CS540 Introduction to Artificial Intelligence

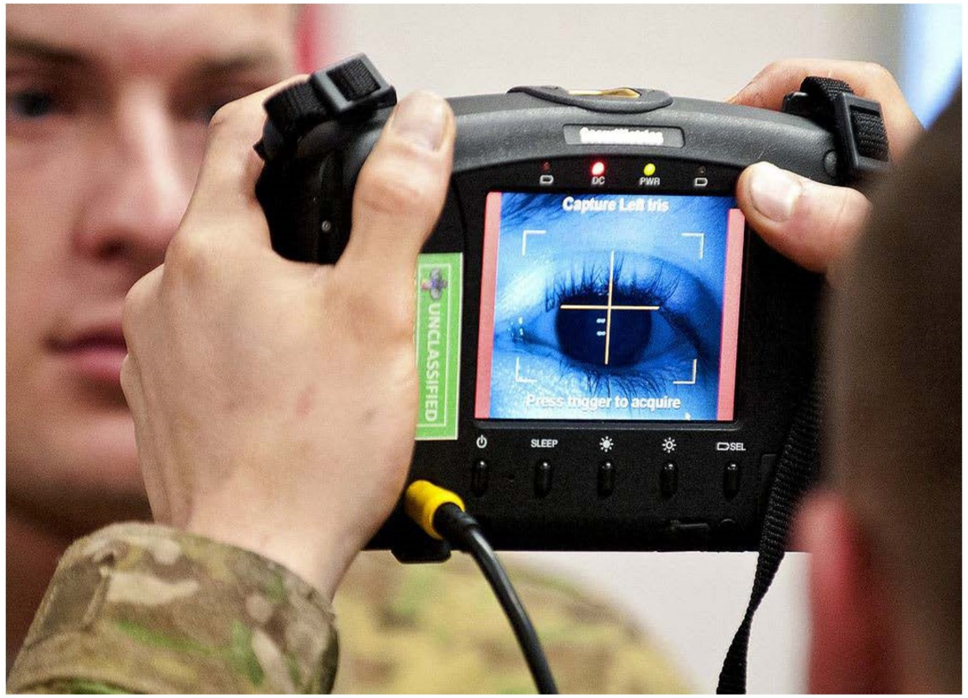
Dandi Chen
dandi.chen@wisc.edu
6/12/2019

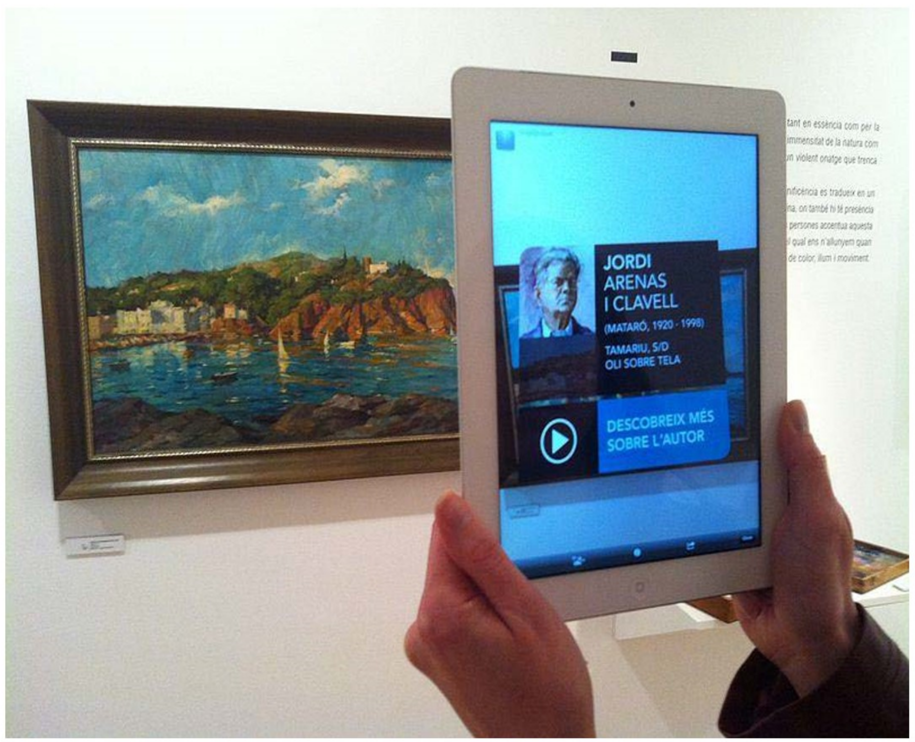
Outline

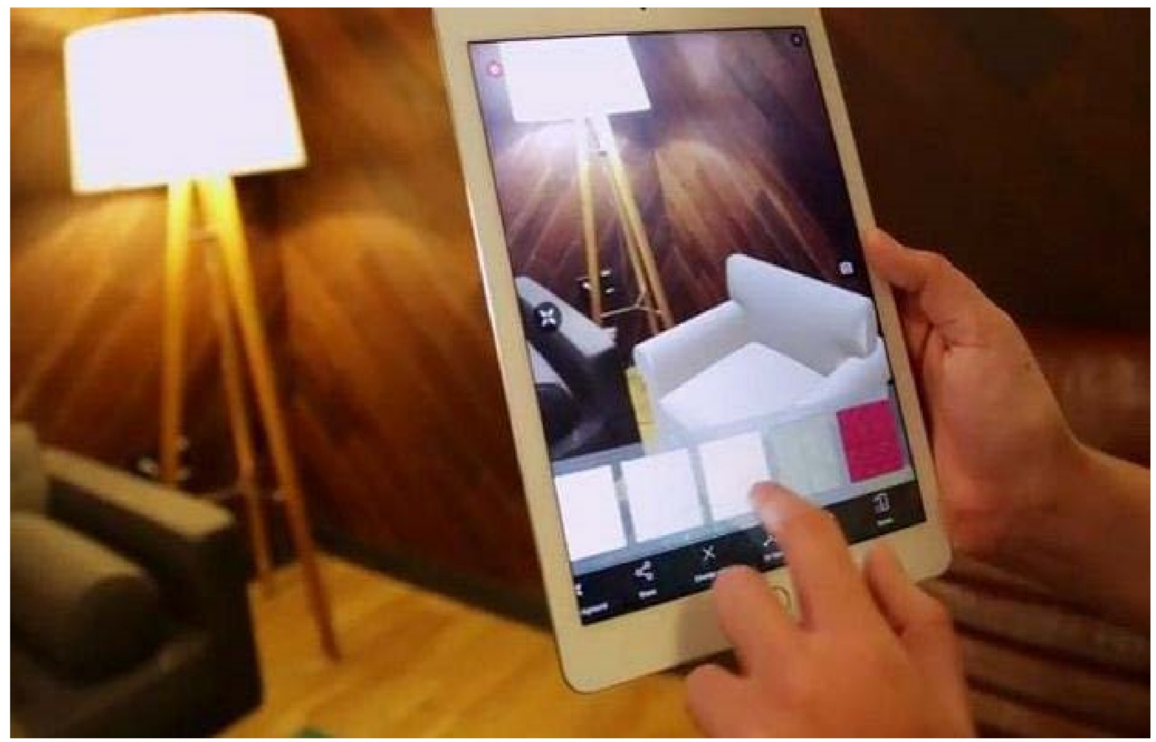
- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications

Outline

- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications







Slides from Fei-Fei Li & Justin Johnson & Serena Yeung

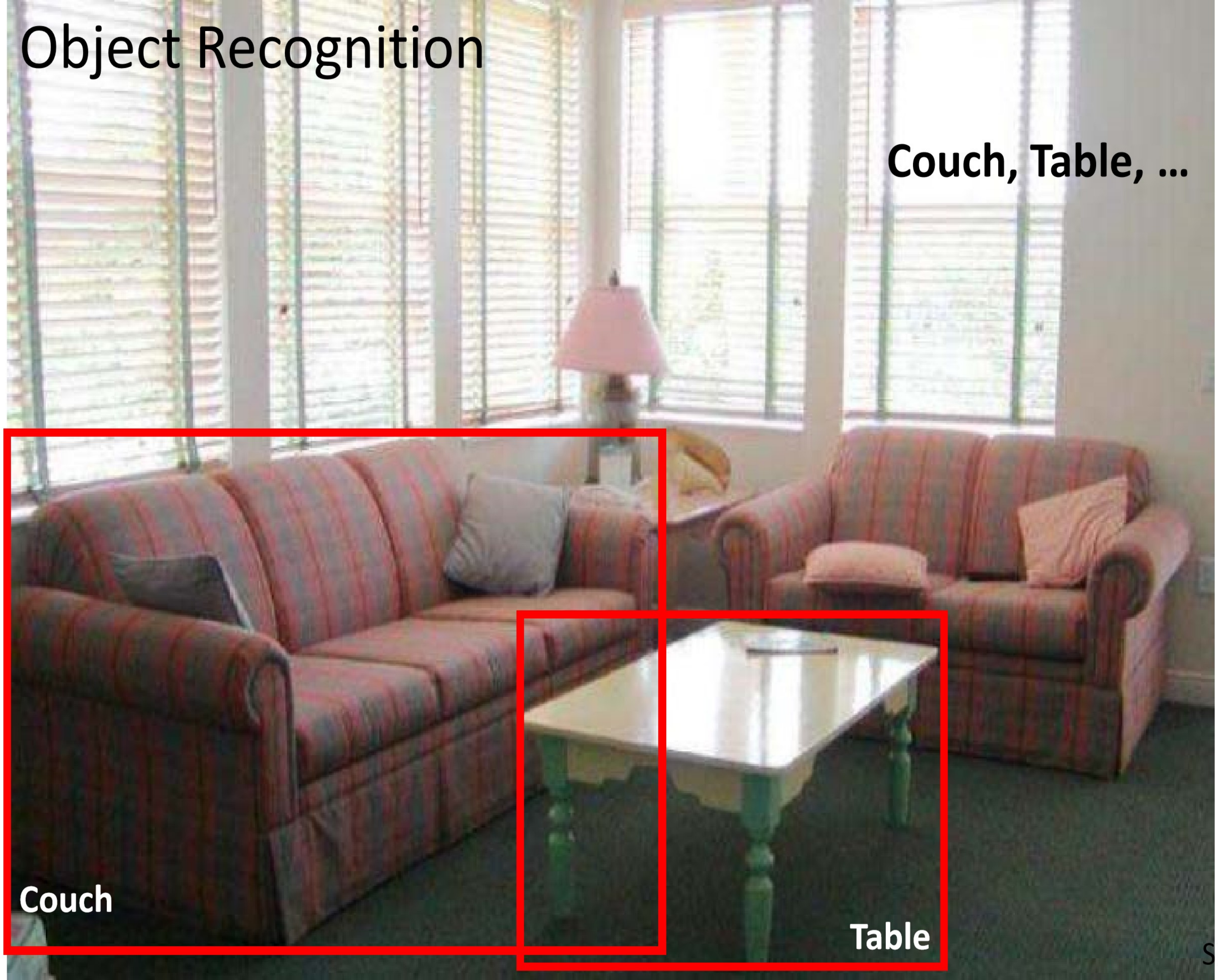
What do humans care about?

Slides from Yin Li

Image Classification/Scene Recognition

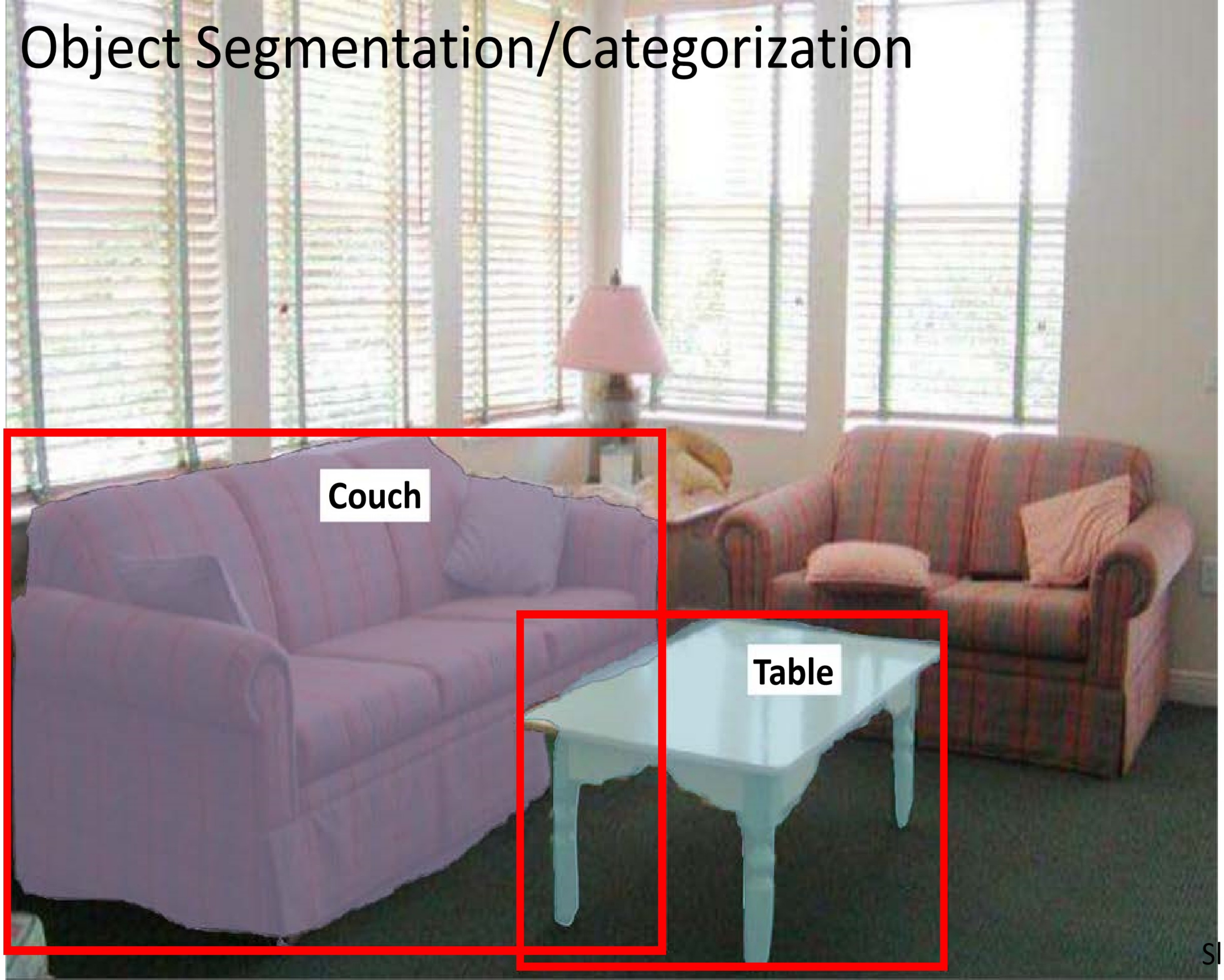
Slides from Yin Li

Object Recognition



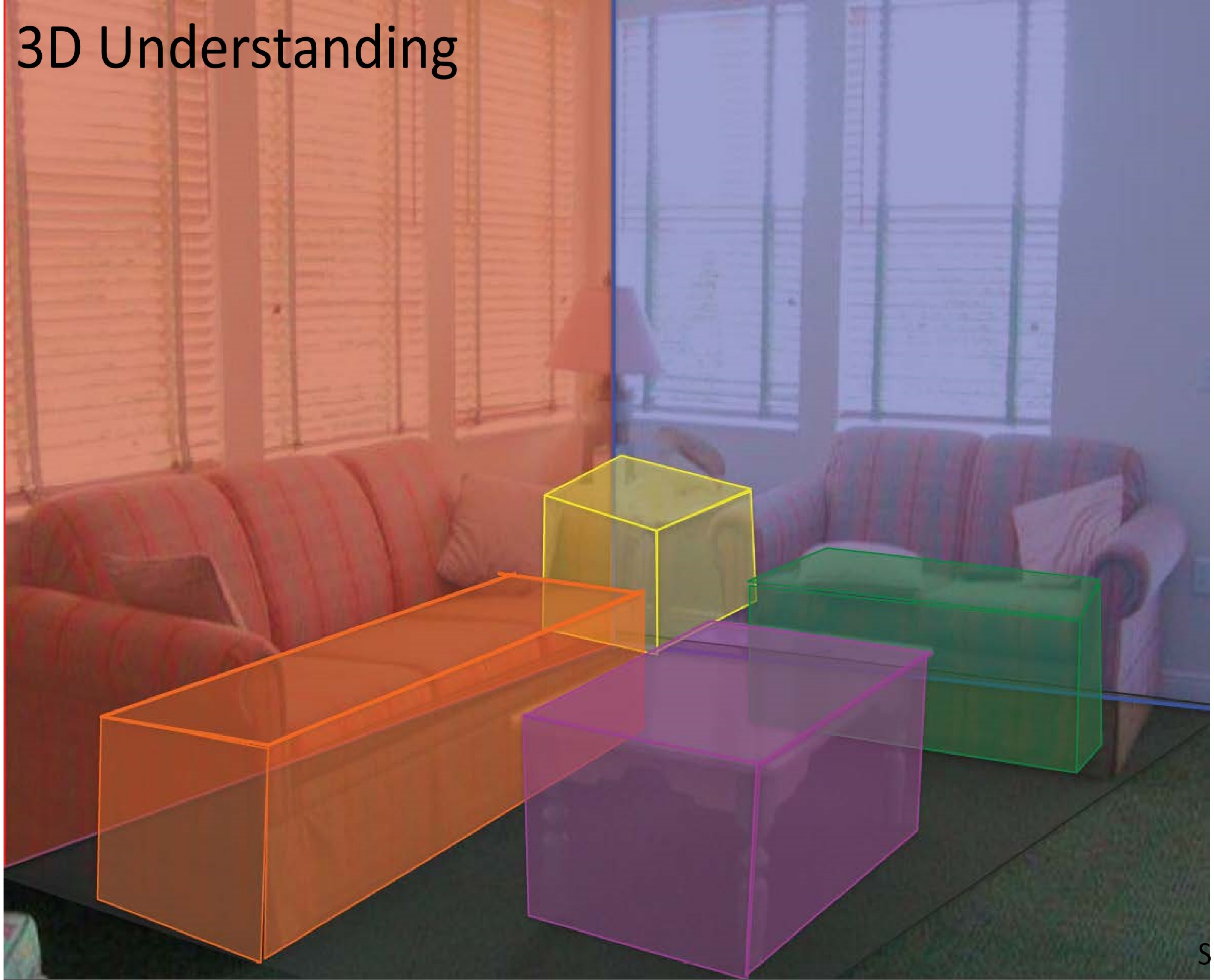
Slides from Yin Li

Object Segmentation/Categorization



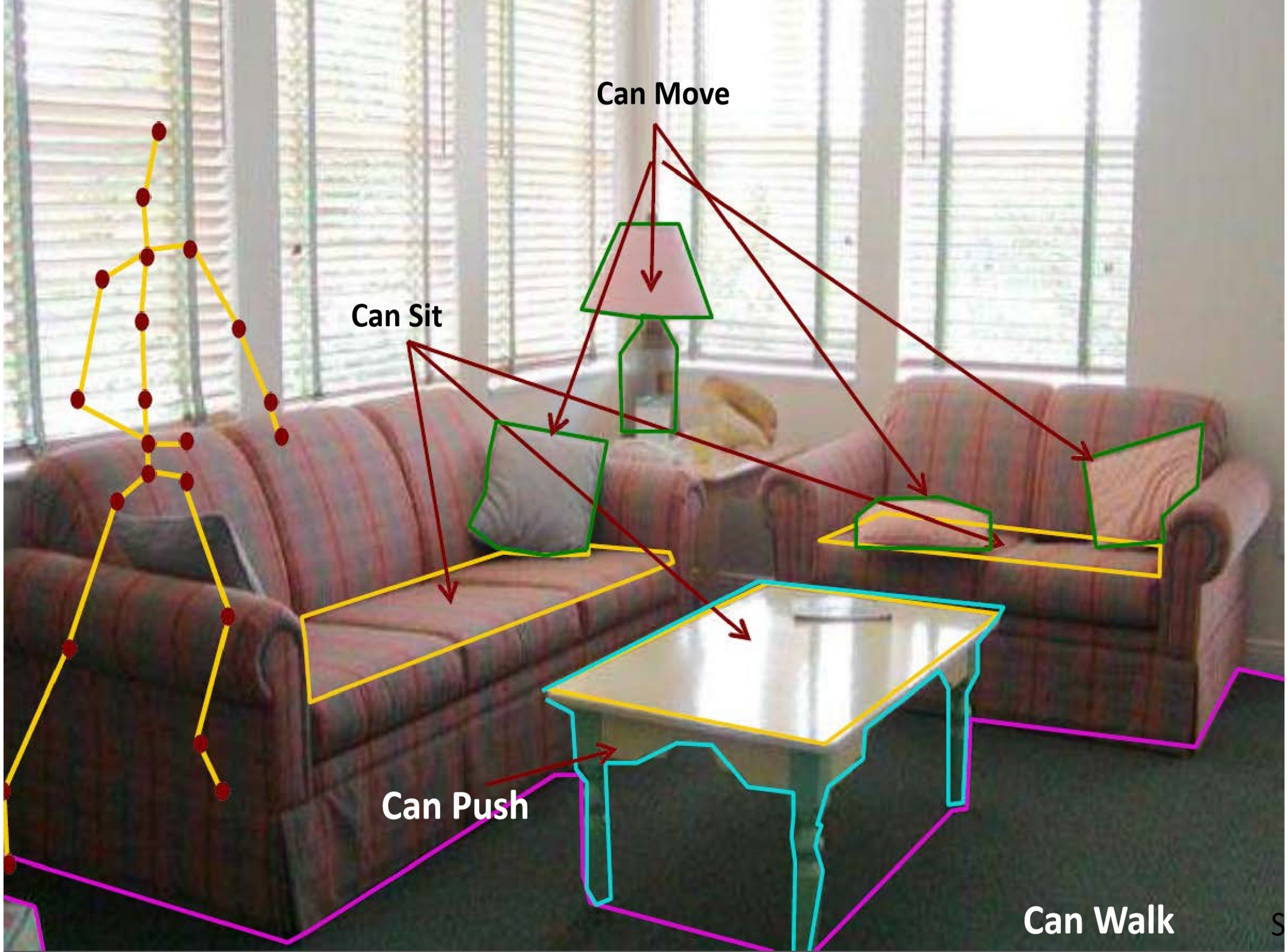
Slides from Yin Li

3D Understanding



Slides from Yin Li

Functional Understanding



Slides from Yin Li

Overview

- Three stages of Computer Vision
  - Low-level: pixels
    - Edges, texture, regions…
  - Mid-level: features
    - Geometry, motion…
  - High-level: semantics
    - Objects, events, scenes…

