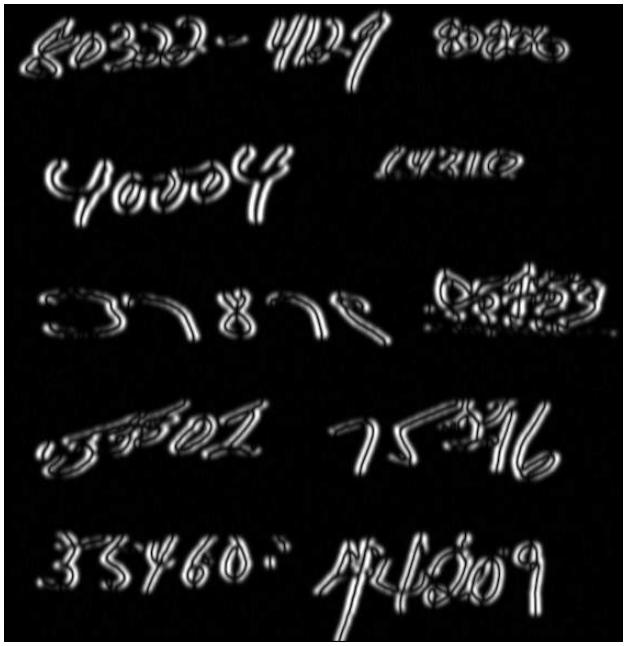
Filters (Convolutions)

Input



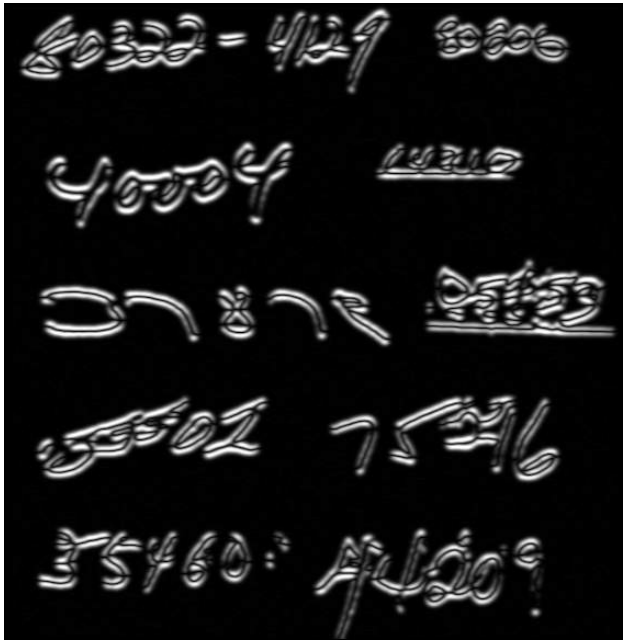
-1	0	+1
-2	0	+2
-1	0	+1

x filter

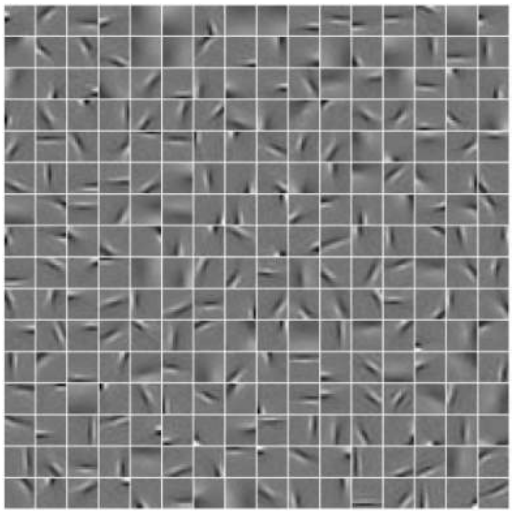


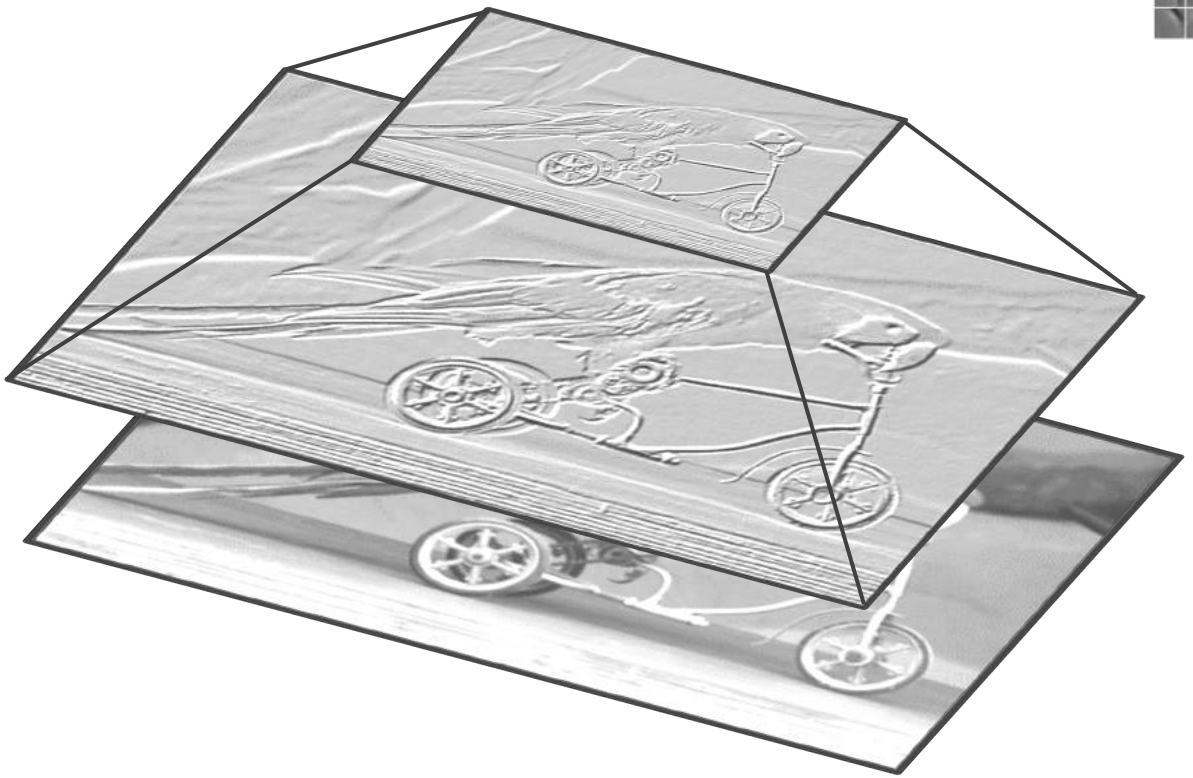
+1	+2	+1
0	0	0
-1	-2	-1

y filter



1989



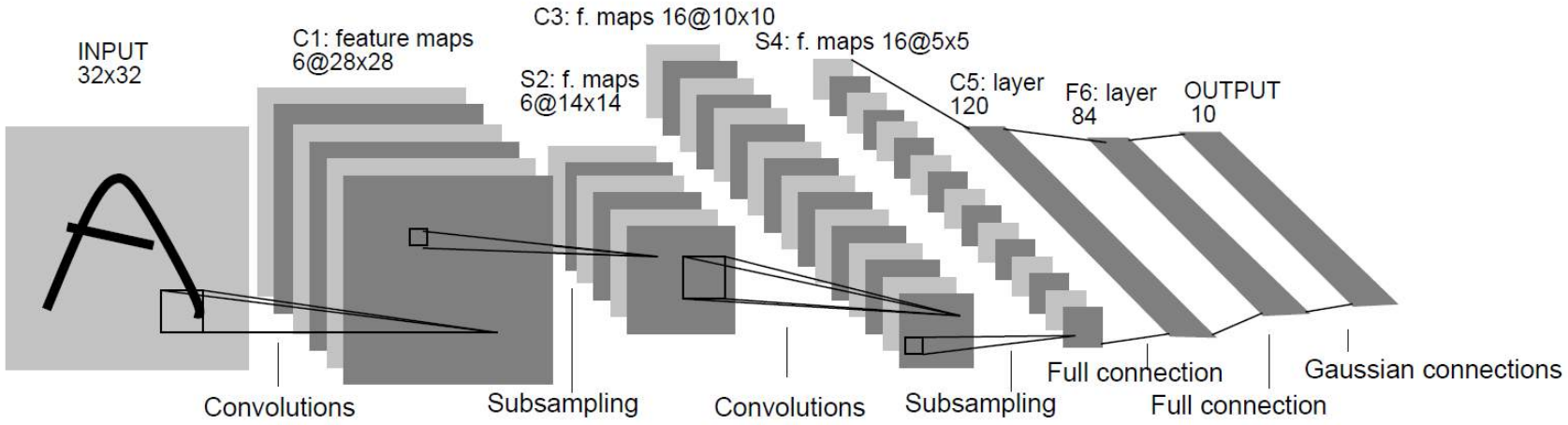


Pooling

Convolution
+ Nonlinear function

Image

1989

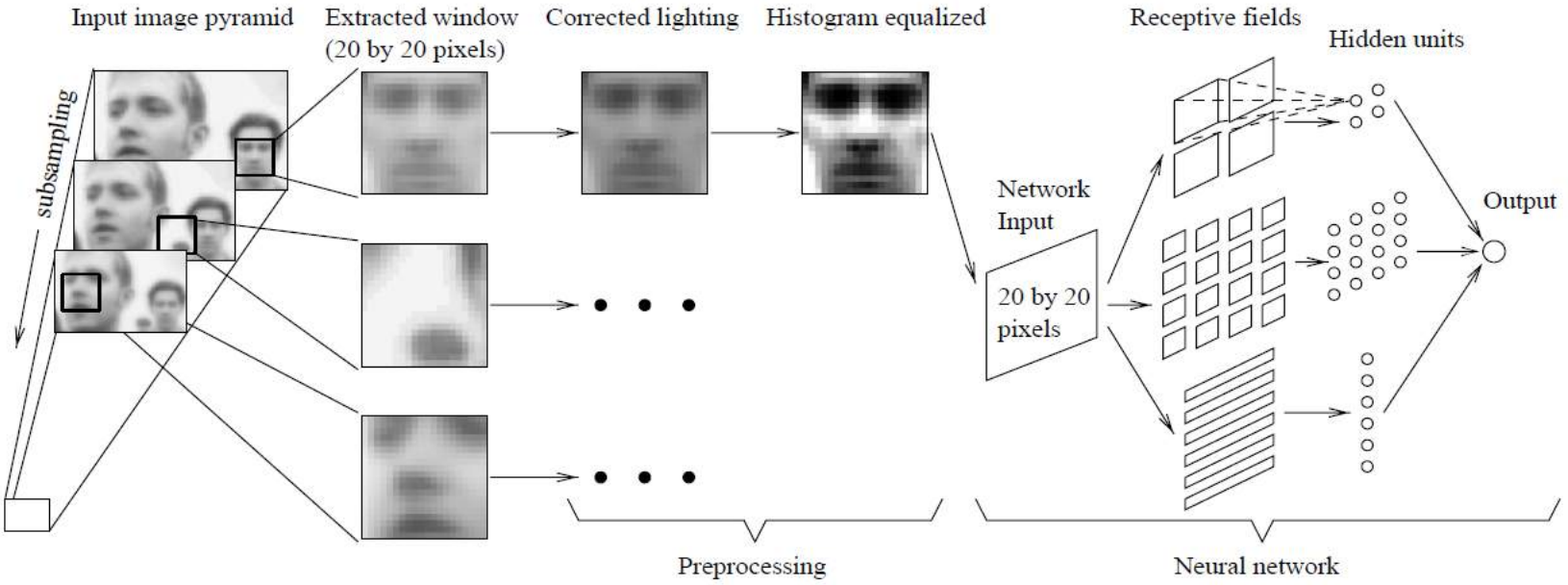


All filter weights are learned from data

Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

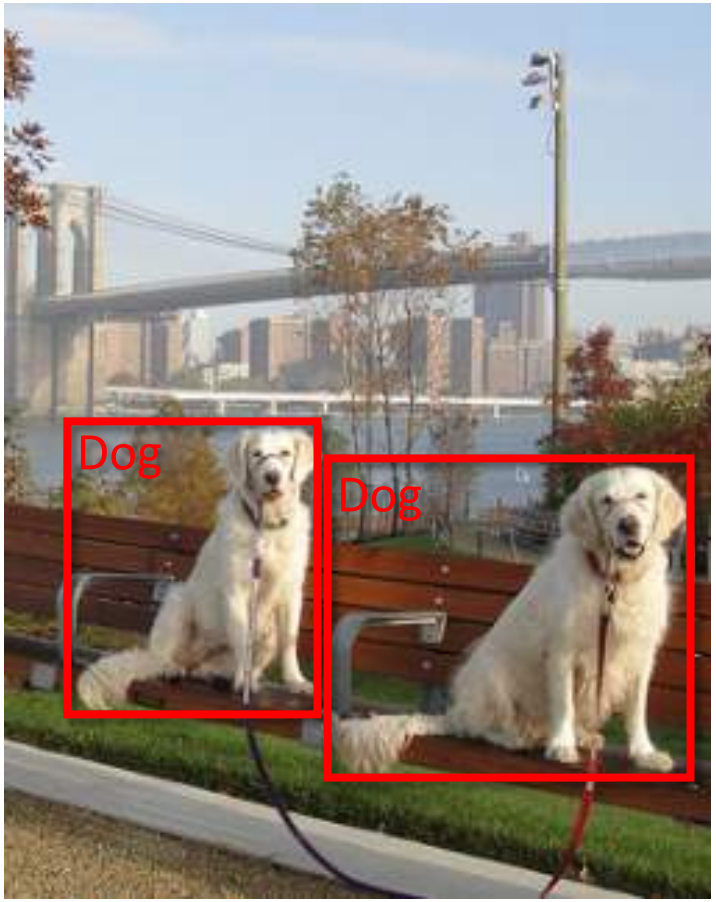
1998

Faces



Neural Network-Based Face Detection, Rowley at al., PAMI 1998

Classification vs. Detection



Limits

Numbers worked great, and so did faces…

Why did other categories fail?

