max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Introduction
@ April 10, 2019

Bernt Schiele & Mario Fritz

www.mpi-inf.mpg.de/hlcv/

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Computer Vision and Multimodal Computing Group @ Max-Planck-Institute for Informatics

Gerard Pons-Moll
Real Virtual Humans

Zeynep Akata
Multimodal Deep Learning
U Amsterdam

Paul Swoboda
Combinatorial Vision Group

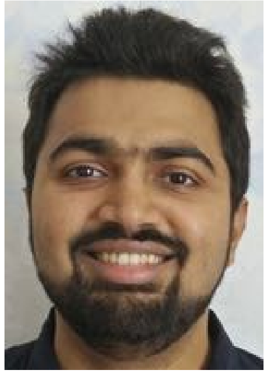
Mario Fritz
Scalable Learning & Perception
CISPA Helmholtz Center i.G.

Bernt Schiele
Computer Vision

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Computer Vision

- Lecturer:
  - Bernt Schiele (schiele@mpi-inf.mpg.de)
  - Mario Fritz (mfritz@mpi-inf.mpg.de)
- Assistants:
  - Yang He (yang@mpi-inf.mpg.de)
  - Rakshith Shetty (rshetty@mpi-inf.mpg.de)



- Language:
  - English
- mailing list for announcements etc.
  - send email (see instructions on the web)
    Rakshith Shetty <rshetty@mpi-inf.mpg.de>

Lecture & Exercise

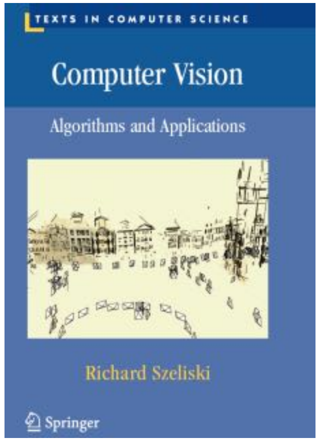
- Officially: 2V (lecture) + 2Ü (exercise)
  - Lecture: Wed: 10:15am - 12pm (room 024)
  - Exercise: Mon: 10:15am - 12pm (room 024)

- typically 1 exercise sheet every 1-2 weeks
  - part of the final grade
  - some pencil and paper, mostly practical including a project
  - larger project in second half of lecture
    - we/you propose projects, mentoring, final presentation
- 1. exercise is Python tutorial

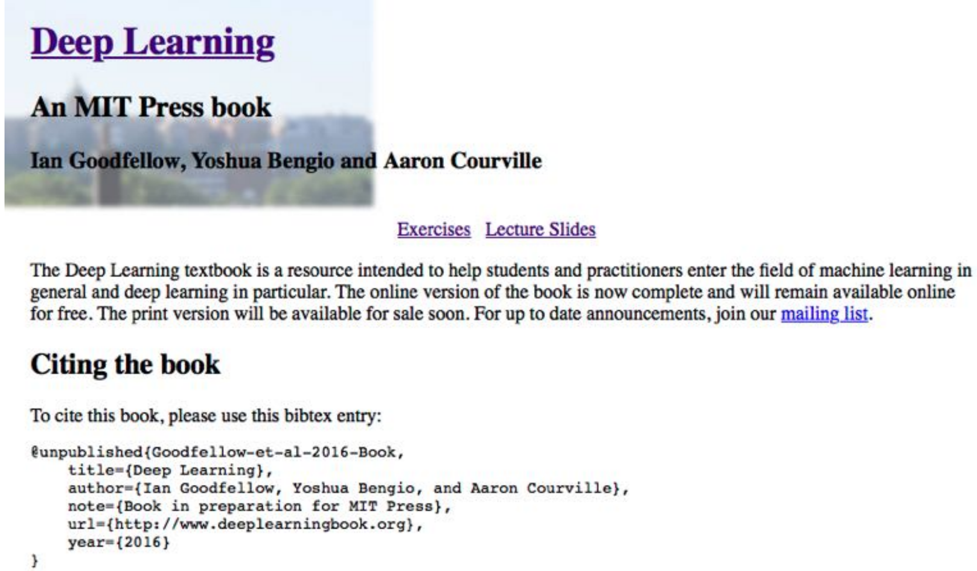
- Exam
  - oral exam (grading 50% oral exam and 50% exercises)
  - after the SS - there will be proposed dates

Material

- For "non-deep-learning" parts of the lecture:
  - available online http://szeliski.org/Book



- Background on deep learning: Deep Learning Book
  - available online http://deeplearning.org



Deep Learning

An MIT Press book

Ian Goodfellow, Yoshua Bengio and Aaron Courville

Exercises Lecture Slides

The Deep Learning textbook is a resource intended to help students and practitioners enter the field of machine learning in general and deep learning in particular. The online version of the book is now complete and will remain available online for free. The print version will be available for sale soon. For up to date announcements, join our mailing list.

Citing the book

To cite this book, please use this bibtex entry:

```
@unpublished{Goodfellow-et-al-2016-Book,
    title={Deep Learning},
    author={Ian Goodfellow, Yoshua Bengio, and Aaron Courville},
    note={Book in preparation for MIT Press},
    url={http://www.deeplearningbook.org},
    year={2016}
}
```

Why Study Computer Vision

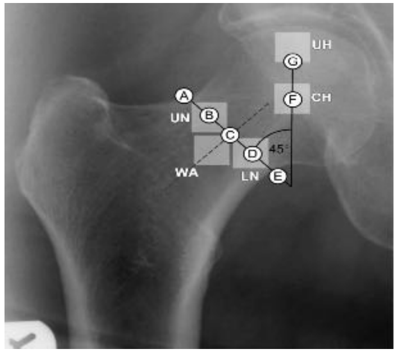
- Science
  - Foundations of perception. How do WE as humans see?
  - computer vision to explore “computational model of human vision”

- Engineering
  - How do we build systems that perceive the world
  - computer vision to solve real-world problems
    (e.g. self-driving cars to detect pedestrians)

- Applications
  - medical imaging (computer vision to support medical diagnosis, visualization)
  - surveillance (to follow/track people at the airport, train-station, ...)
  - entertainment (vision-based interfaces for games)
  - graphics (image-based rendering, vision to support realistic graphics)
  - car-industry (lane-keeping, pre-crash intervention, …)
  - …

Some Applications

- License Plate Recognition
  - London Congestion Charge
  - http://www.cclondon.com/imagingandcameras.html
  - http://en.wikipedia.org/wiki/London_congestion_charge
- Surveillance
  - Face Recognition
  - Airport Security (People Tracking)
- Medical Imaging
  - (Semi-)automatic segmentation and measurements
- Autonomous Driving & Robotics



More Applications

- Vision on Cellphones:
  - e.g. Google Goggles
- Vision for Interfaces:
  - e.g. Microsoft Kinect
- Reconstruction

Google Goggles
Use pictures to search the web.

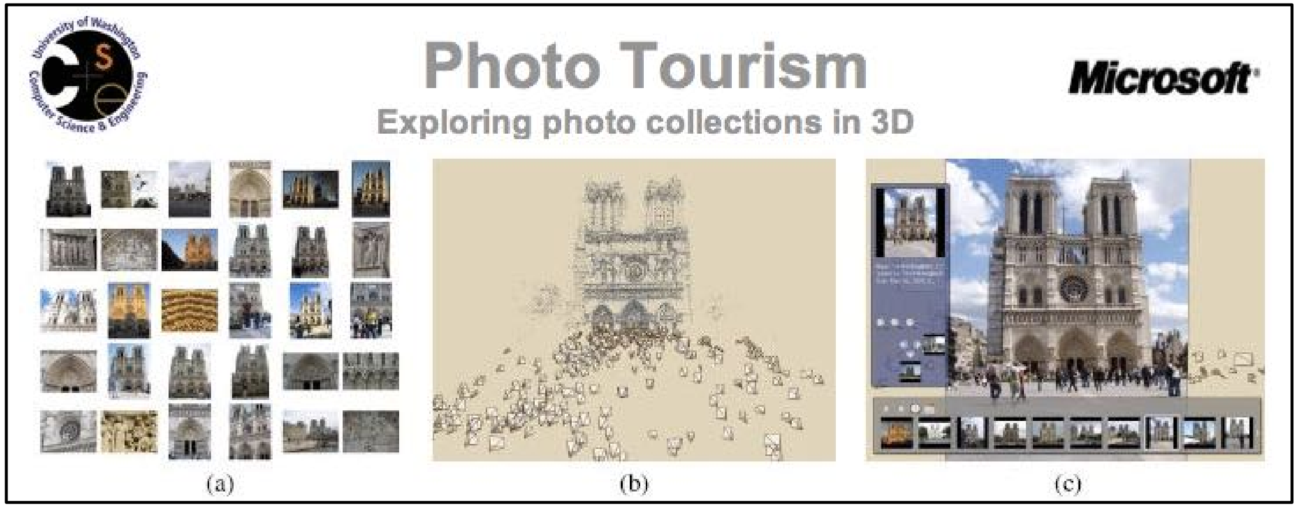
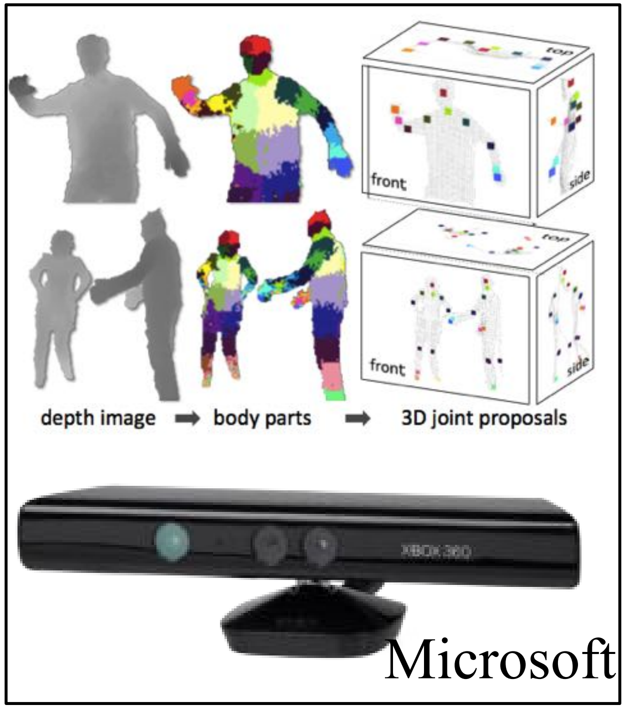


Photo Tourism
Exploring photo collections in 3D
Microsoft
University of Washington Computer Science & Engineering
(a)
(b)
(c)



Goals of today’s lecture

- First intuitions about
  - What is computer vision?
  - What does it mean to see and how do we (as humans) do it?
  - How can we make this computational?

- Applications & Appetizers

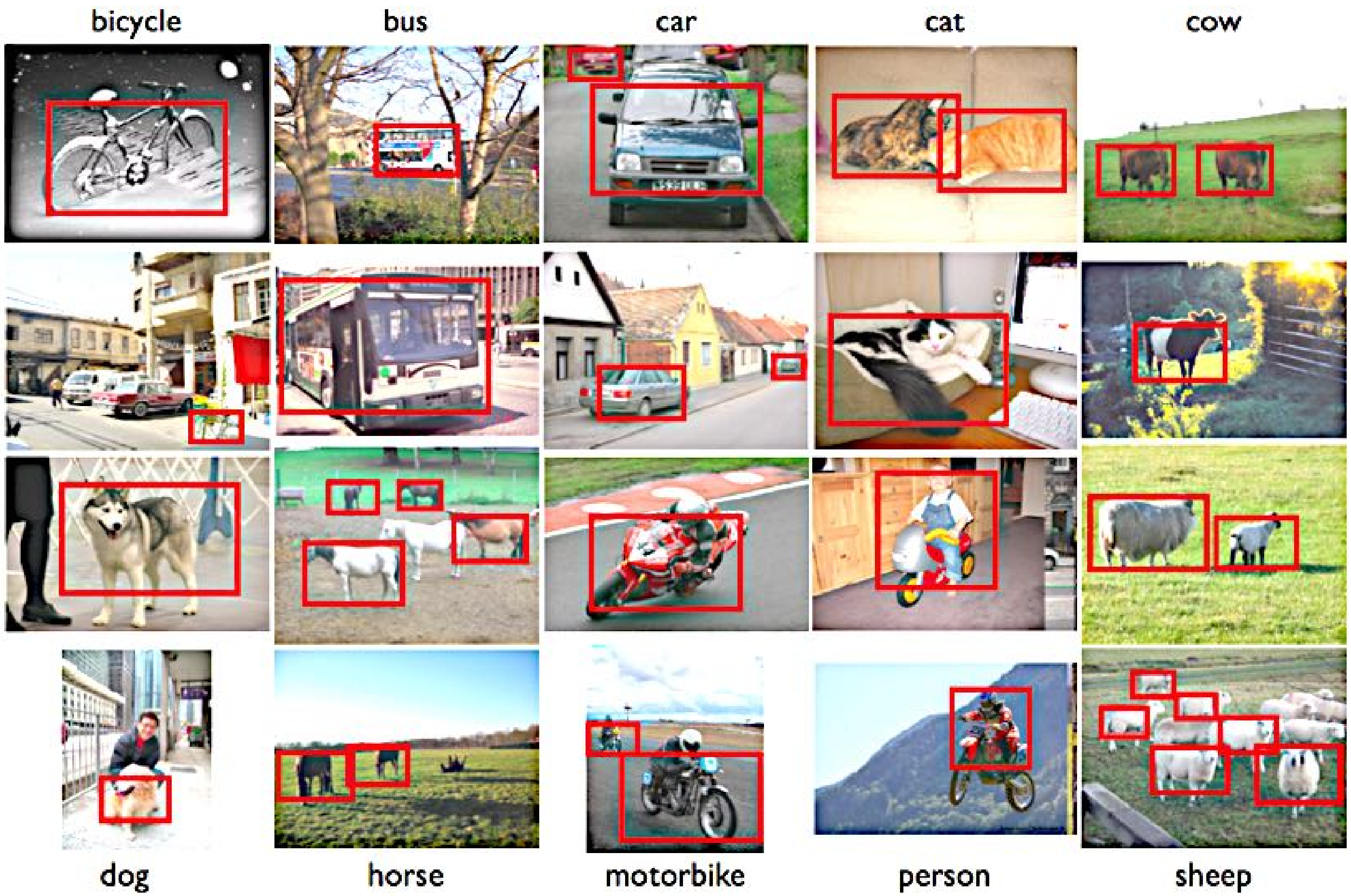
- Role of Deep Learning
  - with several slides taken from Fei-Fei Li, Justin Johnson, Serena Yeung @ Stanford

- 2 case studies:
  - Recovery of 3D structure
    - slides taken from Michael Black @ Brown University / MPI Intelligent Systems
  - Object Recognition
    - intuition from human vision...