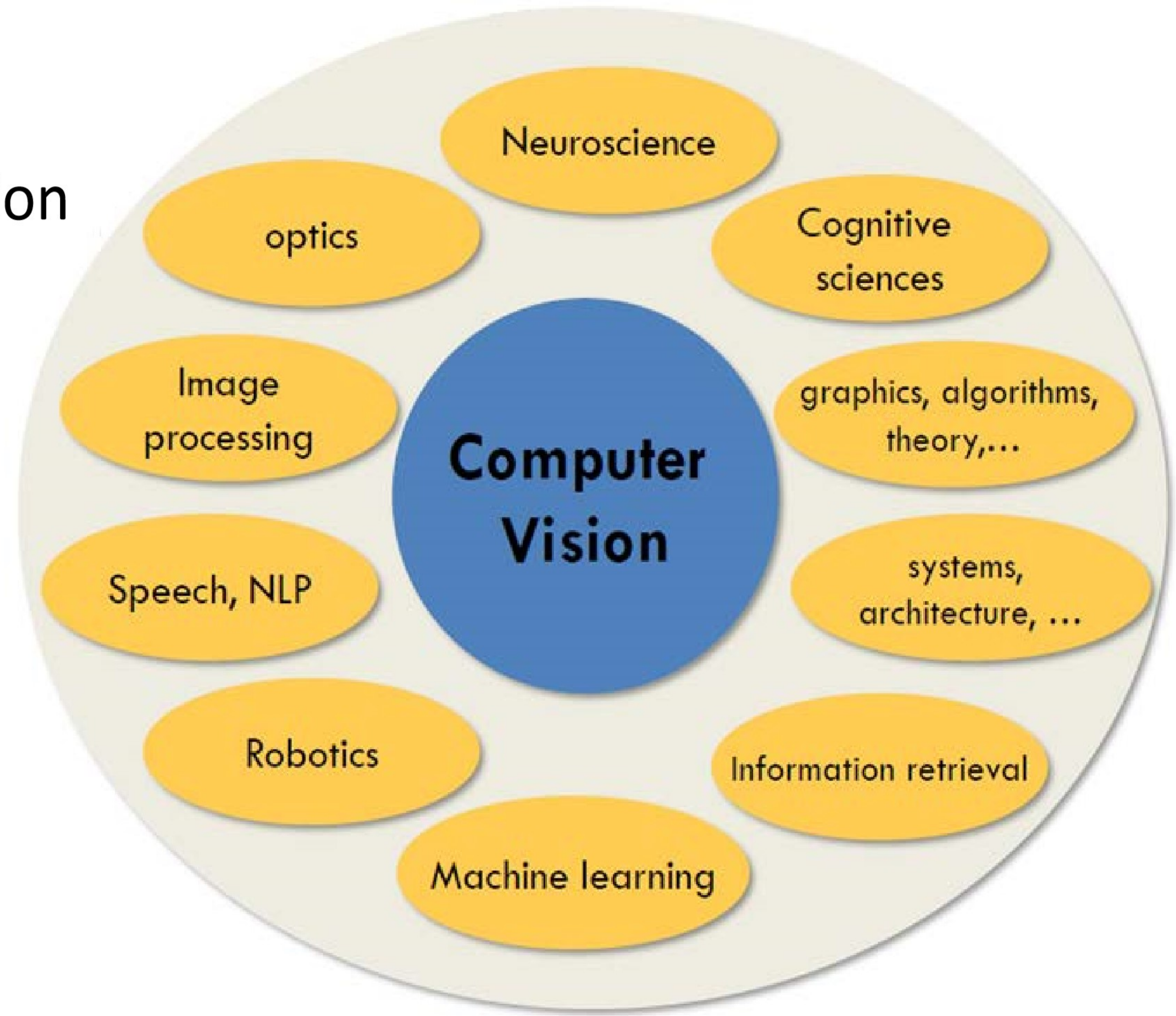


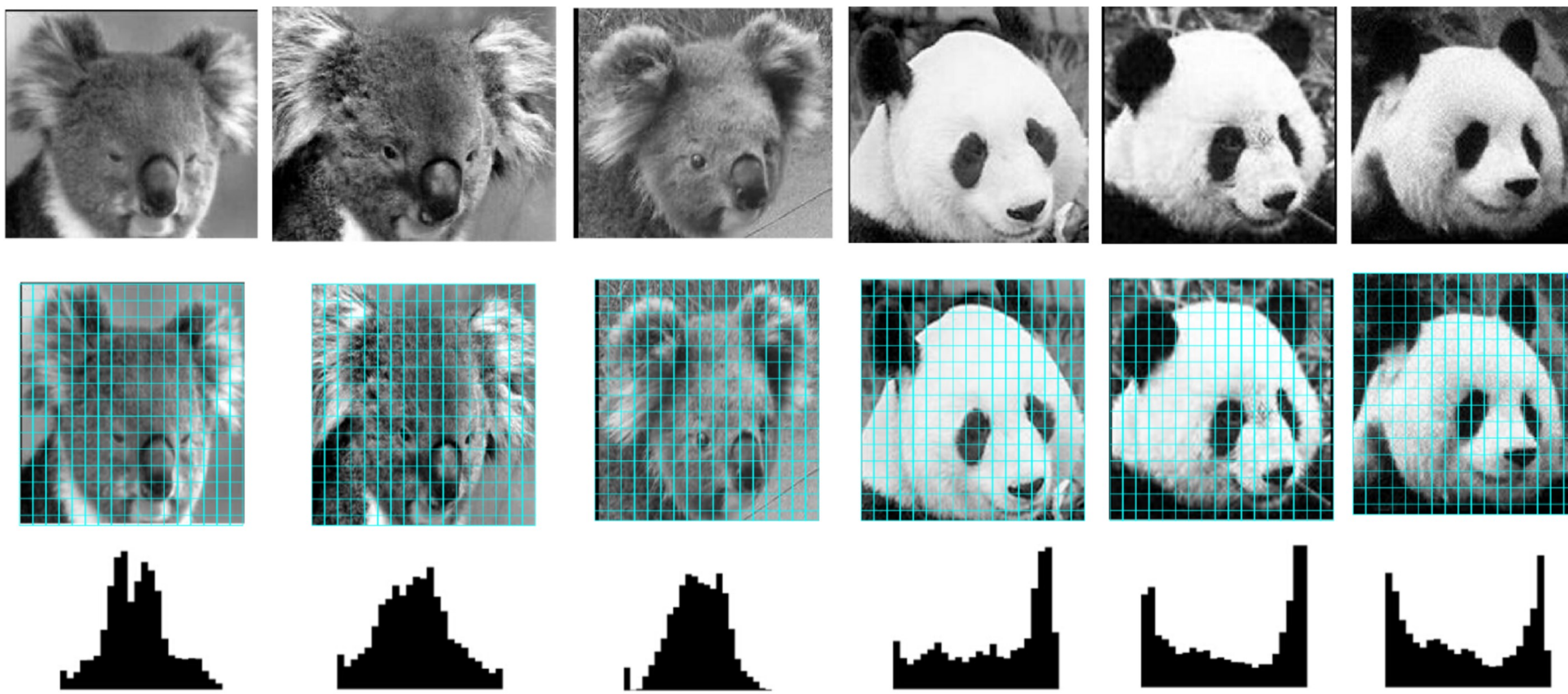
Image from Fei-Fei Li & Justin Johnson & Serena Yeung

Outline

- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications

Representations

- Global appearance
  - Grayscale/color histogram
  - Pixel intensities



pixel

x

histogram

x

pixels with color

Image from Kristen Grauman & Bastian Leibe

Representations

- Gradient-based
  - Edges
  - Contours
  - (Oriented) intensity gradients

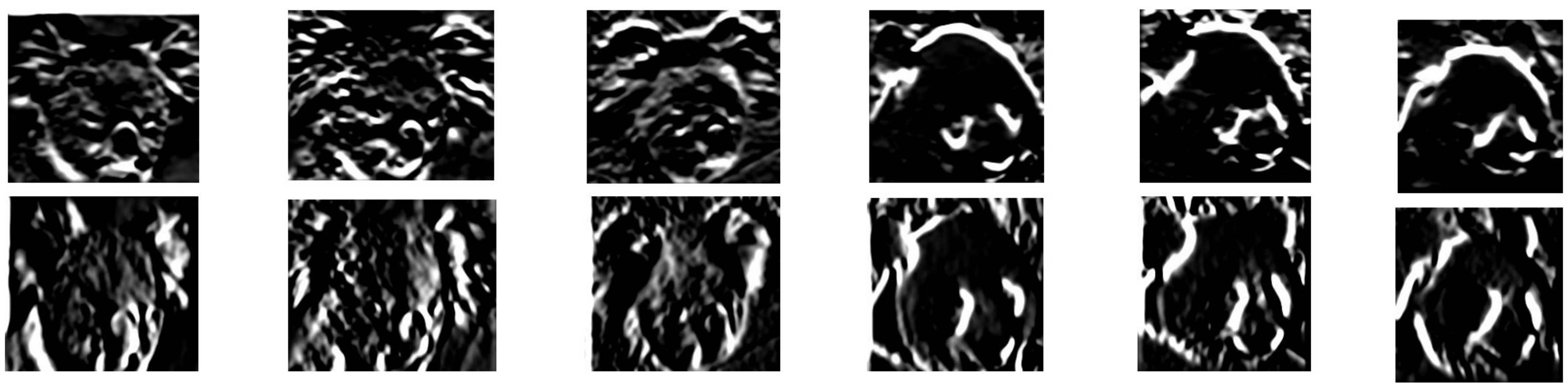
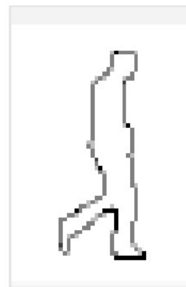


Image from Kristen Grauman & Bastian Leibe

Representations

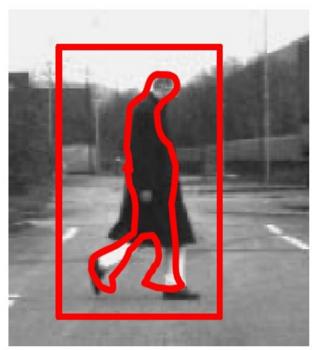
- Gradient-based: Chamfer matching

Input image



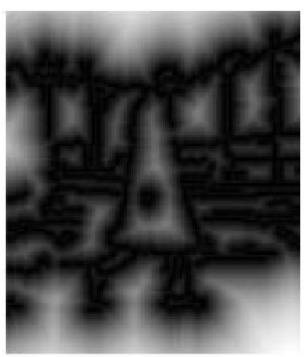
Template shape

Edges detected



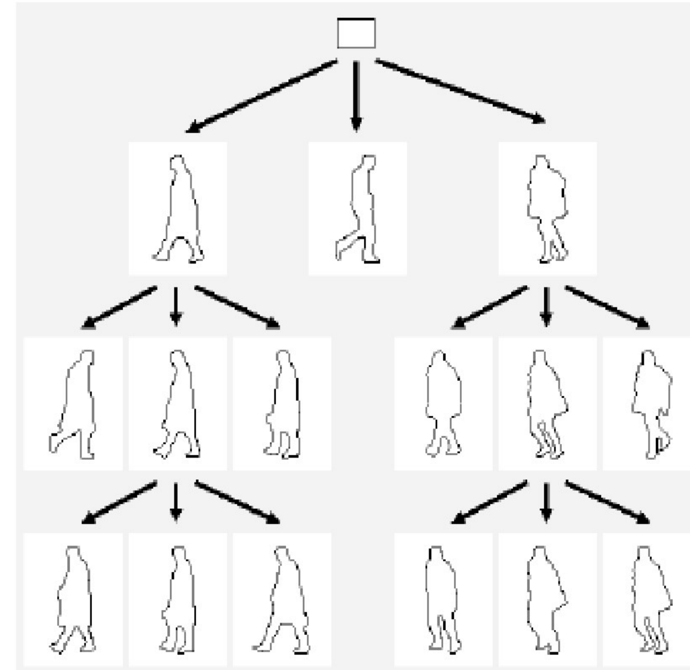
Best match

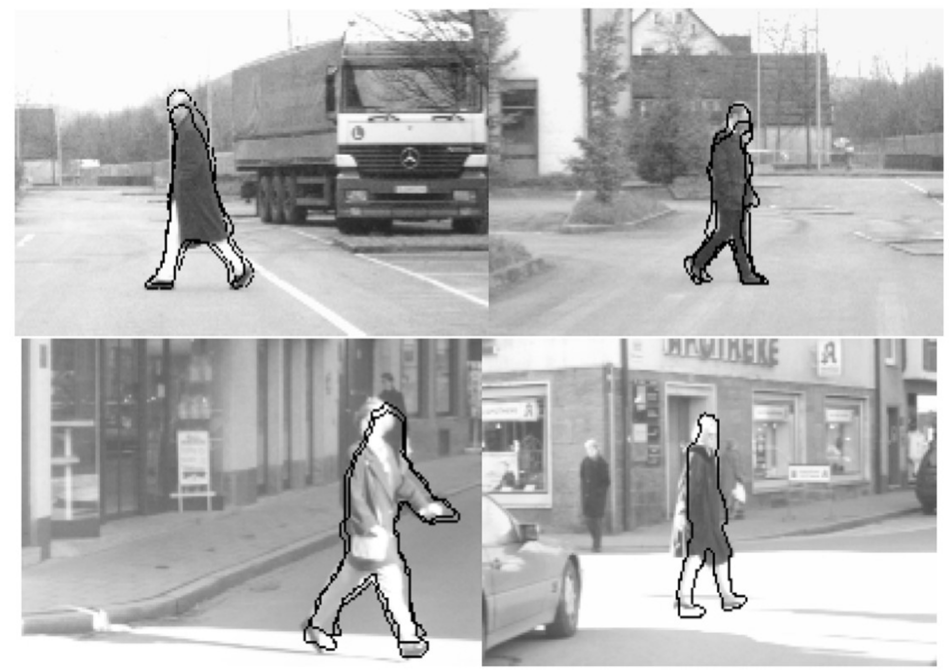
Distance transform



$$D(T,I) = \frac{1}{|T|}\sum_{t\in T} d_I(t)$$

Hierarchy of pedestrian shapes



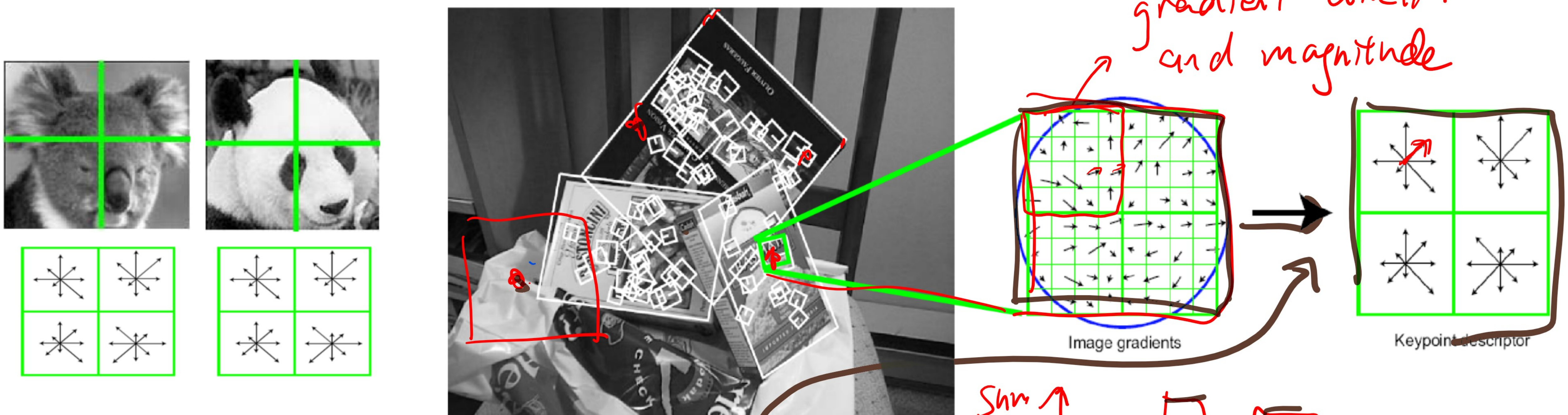


Gavrila, Dariu, et cl. "Real-time object detection for" smart" vehicles." ICCV 1999.

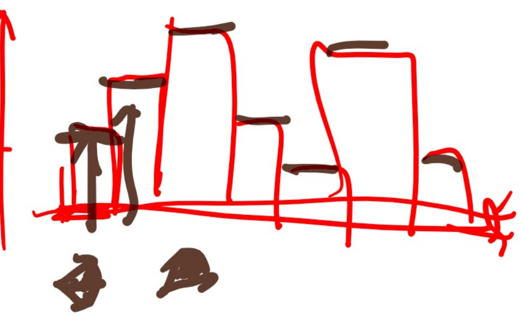
Representations

- Gradient-based: scale-invariant feature transform (SIFT)

gradient direction and magnitude



Sum magnitude



angle

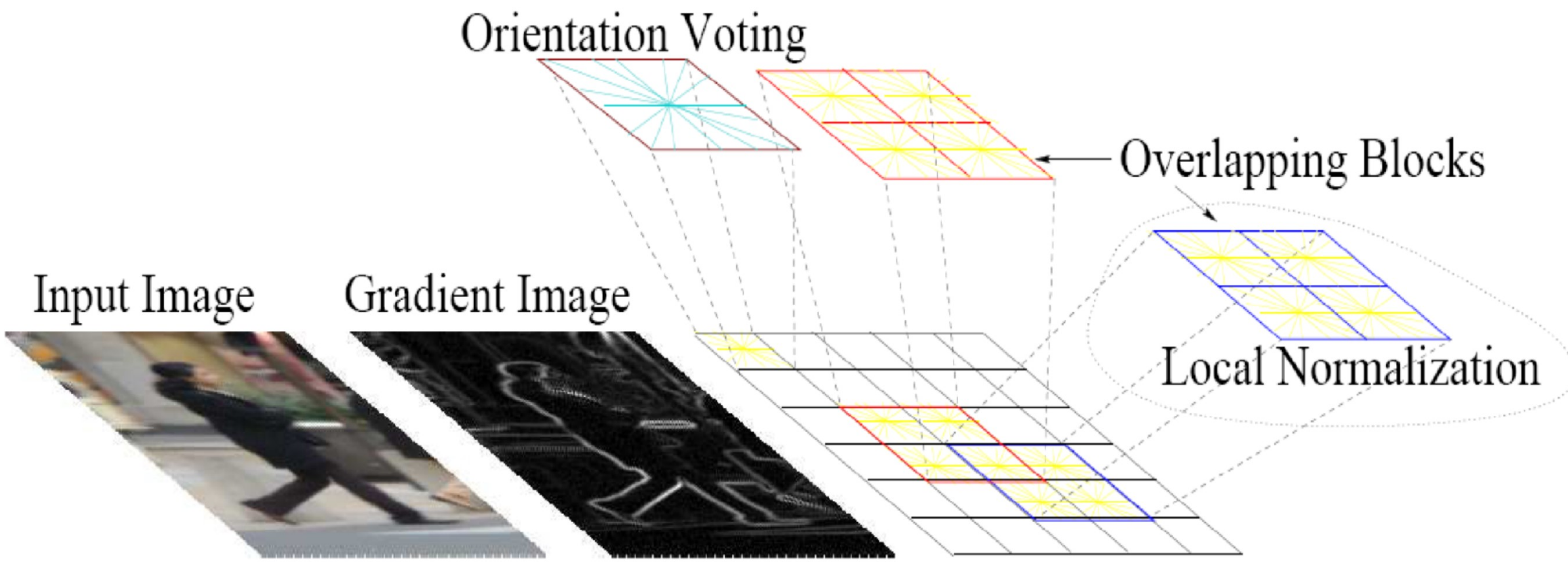
add up the magnitude of gradient facing similar directions.

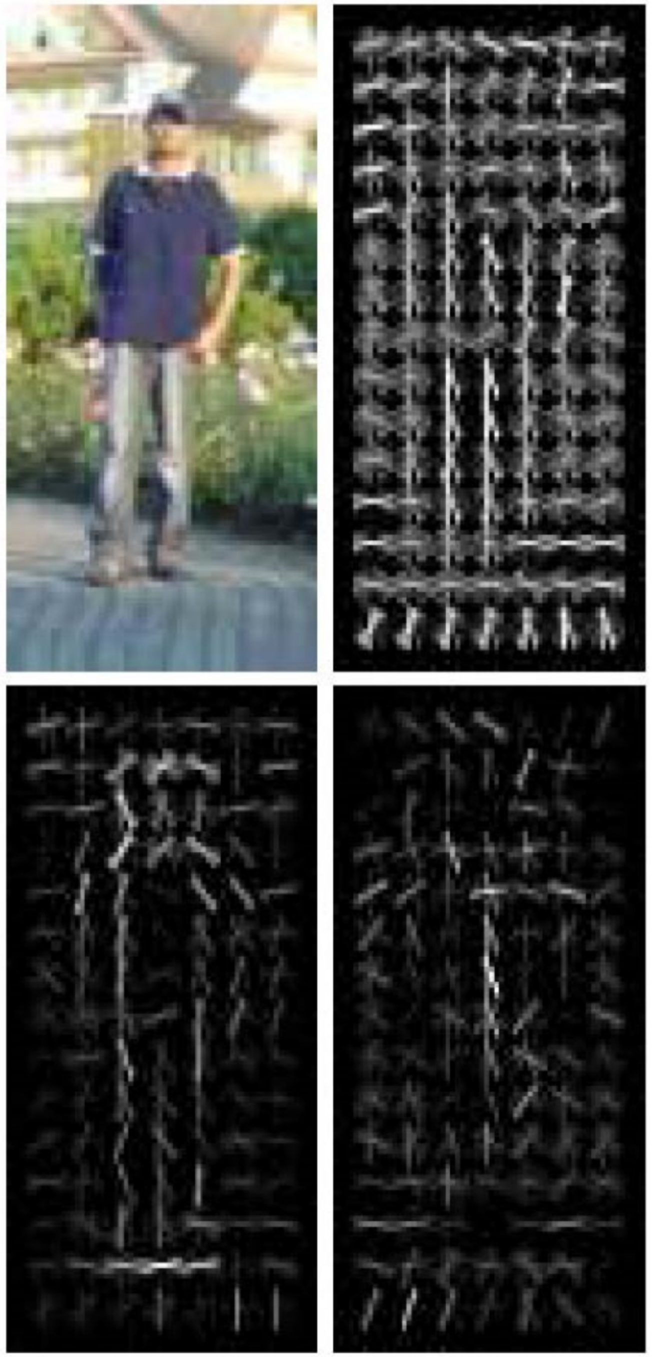
Lowe, David G. "Object recognition from local scale-invariant features." ICCV 1999.

x key point 1 key point 2

Representations

- Gradient-based: histograms of oriented gradients (HOG)





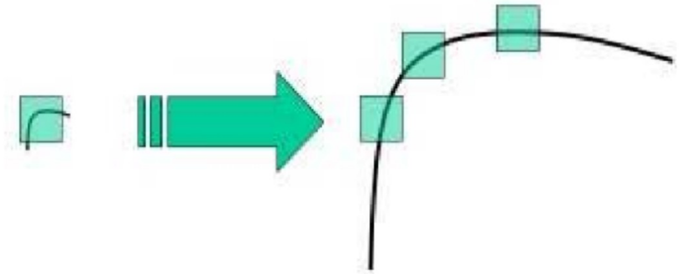
Dalal, Navneet, et al. “Histograms of oriented gradients for human detection.” CVPR 2005.