2 main reasons… (2012)

Late 90s - early 2000s

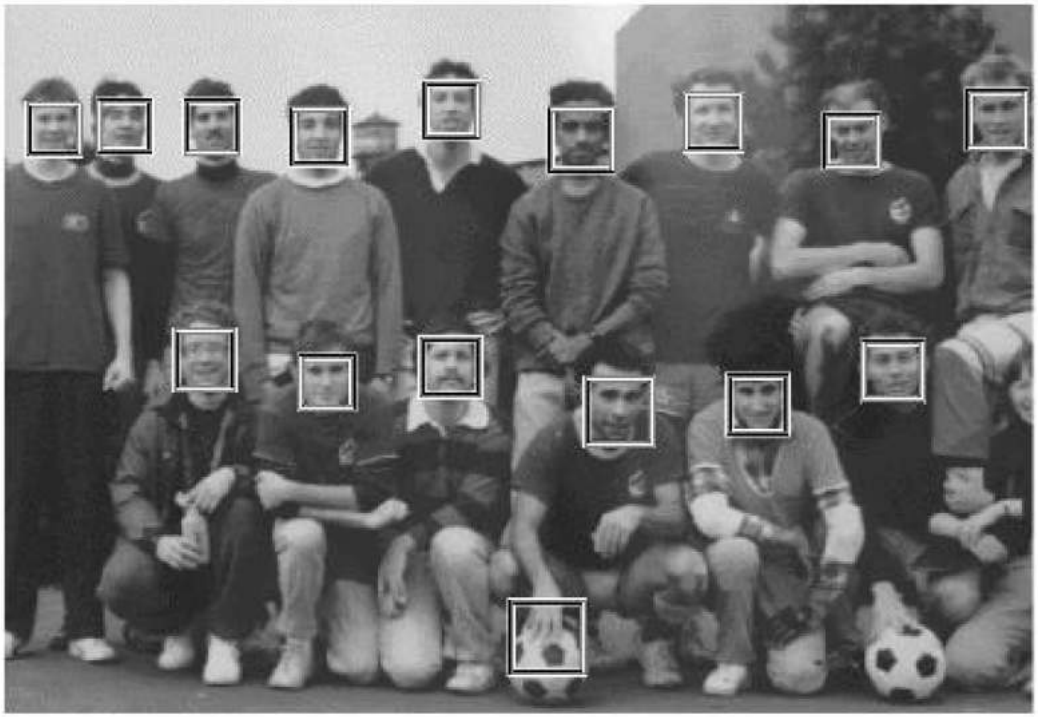
- Data to build observation models
- Data to build priors about the visual world
- Use the models and prior information to infer

Machine Learning based
methods in Computer Vision!

2001

Sliding window in real time!

Boosting + Cascade = Speed



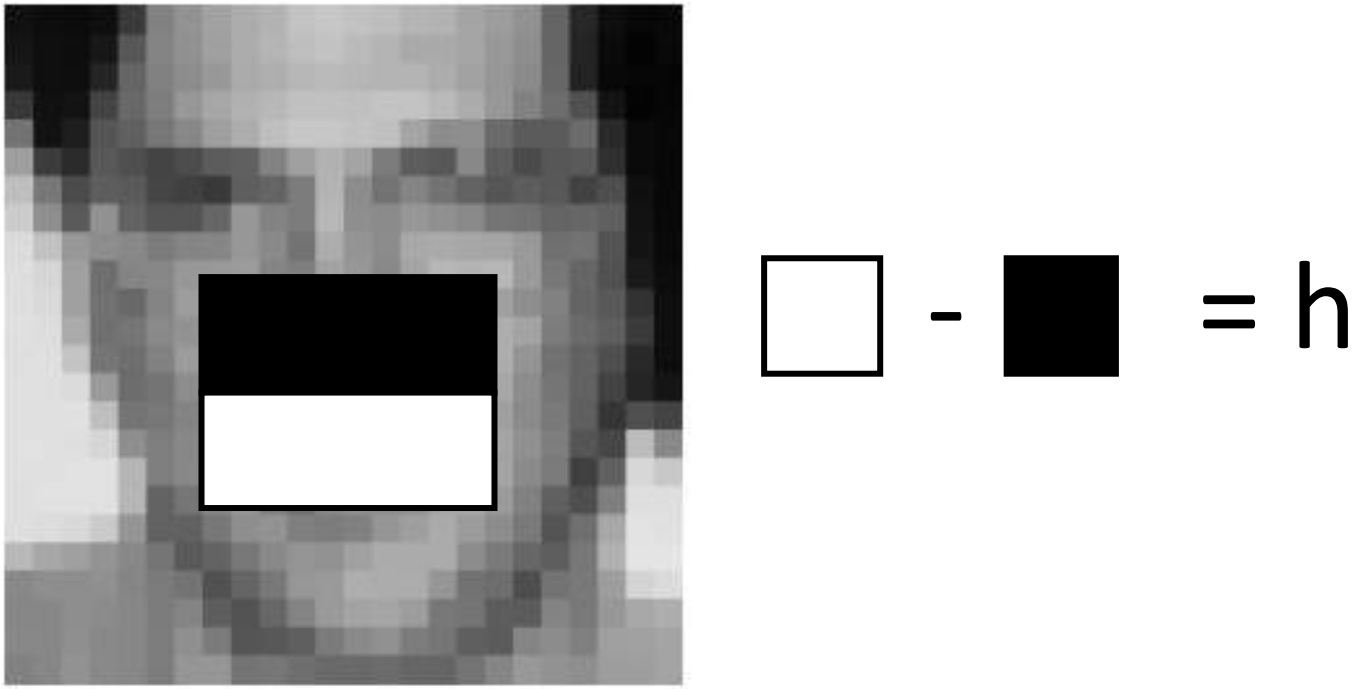
Rapid Object Detection using a Boosted Cascade of Simple Features, Viola and Jones, CVPR 2001

Sliding windows
& Classifier each window

Why did it work?

Why did it work?

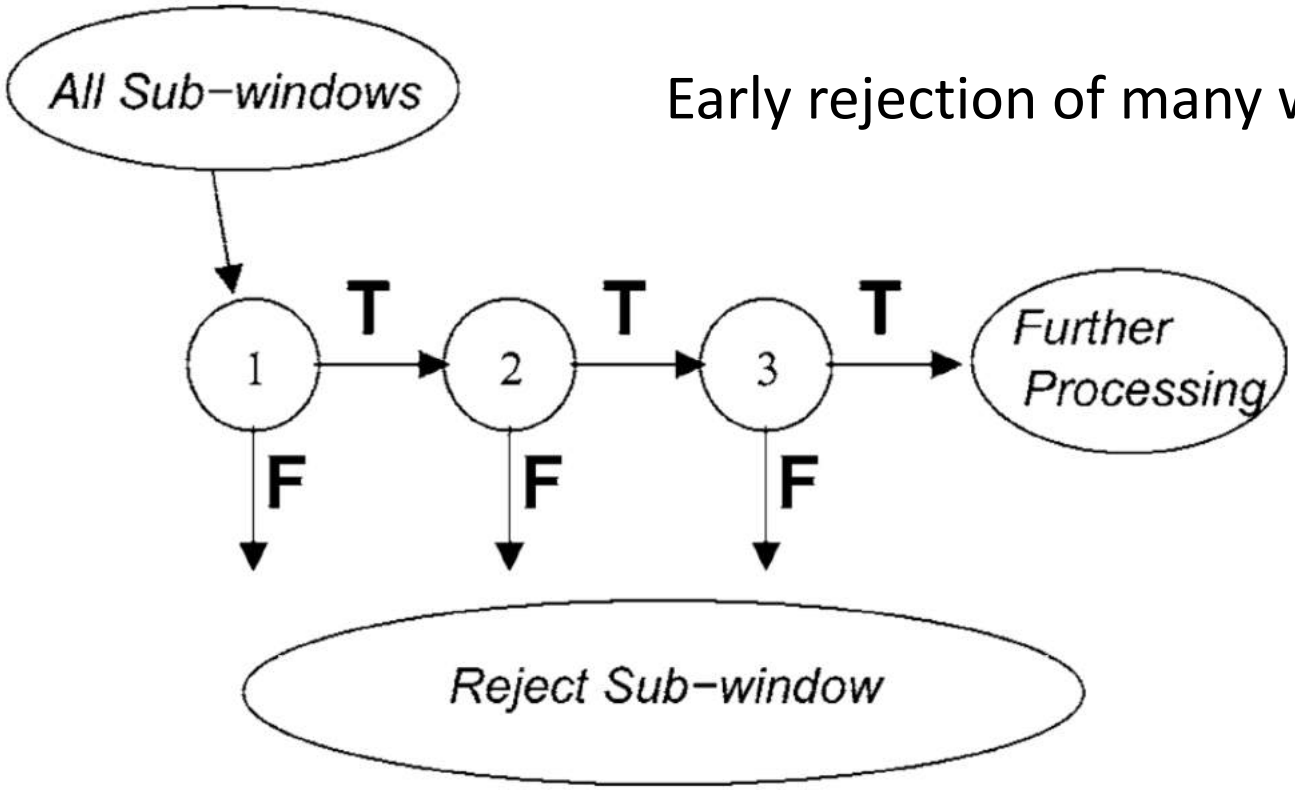
- Simple features (Haar wavelets)



Integral images + Haar wavelets = fast

Why did it work?

- Cascaded Boosting



Early rejection of many windows

Why did it fail?

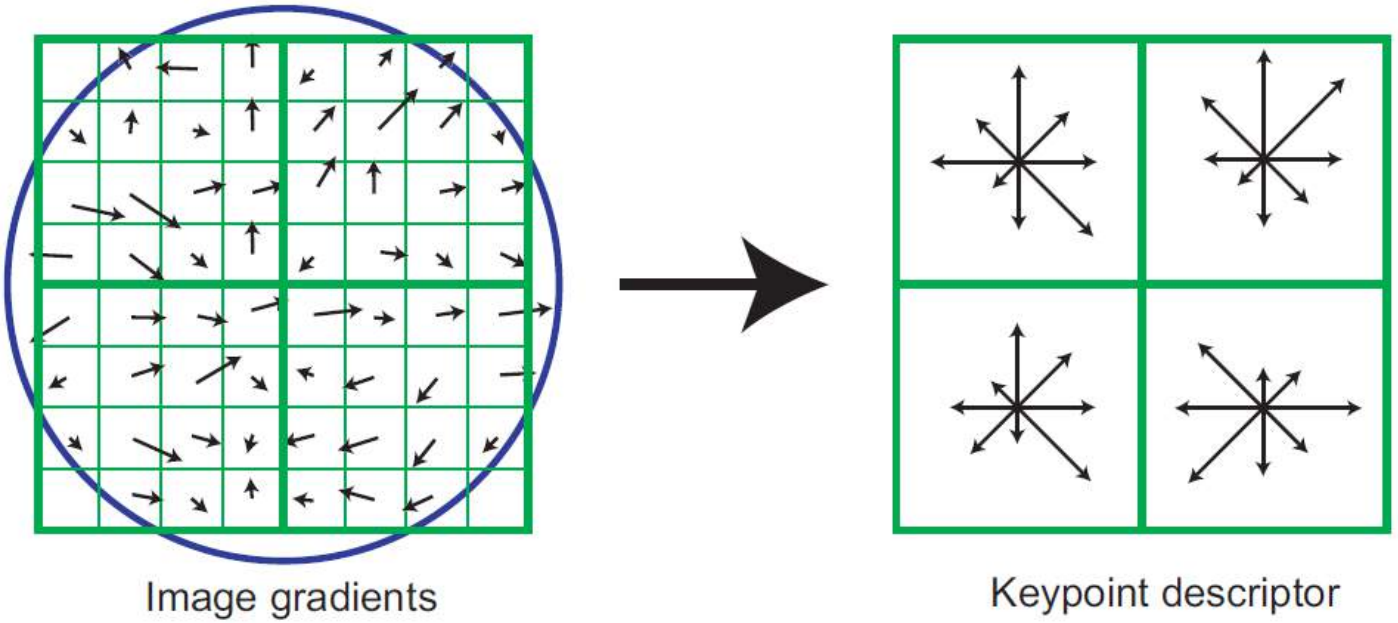
1999 SIFT (Scale Invariant Feature Transform)

No more sliding windows (interest points)

Better features (use more computation)

Object Recognition from Local Scale-Invariant Features,
Lowe, ICCV 1999.

Better descriptor:



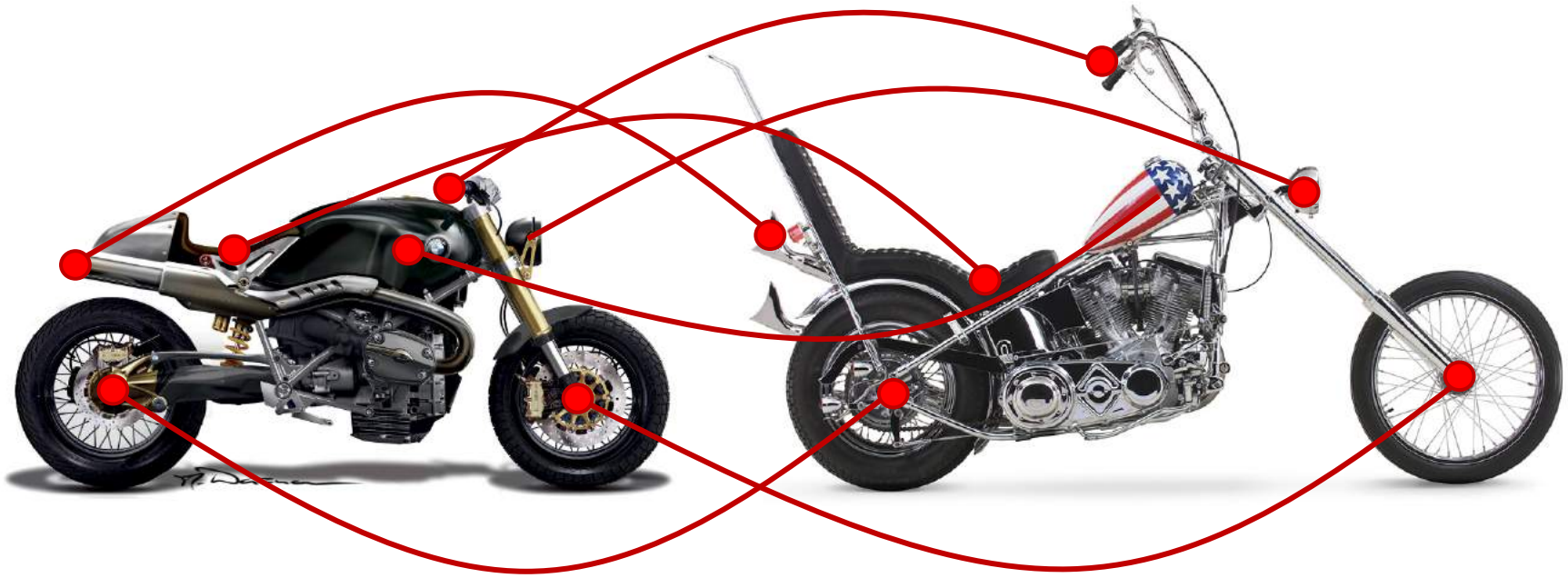
Distinctive image features from scale-invariant keypoints, Lowe, *IJCV* 2004

What worked

“Object instance” recognition (flat, textured objects)