Applications & Appetizers

... work from our group

Detection & Recognition of Visual Categories



Challenges:

- multi-scale
- multi-view
- multi-class

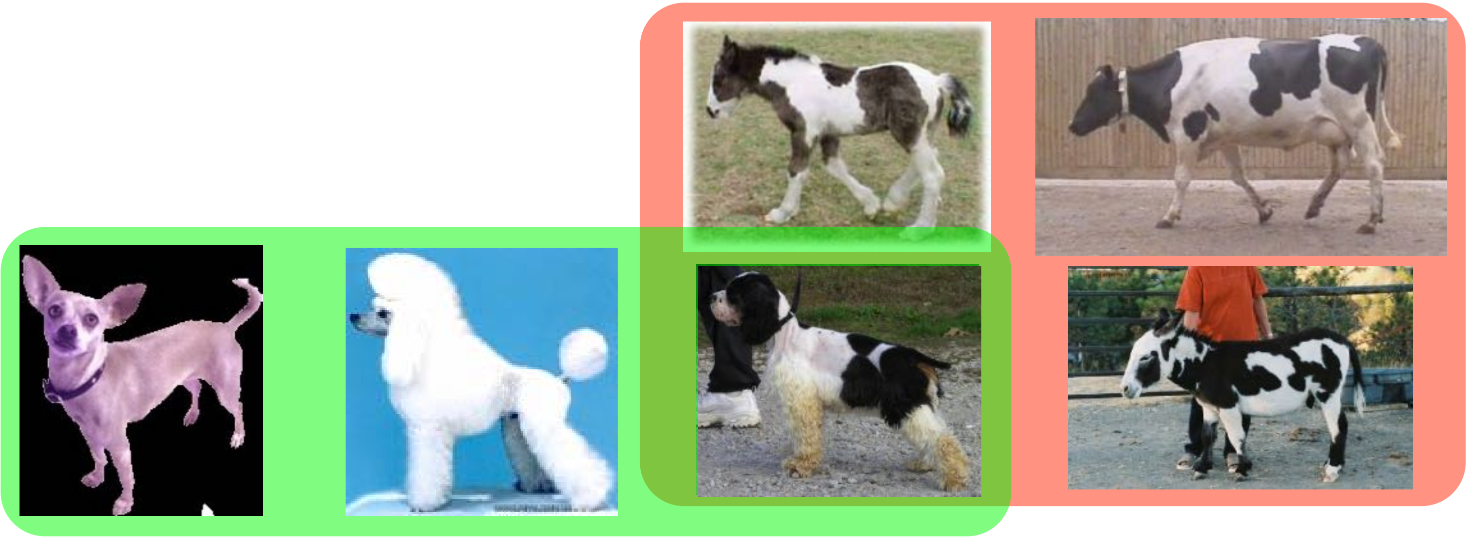
- varying illumination
- occlusion
- cluttered background