Scale-Invariant Feature Transform (SIFT)

key points

1) Scale-space Extrema Detection

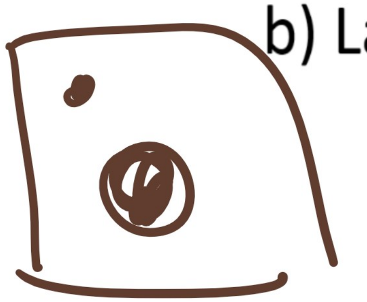
a) Blob detector: Laplacian of Gaussian with various σ



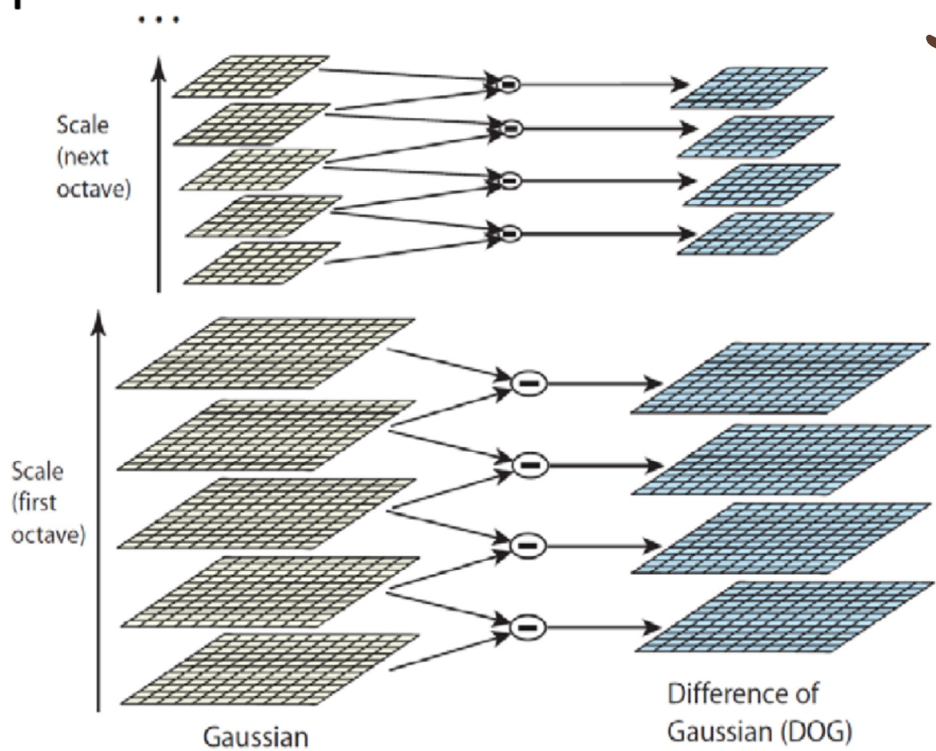
b) Laplacian of Gaussian -> Difference of Gaussian

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

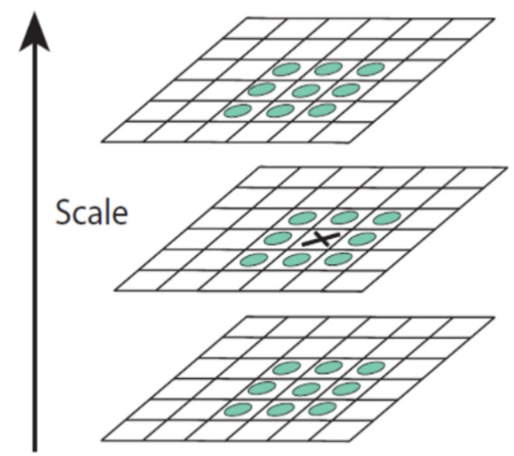


Gaussian Pyramid



$\frac{1}{4}$ image

original image



26 neighbors in 3×3 regions

Lowe, David G. "Distinctive image features from scale-invariant keypoints." IJCV 2004.

Scale-Invariant Feature Transform (SIFT)

2) Orientation Assignment

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$
$$\theta(x,y) = \tan^{-1}(L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))$$

- Assign orientations to keypoints to achieve invariance for image rotation (magnitudes and orientations)

all keypoint direction is the same

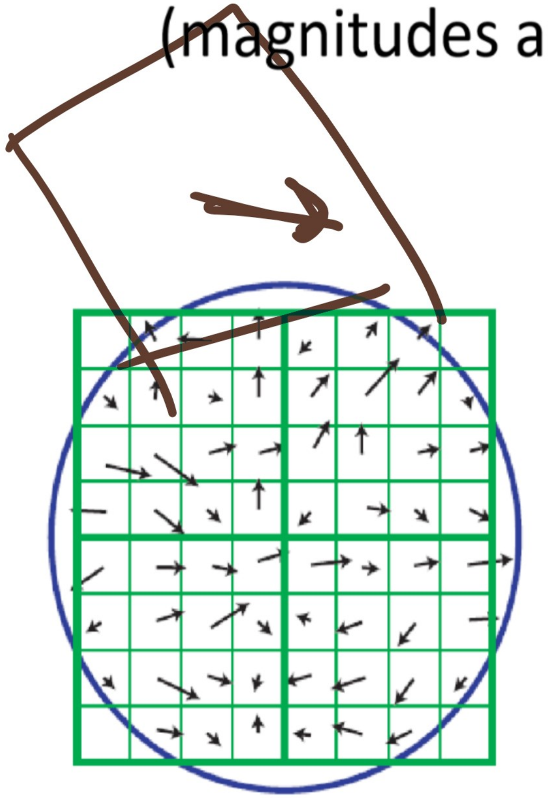
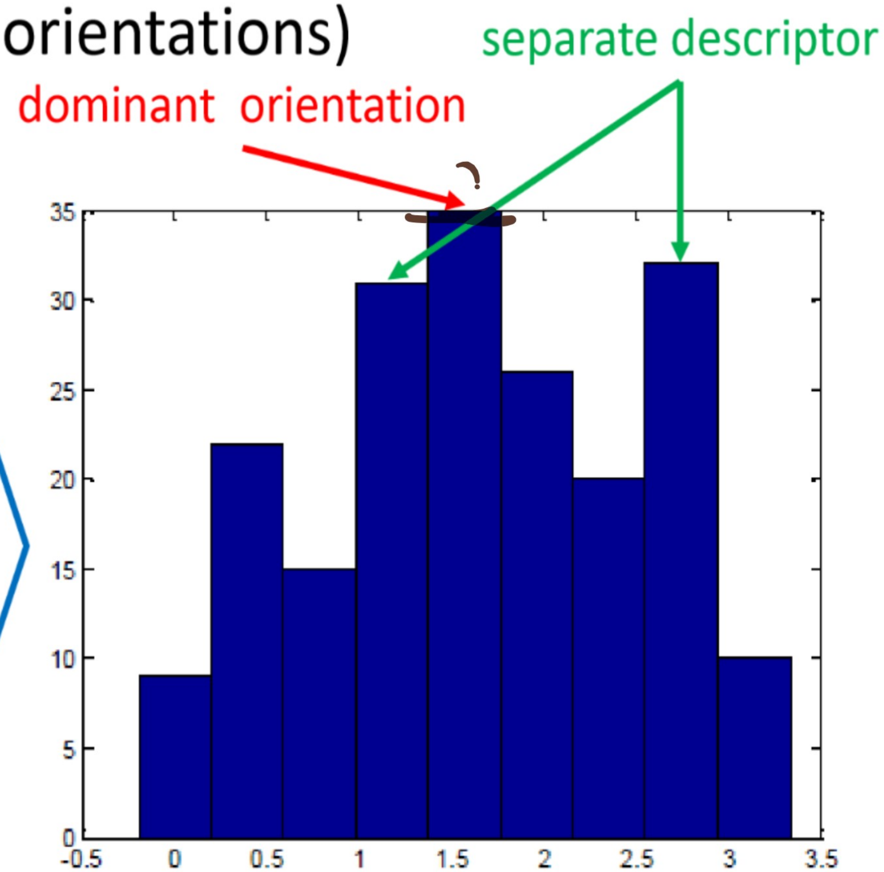


Image gradients



Dominant orientation: keypoint orientation

If multiple peaks or histogram entries more than 0.8 x peak, create a separate descriptor for each orientation.

Histogram of gradient orientation: the bin-counts are weighted by gradient magnitudes and a Gaussian weighting function. Usually, 36 bins are chosen covering 360 degrees.

Lowe, David G. "Distinctive image features from scale-invariant keypoints." IJCV 2004.

Scale-Invariant Feature Transform (SIFT)

3) Keypoint Descriptor

a) Define a small region around the keypoint.

b) Divide it into n×n cells (usually n= 2). Each cell is of size 4×4.

c) Build a gradient orientation histogram in each cell. 8 orientation -> 4×4×8 = 128 dim

d) Assign dominant orientation to the keypoint.

2x2

overlap

or

8 direction

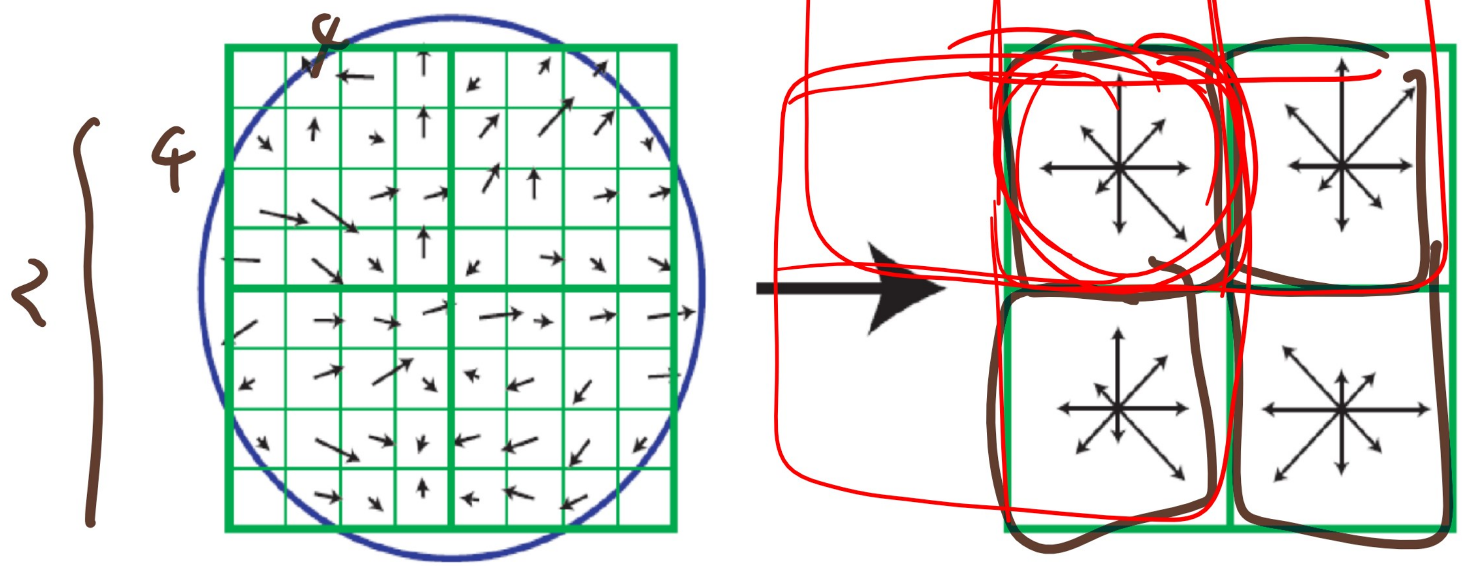


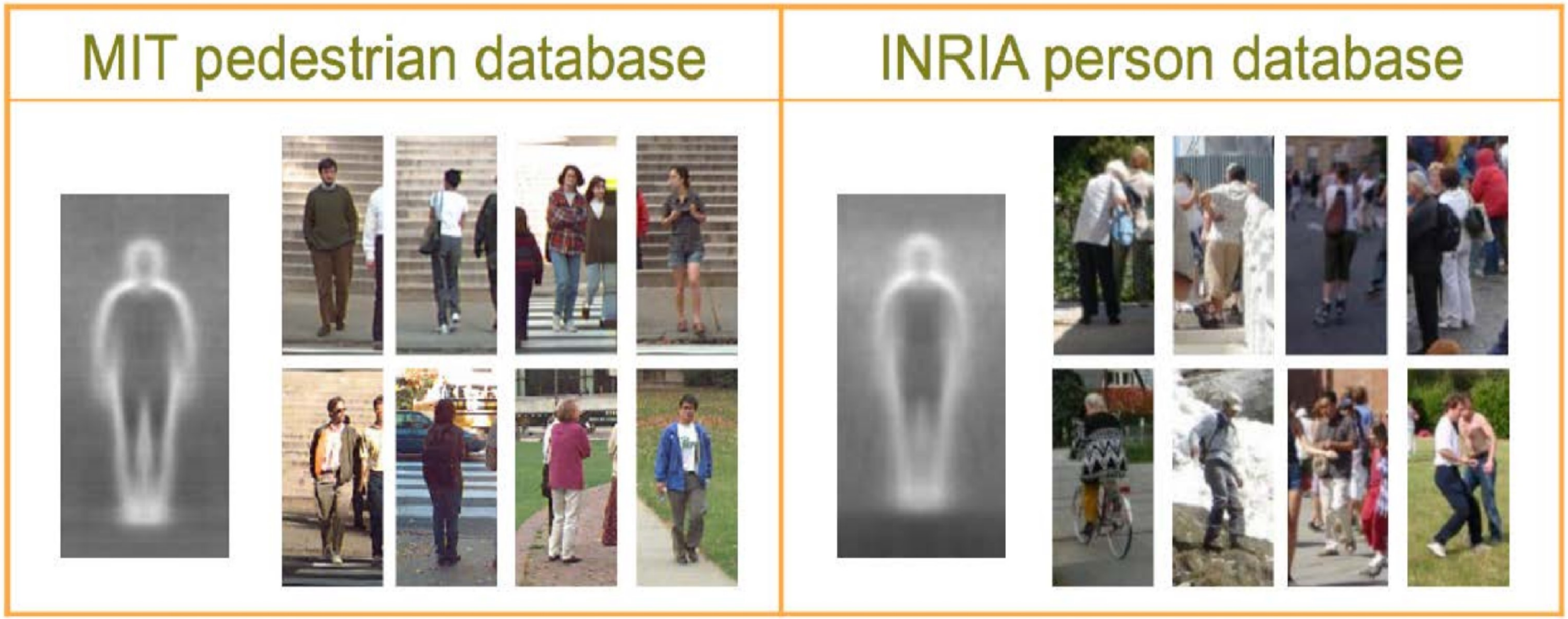
Image gradients Keypoint descriptor

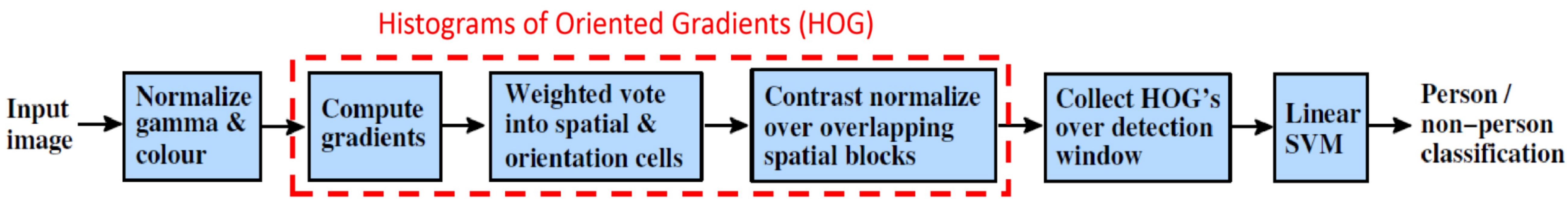
Lowe, David G. "Distinctive image features from scale-invariant keypoints." IJCV 2004.

$X_i \rightarrow$ 128

e.g.

Histograms of Oriented Gradients (HOG)





Dalal, Navneet, et al. "Histograms of oriented gradients for human detection." CVPR 2005.

Histograms of Oriented Gradients (HOG)

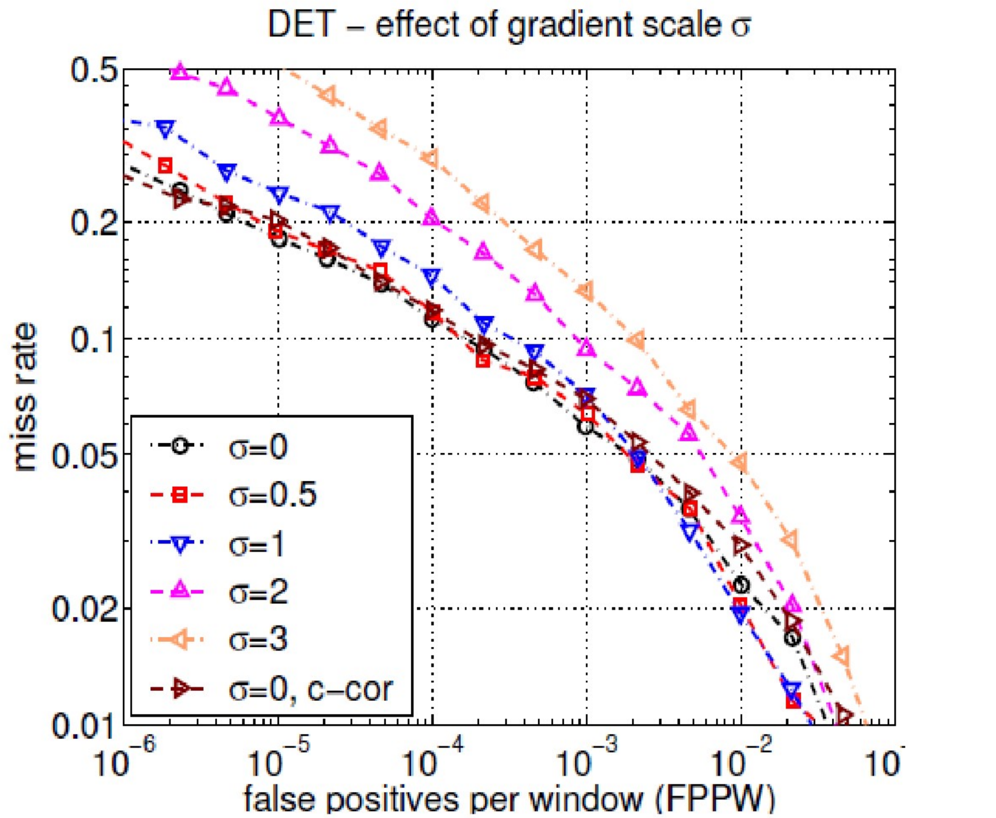
1) Compute gradients. The gradient of an image is defined as the change in pixel intensity due to the change in the location of the pixel.

Dalal, Navneet, et al. "Histograms of oriented gradients for human detection." CVPR 2005.

Histograms of Oriented Gradients (HOG)

1) Compute gradients: [-1, 0, 1] & σ = 0 – best performance

Mask Type	1D centered	1D uncentered	1D cubic-corrected	2x2 diagonal	3x3 Sobel
Operator	[-1, 0, 1]	[-1, 1]	[1, -8, 0, 8, -1]	$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$
Miss rate at 10^{-4} FPPW	11%	12.5%	12%	12.5%	14%



*σ = 0: no Gaussian smoothing.

Dalal, Navneet, et al. “Histograms of oriented gradients for human detection.” CVPR 2005.

Histograms of Oriented Gradients (HOG)

2) Weighted vote into spatial & orientation cells

a) Divide gradient image into non-overlapping cells. Each cell is typically 8×8 pixels.

b) Similar to SIFT, compute histogram of orientations in each cell.

c) Check best number of bins.

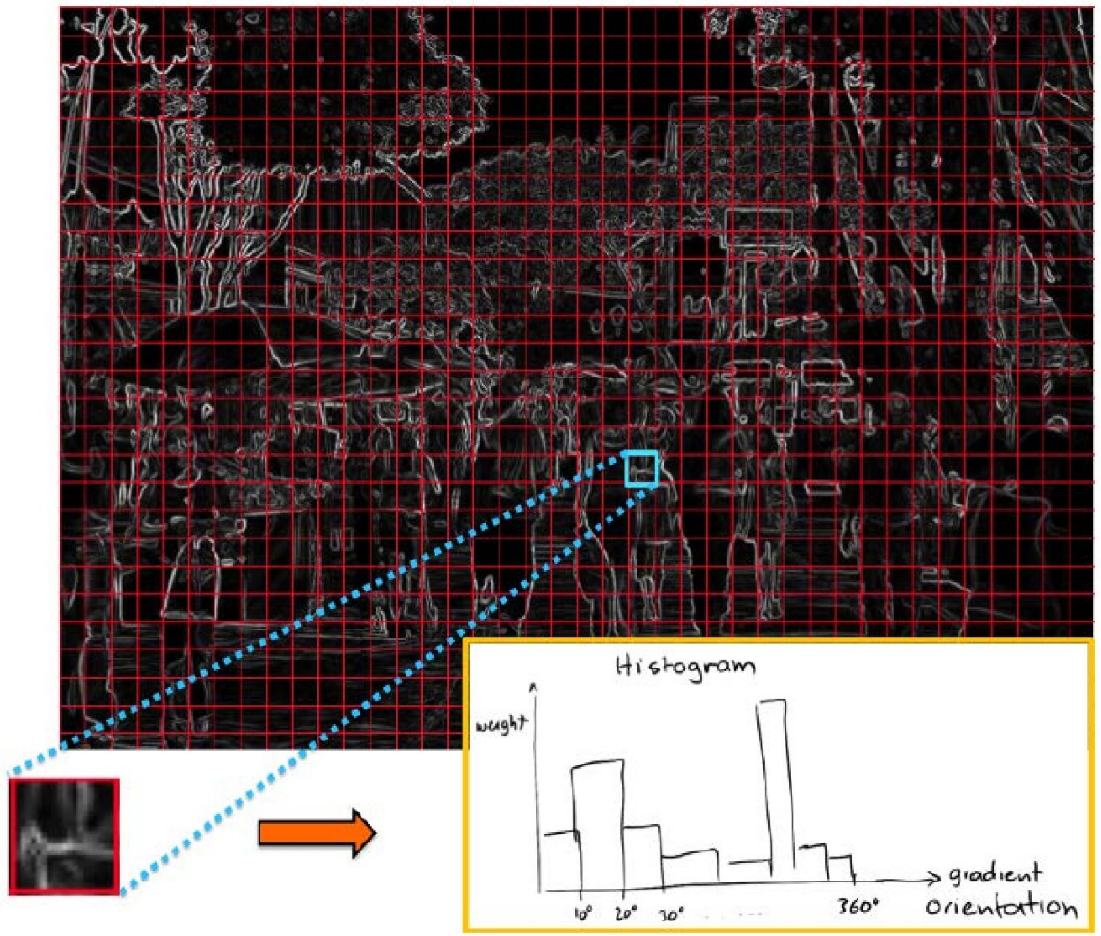


Image from Sanja Fidler

Histograms of Oriented Gradients (HOG)

2) Weighted vote into spatial & orientation cells

a) Divide gradient image into non-overlapping cells. Each cell is typically 8×8 pixels.

b) Similar to SIFT, compute histogram of orientations in each cell.

c) Check best number of bins.