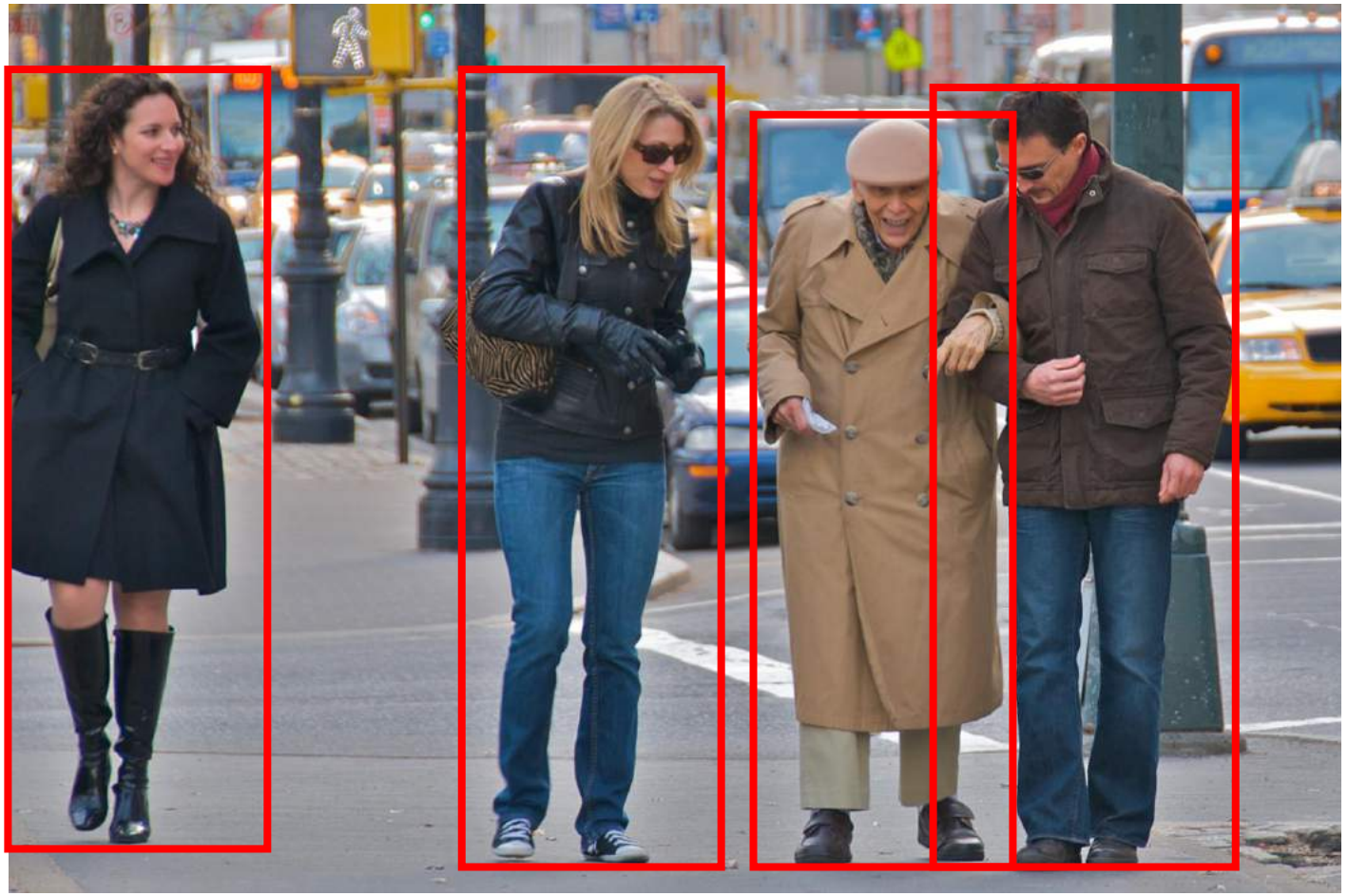
Why it failed …

Interest points don't work for category recognition



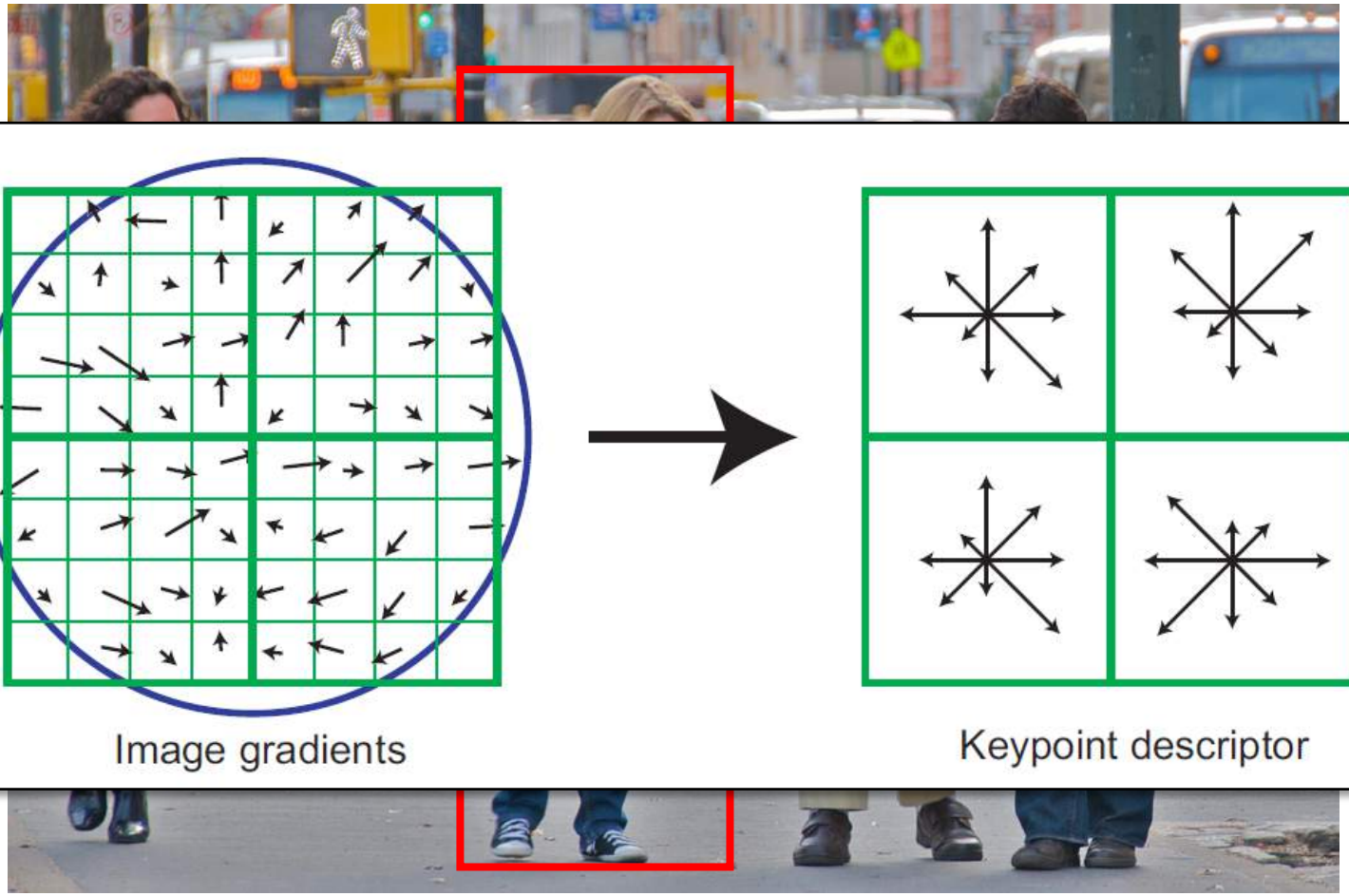
Object Class Recognition by Unsupervised Scale-Invariant Learning, Fergus et al., *CVPR* 2003.

2005 HOG (histograms of oriented gradients)

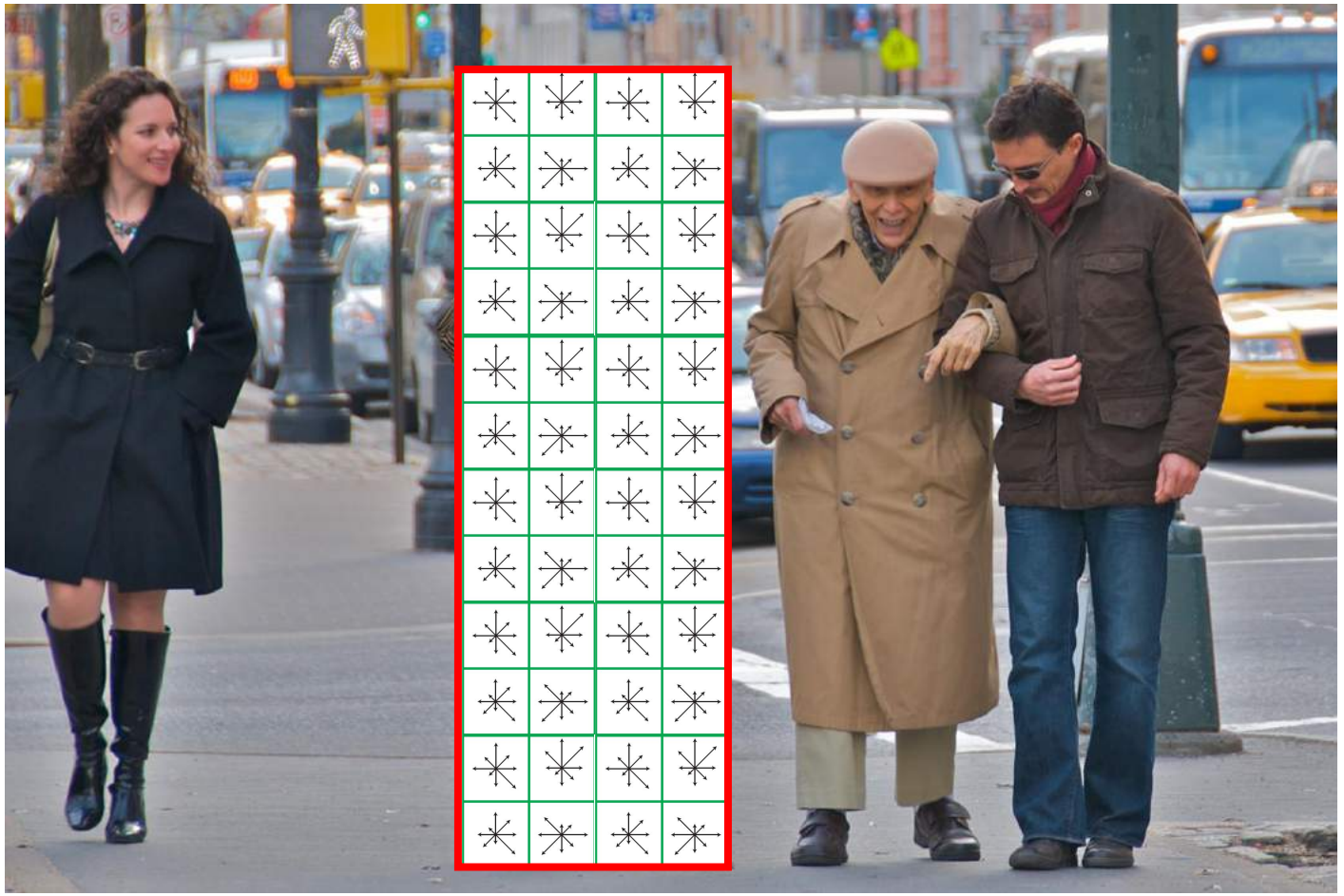


Histograms of oriented gradients for human detection, Dalal and Triggs, CVPR 2005.

2005 HOG (histograms of oriented gradients)



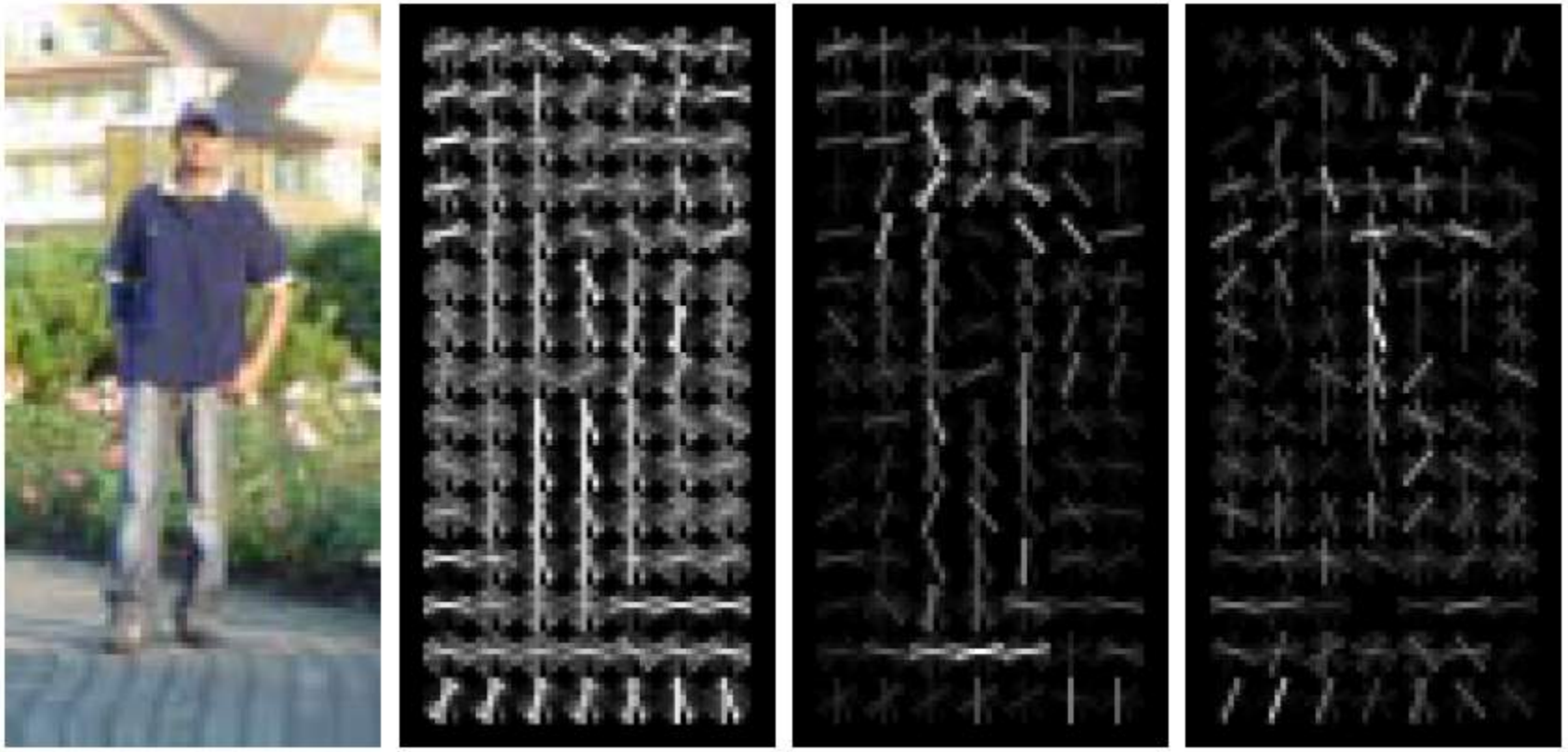
2005 HOG (histograms of oriented gradients)