- articulation
- high intraclass variance
- low interclass variance

Challenges of Visual Categorization

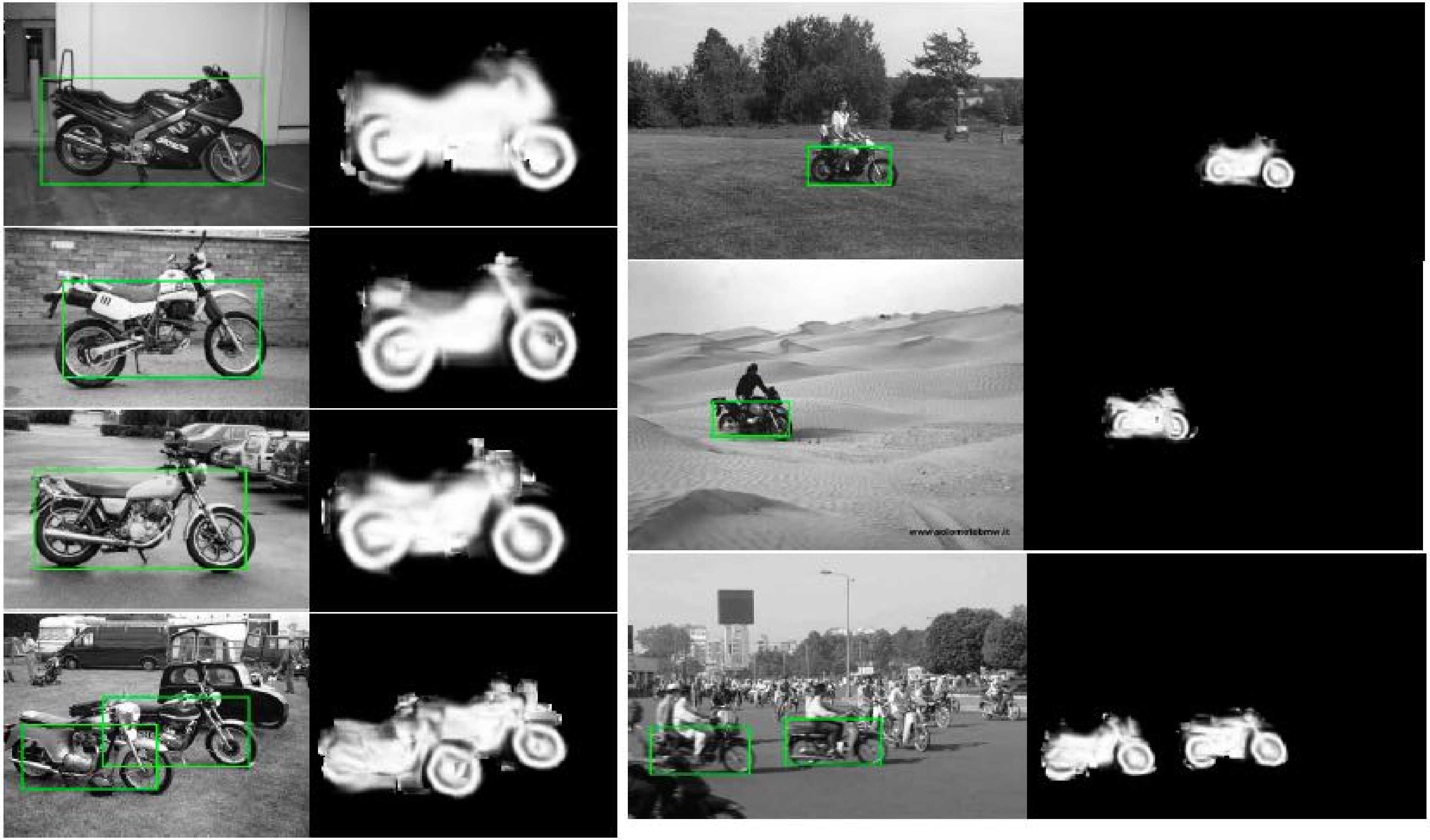
- high intra-class variation

- low inter-class variation

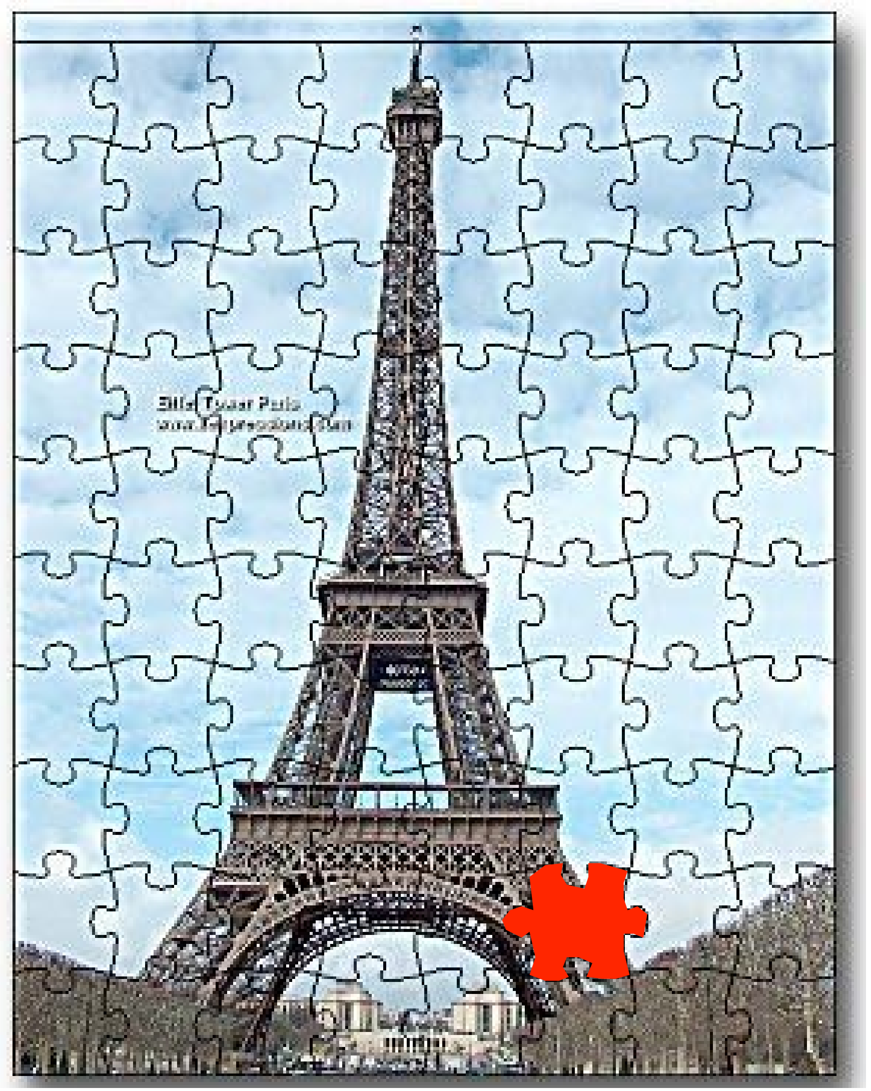


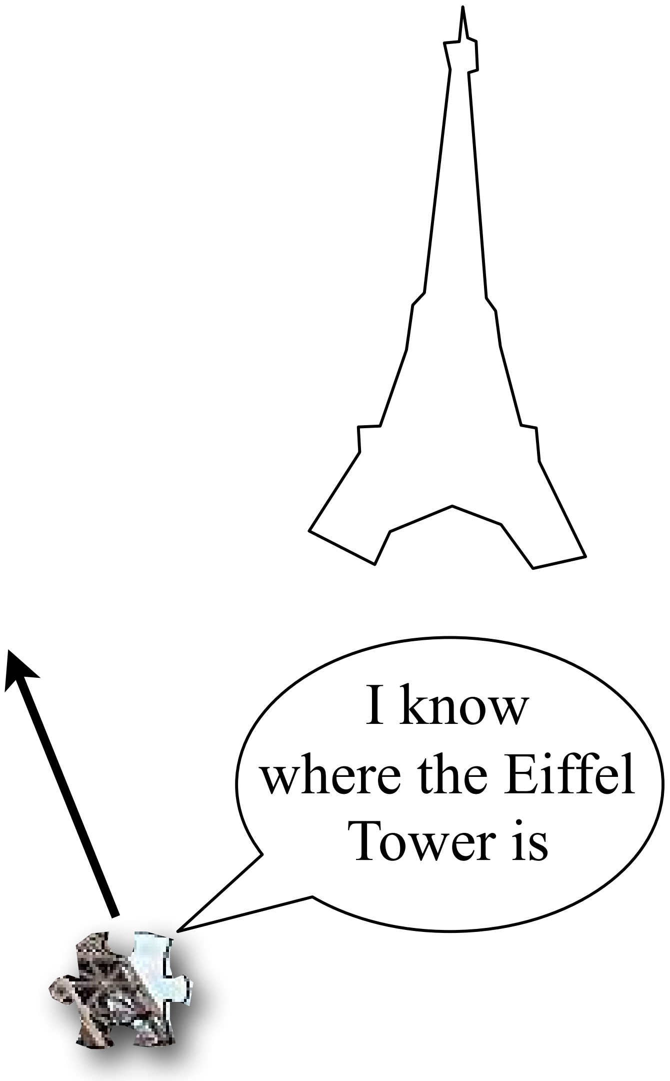
- high intra-class variation

Sample Category: Motorbikes



Basic Idea

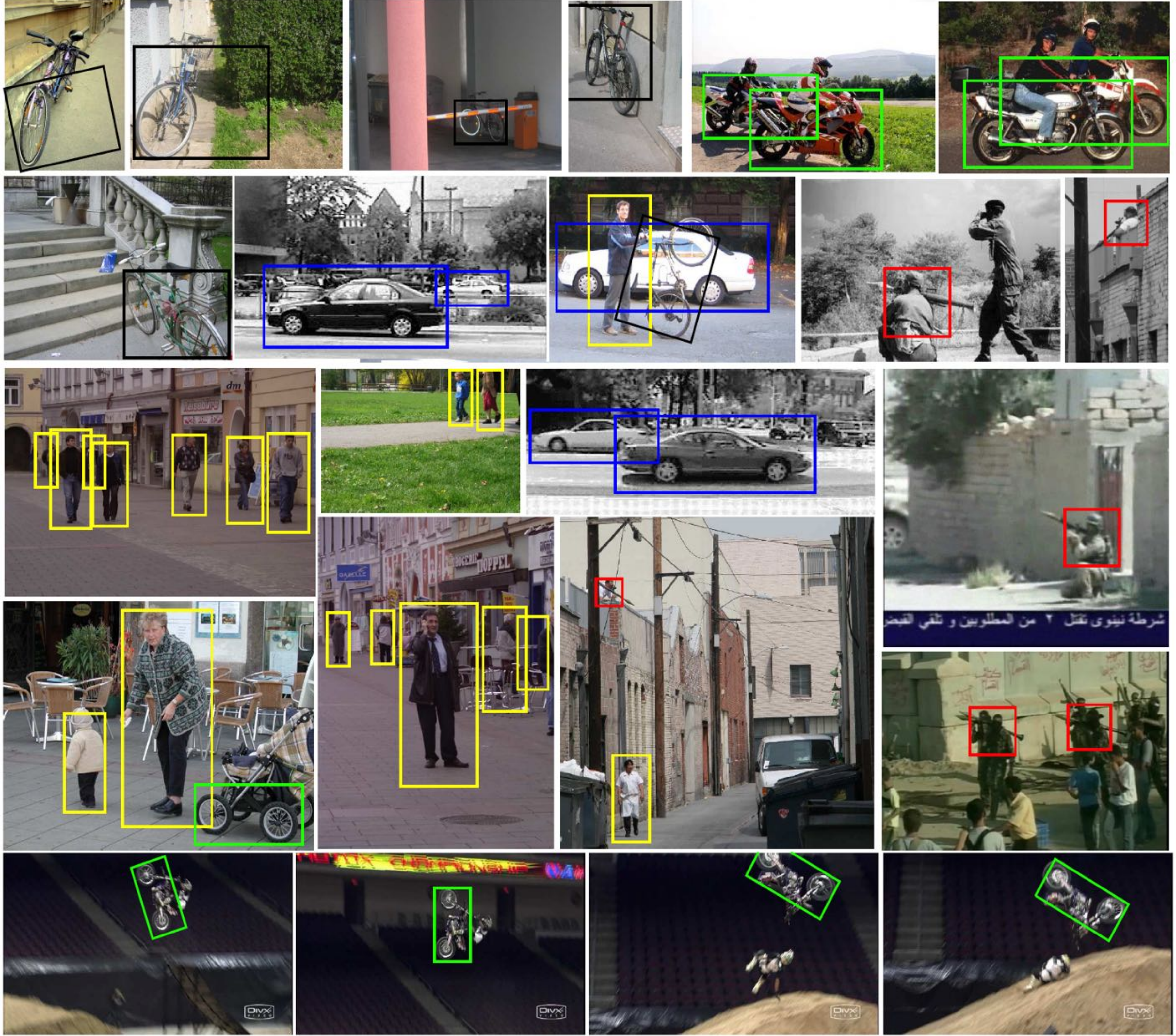




I know where the Eiffel Tower is

global

local

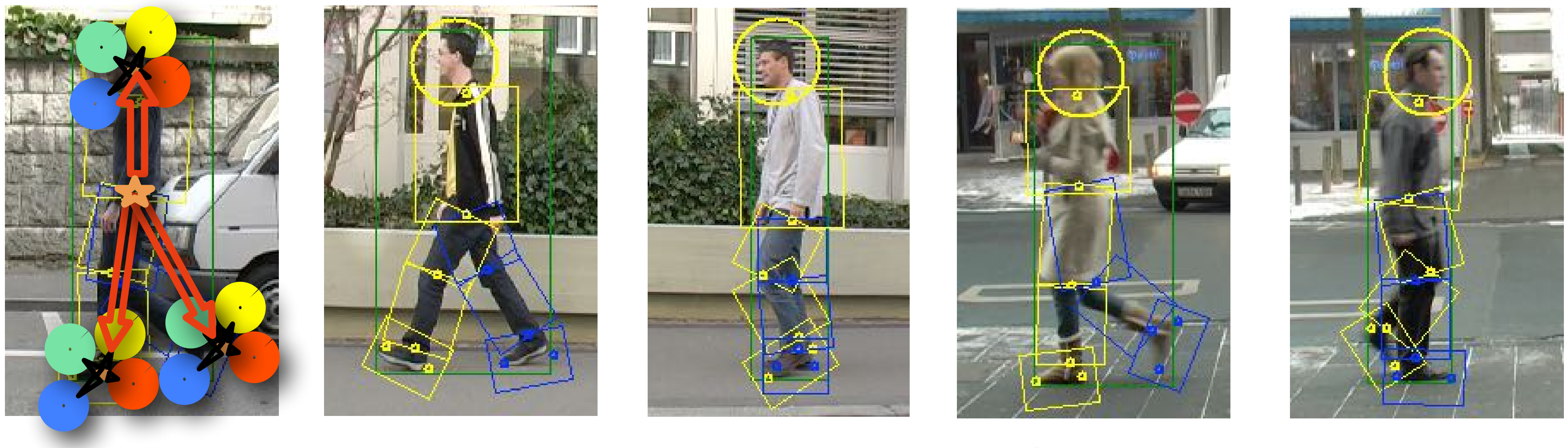

شرطة نينوى تقتل ٢ من المطلوبين و تلقي القبض
dm
DOPPEL
GAZELLE
DivX
DivX
DivX
DivX

Video...

Articulation Model

- Assume uniform position prior for the whole body
- Learn the conditional relation between part position and body center from data:

$$p(L|a) = p(\mathbf{x}^o) \prod_{i=1}^{N} p(\mathbf{x}^i|\mathbf{x}^o, a)$$



400 annotated training images

Modeling Body Dynamics

- Visualization of the hierarchical Gaussian process latent variable model (hGPLVM)

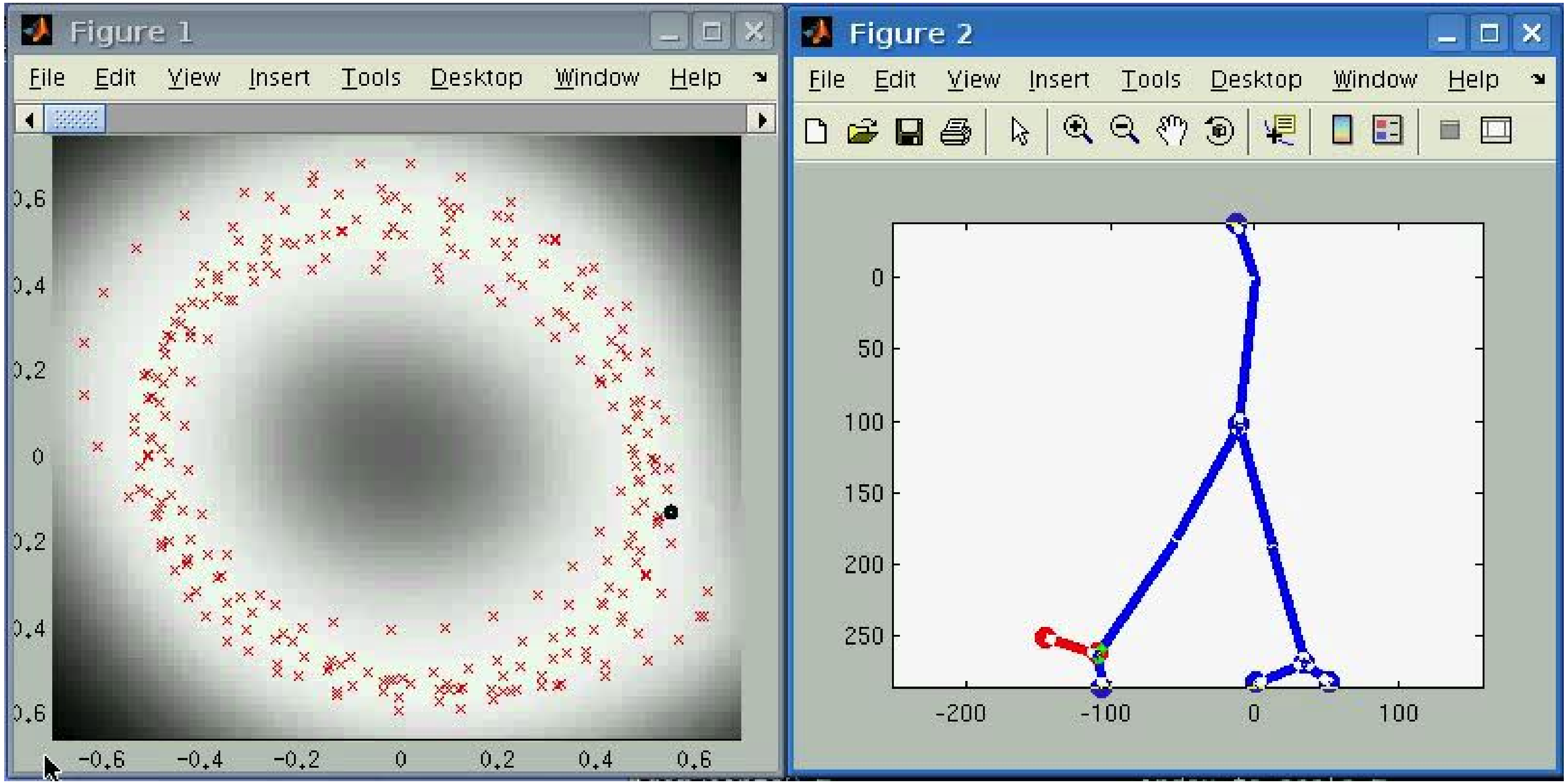
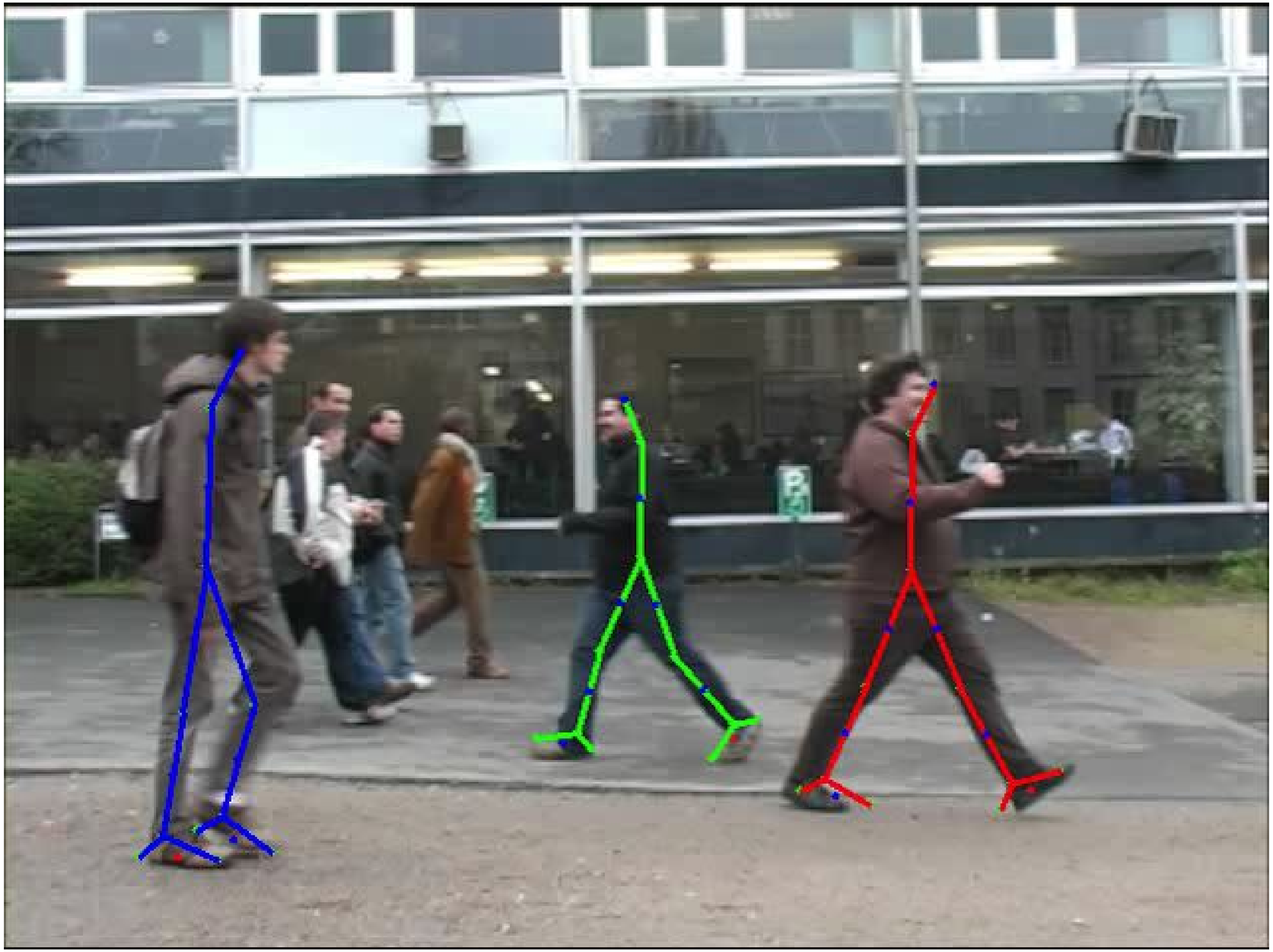
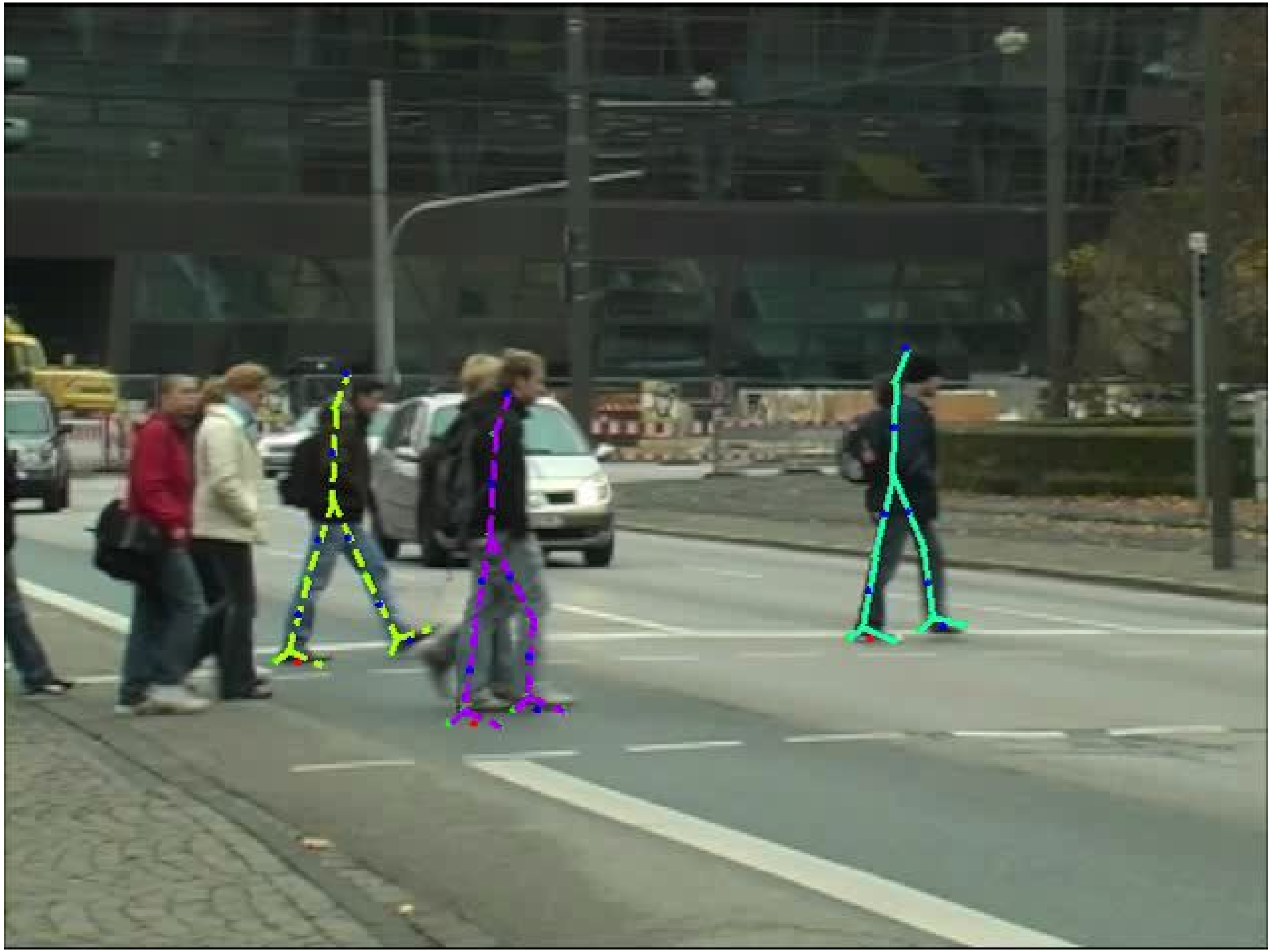
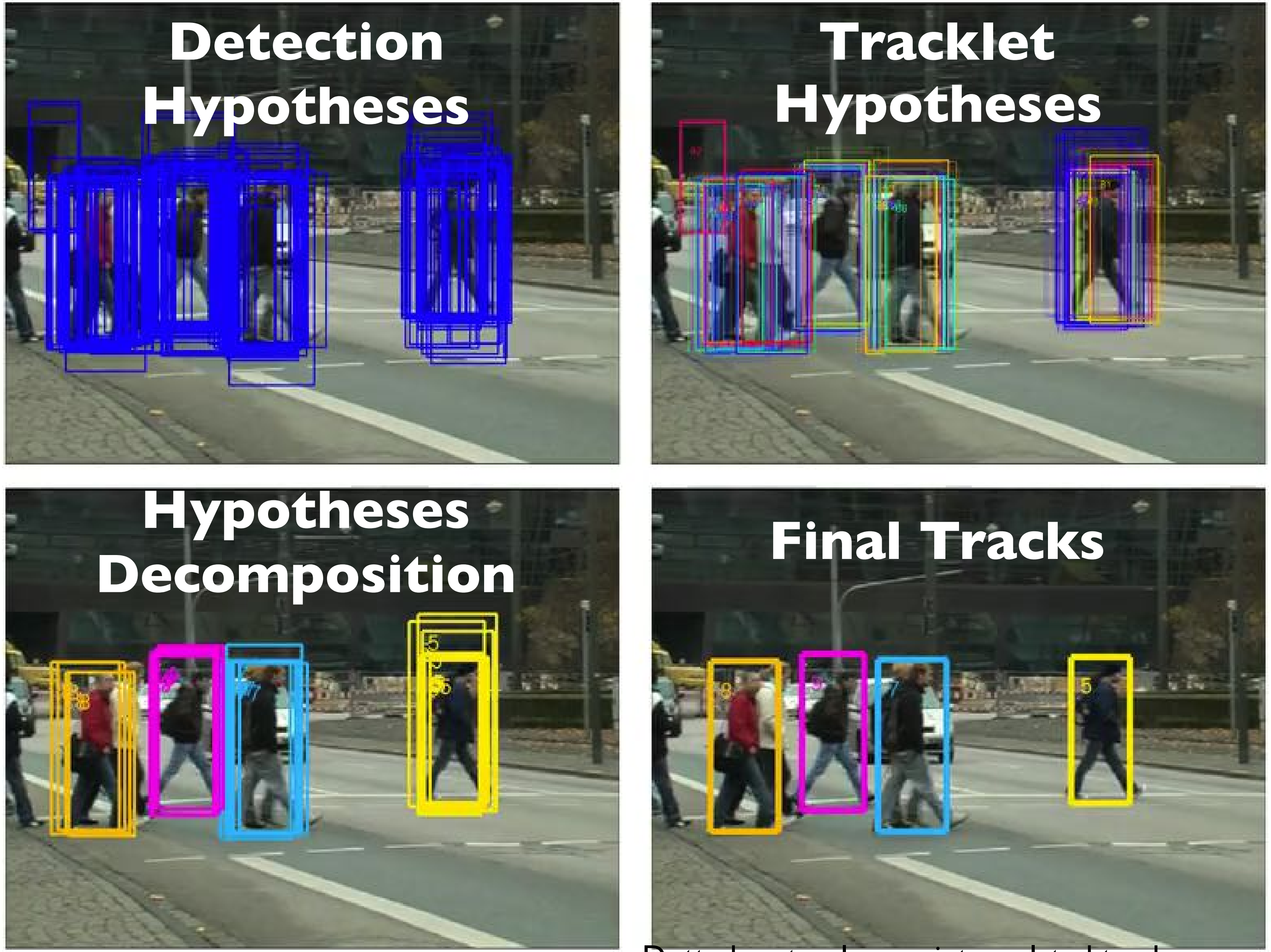