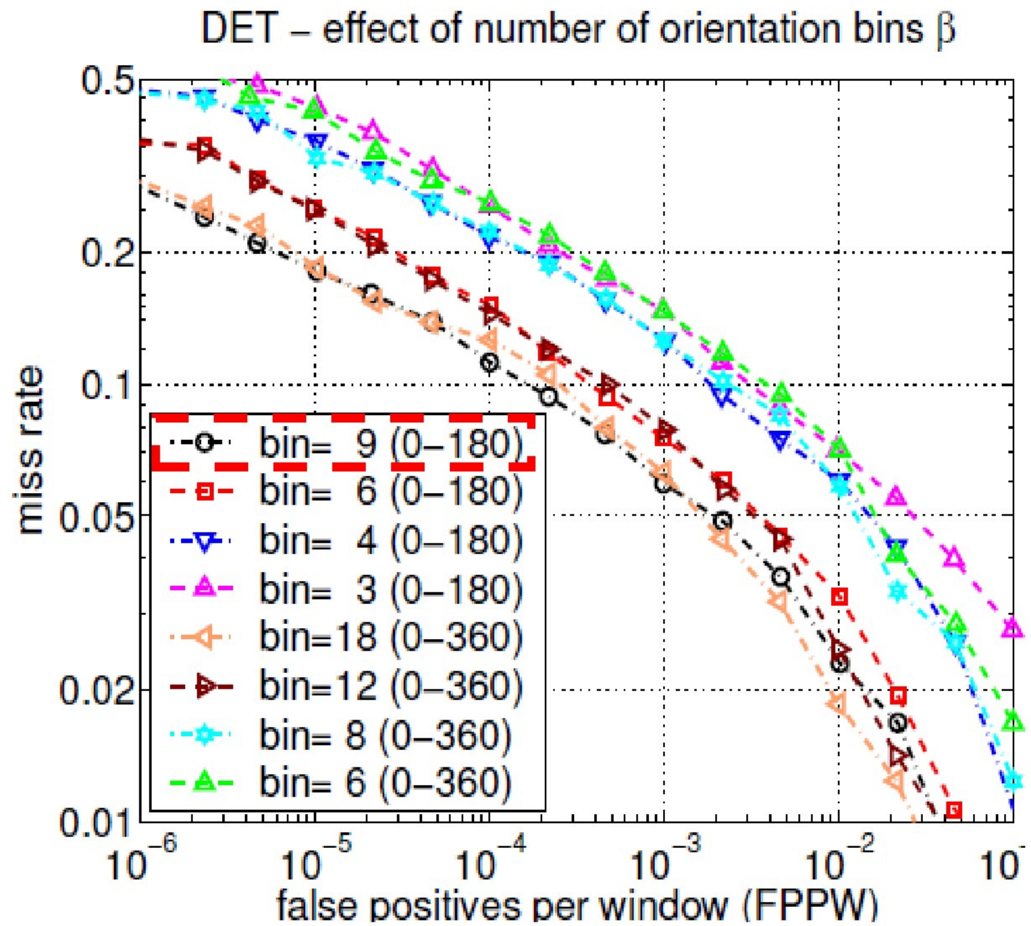
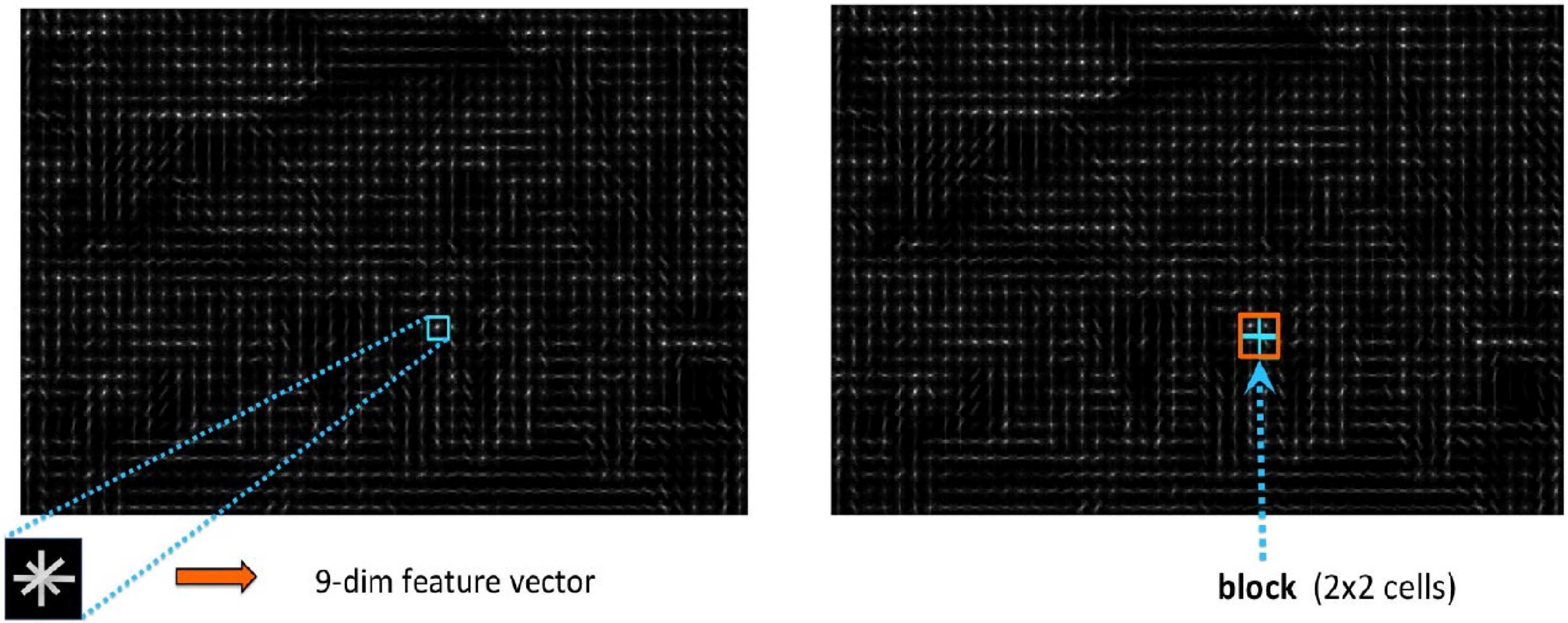


Image from Sanja Fidler

Histograms of Oriented Gradients (HOG)

2) Weighted vote into spatial & orientation cells



Note: all the orientations that are present in the cell are plotted.

Image from Sanja Fidler

Histograms of Oriented Gradients (HOG)

3) Contrast normalize over overlapping spatial blocks

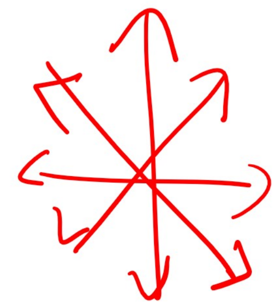
a) L_2 block normalization: $\boldsymbol{v} \to \boldsymbol{v}/\sqrt{\|\boldsymbol{v}\|_2^2 + \varepsilon^2}$
b) Final descriptor for each cell
c) Normalization per window

Since each cell is in 4 blocks, we have 4 different normalizations, and we make each one into separate features.

Dalal, Navneet, et al. "Histograms of oriented gradients for human detection." CVPR 2005.

Histograms of Oriented Gradients (HOG)

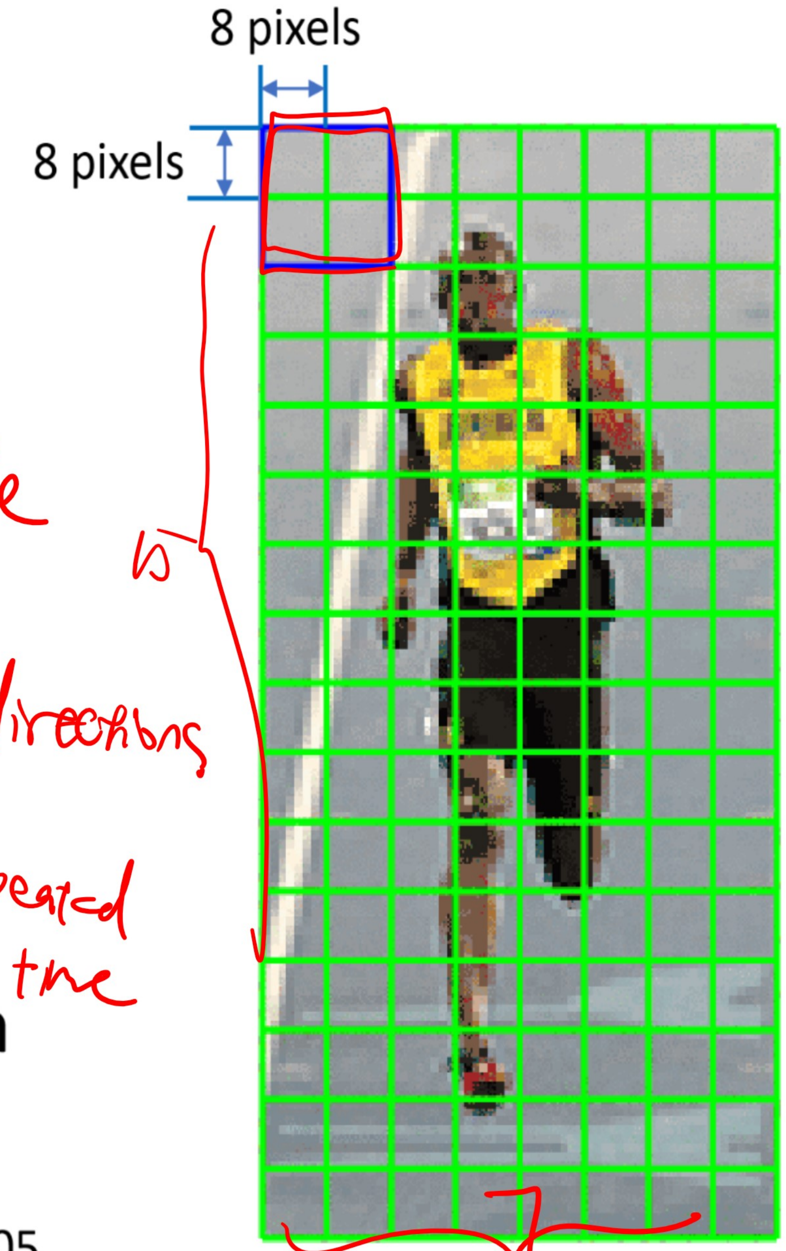
e.g. image patch = 64×128 pixels



- each cell - 16×16 pixels
- each block – 2×2 cells
  - 9 dim/cell * 4 cells = 36 dim/block
- Step size - 8×8 pixels
  - 64/8×128/8 = 128 grids
  - 7 horizontal block, 15 vertical block
- Feature for this patch: $\underbrace{9}_{\text{9 bin direction}} \times 4 \times 7 \times 15 = 3780$ dim

add magnitude for gradient with similar directions

each cells repeated 4 time



Dalal, Navneet, et al. "Histograms of oriented gradients for human detection." CVPR 2005.

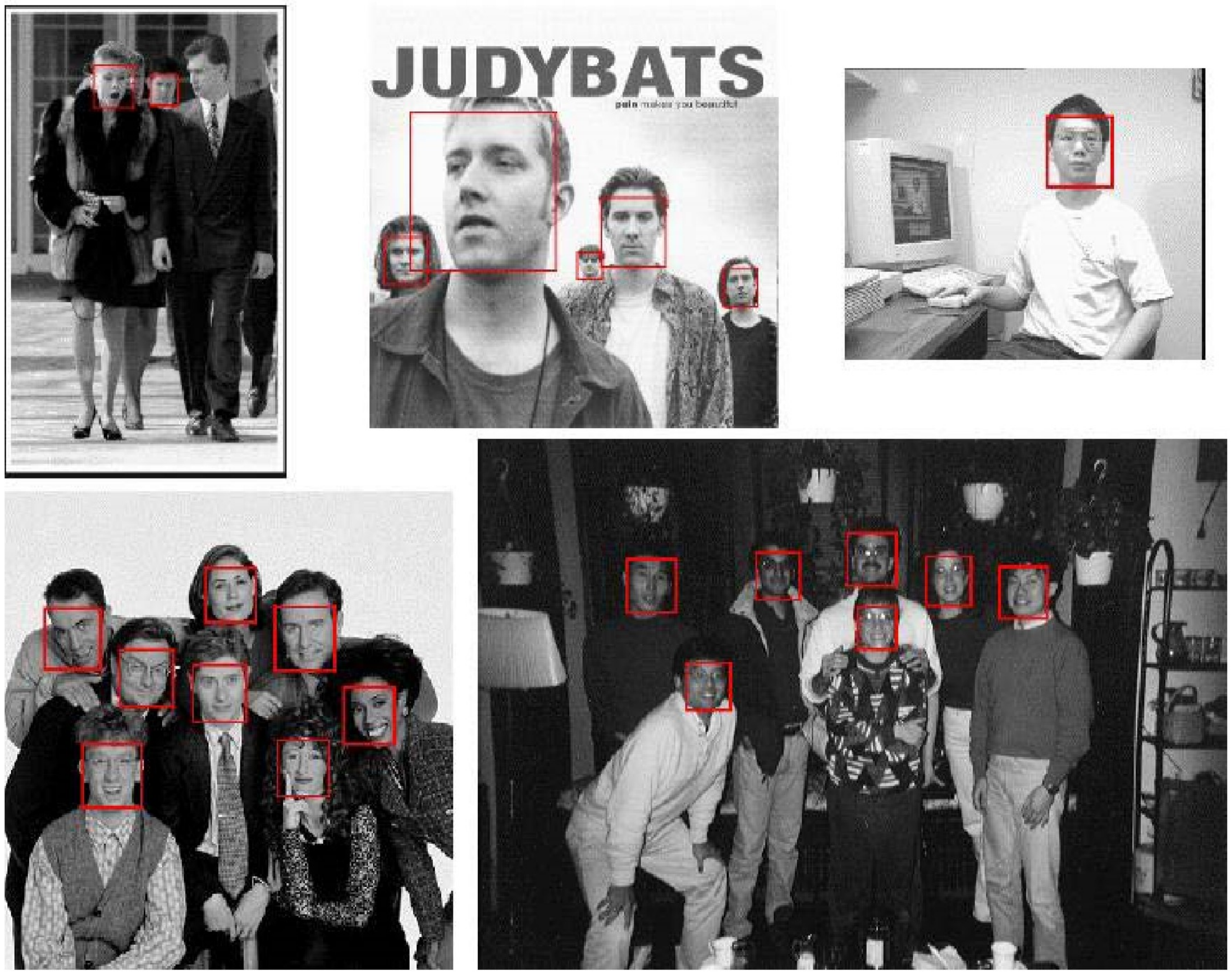
$x_i \rightarrow$

3780

Outline

- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications

Face Detection



Robust:
- High true-positive(tp) rate
- Low false-positive(fp) rate

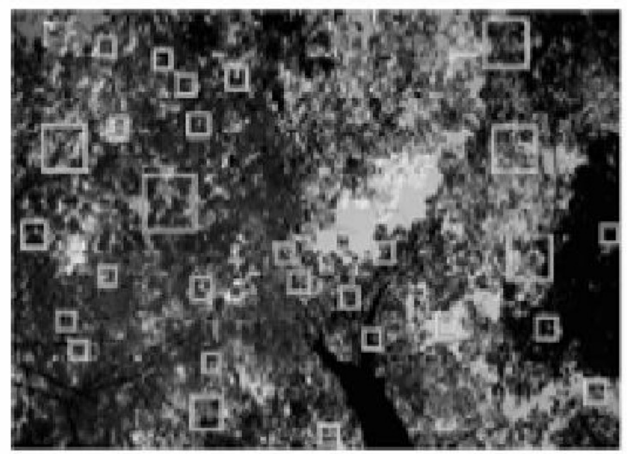
Real-time:
- At least 2 frames per sec

Detection:
- Faces v.s. non-faces

*tp: groundtruth – pos, prediction – pos
*fp: groundtruth – neg, prediction - pos

Viola, Paul, et al. "Rapid object detection using a boosted cascade of simple features." CVPR 2001.

How to Represent a Face?



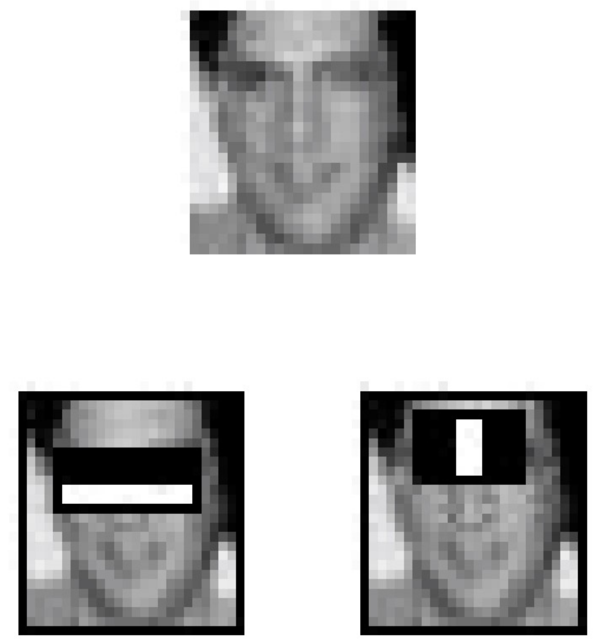
Slides from Paul Viola & Michael Jones

Feature Extraction

- Can a simple feature (i.e. a value) indicate the existence of a face?

- All faces share some similarities.
  - The eyes region is darker than the upper-cheeks.
  - The nose bridge region is brighter than the eyes.

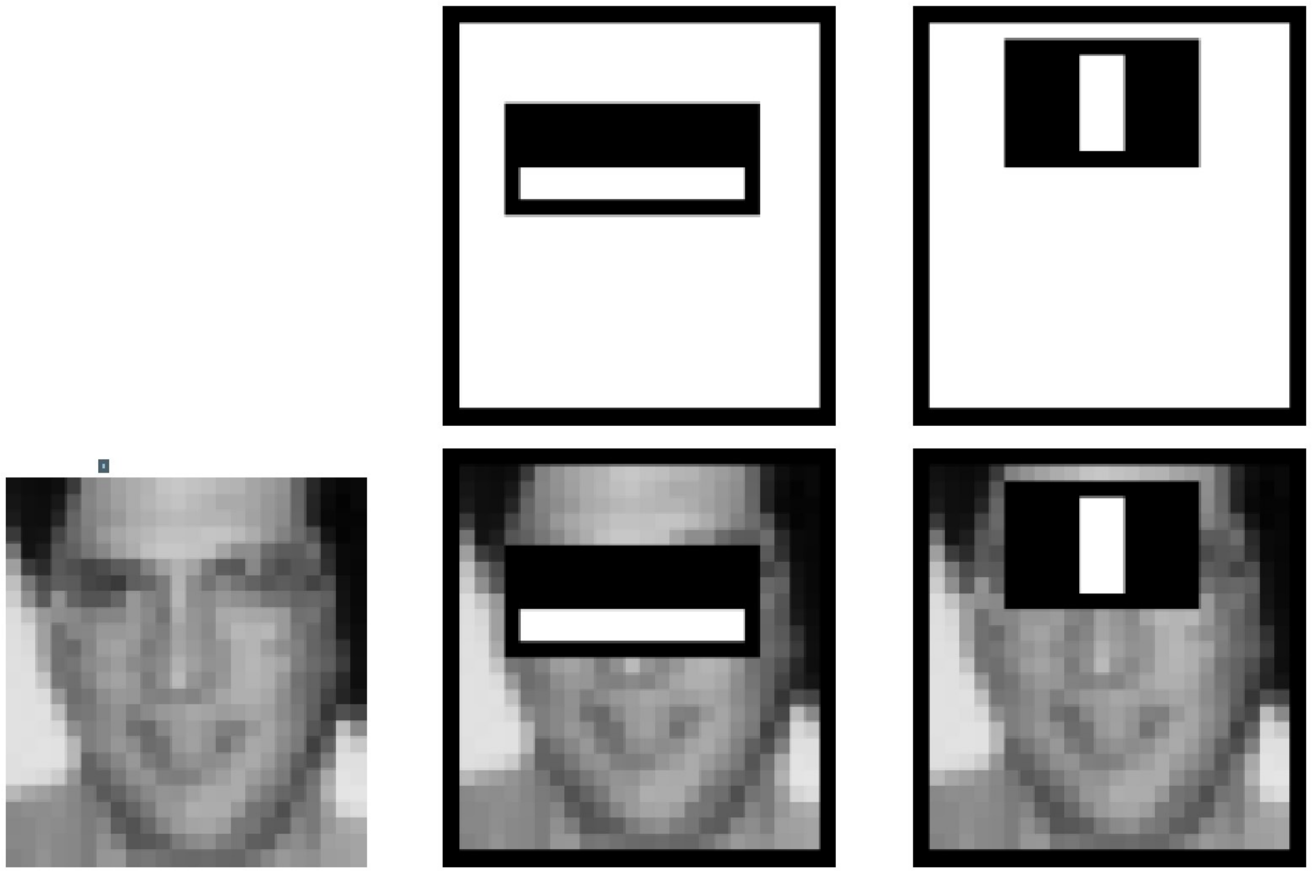
- Encode domain knowledge
  - Location - Size: eyes & nose bridge region
  - Value: darker / brighter

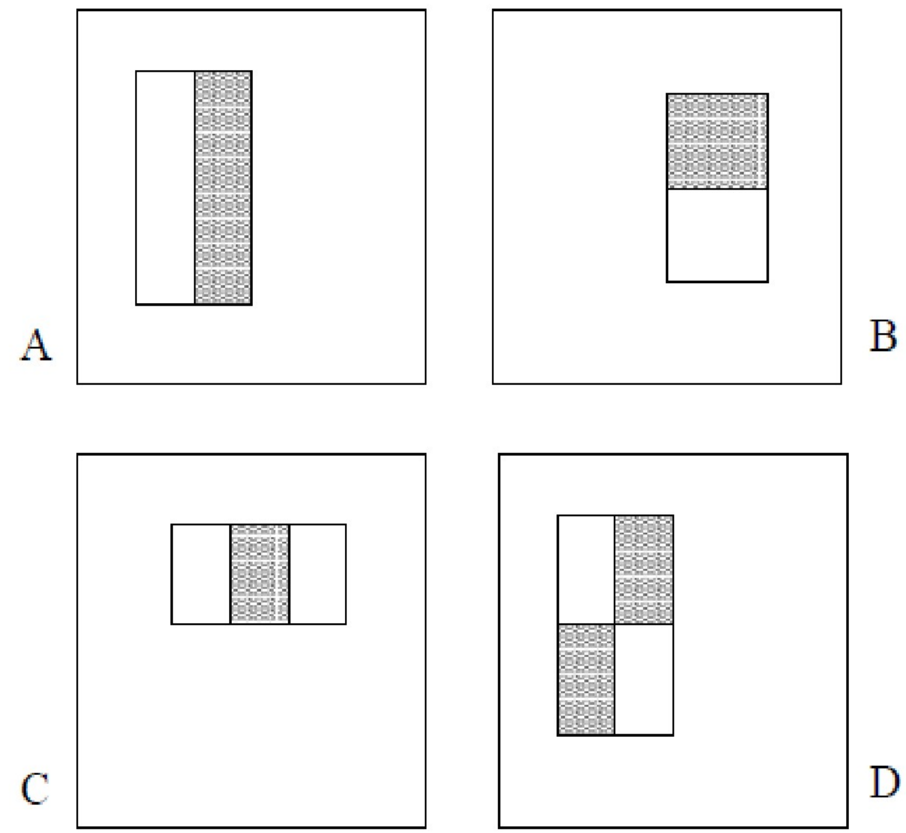


Slides from Kostantina Palla & Alfredo Kalaitzis

Feature Extraction

- Rectangle Features
  - value = ∑ (pixels in black area)- ∑ (pixels in white area)