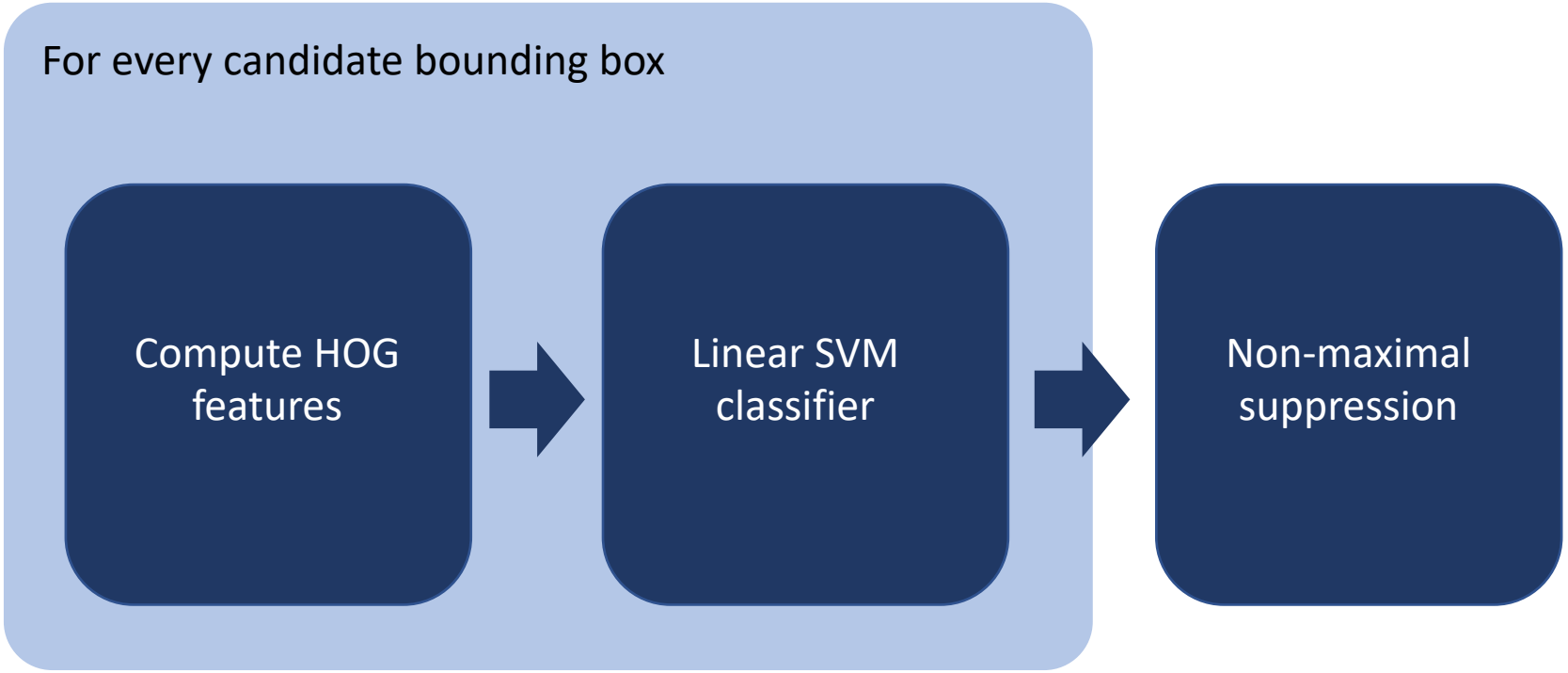
Histograms of oriented gradients for human detection, Dalal and Triggs, *CVPR* 2005.

2005 HOG (histograms of oriented gradients)

Presence > Magnitude



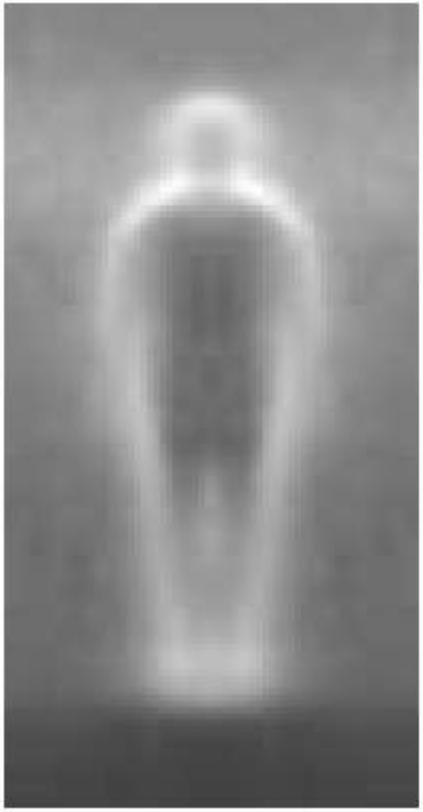
2005 HOG (histograms of oriented gradients)



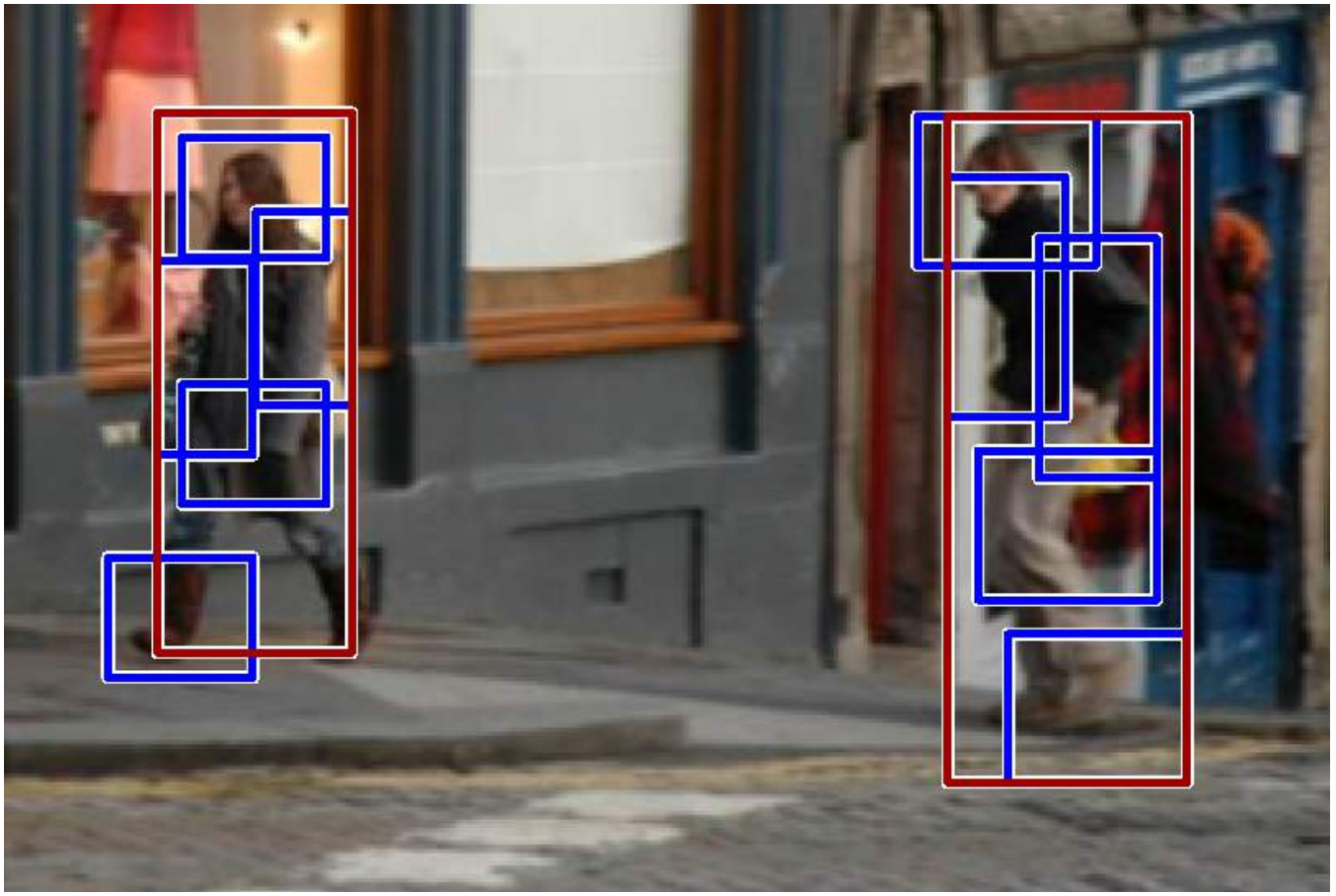
Why it worked

- Even noisy boundaries give us good cues about the object
- Hard negative mining
- Computers are fast enough

Why it failed

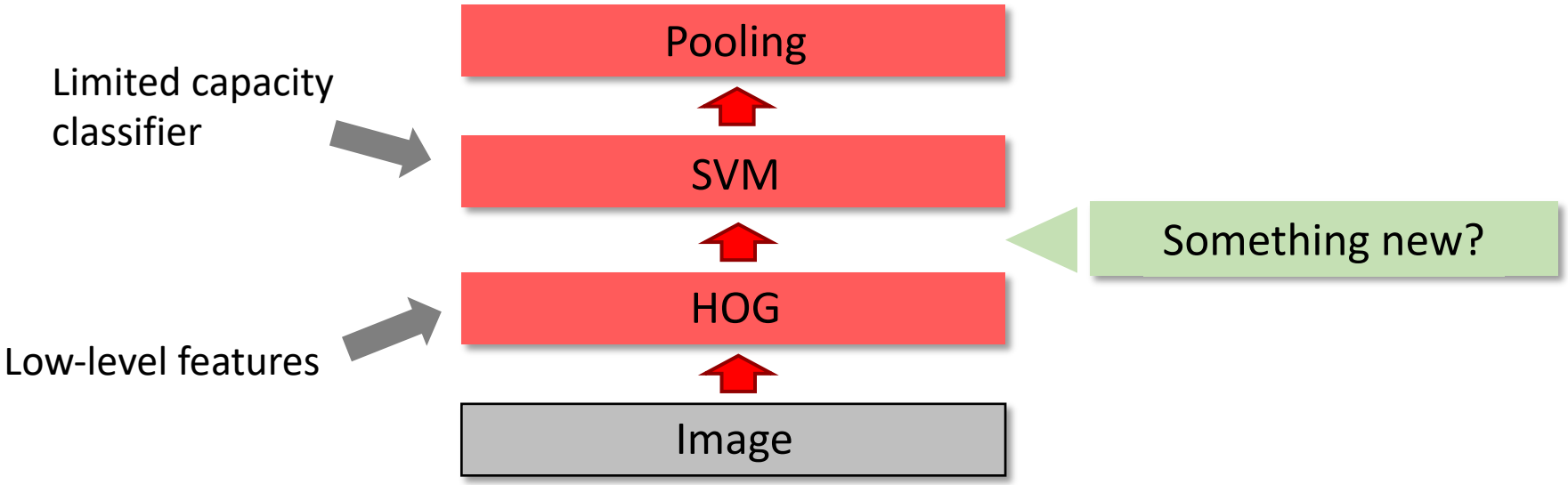


2008 DPM (Deformable parts model)



Object Detection with Discriminatively Trained Part Based Model,
Felzenszwalb, Girshick, McAllester and Ramanan, *PAMI*, 2010

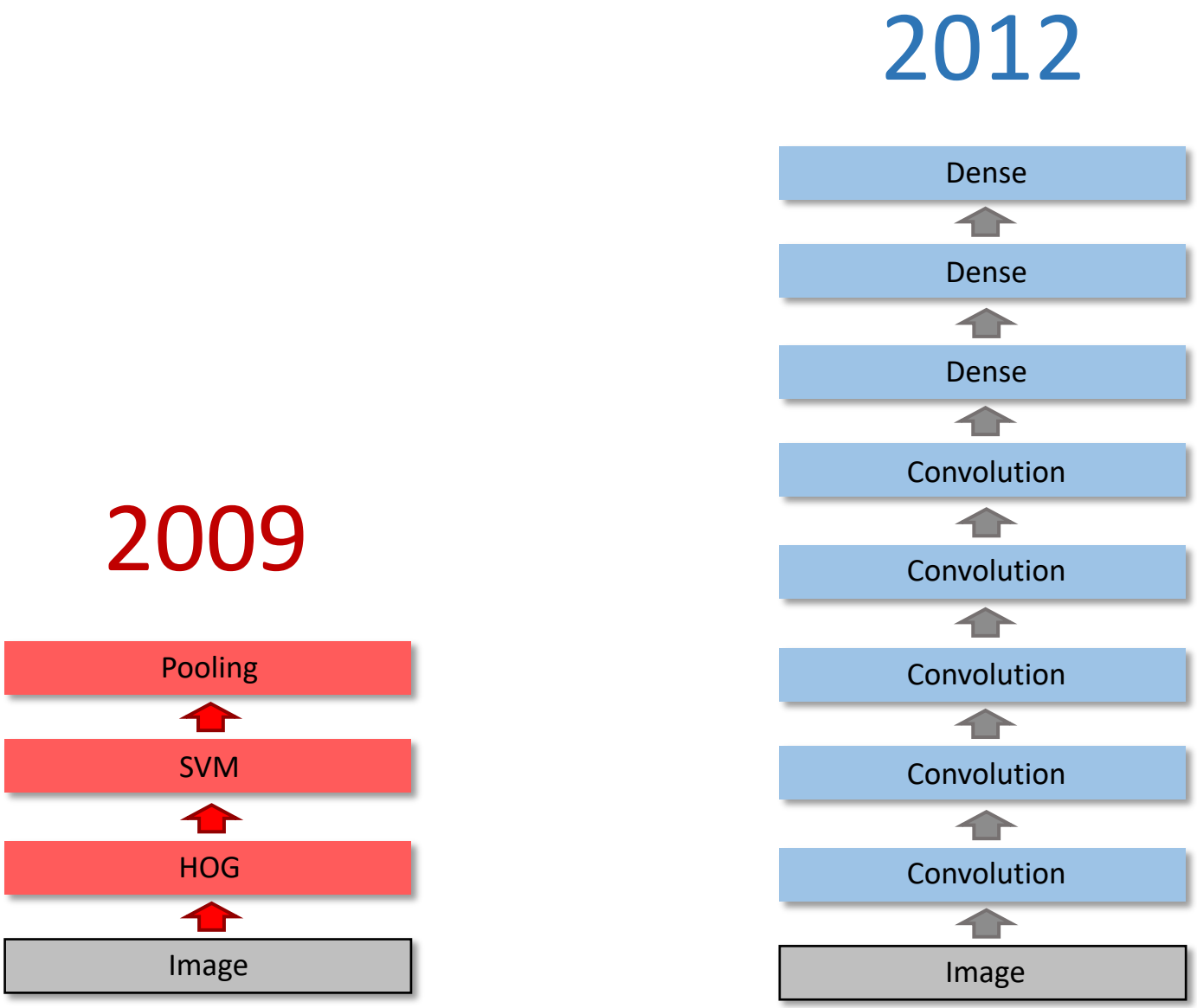
HOG + DPM


Pooling
Limited capacity classifier
SVM
Something new?
HOG
Low-level features
Image

Where are we now?