Figure 1
File Edit View Insert Tools Desktop Window Help
0.6
0.4
0.2
0
0.2
0.4
0.6
-0.6 -0.4 -0.2 0 0.2 0.4 0.6
Figure 2
File Edit View Insert Tools Desktop Window Help
0
50
100
150
200
250
-200 -100 0 100





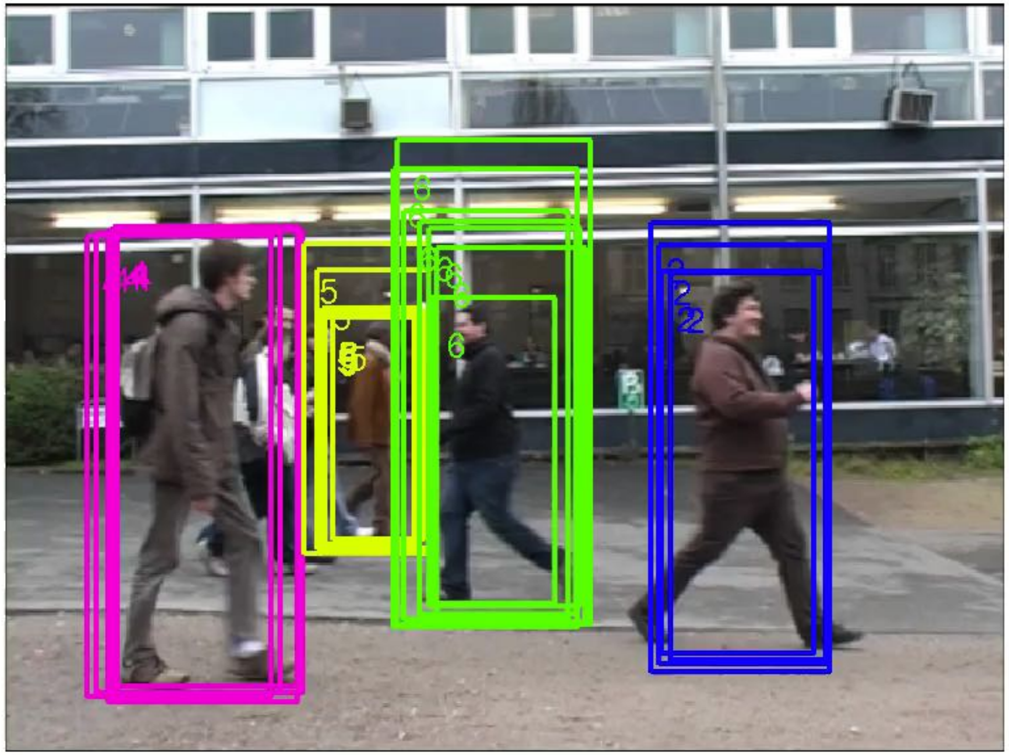
Our Subgraph Multicut Tracking Results



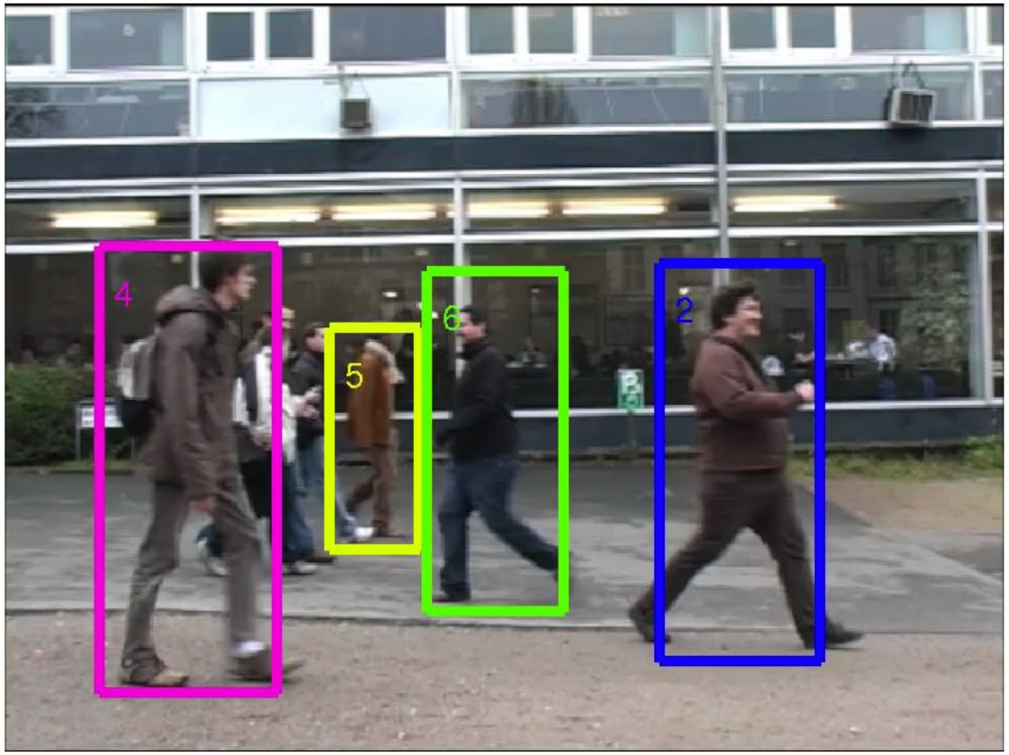
Dotted rectangles are interpolated tracks.

More Results

Dotted rectangles are interpolated tracks.