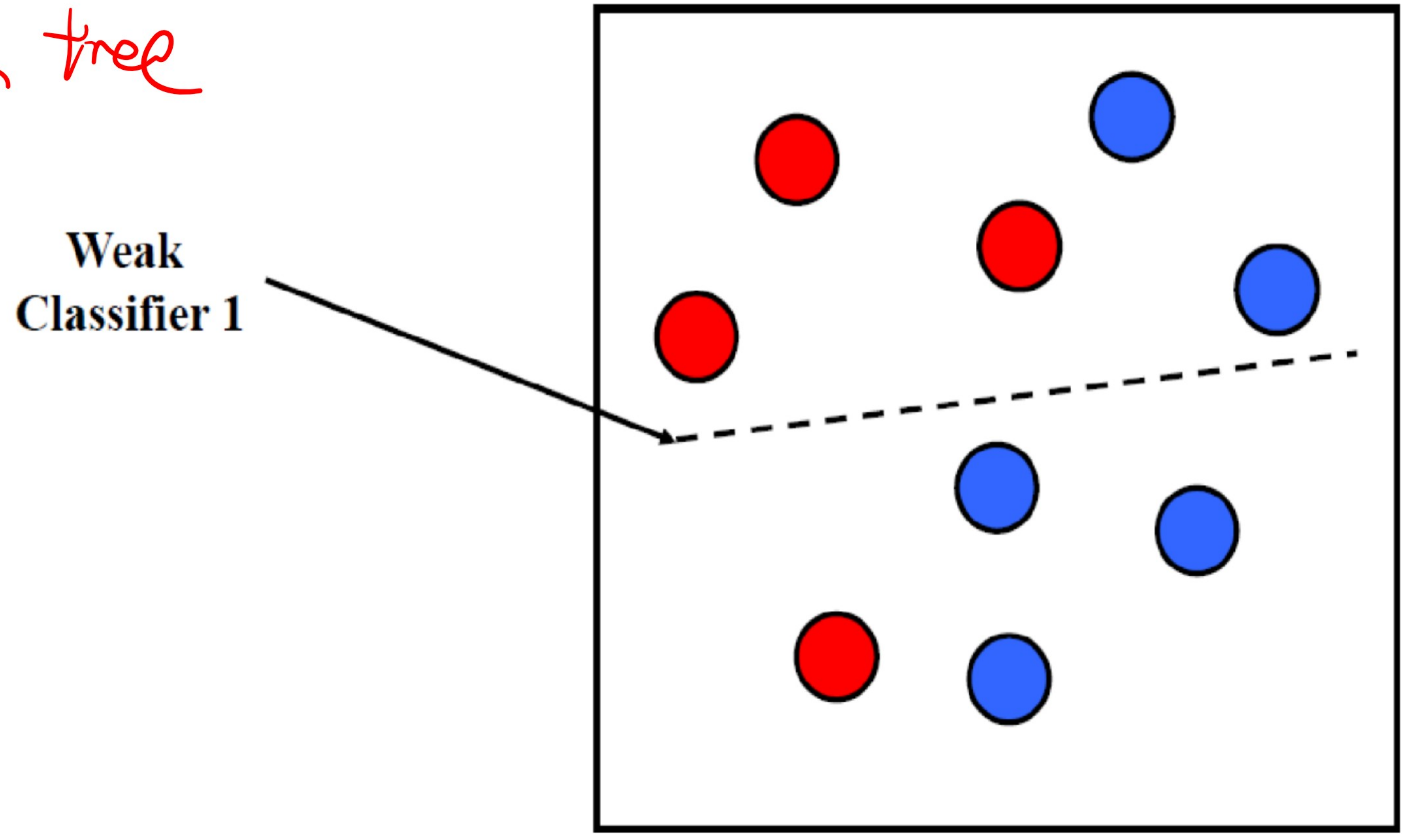
Slides from Kostantina Palla & Alfredo Kalaitzis

Huge "Library" of Filters

Slides from Kostantina Palla & Alfredo Kalaitzis

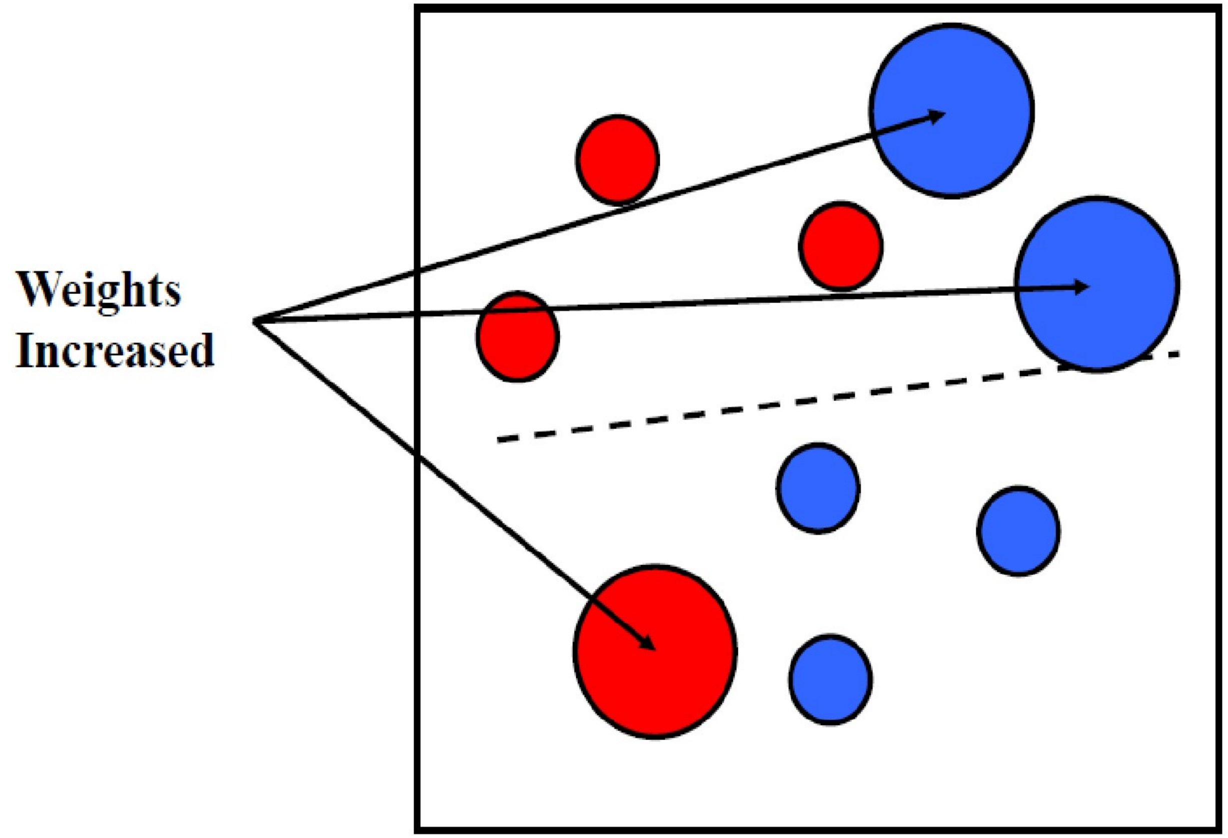
AdaBoost: Intuition

Decision tree

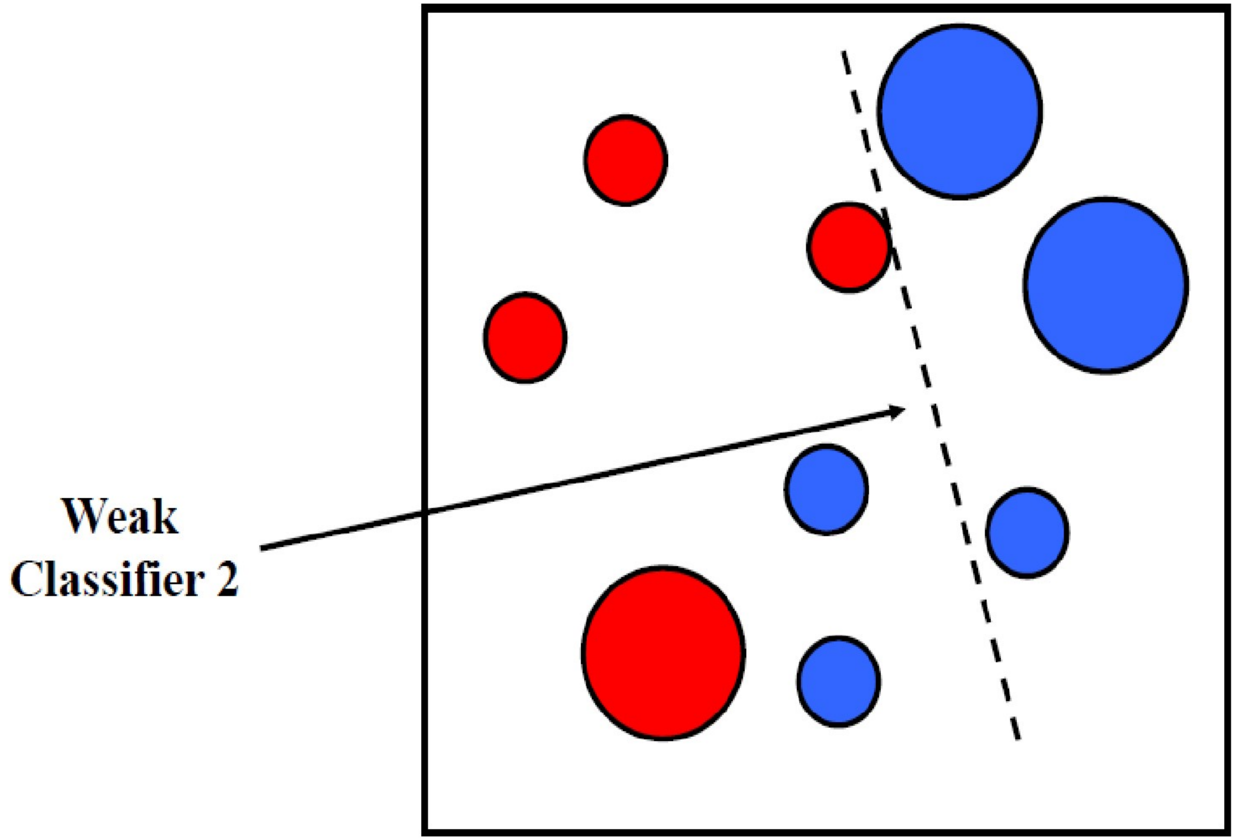


Slides from Robert Fergus

AdaBoost: Intuition

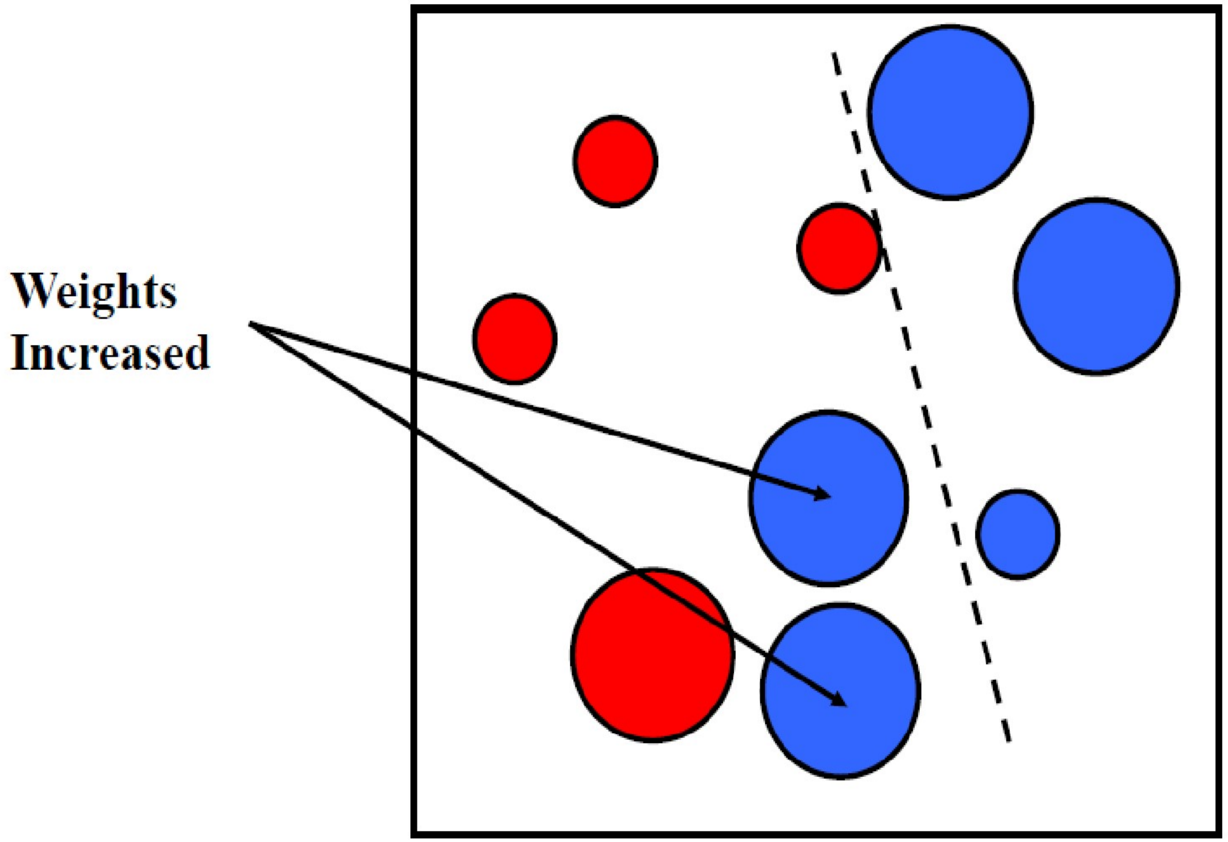


AdaBoost: Intuition



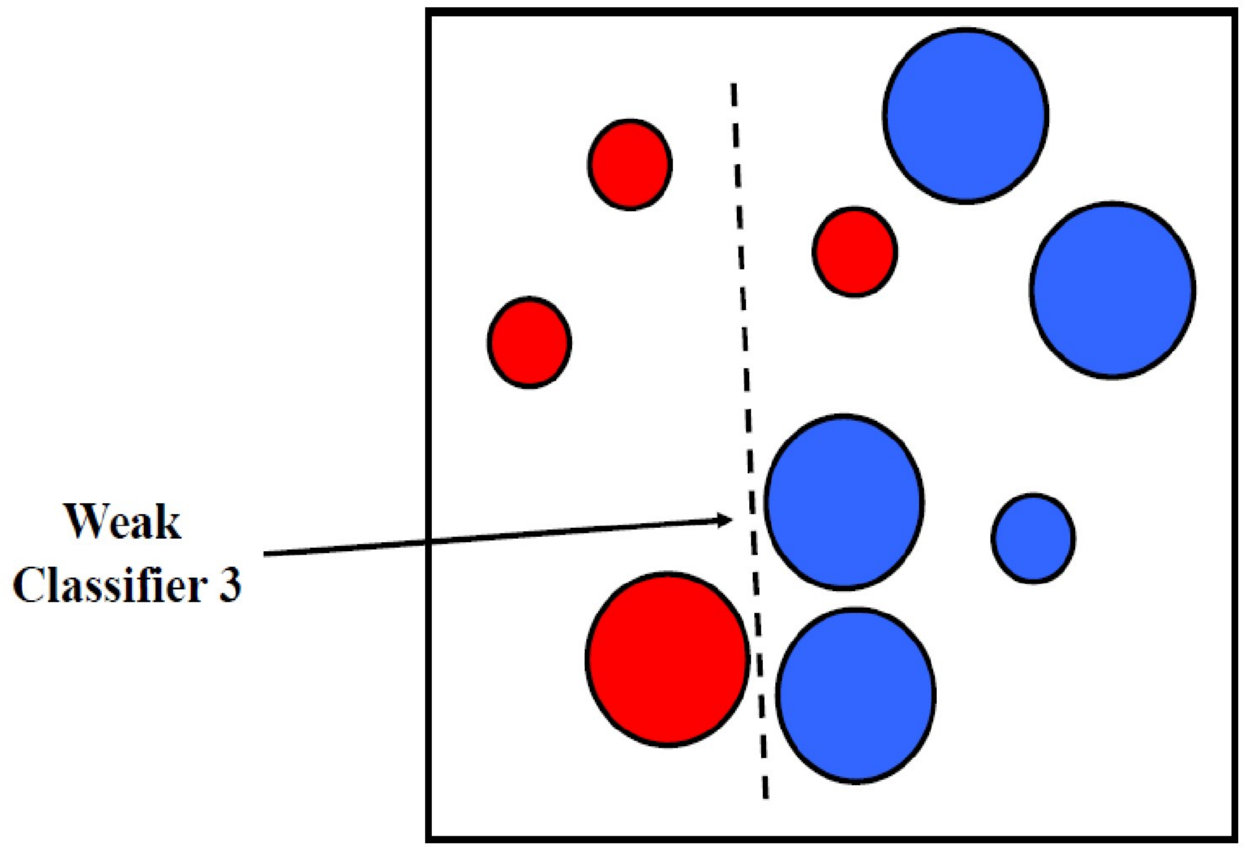
Slides from Robert Fergus

AdaBoost: Intuition



Slides from Robert Fergus

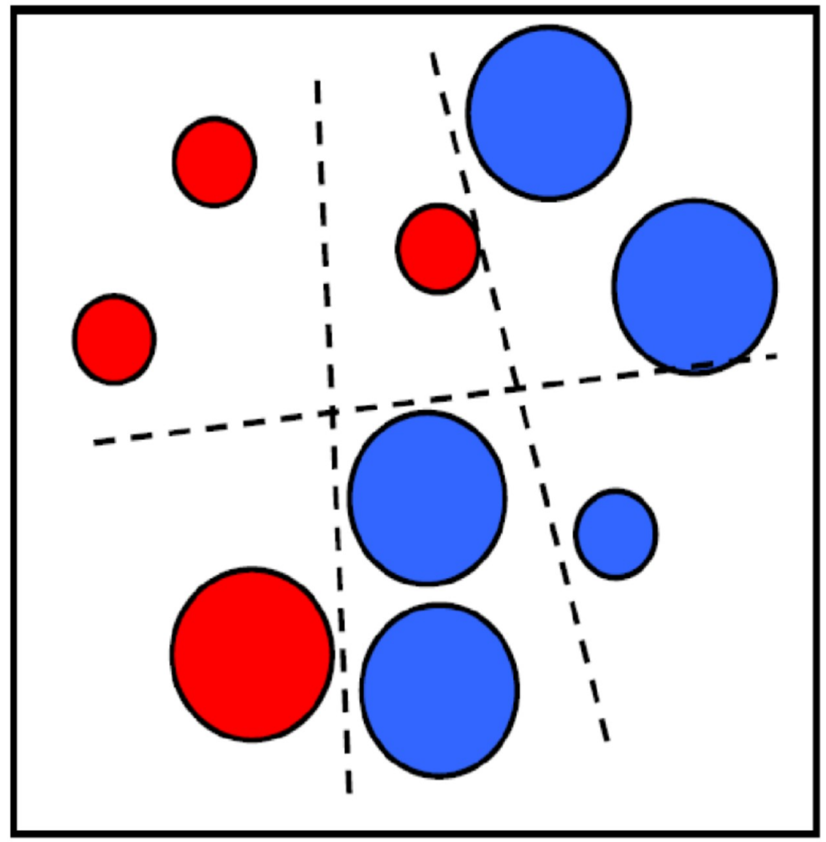
AdaBoost: Intuition



Slides from Robert Fergus

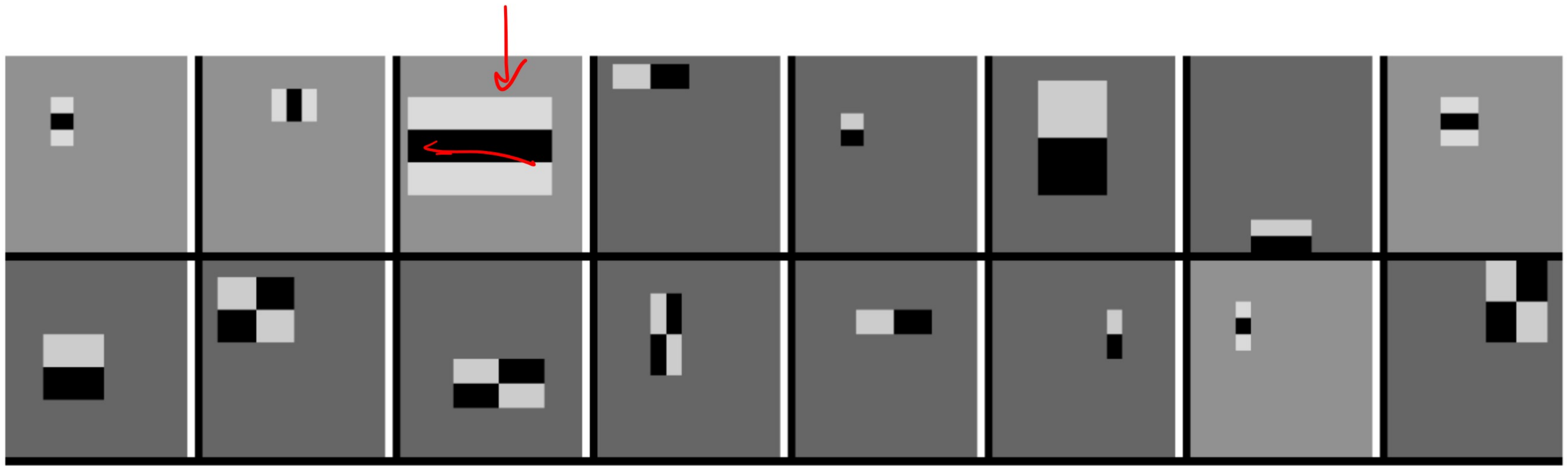
AdaBoost: Intuition

Final classifier is linear combination of weak classifiers



Slides from Robert Fergus

Learned Features

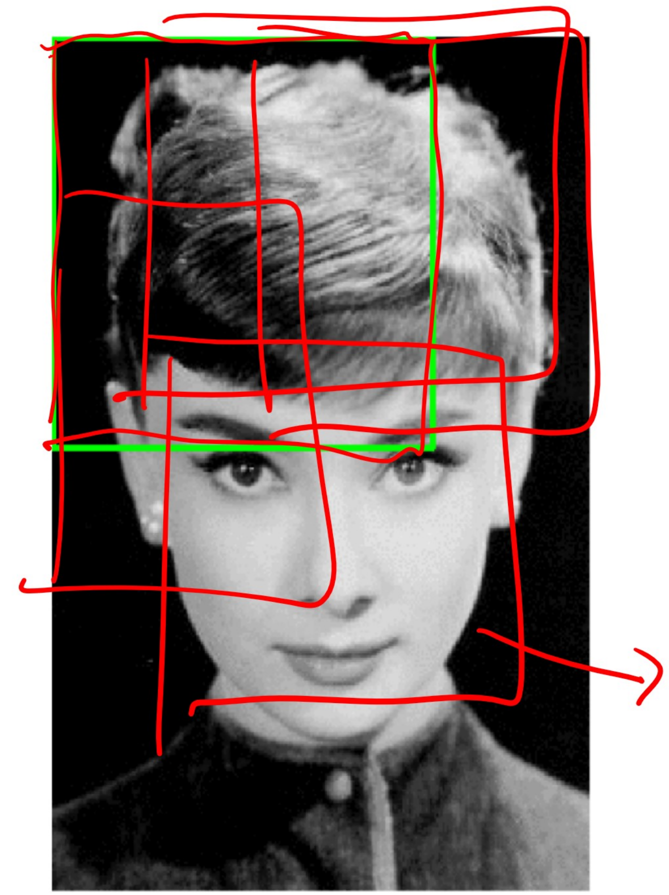


Slides from David Lowe

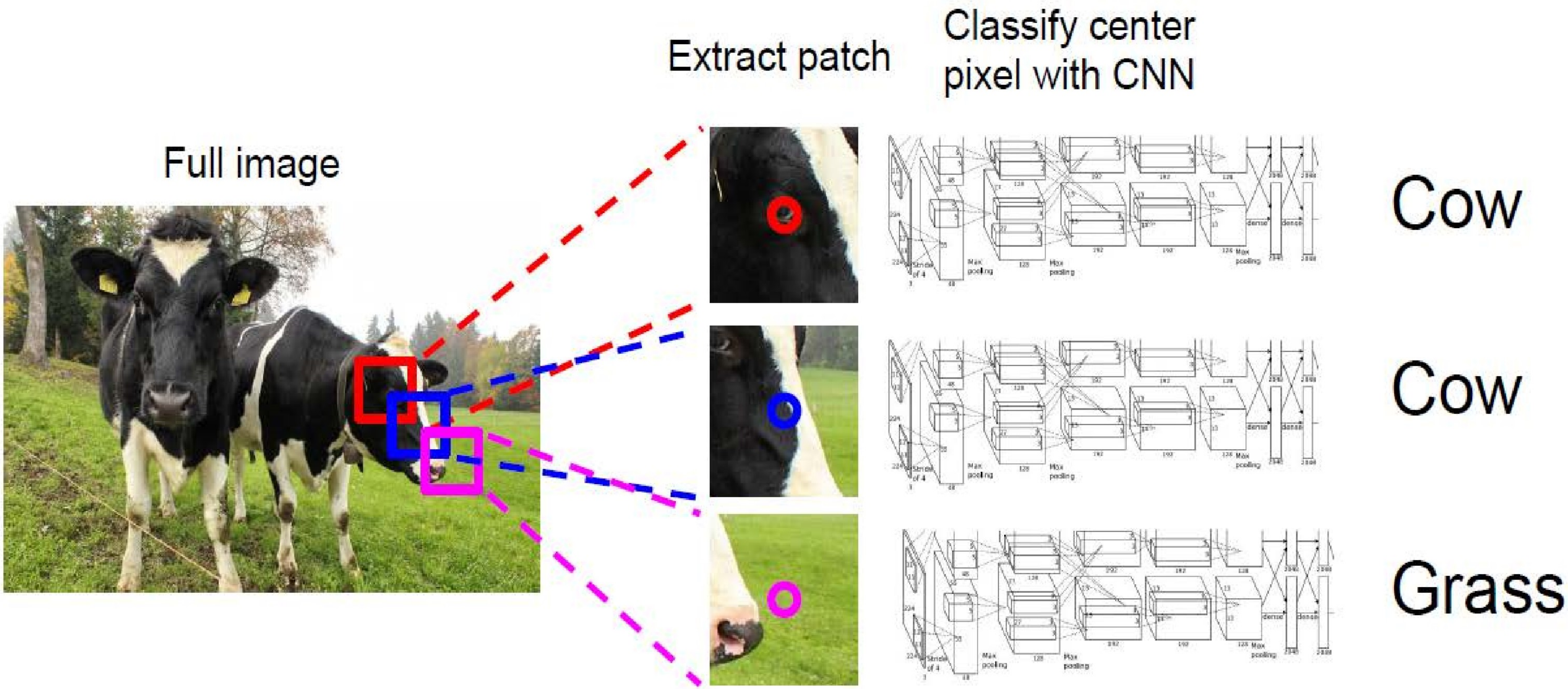
Sliding Window

Create fixed features

- Sliding window
  - A rectangular region
  - Fixed width and height
  - "Slides" across an image
  - Overlap v.s. non-overlap
- For each window
  - Apply binary classification: face v.s. non-face
- Goal: localization