A quick intro to deep learning

Algorithms


2012
Dense
Dense
Dense
Convolution
Convolution
Convolution
Convolution
Convolution
Image
2009
Pooling
SVM
HOG
Image

2004-2006 Caltech 101 and 256







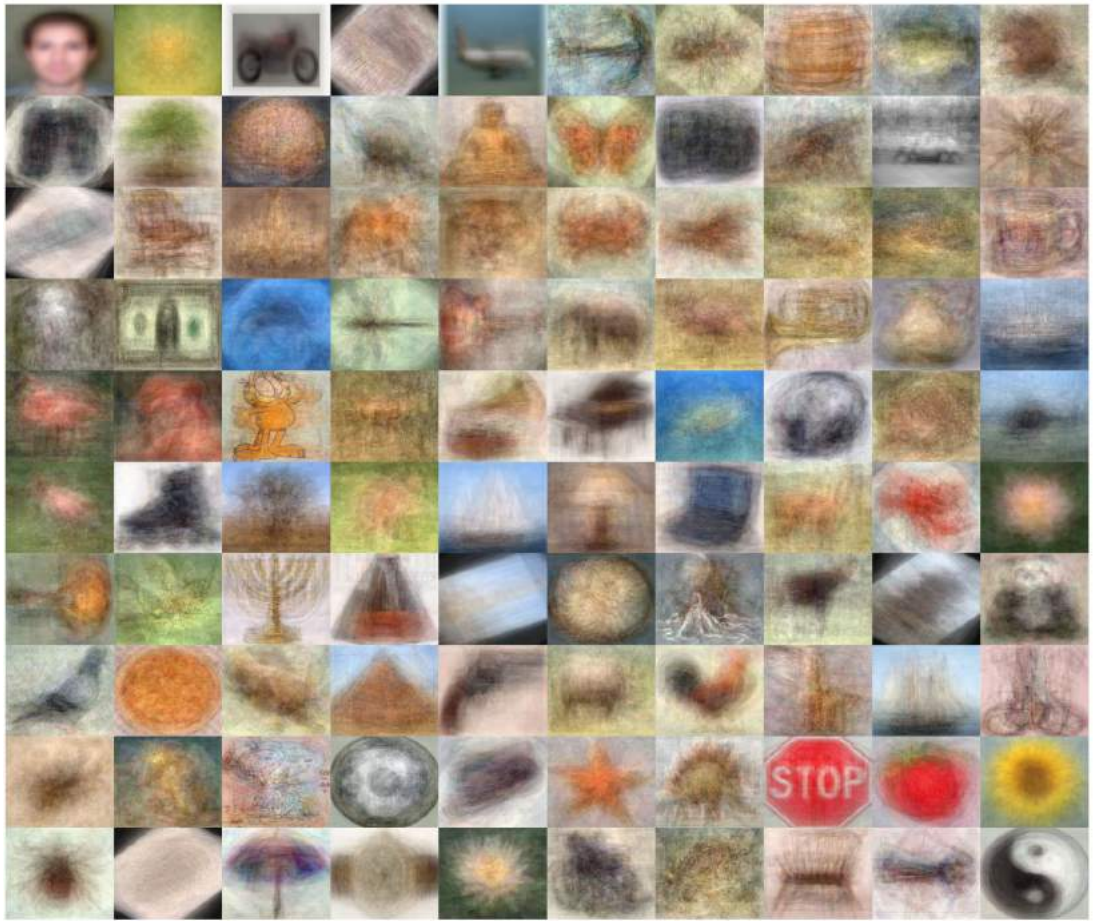







2006

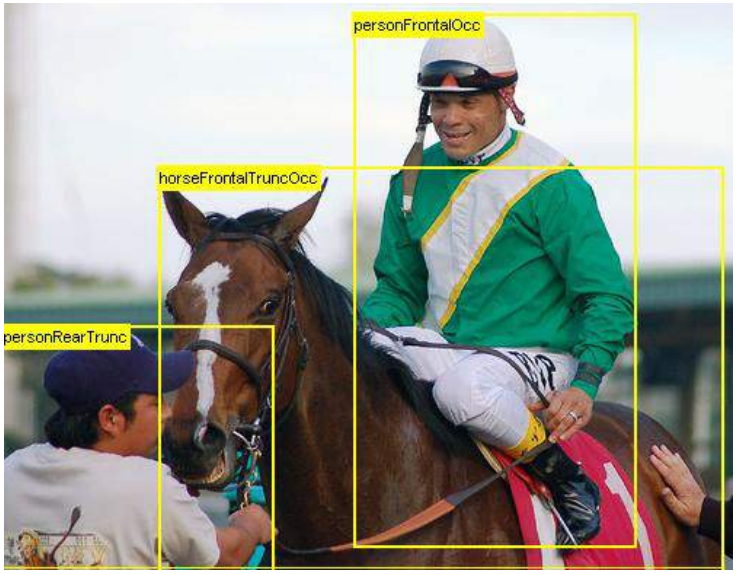
15-30 training images, up to ~70% accuracy.

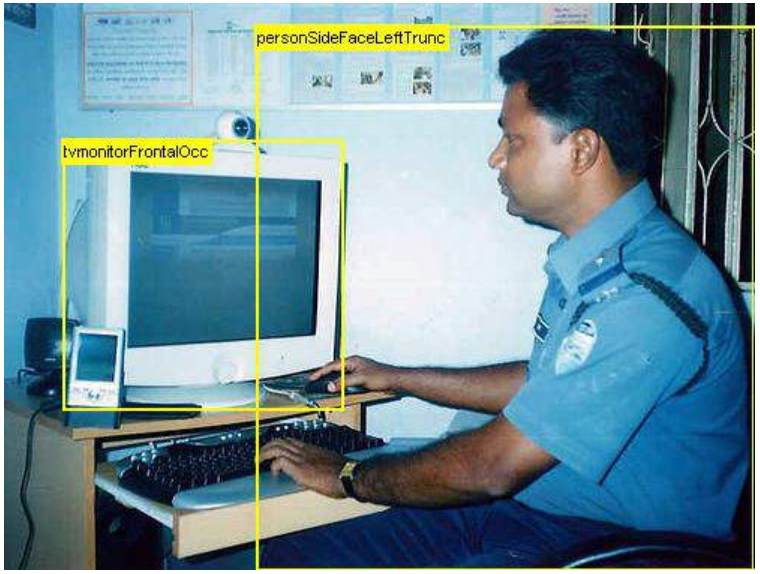


Antonio Torralba

2007 PASCAL VOC

20 classes

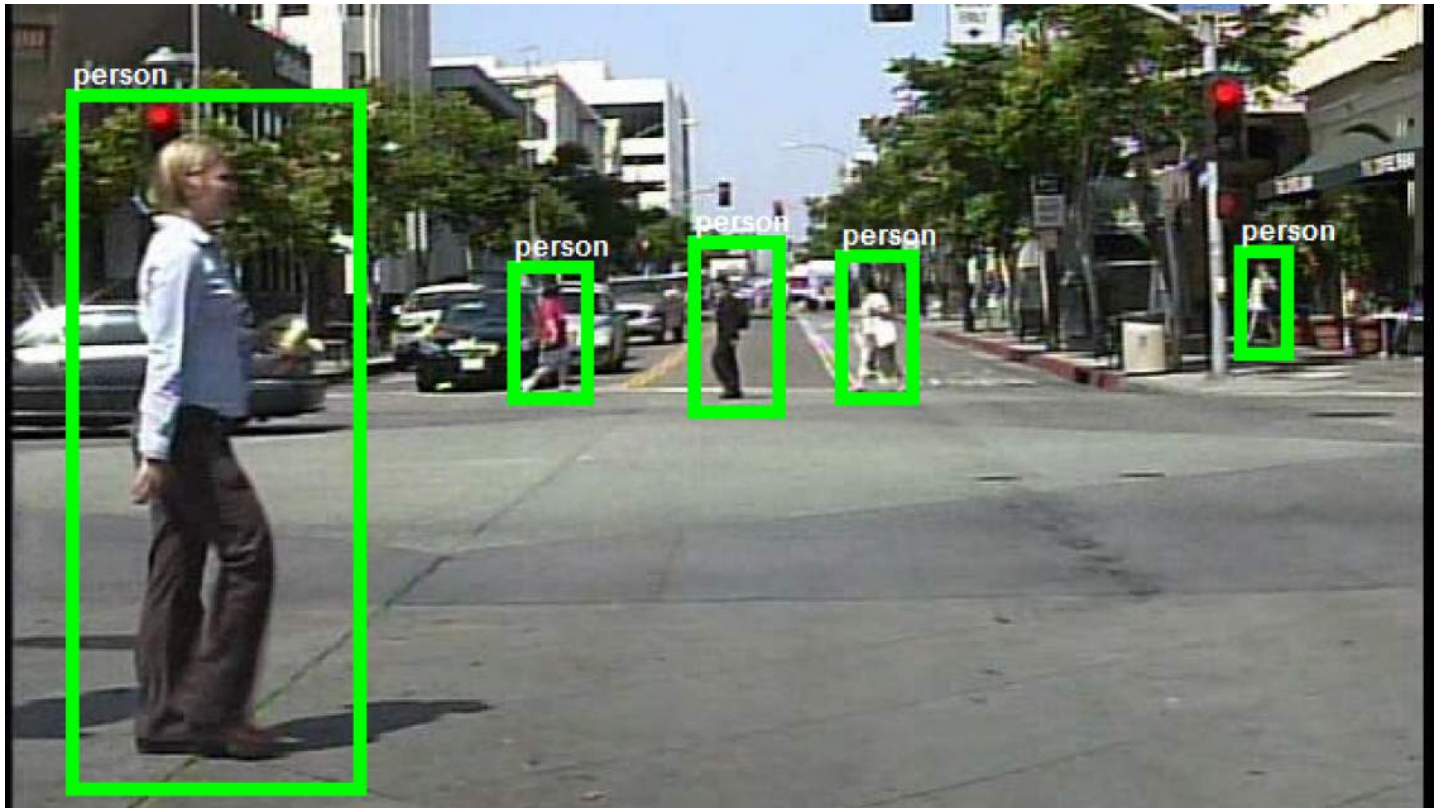




The PASCAL Visual Object Classes (VOC) Challenge, Everingham, Van Gool, Williams, Winn and Zisserman, *IJCV*, 2010

2009 Caltech Pedestrian

1 class, lots of instances.



Pedestrian Detection: An Evaluation of the State of the Art, Dollár, Wojek, Schiele and Perona, *PAMI*, 2012

2009 ImageNet

22K categories, 14M images

ImageNet: A Large-Scale Hierarchical Image Database,
Deng, Dong, Socher, Li, Li and Fei-Fei, *CVPR,* 2009

2010 SUN

908 scene categories

Beer garden

SUN Database: Large-scale Scene Recognition from Abbey to Zoo
Xiao, Hays, Ehinger, Oliva, and Torralba, *CVPR*, 2010.

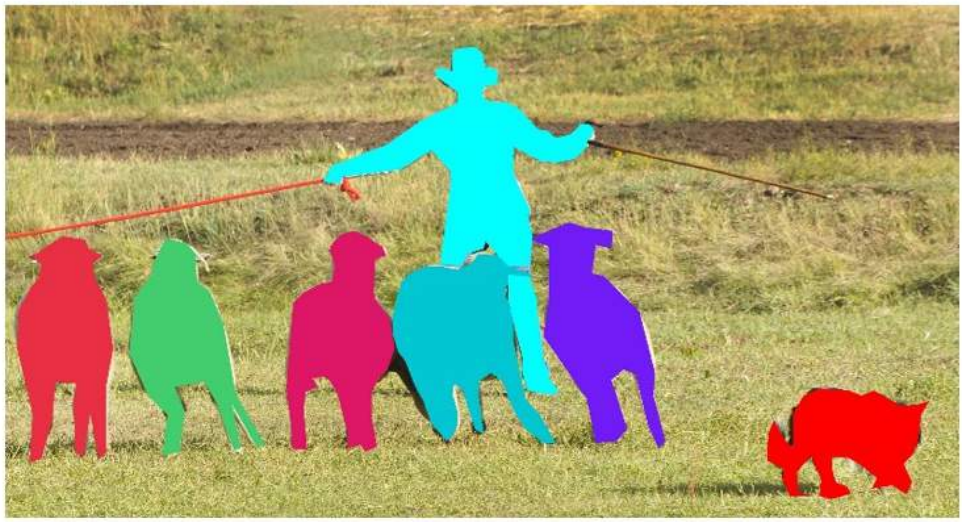
2014 COCO

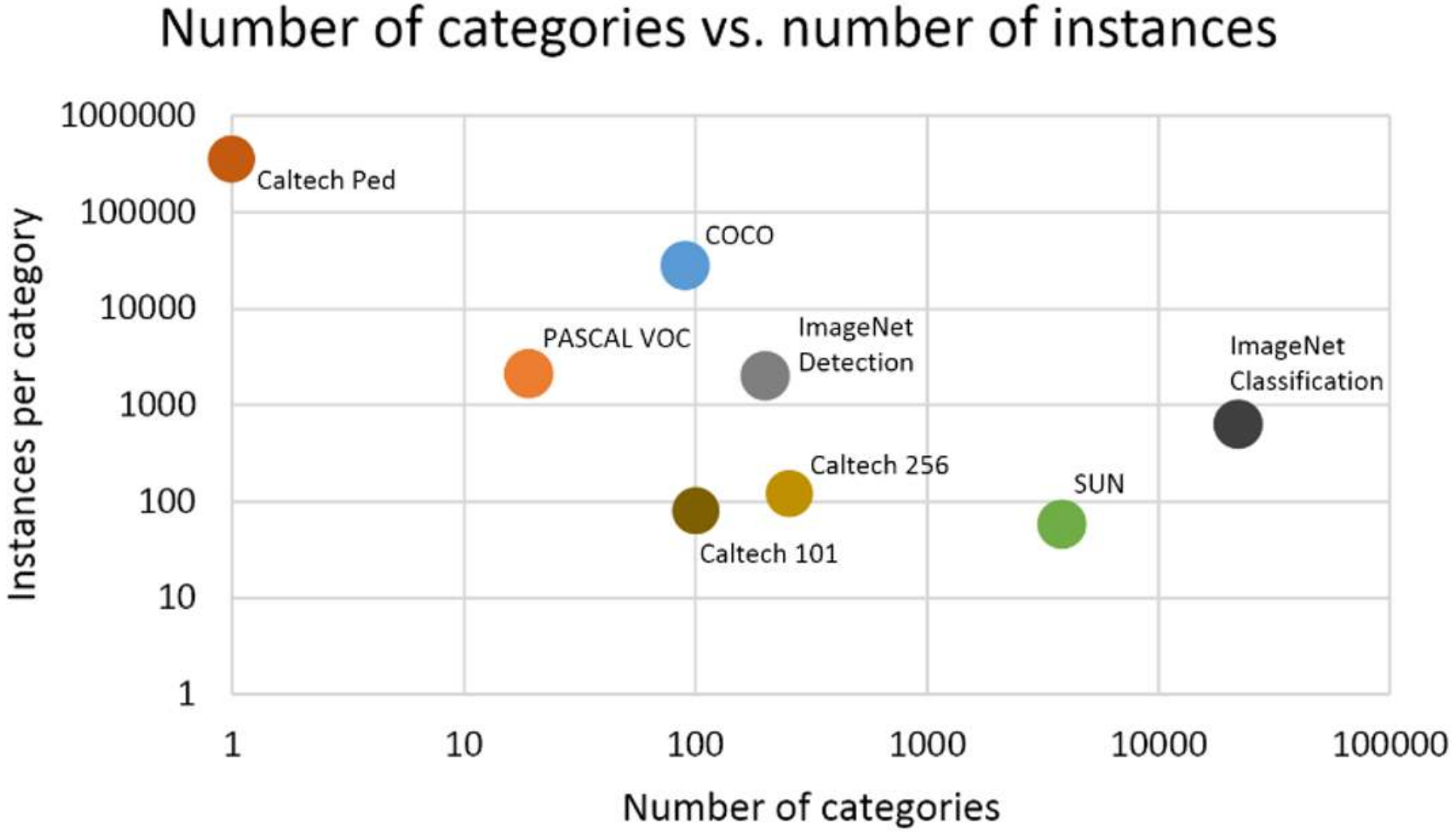
80 object categories with instance masks