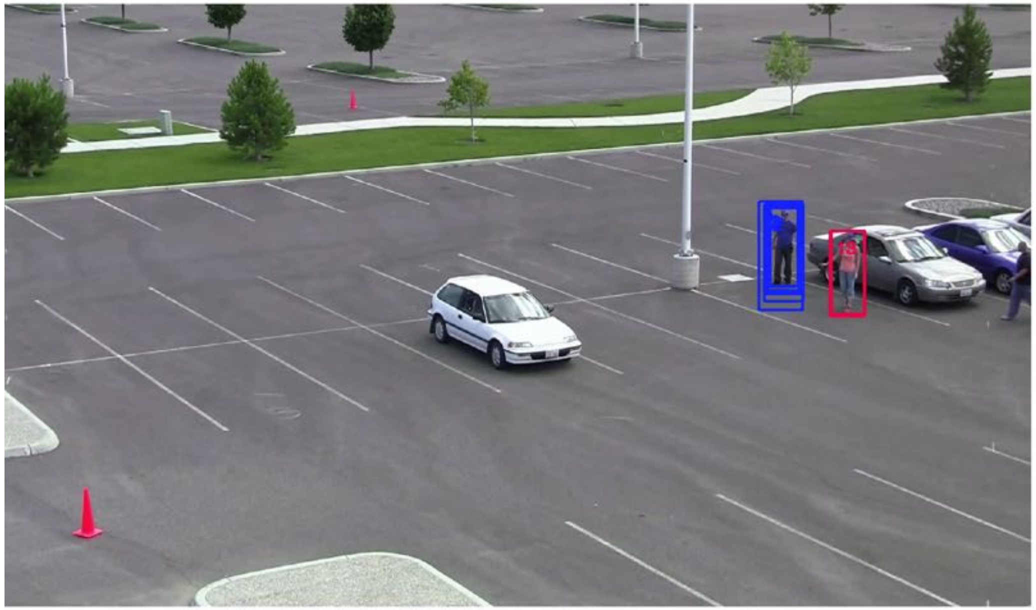
Decompositions (clusters)



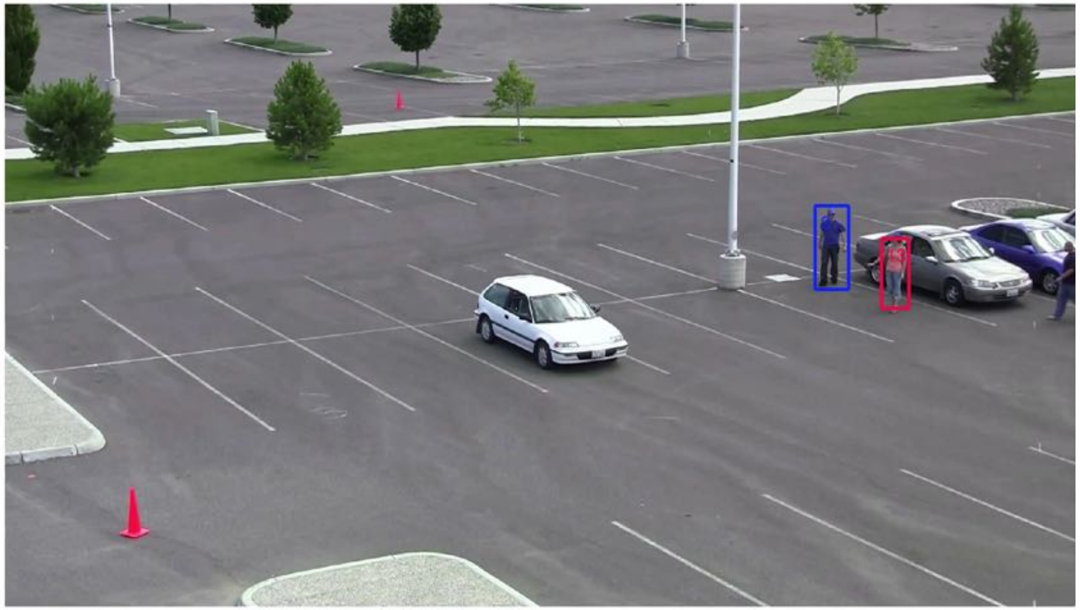
Tracks

More Results

Dotted rectangles are interpolated tracks.



Decompositions (clusters)



Tracks

**Deep Learning
have become an important tool
for object recognition**

(and other computer vision tasks)

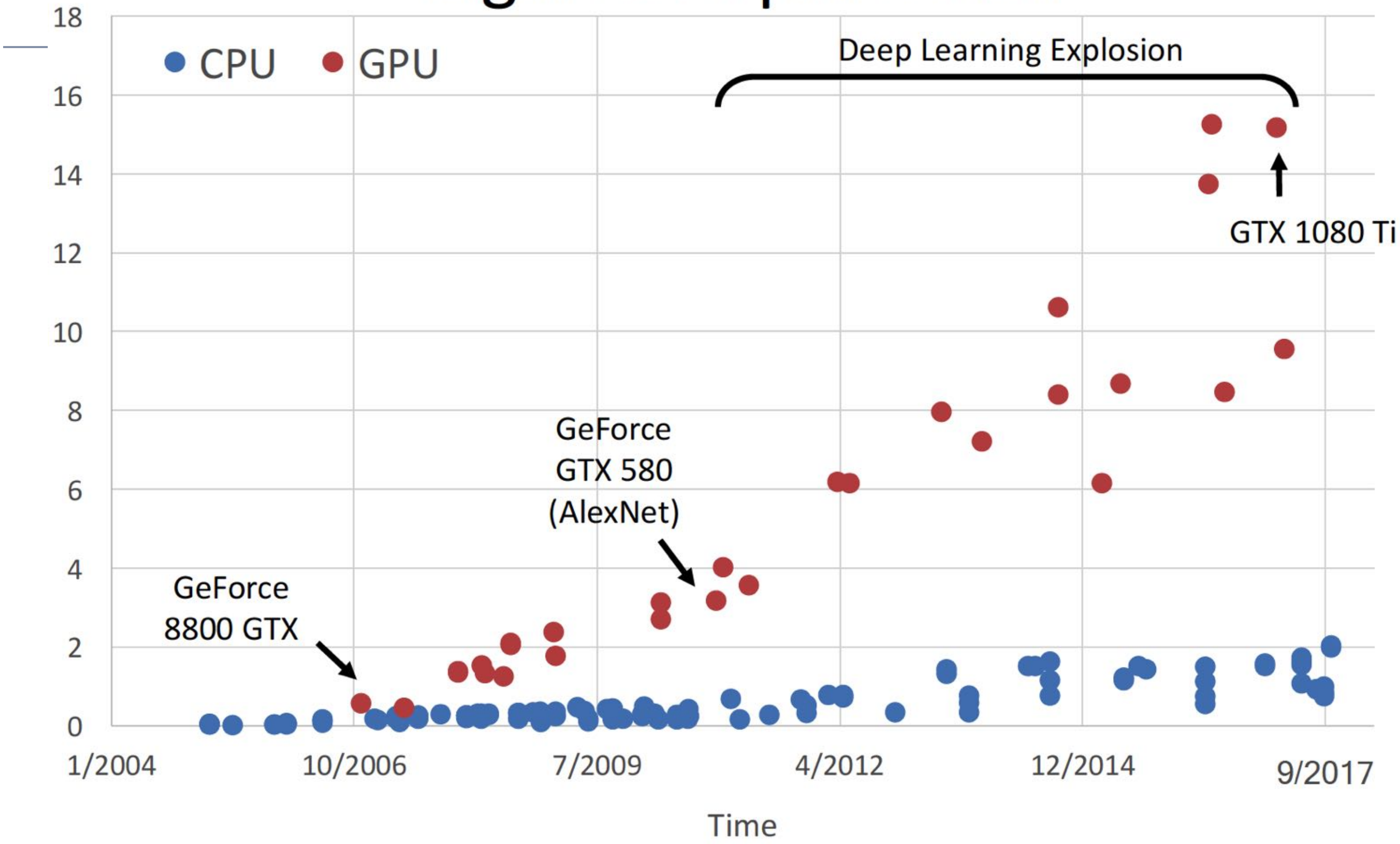
Let's briefly discuss CNNs
(Convolutional Neural Networks)

Ingredients for Deep Learning

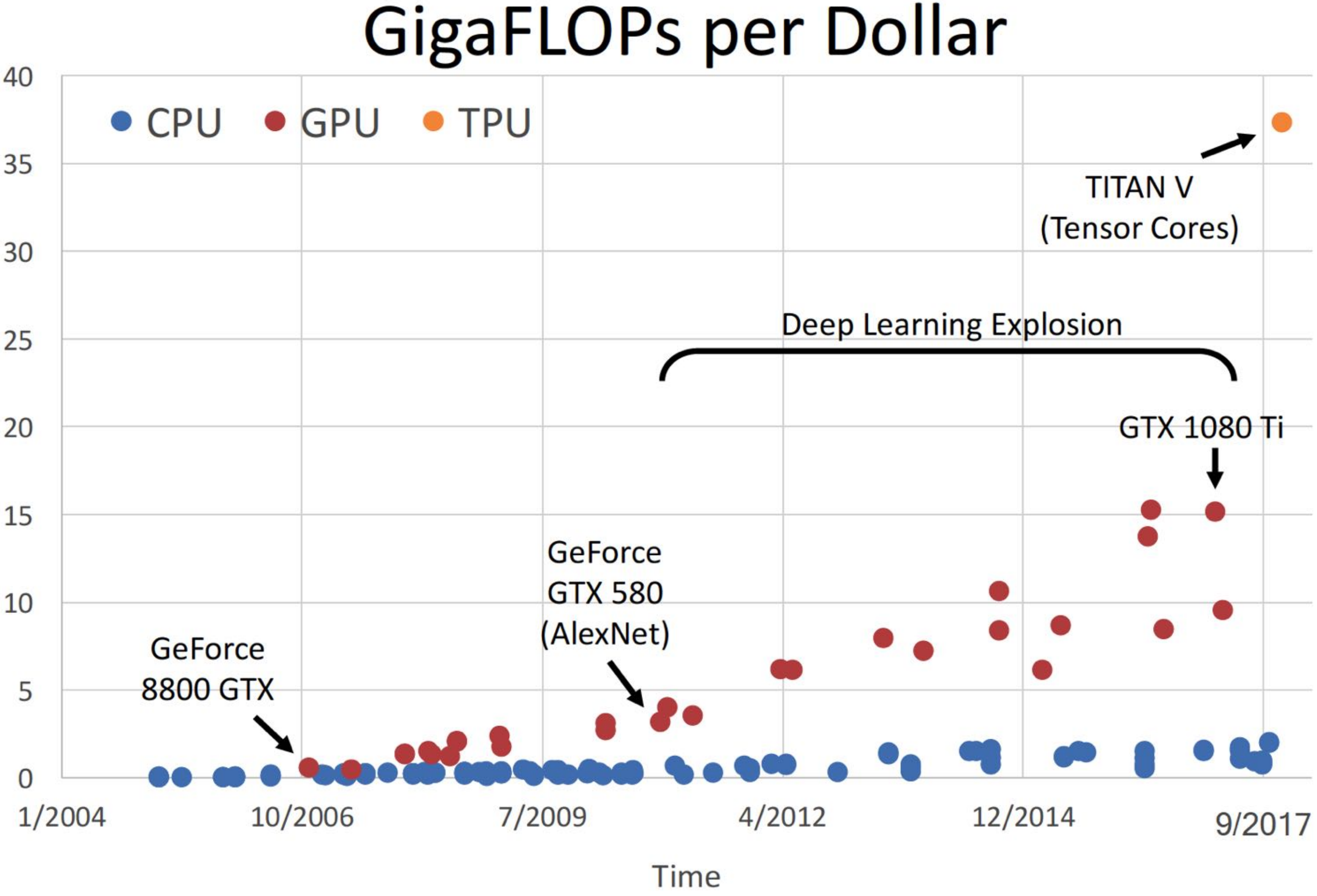
Algorithms
Data
Computation
11
48
55
27
128
224
5
13
3
27
11
55
192
128
dense
1000
2048
Max
GEFORCE GTX