Sliding Window: Semantic Segmentation

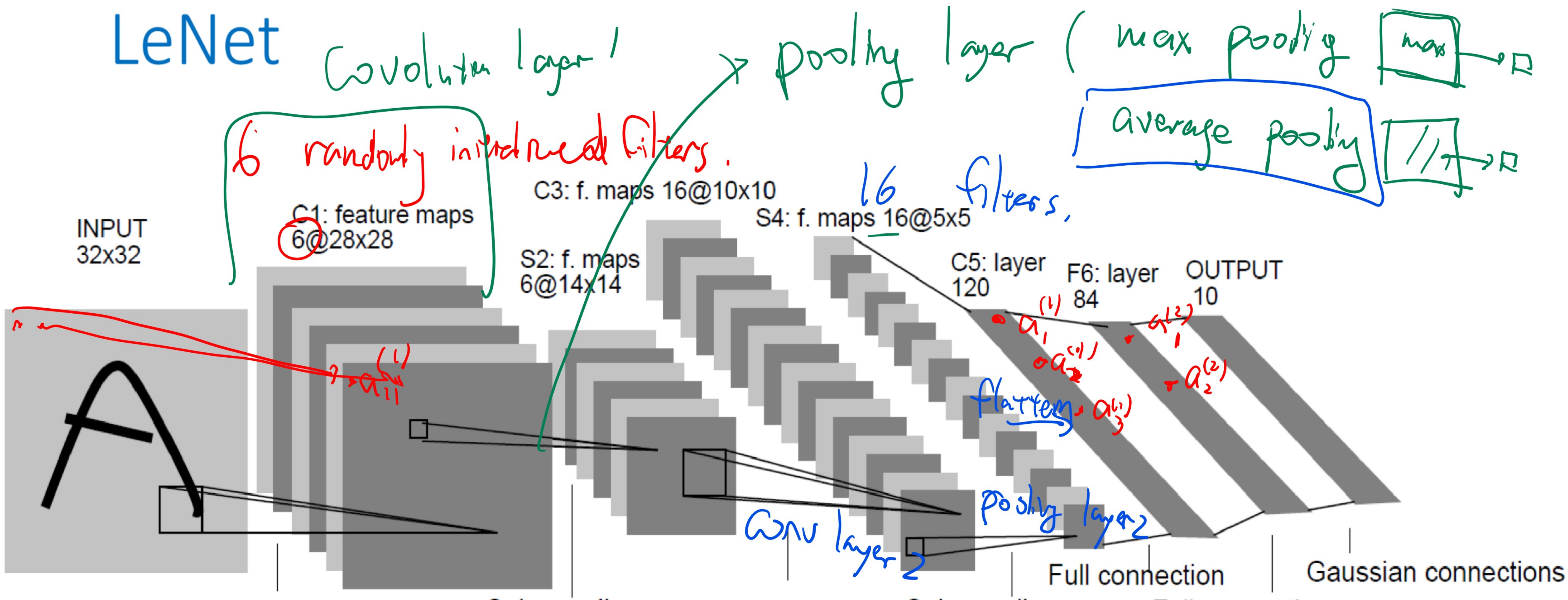


Farabet, Clement, et al., "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro, Pedro HO, et al, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Outline

- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications

LeNet



5×5 filter

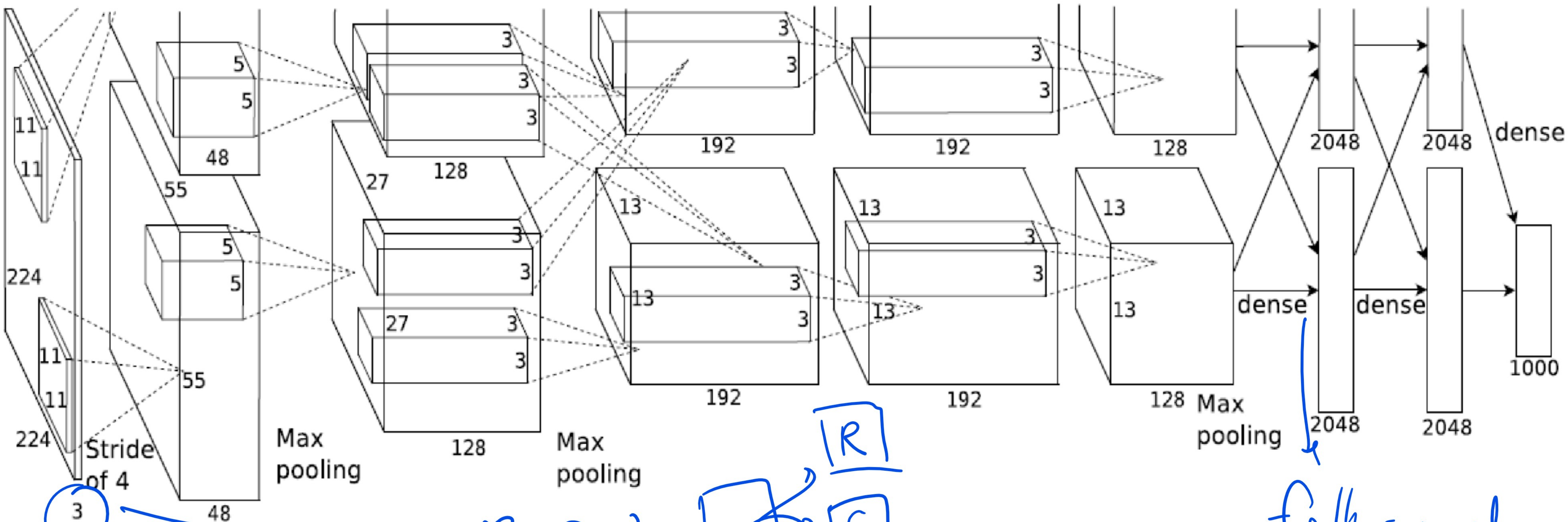
[w11 w12 ...] unknown, randomly initialized, training

*Feature map = activation map: the output activations for a given filter.

*Subsampling: local averaging, reducing the resolution of the feature map.

LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proc. Of the IEEE (1998).

AlexNet



grayscale

R, G, B

R

G

B

fully connected

Architecture:

conv1 -> max pool1 -> norm1 -> conv2 -> max pool2 -> norm2 -> conv3 -> conv4 -> conv5 -> max pool3 -> fc6 -> fc7 -> fc8

Krizhevsky, Alex, et al. "Imagenet classification with deep convolutional neural networks." NIPS 2012.

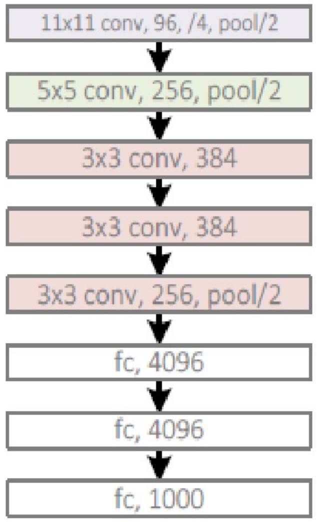
Slides from Fei-Fei Li & Justin Johnson & Serena Yeung

Revolution of Depth

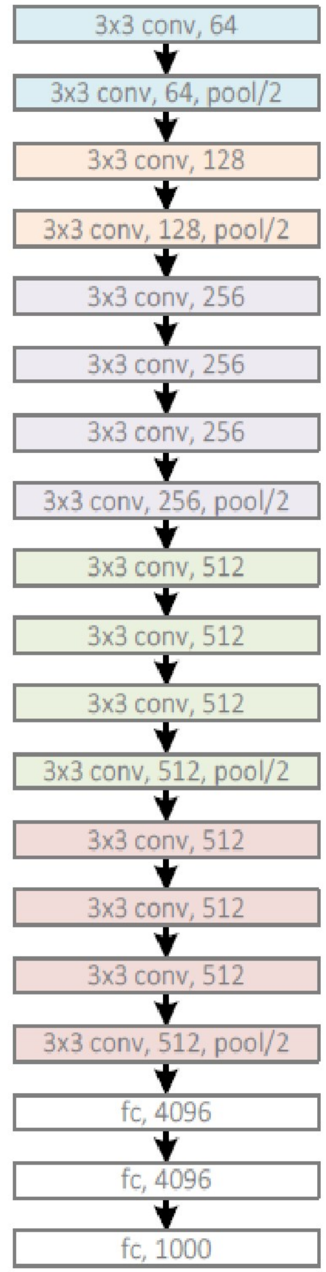
Slides from Kaiming He

Revolution of Depth

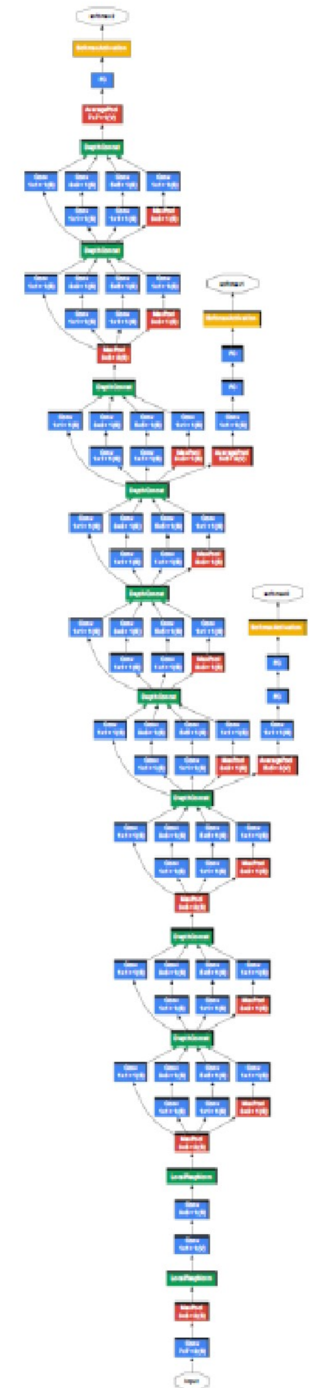
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)



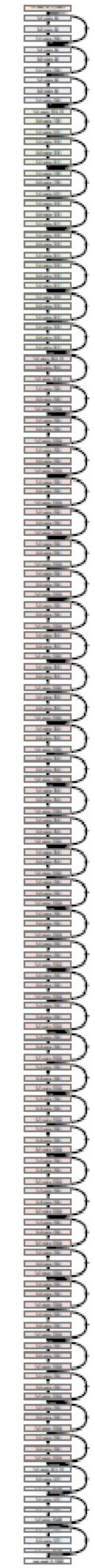
Slides from Kaiming He

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

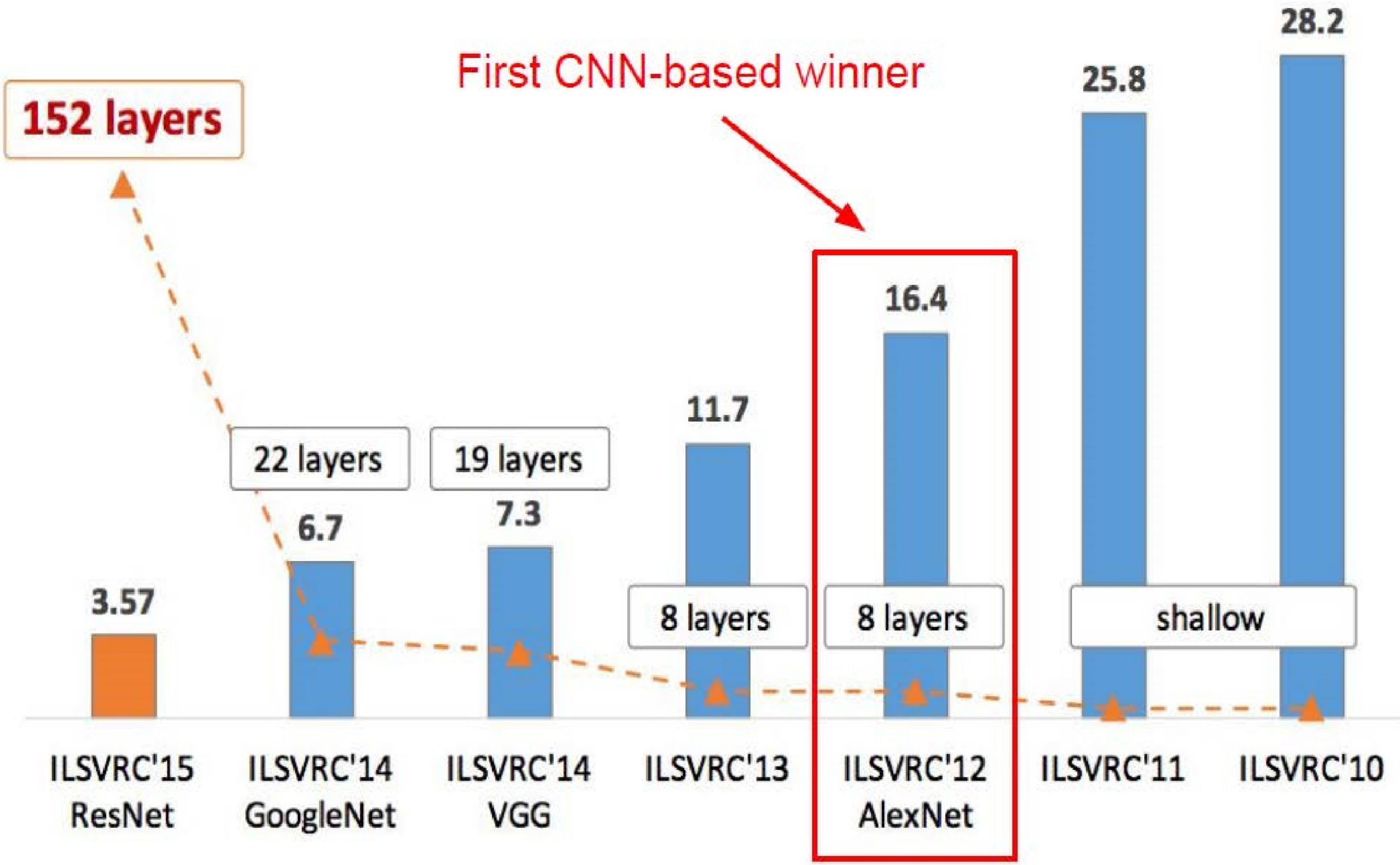
VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)