Detection in Context

Instance segmentation

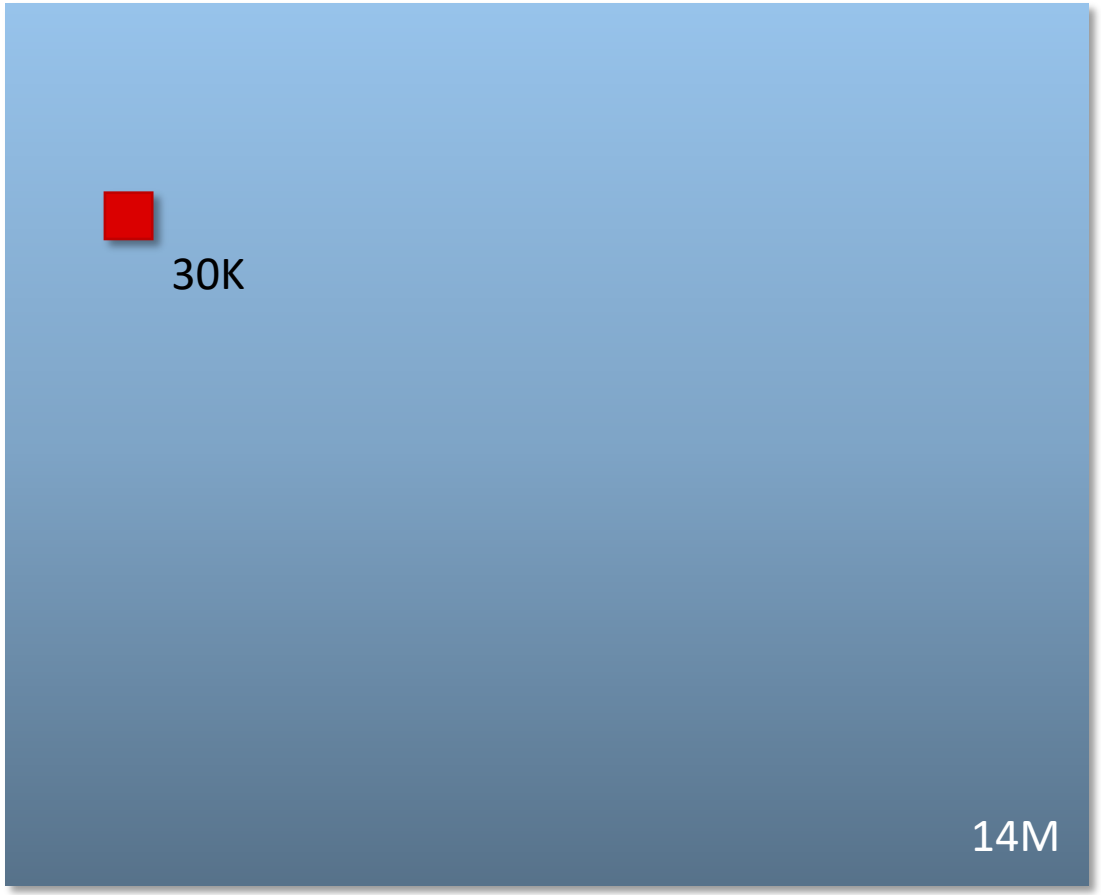
Non-iconic instances




Number of categories vs. number of instances
Instances per category
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
Number of categories
Caltech Ped
COCO
PASCAL VOC
ImageNet Detection
ImageNet Classification
Caltech 256
SUN
Caltech 101

Images

2009 2012

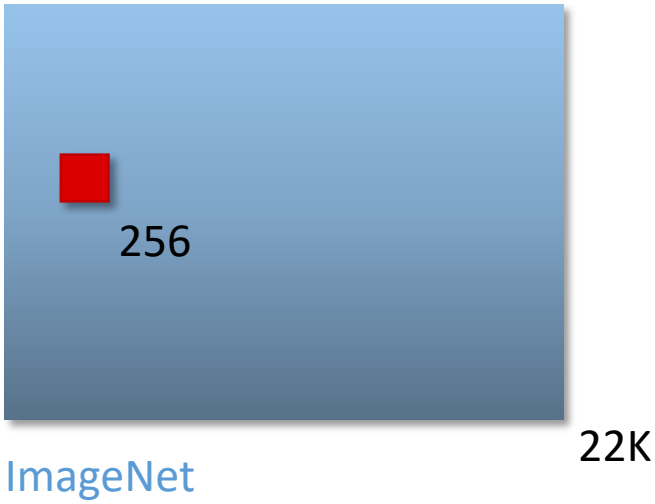


ImageNet

Categories

2009

2012



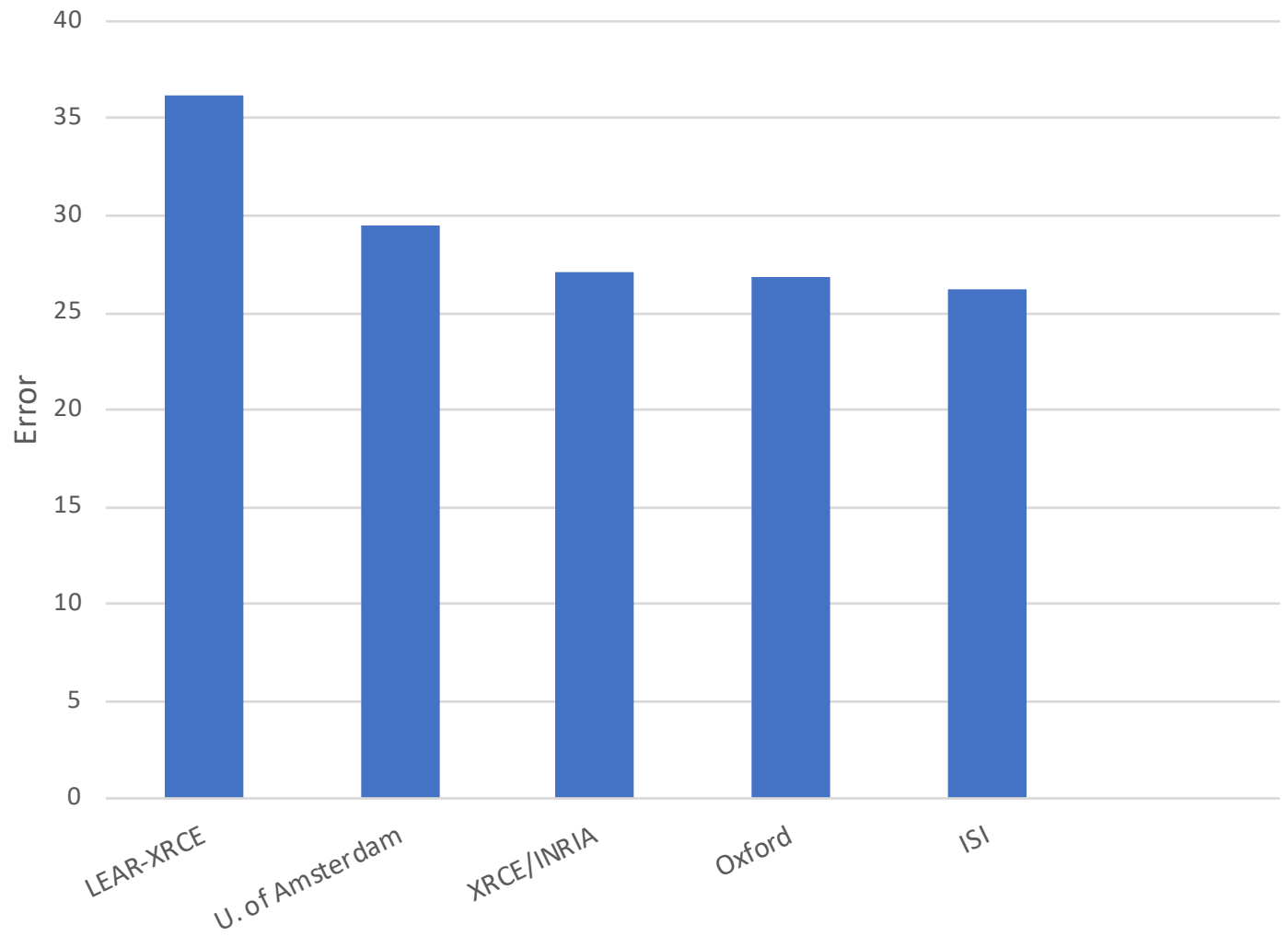
Data

GPUs

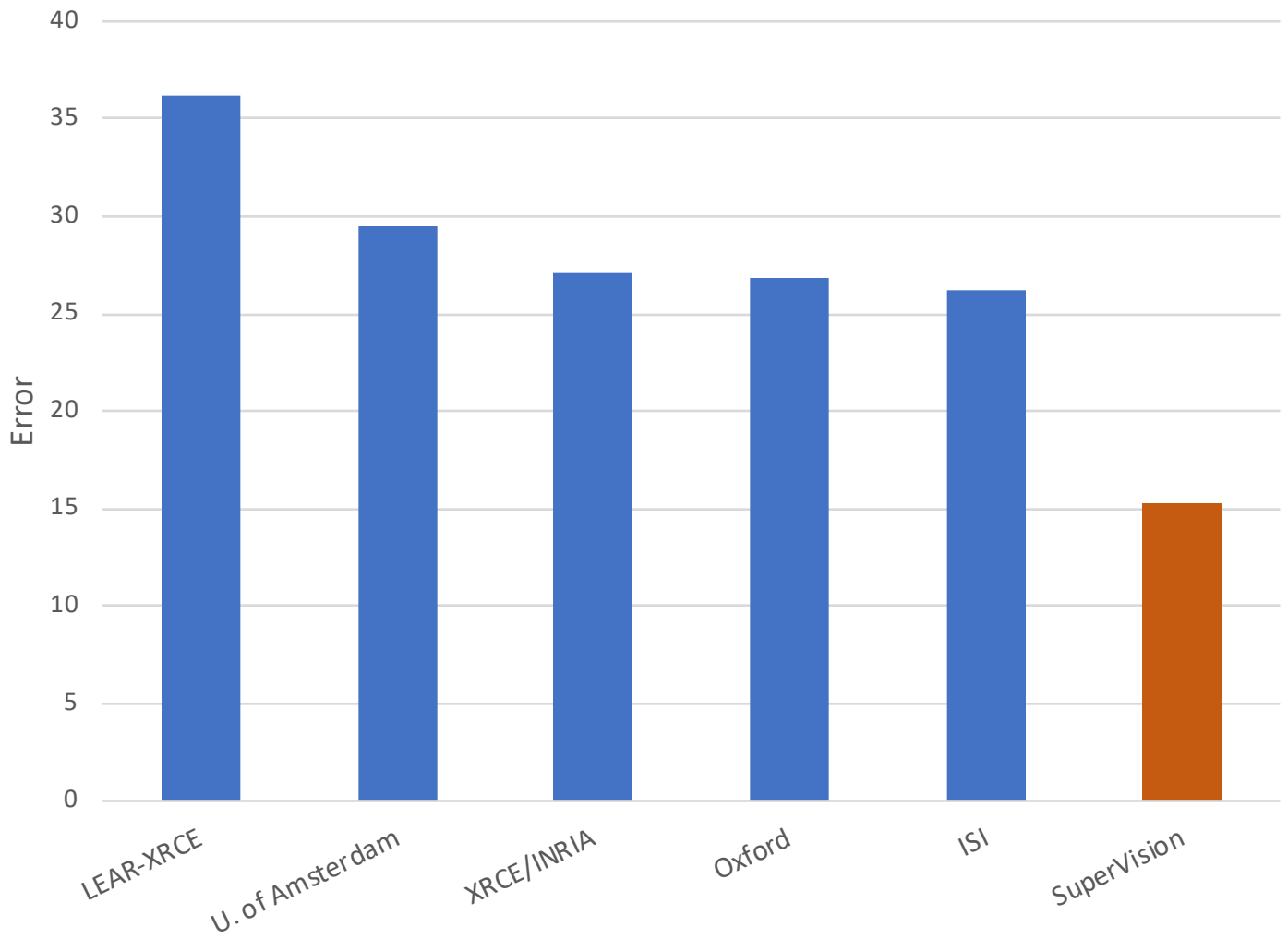
\+ Neural Network

?