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

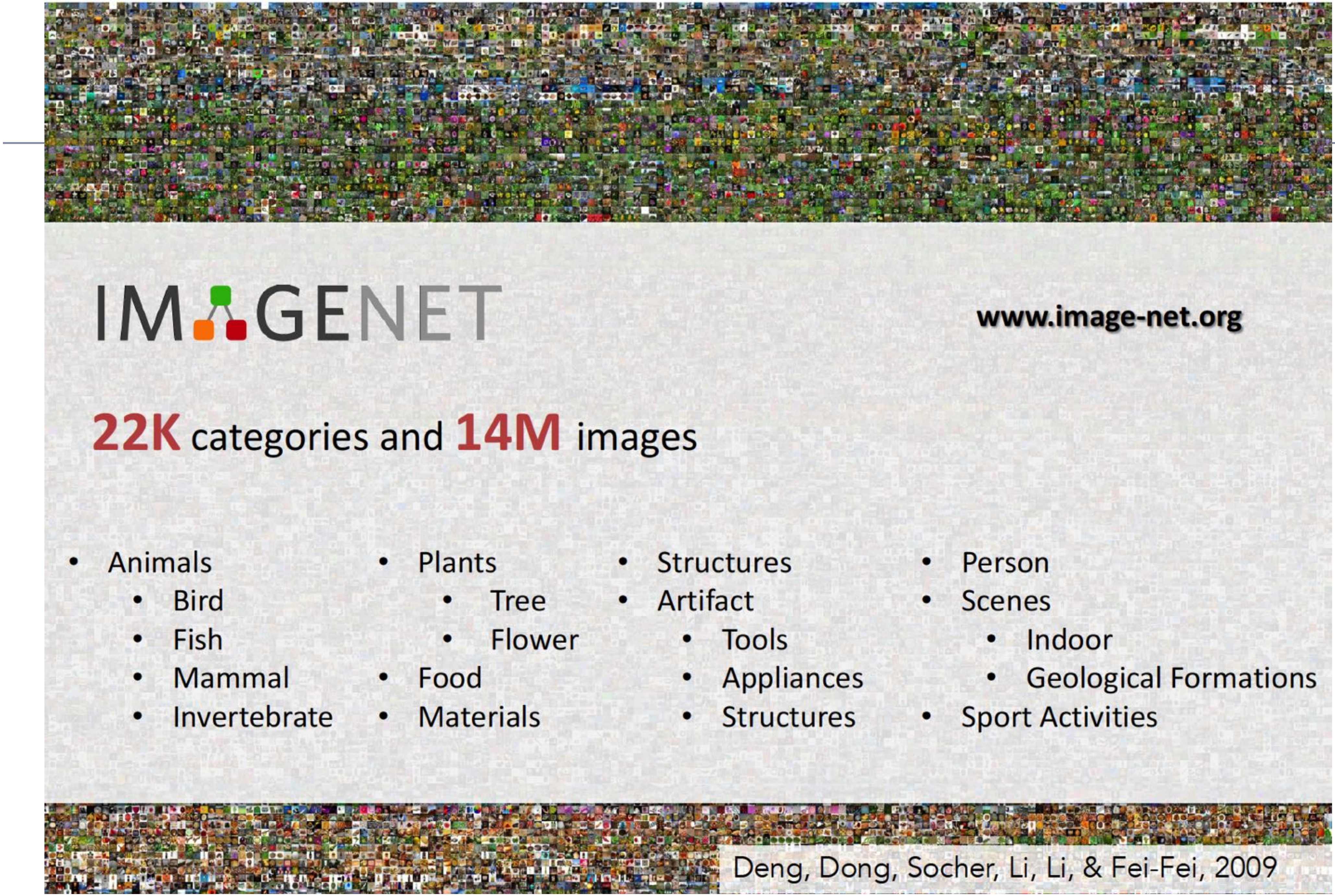
GigaFLOPs per Dollar



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



slide credit: Fei-Fei, Justin Johnson, Serena Yeung



IMAGENET

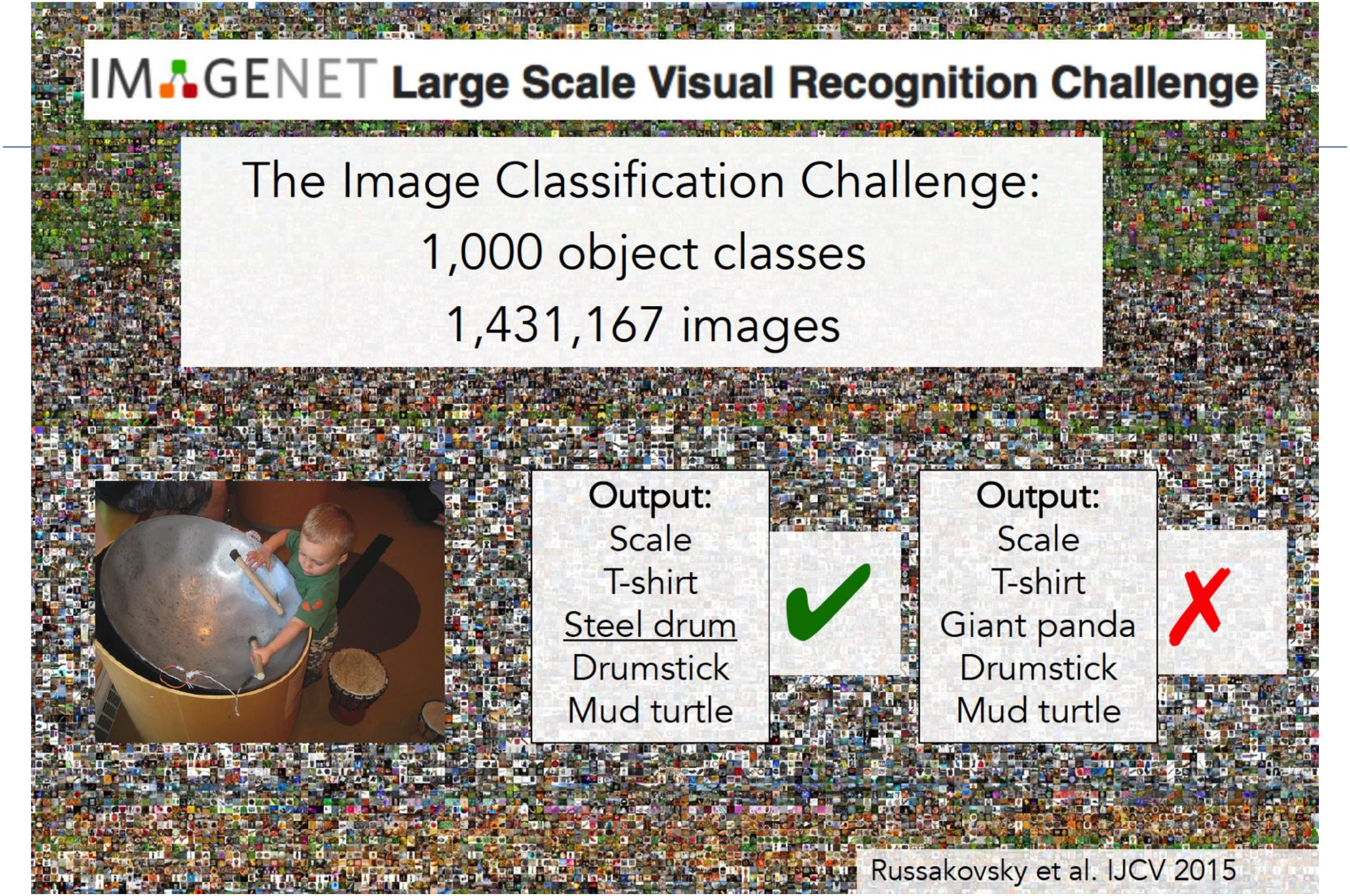
www.image-net.org

22K categories and 14M images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✓

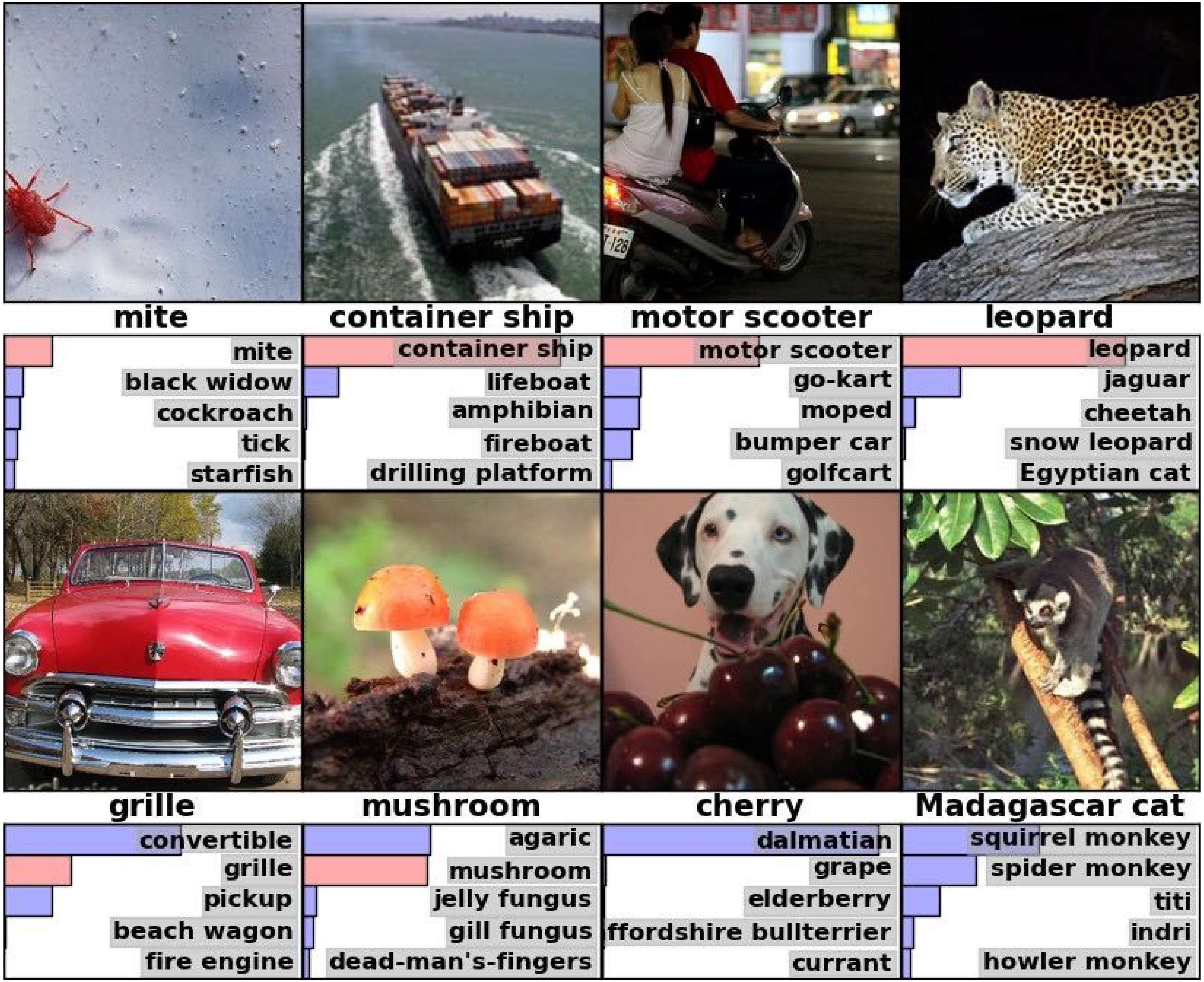
Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