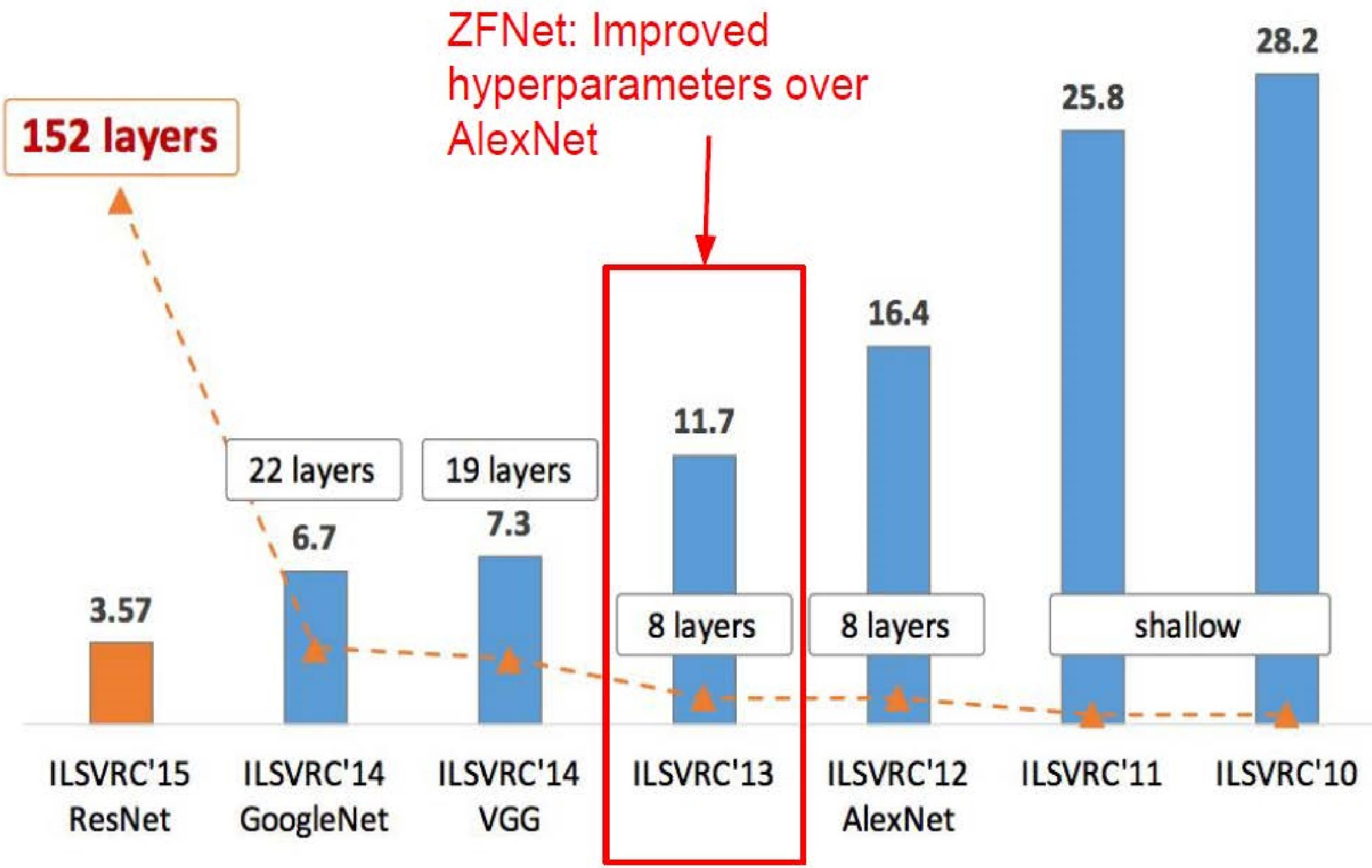
Slides from Kaiming He

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Winners



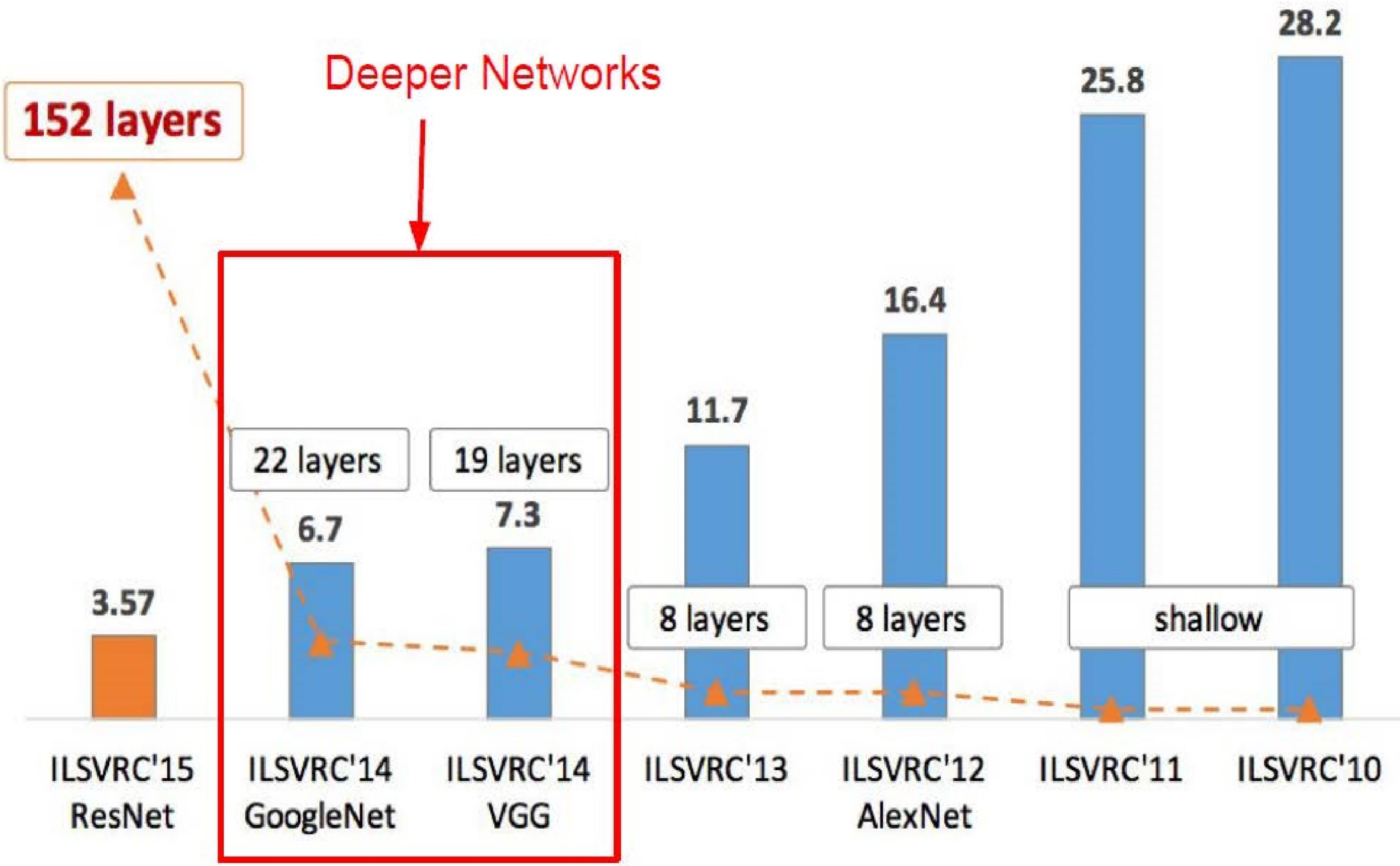
Slides from Kaiming He

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Winners



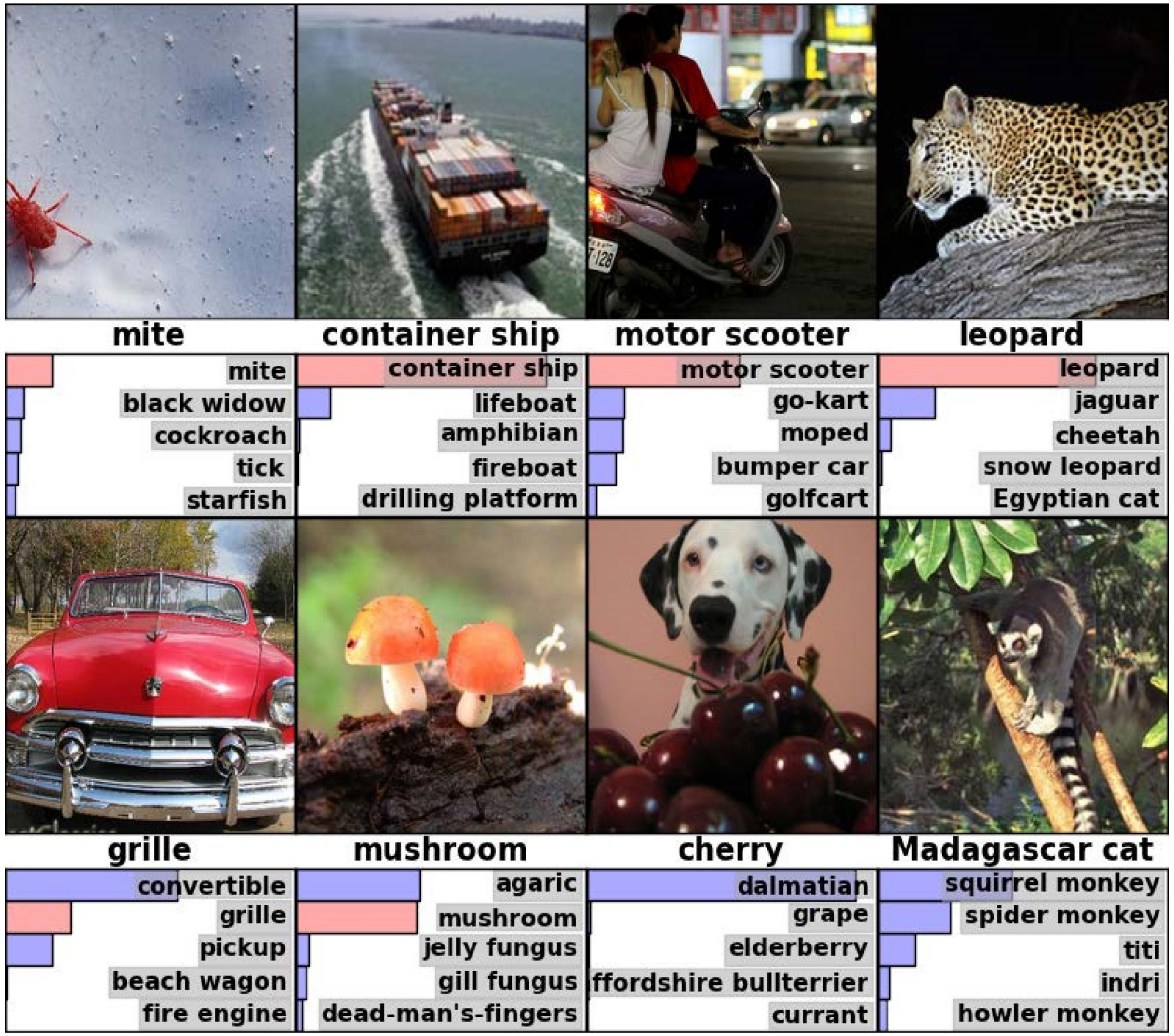
Slides from Kaiming He

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Winners



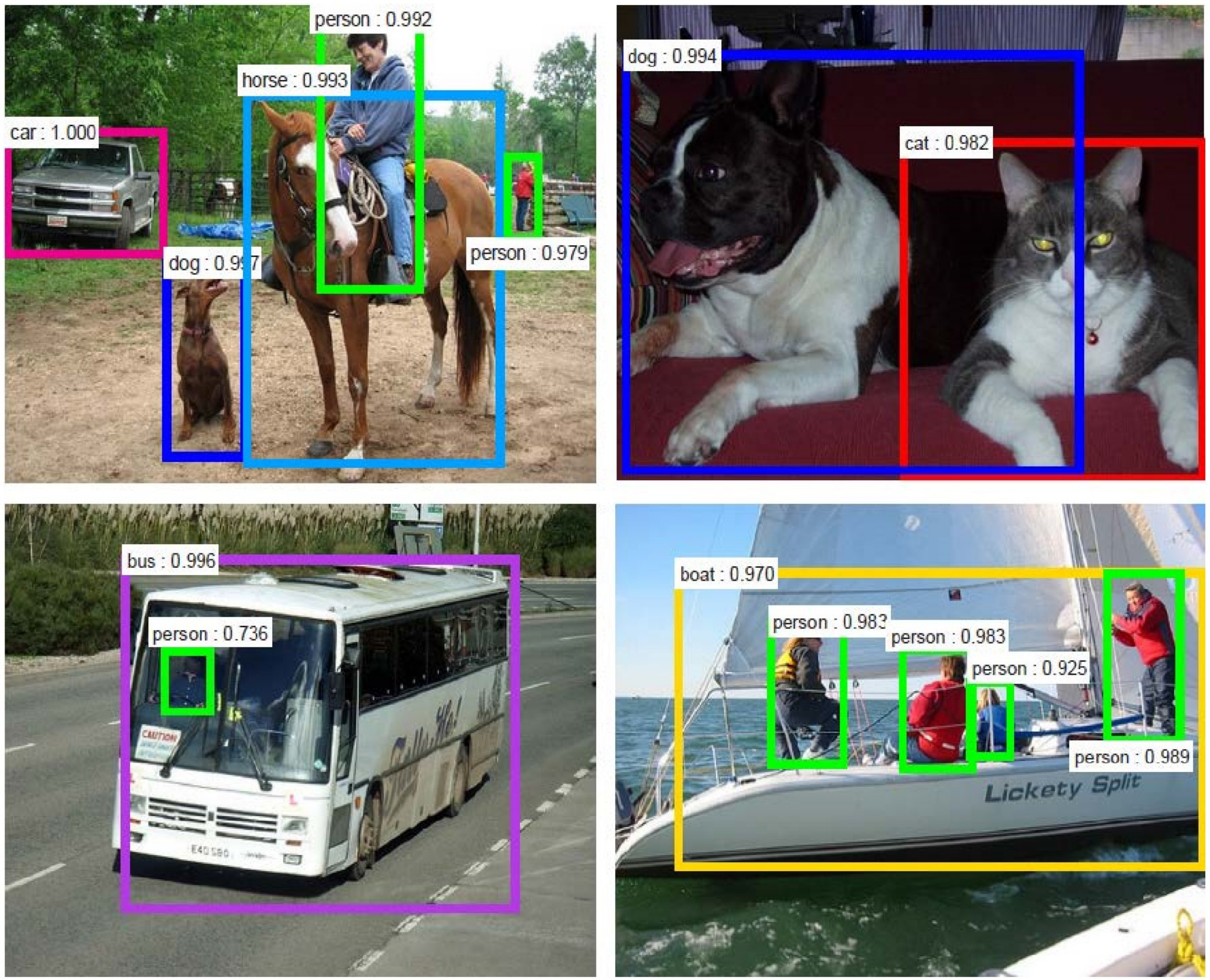
Slides from Kaiming He

Image Classification



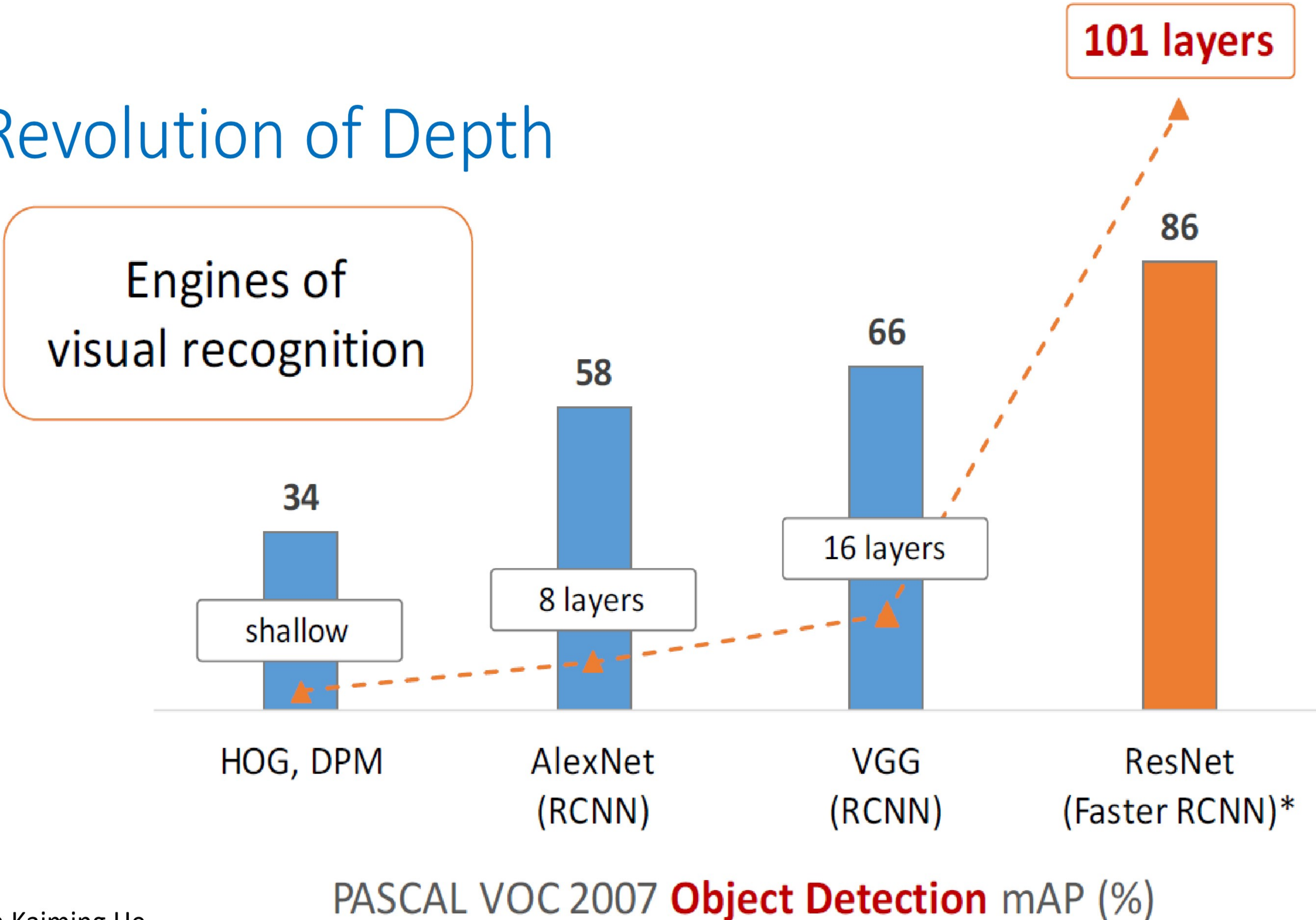
Krizhevsky, Alex, et al. "Imagenet classification with deep convolutional neural networks." NIPS 2012.

Object Detection



Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." NIPS 2015.

Revolution of Depth

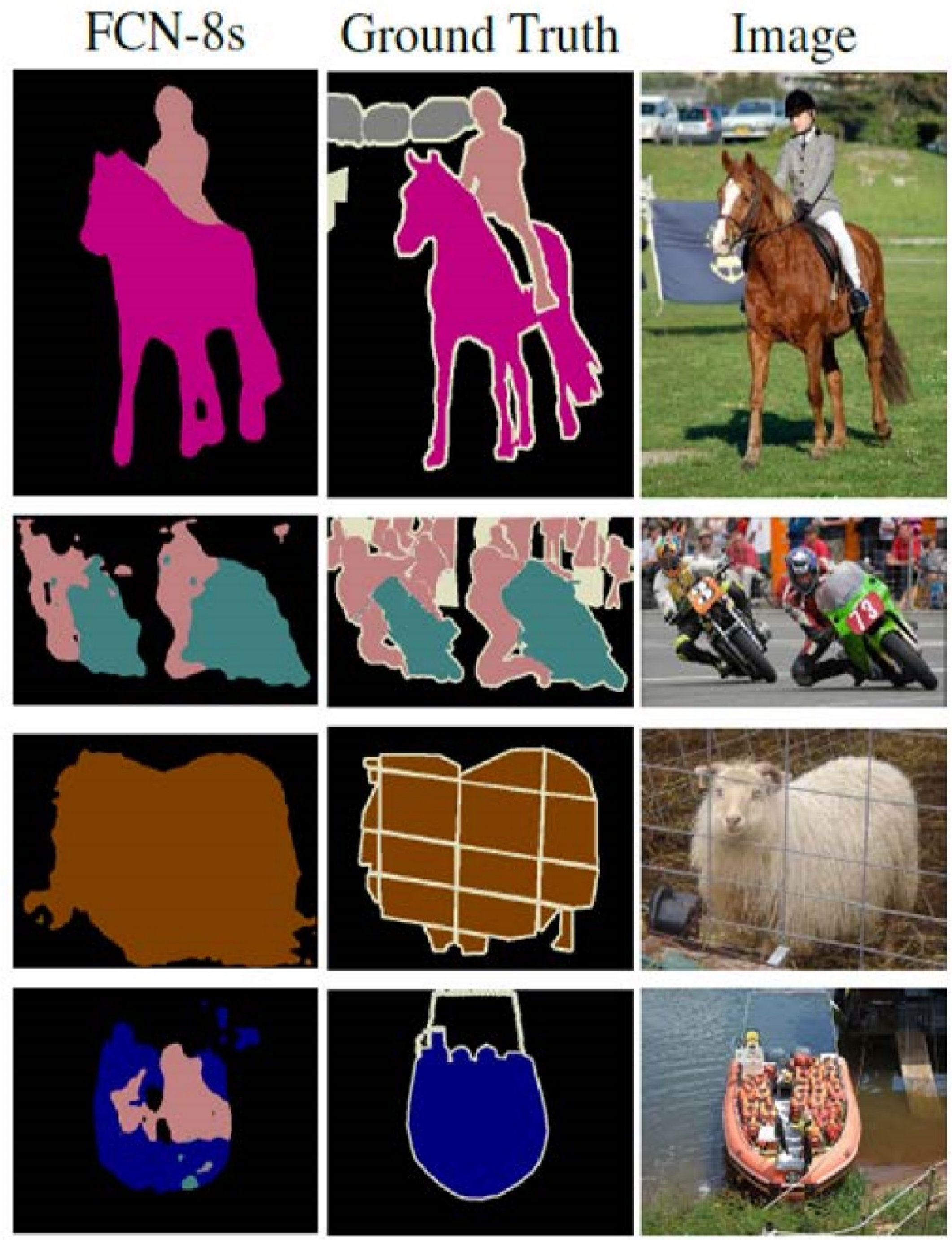


Slides from Kaiming He

Outline

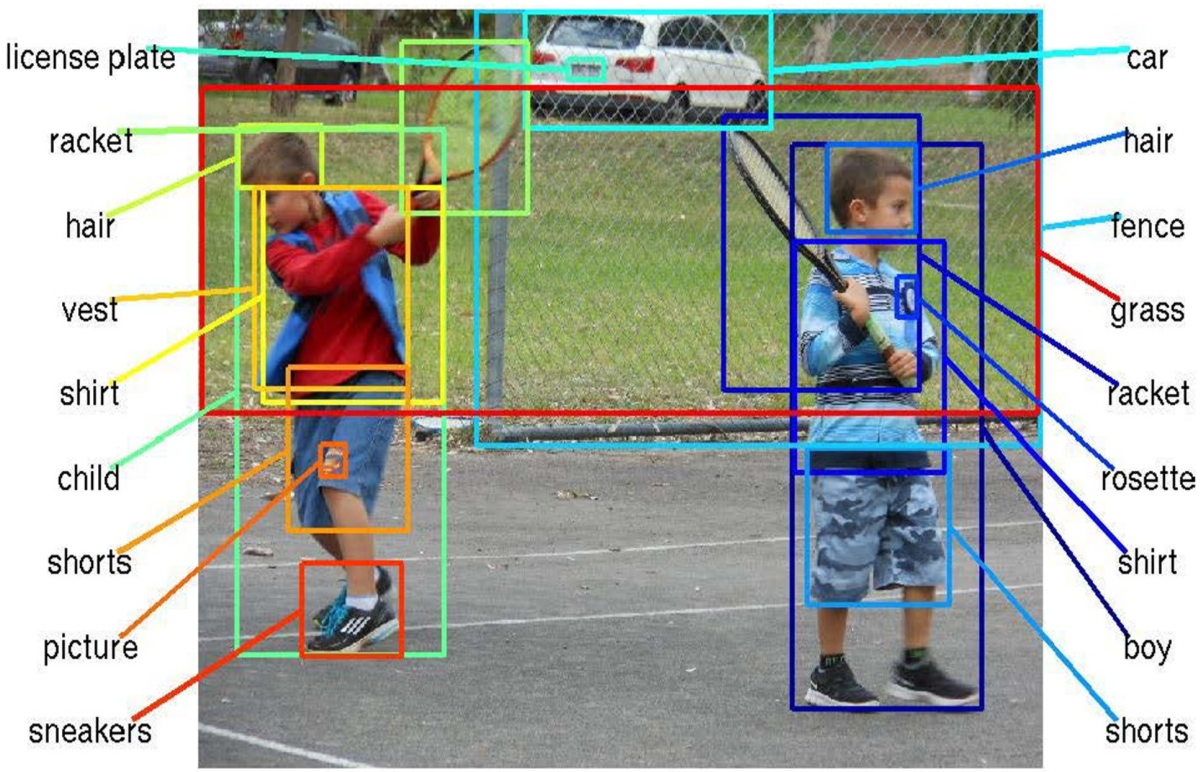
- Computer Vision Overview
- Image Representations - Features
  - SIFT
  - HOG
- Case study: Viola-Jones Face Detector
  - Haar-Like feature
  - AdaBoost
  - Sliding Window
- CNN Architectures
- Appendix: Applications

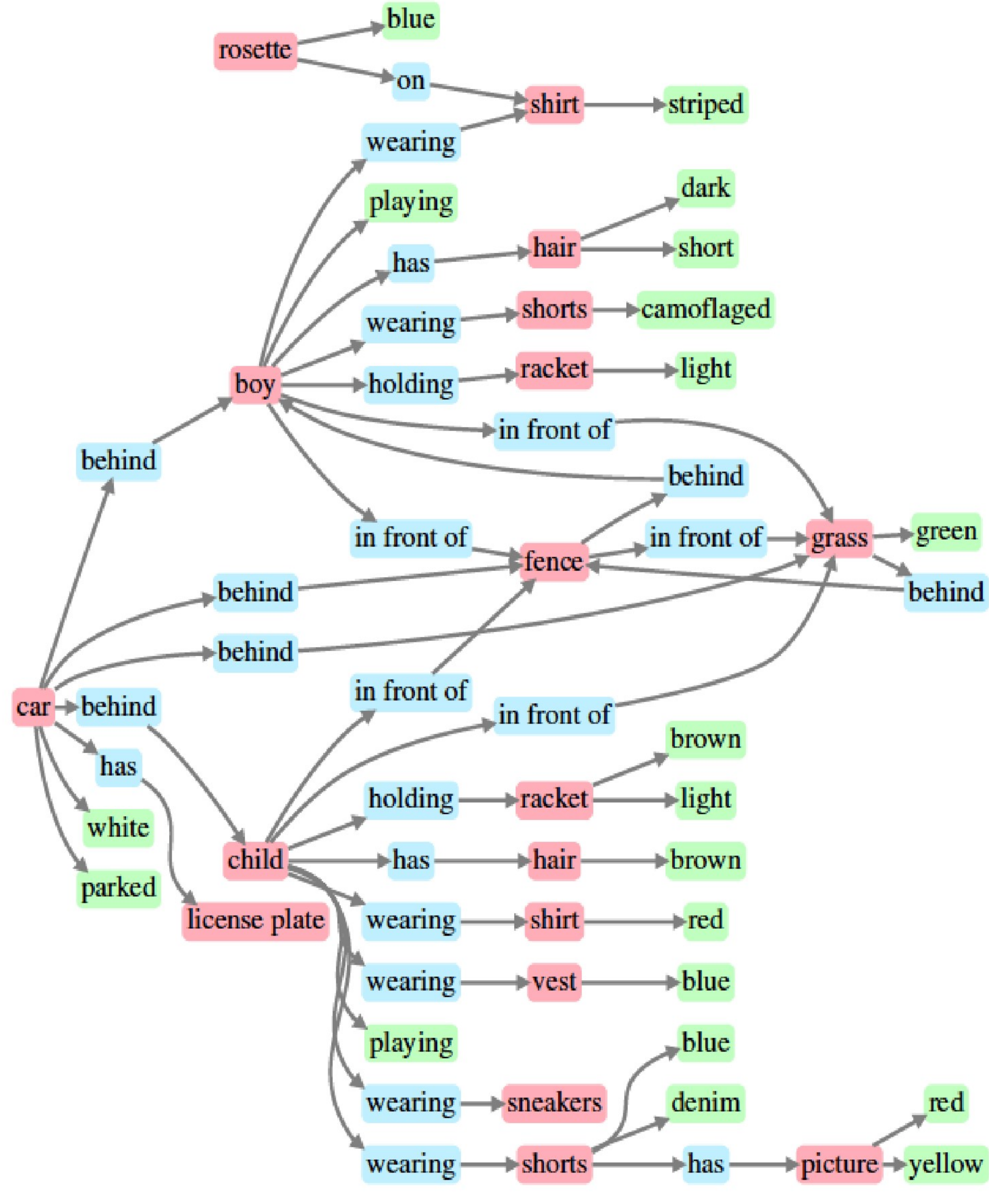
Image Segmentation



He, Kaiming, et al. "Mask r-cnn." ICCV 2017.
Long, Jonathan, et al. "Fully convolutional networks for semantic segmentation." CVPR 2015.