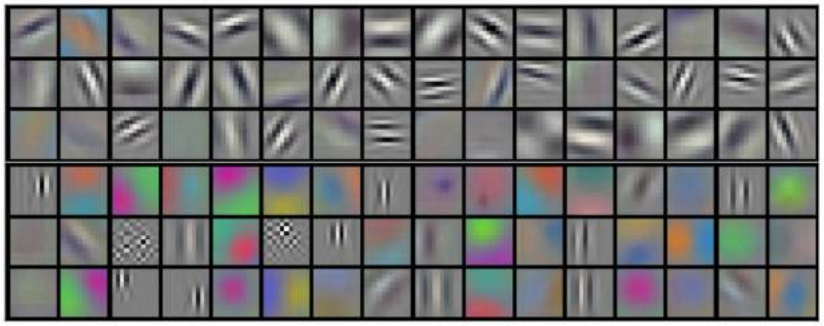
2012 ImageNet 1K (Fall 2012)

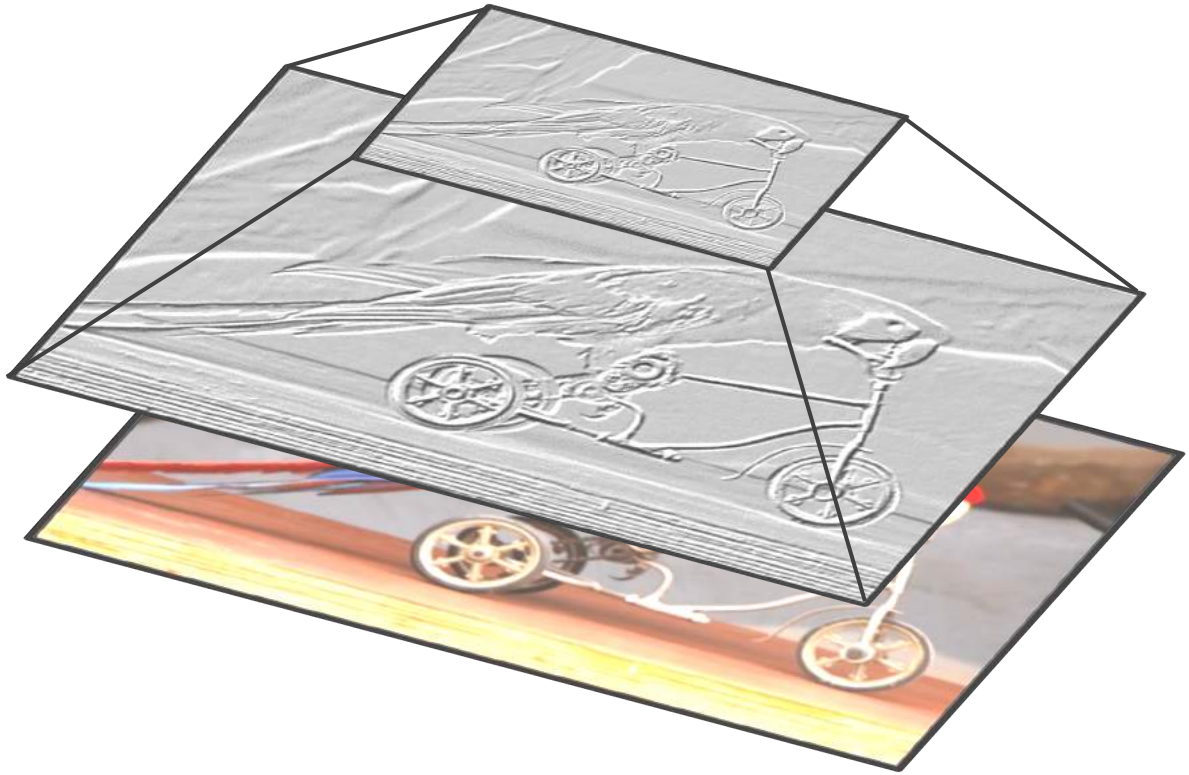


2012 ImageNet 1K (Fall 2012)



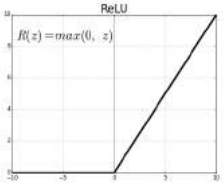
2012





Pooling

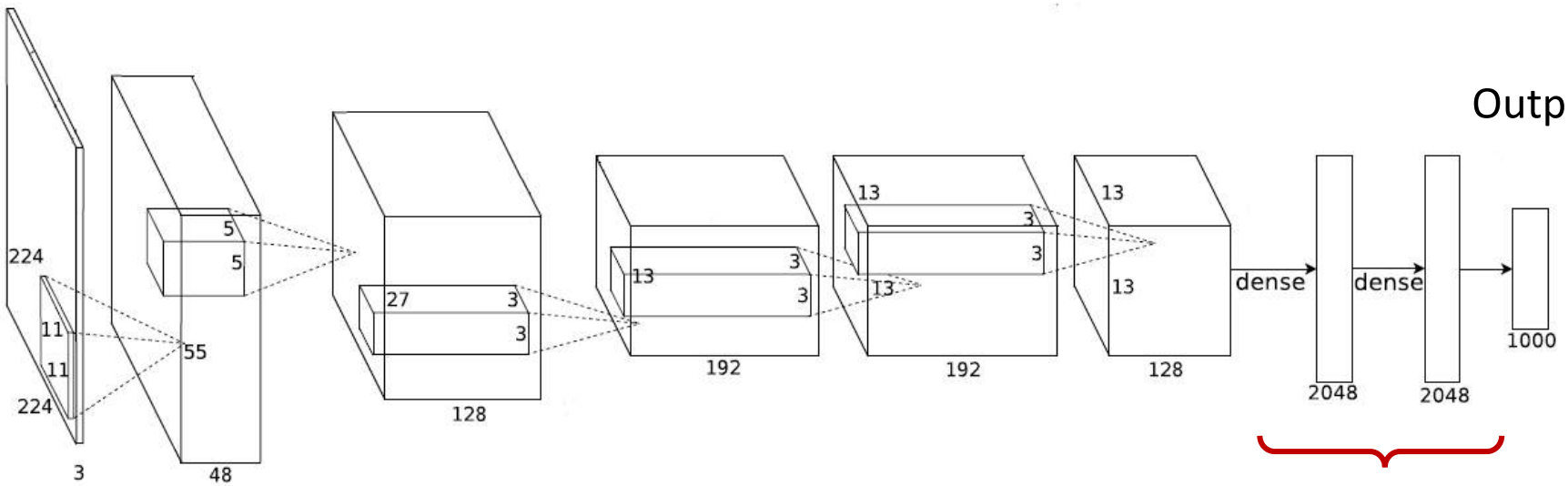
Convolution
+ ReLU



Image

2012 DNNs (deep neural networks)

Image



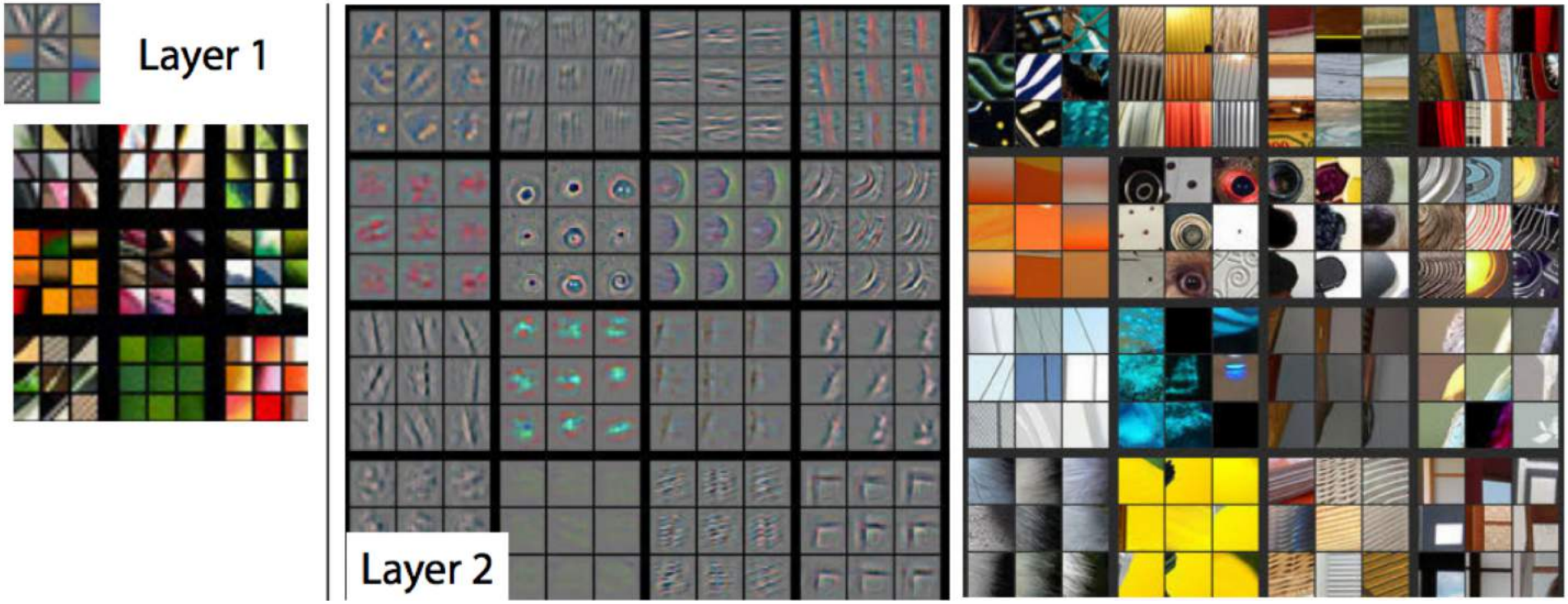
90% parameters

Visualizing and Understanding Convolutional NetworksImagenet Classification with Deep Convolutional Neural Networks, Krizhevsky, Sutskever, and Hinton, NIPS 2012

What are learned …

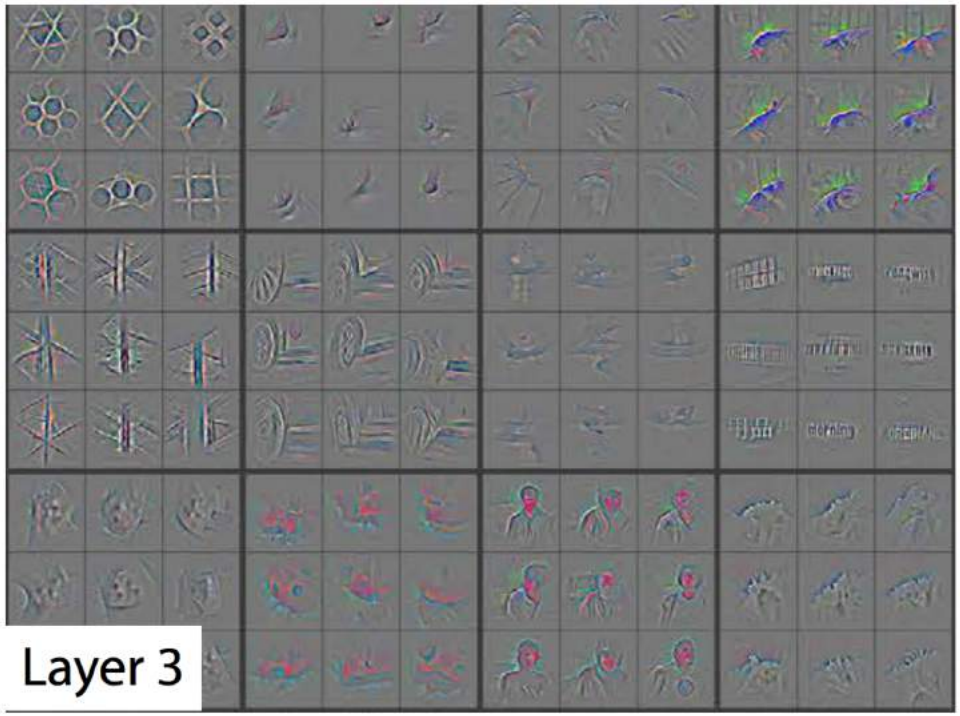
Filters (visualization)

Patches with high responses

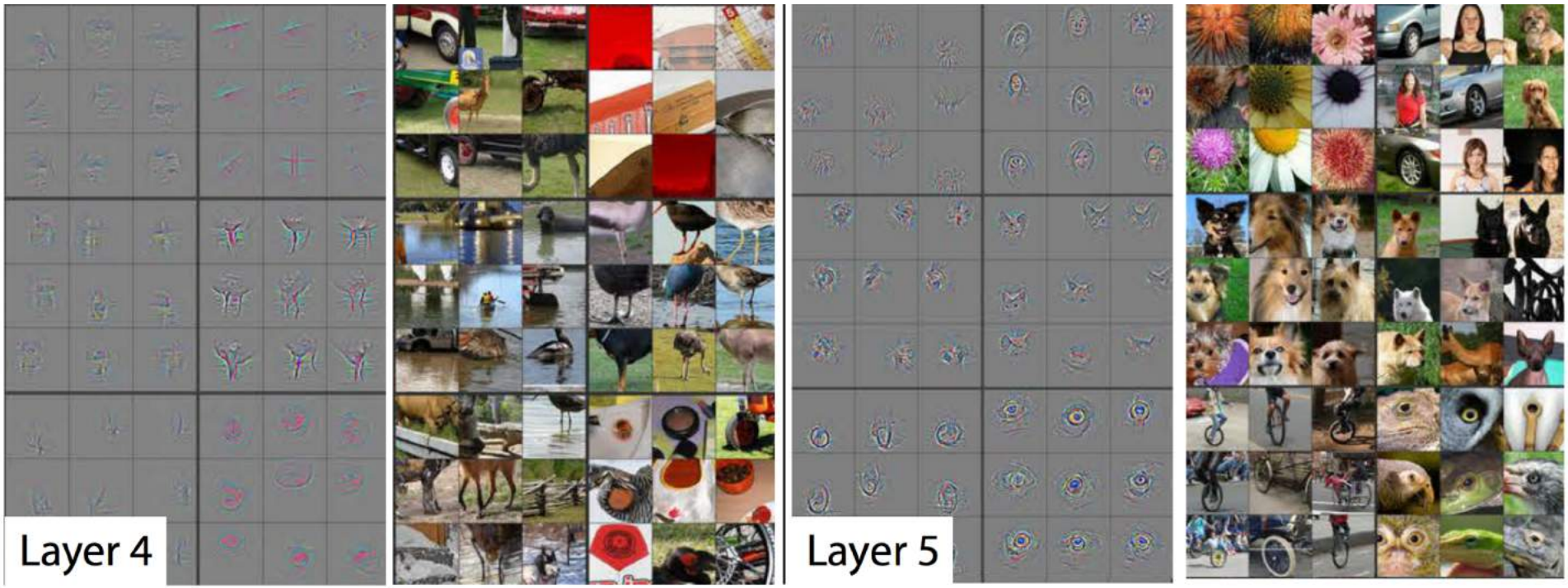


Visualizing and Understanding Convolutional Networks,
Zeiler and Fergus, ECCV 2014