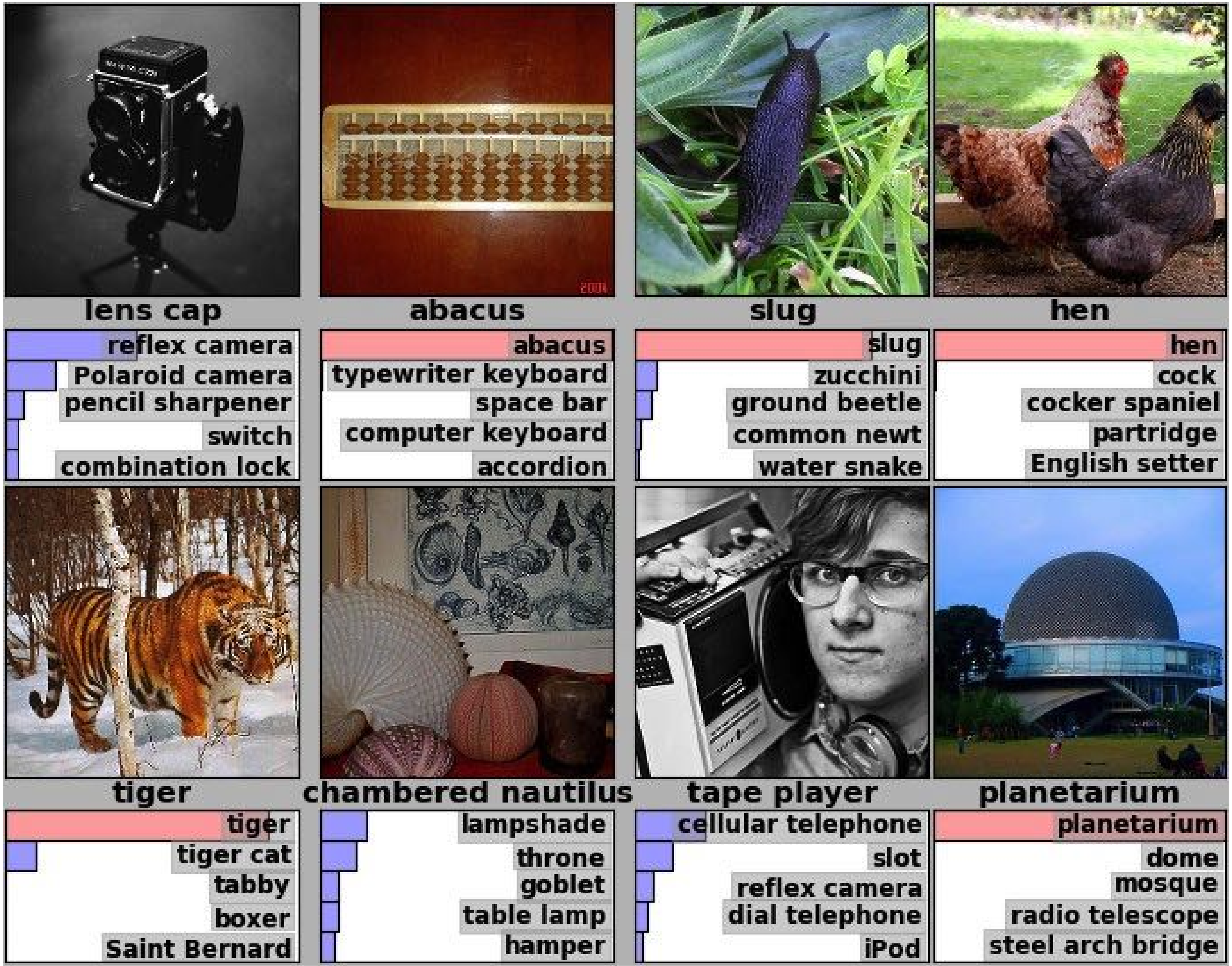
Russakovsky et al. IJCV 2015

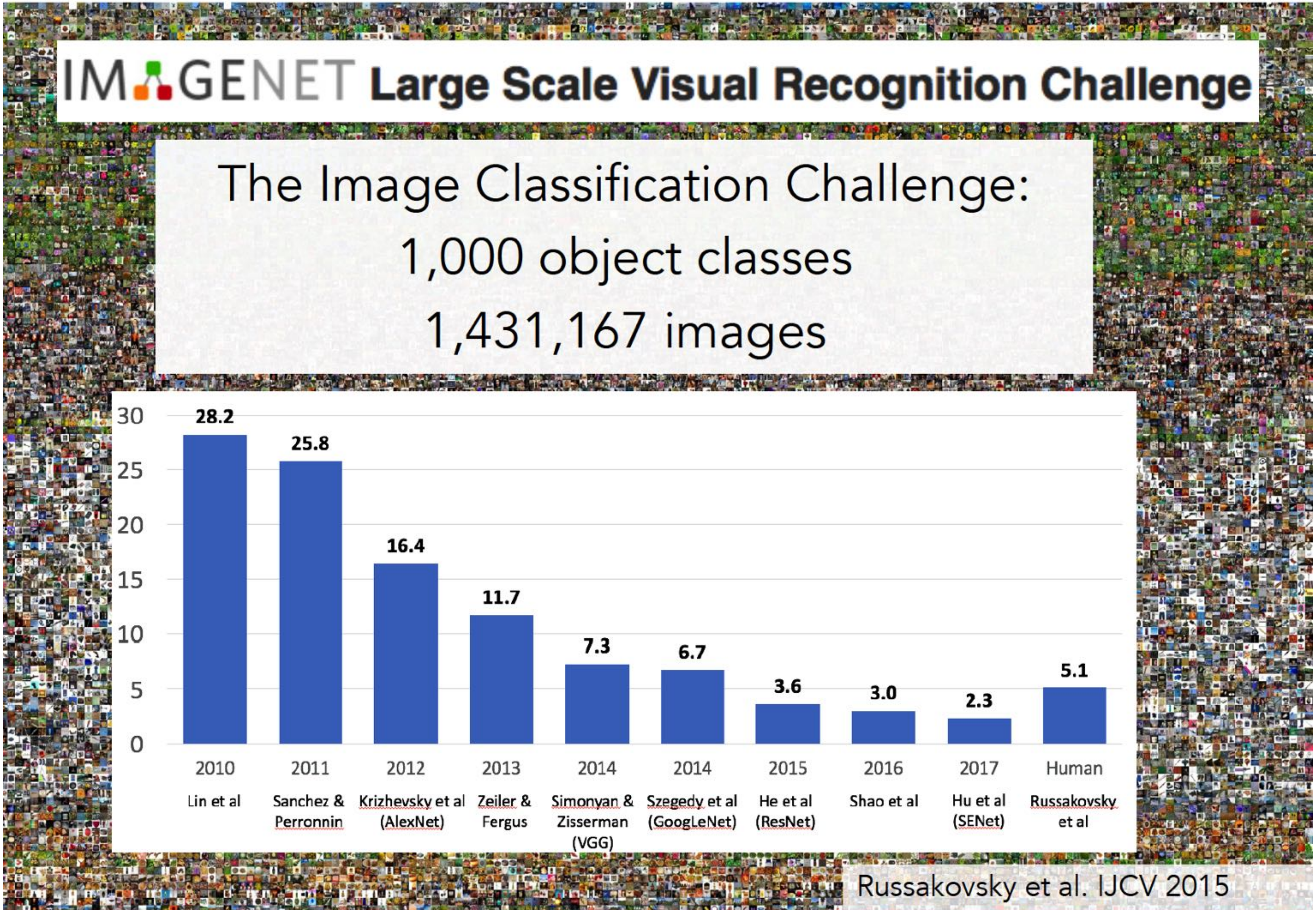
slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Validation classification



Validation classification





IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images

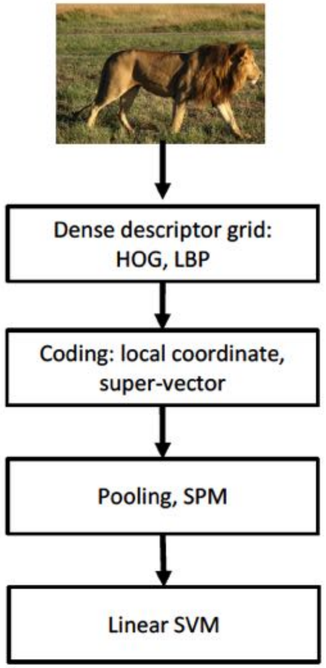
Russakovsky et al. IJCV 2015

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC



Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

Pooling, SPM

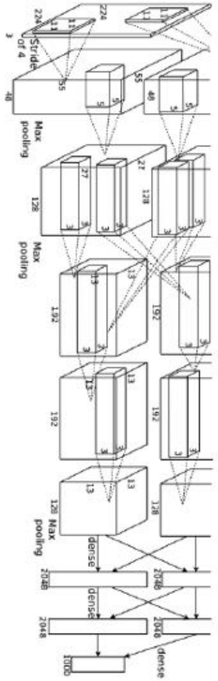
Linear SVM

[Lin CVPR 2011]

Lion image by Swissfrog is licensed under CC BY 3.0

Year 2012

SuperVision

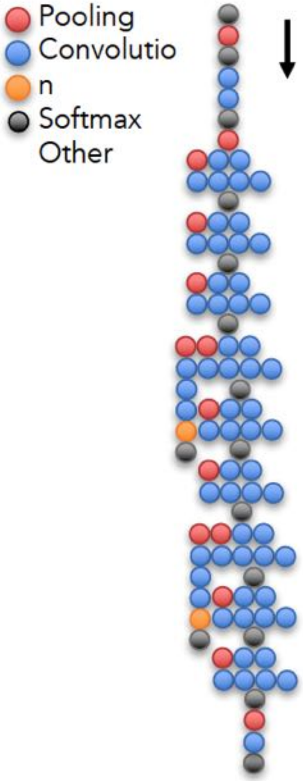


[Krizhevsky NIPS 2012]

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

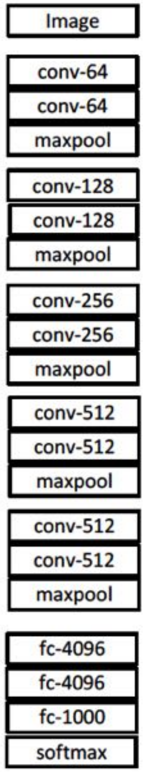
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

VGG



Image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

fc-4096

fc-4096

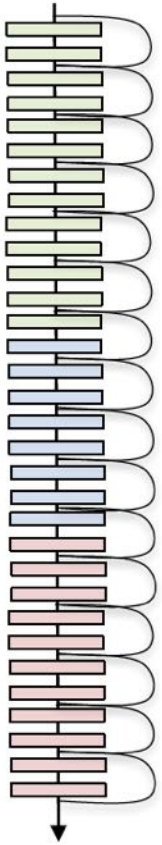
fc-1000

softmax

[Simonyan arxiv 2014]

Year 2015

MSRA



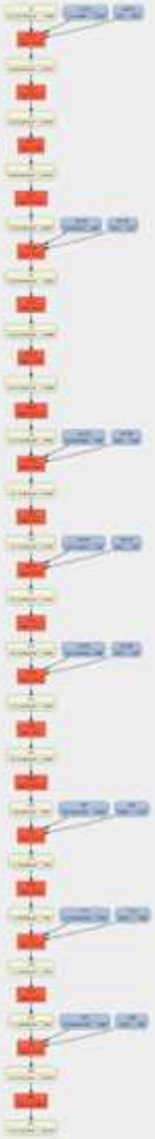
[He ICCV 2015]

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

How deep is enough?

AlexNet (2012)

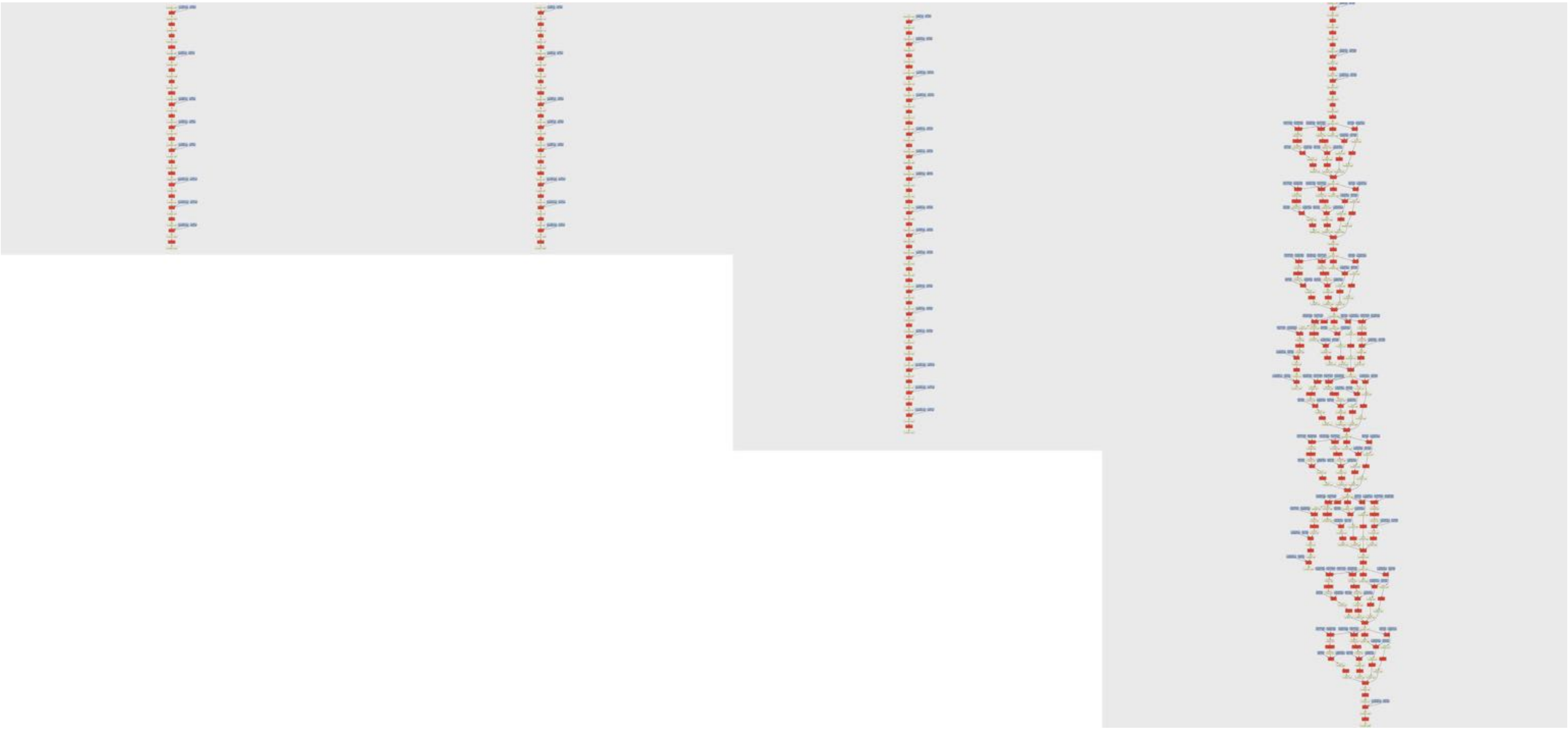
5 convolutional layers



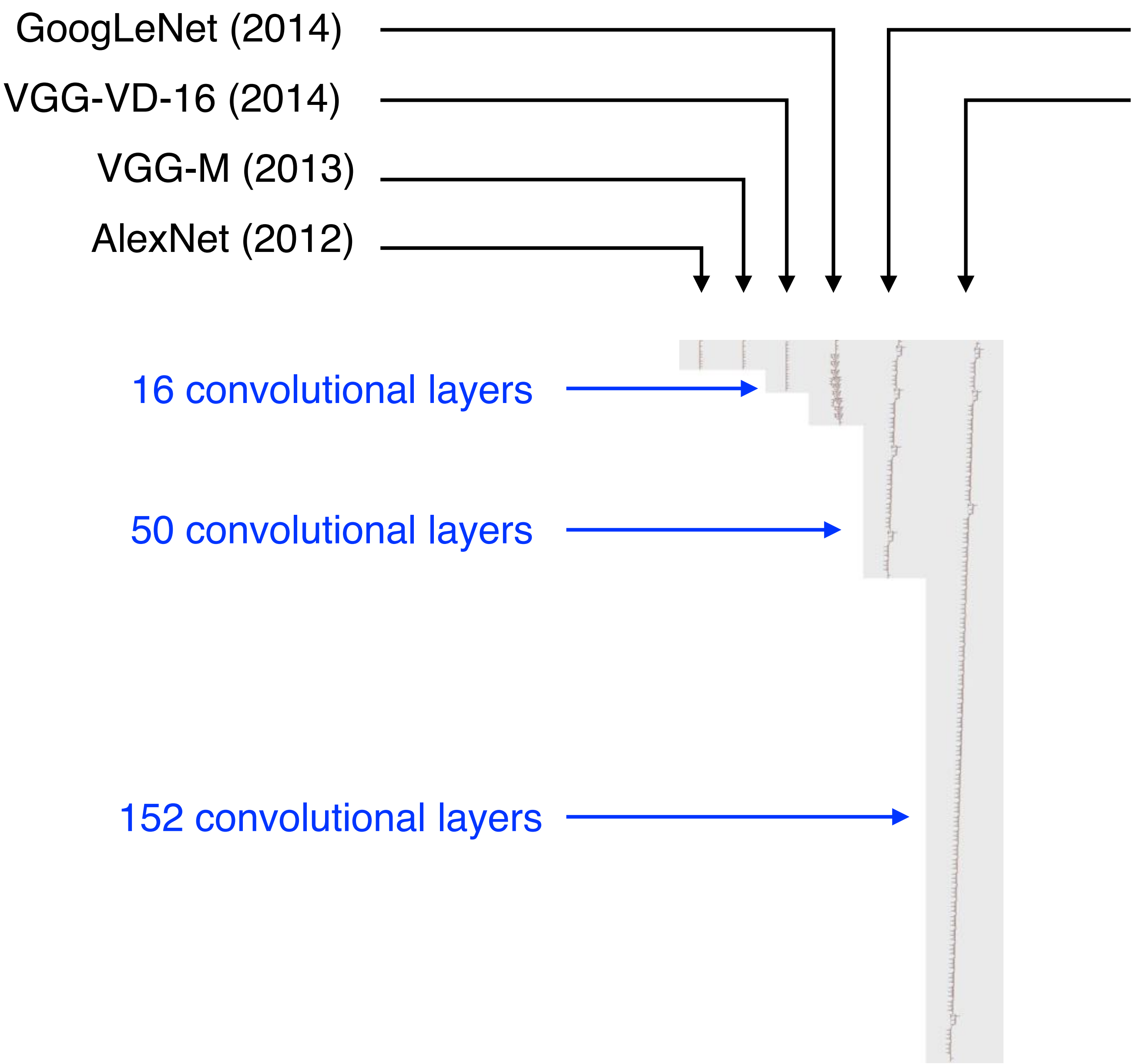
3 fully-connected layers

How deep is enough?