Image Retrieval



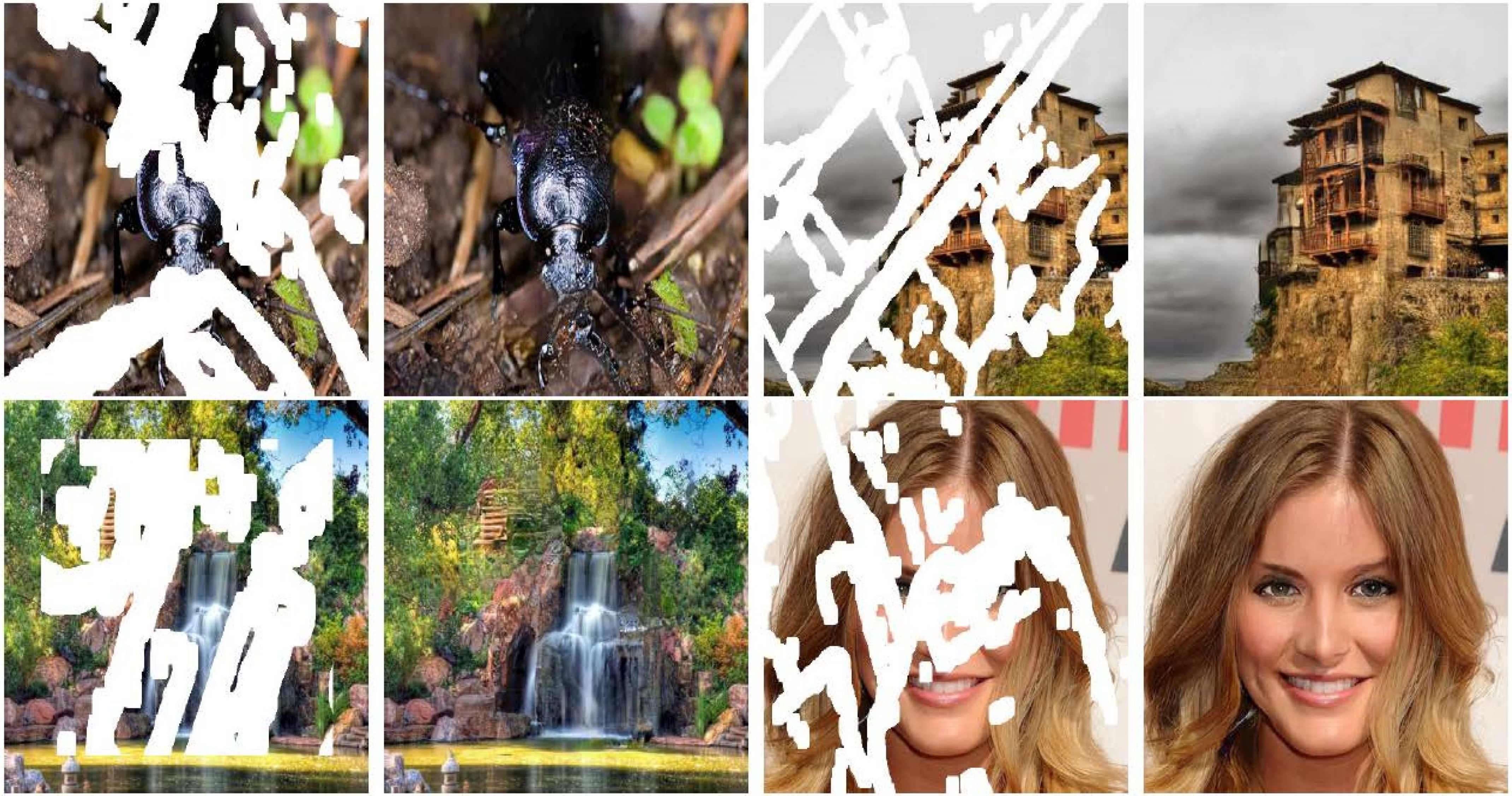


Slides from Fei-Fei Li & Justin Johnson & Serena Yeung

Image Colorization

Zhang, Richard, et al. "Colorful image colorization." ECCV 2016.

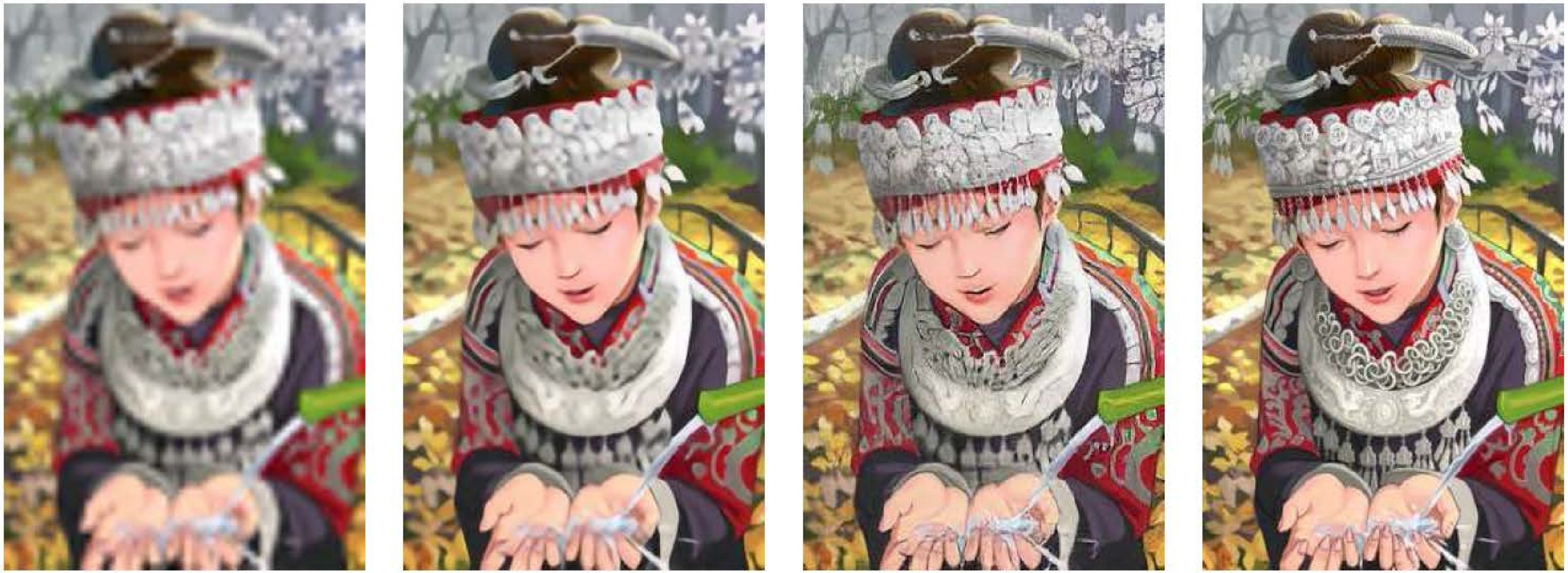
Image Reconstruction



Liu, Guilin, et al. "Image inpainting for irregular holes using partial convolutions.“ ECCV 2018.

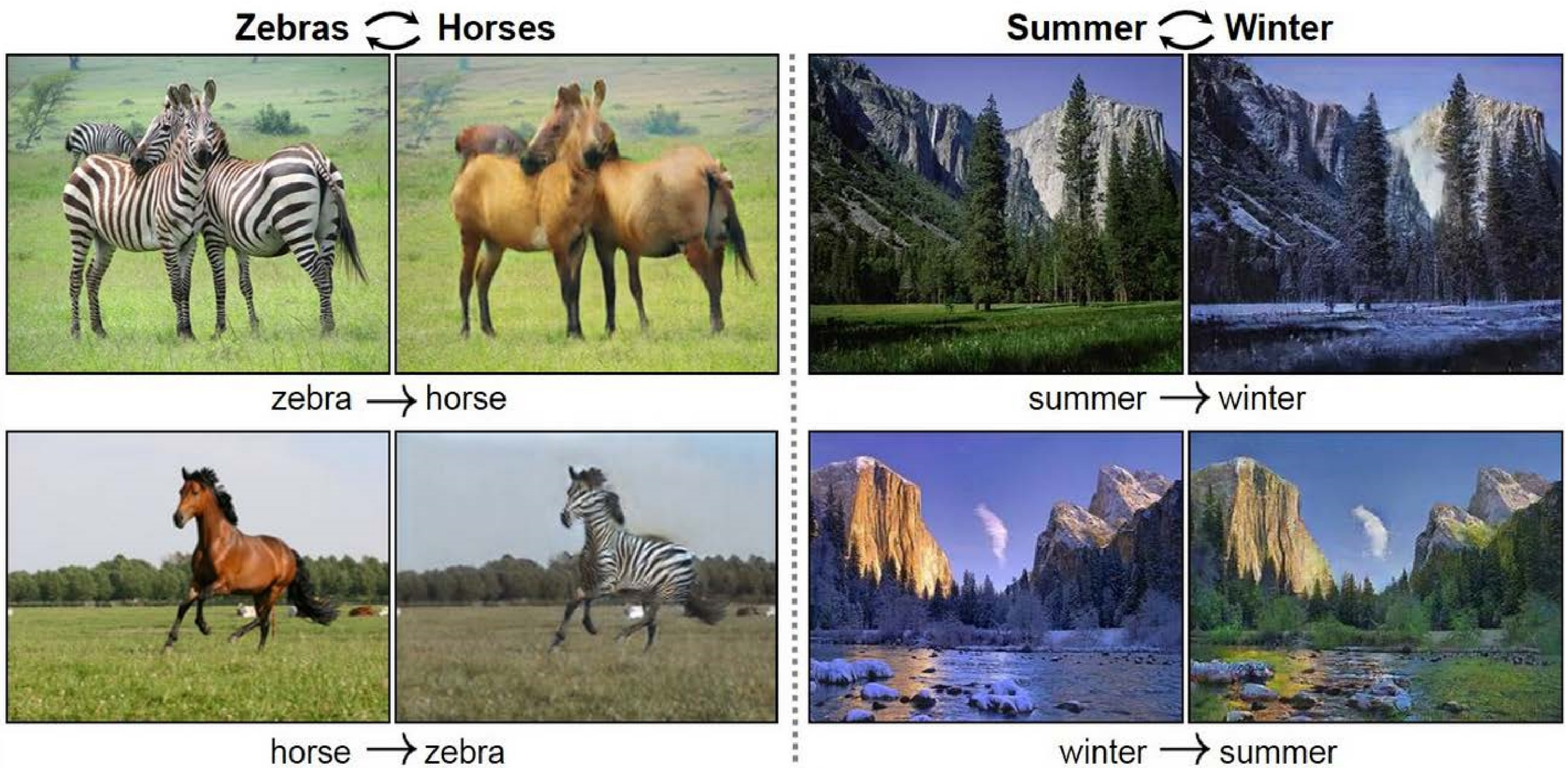
Image Super-Resolution

bicubic (21.59dB/0.6423) | SRResNet (23.53dB/0.7832) | SRGAN (21.15dB/0.6868) | original



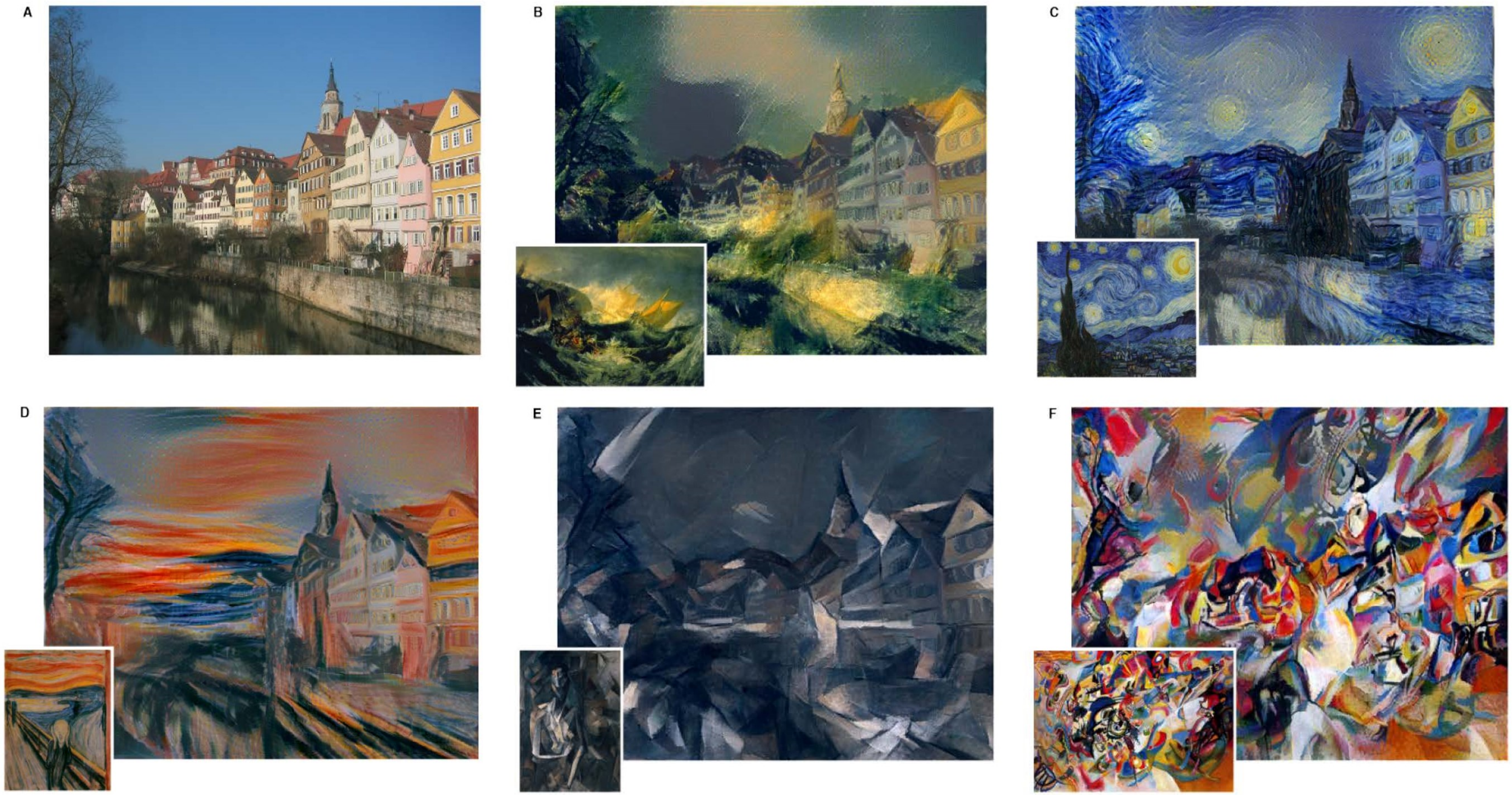
Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." CVPR 2017.

Image Synthesis



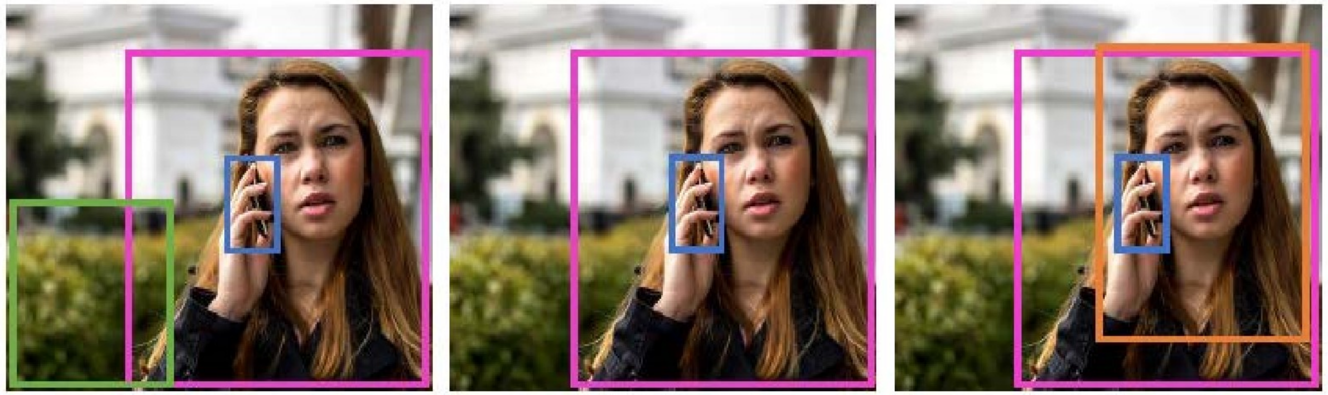
Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." ICCV 2017.

Style Transfer



Gatys, Leon A., et al. "Image style transfer using convolutional neural networks." CVPR 2016.

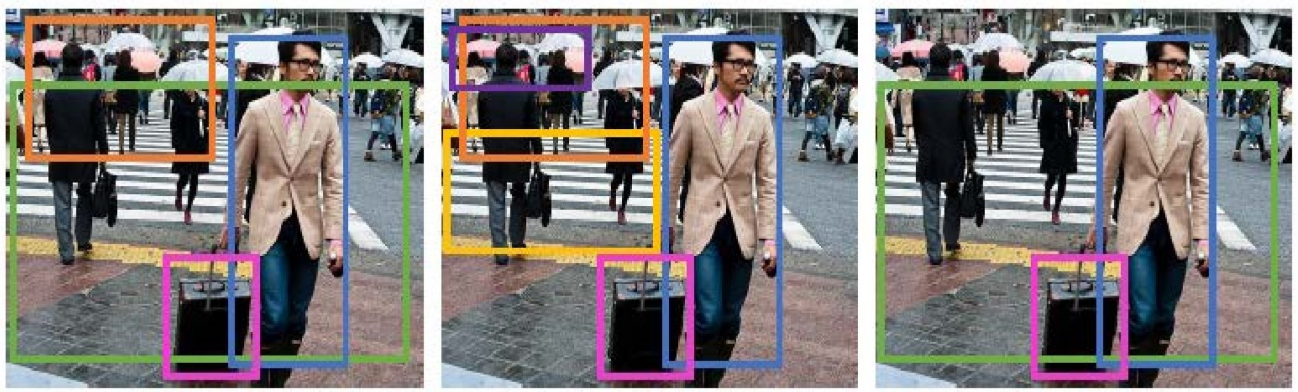
Image Captioning



A woman near bushes on a cell phone.

A young woman looks somber while using a cell phone.

A woman with long hair talking on a cellphone.



A man walks down a city street pulling a suitcase while a lot of other people are walking across the street.

A busy crosswalk with several people carrying umbrellas and a man with luggage.

A man pulling a suitcase across a street.

A group of young people playing a game of frisbee.

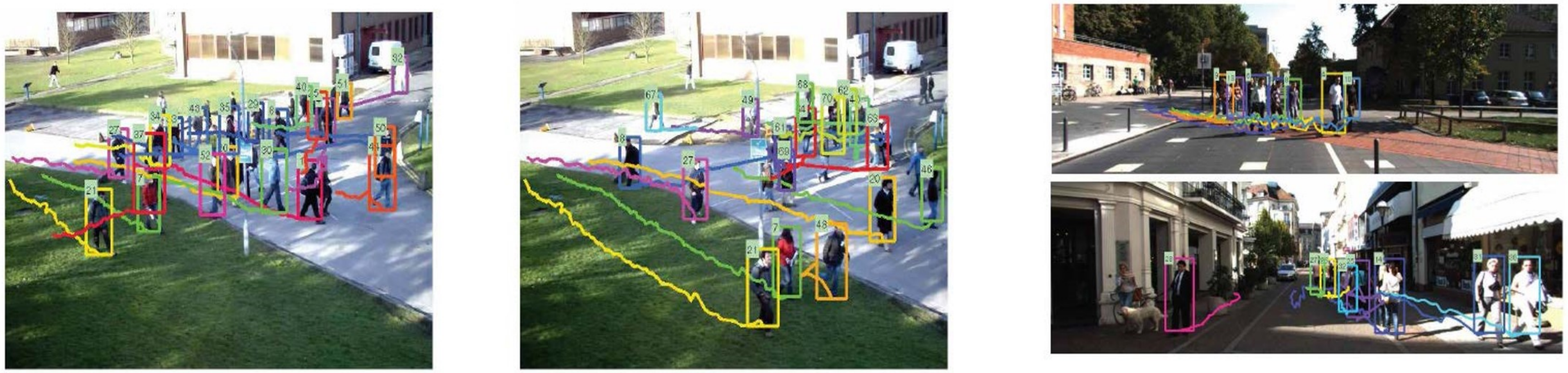
A herd of elephants walking across a dry grass field.

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." CVPR 2015.
Cornia, Marcella, et al. "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions." CVPR 2019.

Visual Question Answering

Goyal, Yash, et al. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." CVPR 2017.

Object Tracking



PETS09-S2L2 #68 PETS09-S2L2 #111 KITTI-16 #90, KITTI-19 #281

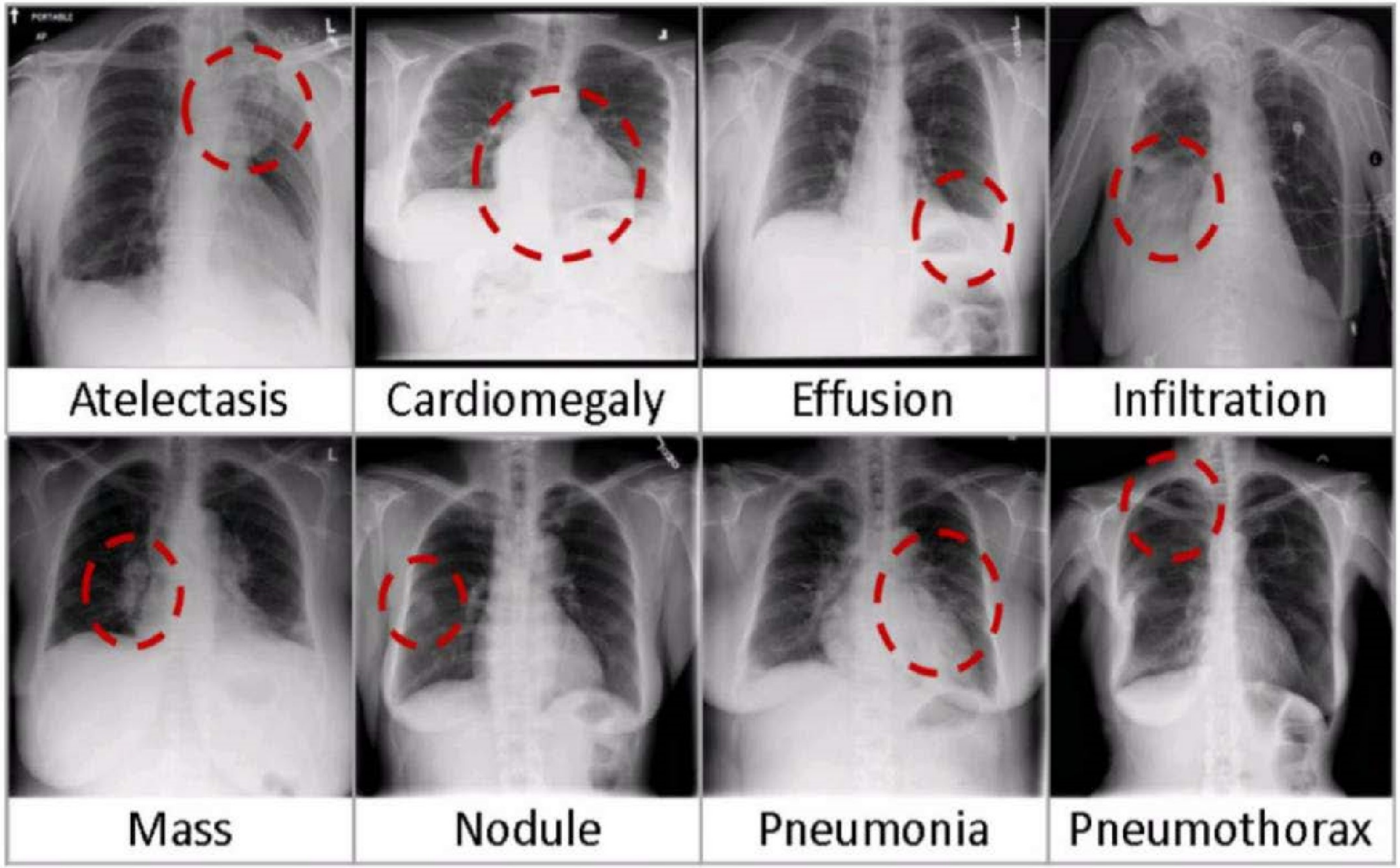
Frame #160 Frame #190 Frame #220

Xiang, Yu, et al. "Learning to track: Online multi-object tracking by decision making." ICCV 2015.
Yun, Sangdoo, et al. "Action-decision networks for visual tracking with deep reinforcement learning." CVPR 2017.

Human Pose Estimation

Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." CVPR 2017.

Medical Image Analysis



Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." CVPR 2017.