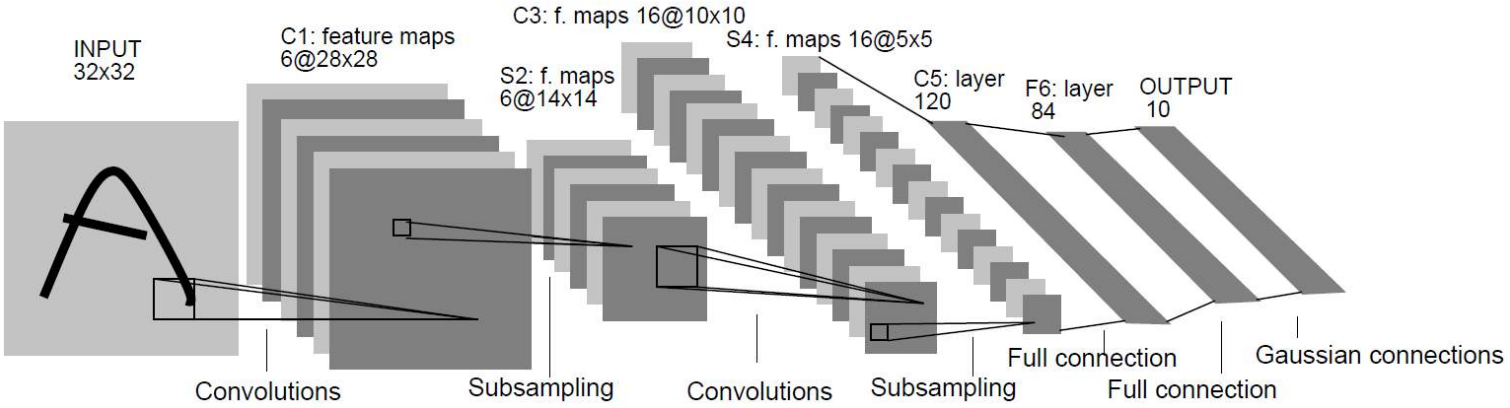
What are learned …



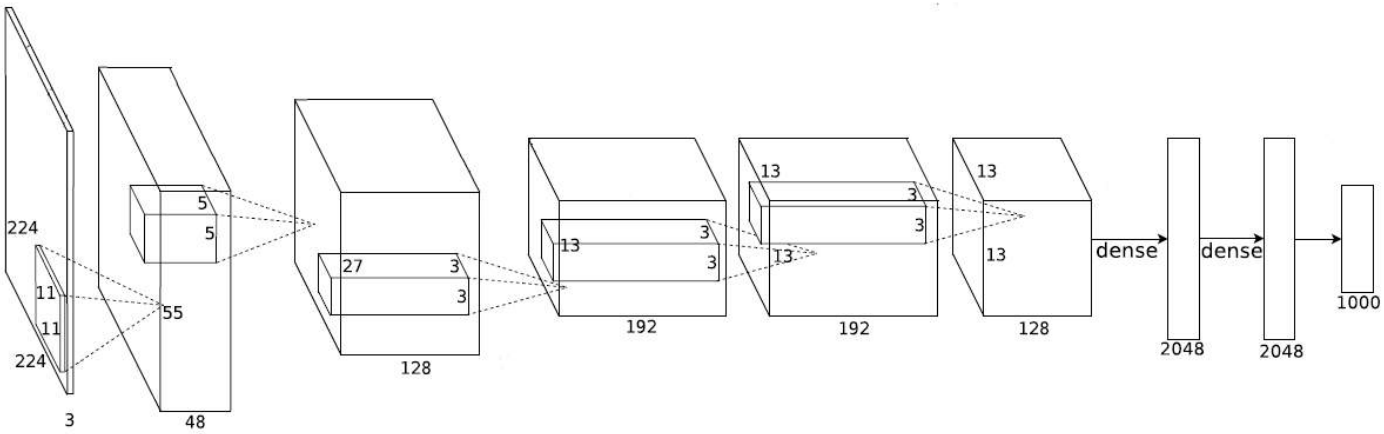
What are learned …



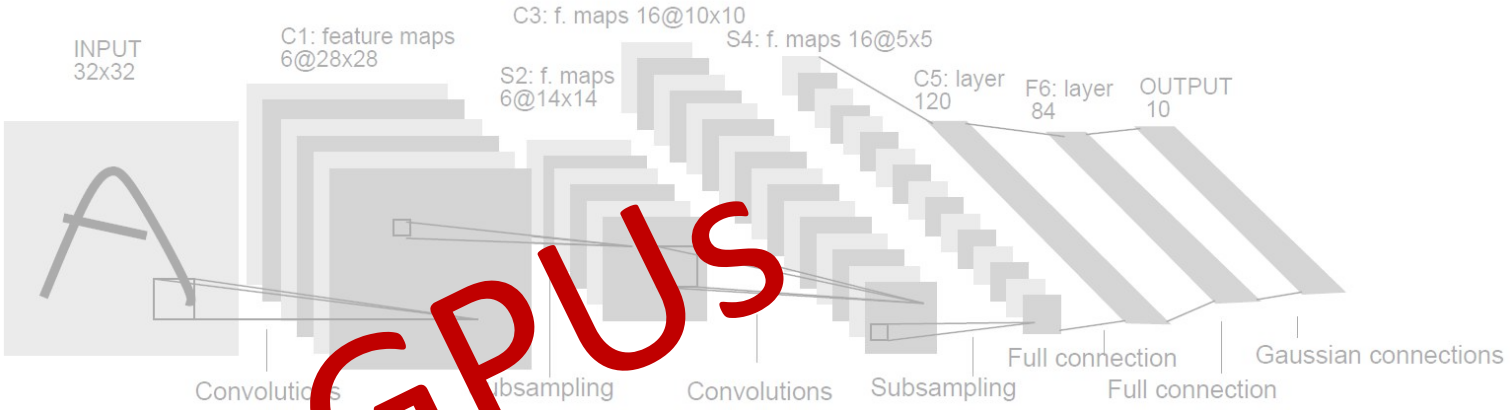
Building the concept of objects progressively
(edge/texture → common shapes → parts → objects)



Gradient-Based Learning Applied to Document Recognition,
LeCun, Bottou, Bengio and Haffner, Proc. of the IEEE, **1998**



Imagenet Classification with Deep Convolutional Neural Networks,
Krizhevsky, Sutskever, and Hinton, NIPS **2012**

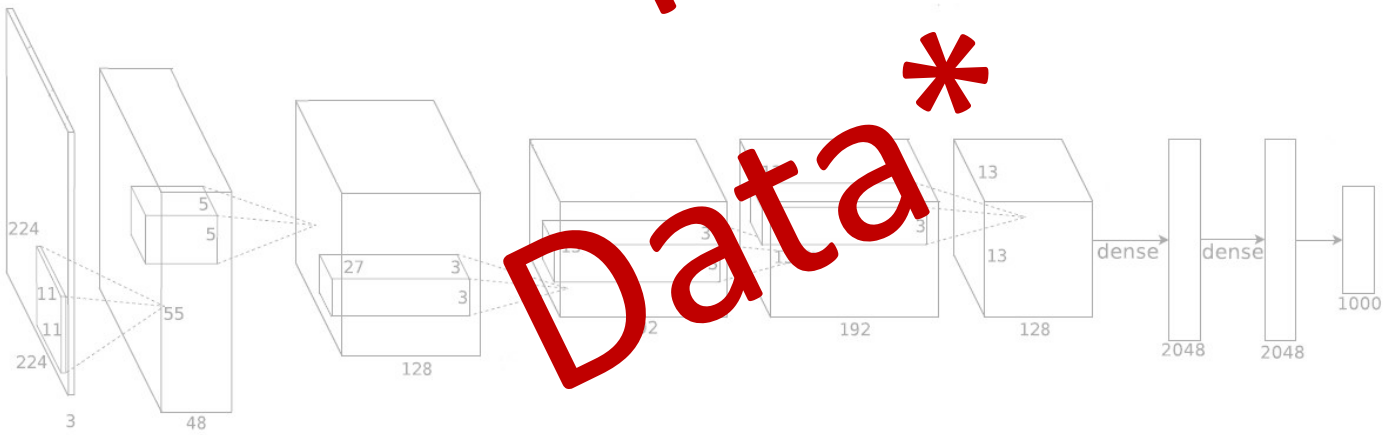


Gradient-Based Learning Applied to Document Recognition,
LeCun, Bottou, Bengio and Haffner, Proc. of the IEEE, **1998**

GPUs

\+

Data*



Imagenet Classificati
Krizhevsky, Sutskever

* Rectified activations and dropout

2012 ~ Now

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

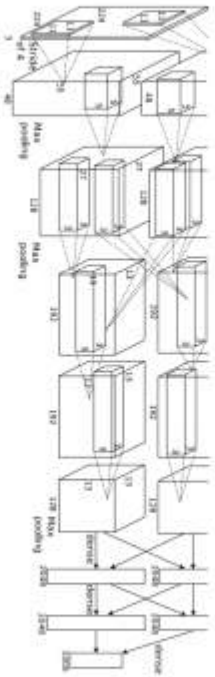
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

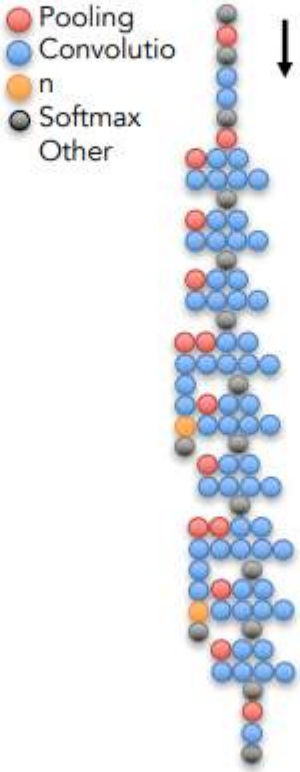
SuperVision



[Krizhevsky NIPS 2012]

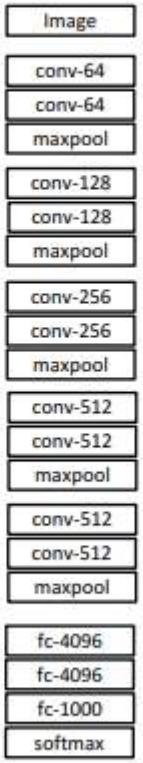
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

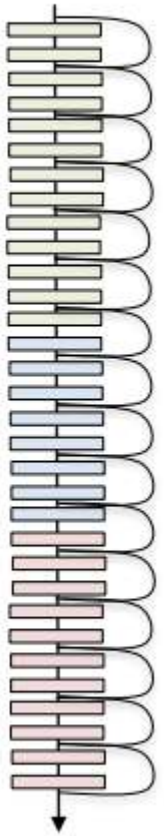
VGG



[Simonyan arxiv 2014]

Year 2015

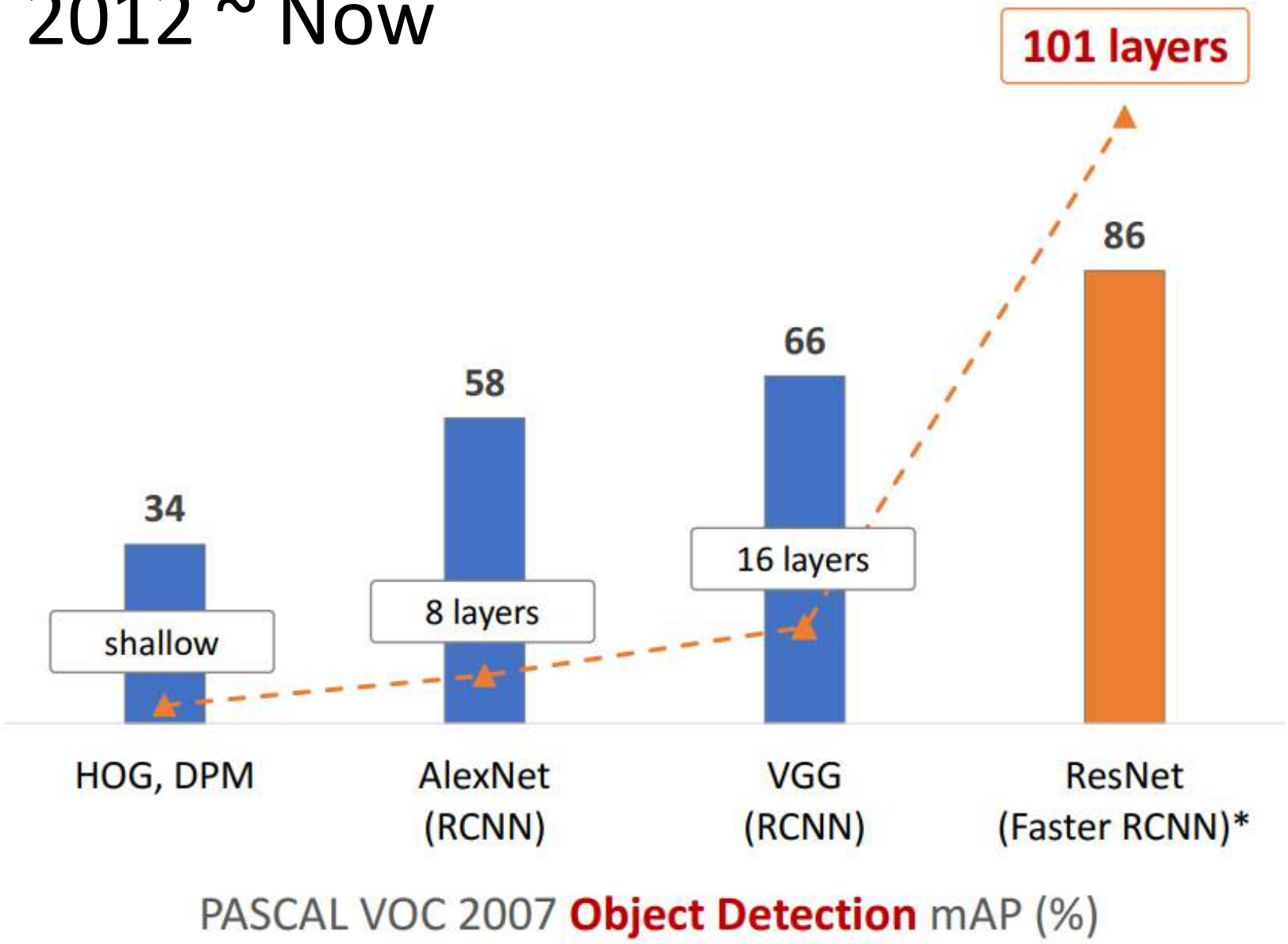
MSRA



[He ICCV 2015]

slide by Fei-Fei, Johnson & Yeung

2012 ~ Now



slide by He

2012 ~ Now

- Architectures
  - AlexNet (ReLU + Dropout)
  - Inception Network
  - Residual Network
- Models
  - CNN / RNN
  - GAN / VAE
- Initialization & Normalization
  - Better initialization
  - Batch Norm
- Optimization
  - Adam, etc …

- Image Classification
- Object Detection
- Semantic Segmentation
- Image Generation
- 3D Understanding
- Visual Question Answering
- Image captioning
- Action recognition
- Medical Image Analysis
- Robotics
- And many many more …

2012 ~ Now

slide by Fei-Fei, Johnson & Yeung

- Data to build observation models
- Data to build priors about the visual world
- Use the models and prior information to infer

Learning based methods in Computer Vision !

Not about developing machine learning models
Focus on using learning models for vision problems!

“What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking.”

-- David Marr, *Vision (1982)*

The end