AlexNet (2012) VGG-M (2013) VGG-VD-16 (2014) GoogLeNet (2014)



How deep is enough?



Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. In Proc. NIPS, 2012.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions*. In Proc. CVPR, 2015.

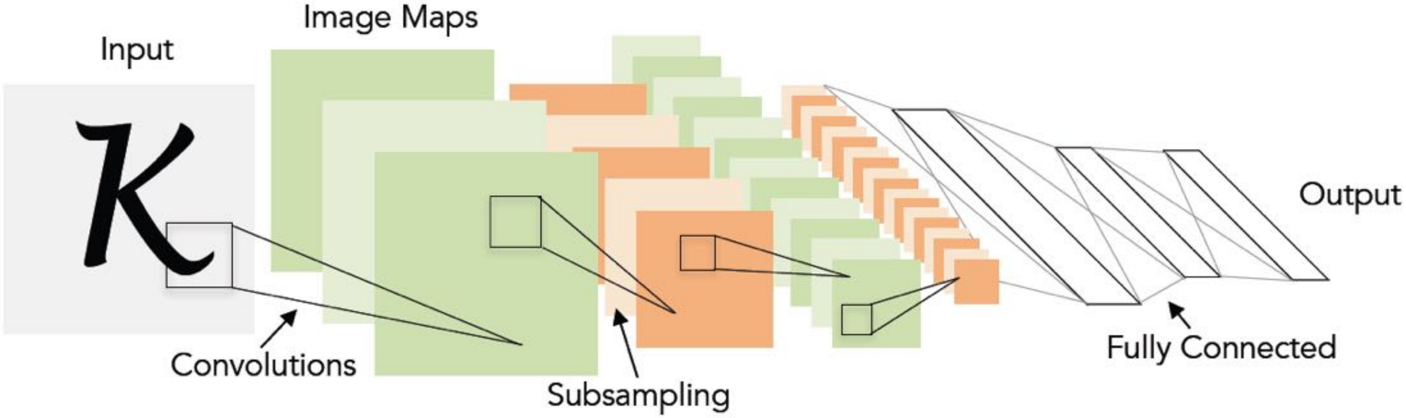
K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. In Proc. ICLR, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In Proc. CVPR, 2016.

Convolutional Neural Networks (CNNs) were not invented overnight...

1998

LeCun et al.



# of transistors

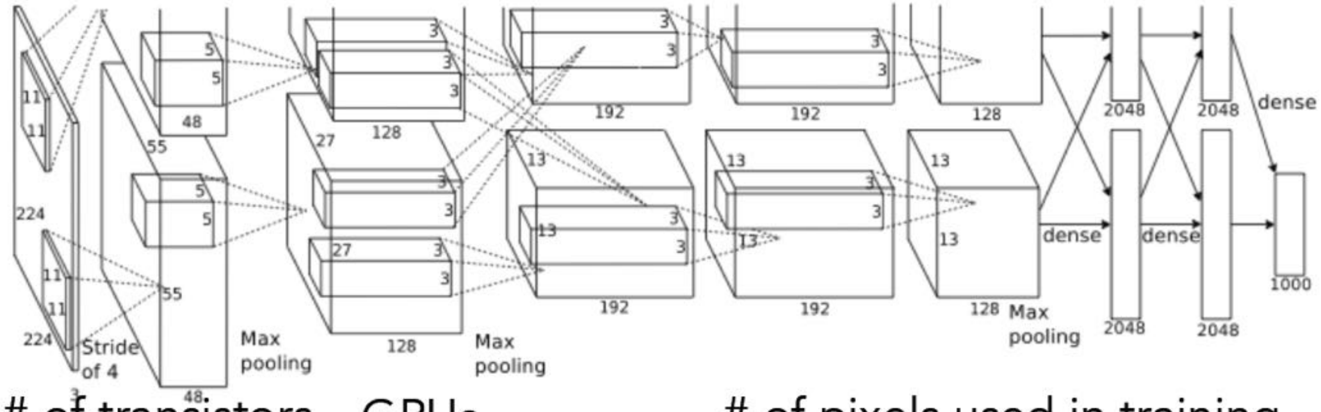
10^6

# of pixels used in training

10^7 NIST

2012

Krizhevsky et al.



# of transistors

10^9

GPUs

# of pixels used in training

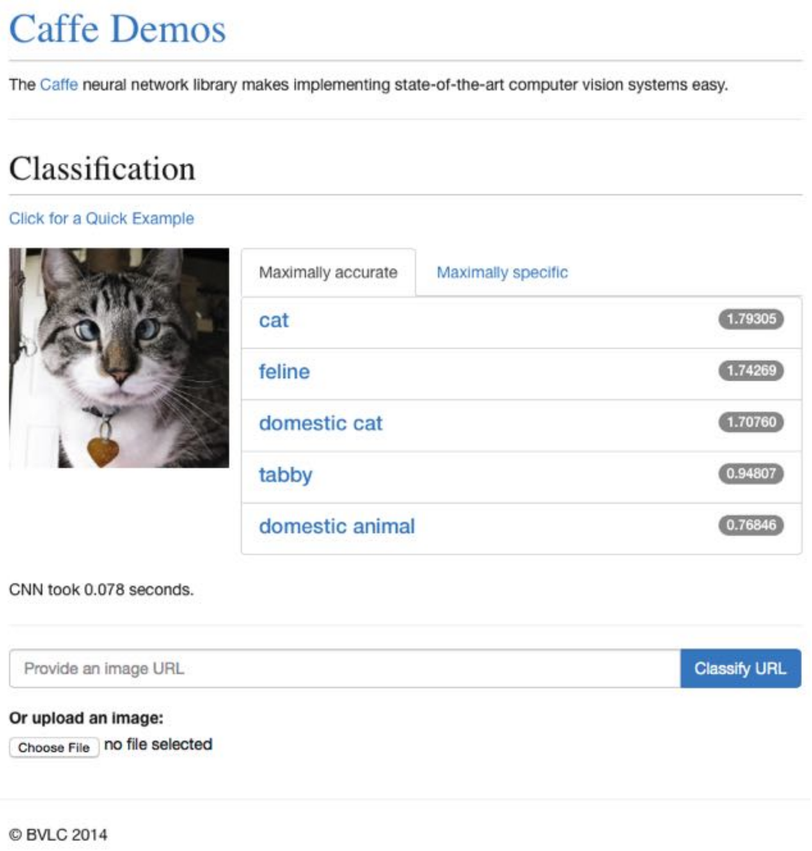
10^{14} IMAGENET

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

Try it out yourself

- Caffe ist an open implementation from the Berkeley Vision Group
  - http://caffe.berkeleyvision.org
  - http://demo.caffe.berkeleyvision.org



Caffe Demos

The Caffe neural network library makes implementing state-of-the-art computer vision systems easy.

Classification

Click for a Quick Example

cat	1.79305
feline	1.74269
domestic cat	1.70760
tabby	0.94807
domestic animal	0.76846

CNN took 0.078 seconds.

Provide an image URL

Or upload an image:

no file selected

© BVLC 2014

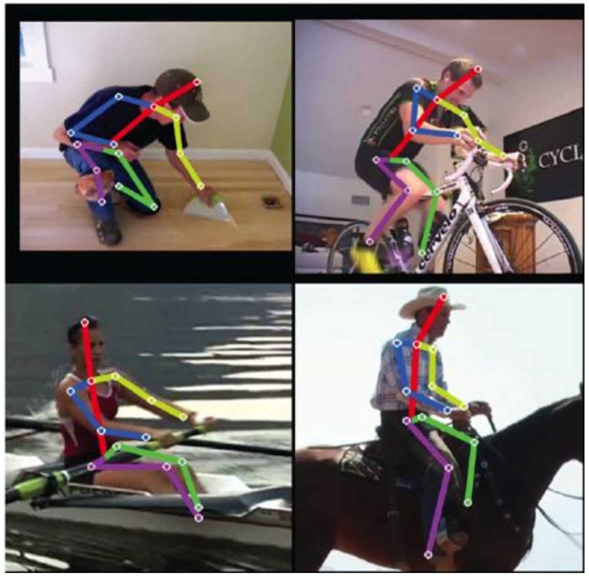
Deep Learning
have become an important tool
for object recognition / image classification

but there exist many other computer vision tasks
where Deep Learning is also an essential ingredient

a few examples...

Human Pose Estimation

- **Single Person Pose Estimation** - two "phases"
  - **Phase 1**: **pictorial structures models** e.g. [Felzenszwalb&Huttenlocher@ijcv05], [Andriluka&al@ijcv11], [Yang&Ramanan@pami13], [Pishchulin&al@iccv13], …
  - **Phase 2**: using **deep learning** e.g. [Thoshev,Szegedy@cvpr14], [Thompson&al@nips14], [Chen&Yuille@nips14], [Carreira&al@cvpr16], [Hu&Ramanan@cvpr16], [Wei&al@cvpr16], [Newell&al@cvpr16], …



[Andriluka,Pishchulin,Gehler,Schiele@CVPR'14]

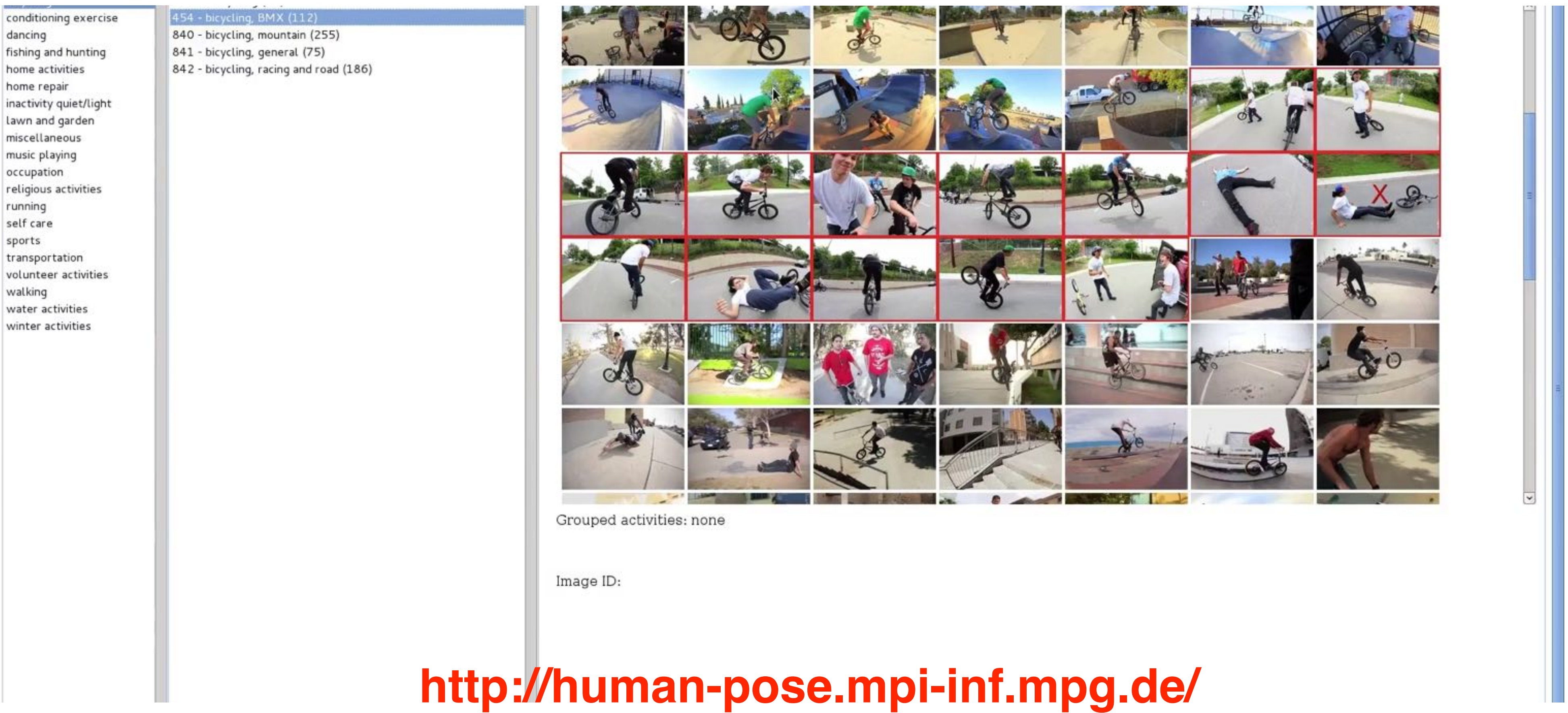
MPII Human Pose Dataset: Dataset demo

- 410 human activities (after merging similar activities)
- over 40,000 annotated poses
- over 1.5M video frames

Activity Categories

Activities

Images



conditioning exercise
dancing
fishing and hunting
home activities
home repair
inactivity quiet/light
lawn and garden
miscellaneous
music playing
occupation
religious activities
running
self care
sports
transportation
volunteer activities
walking
water activities
winter activities

454 - bicycling, BMX (112)
840 - bicycling, mountain (255)
841 - bicycling, general (75)
842 - bicycling, racing and road (186)

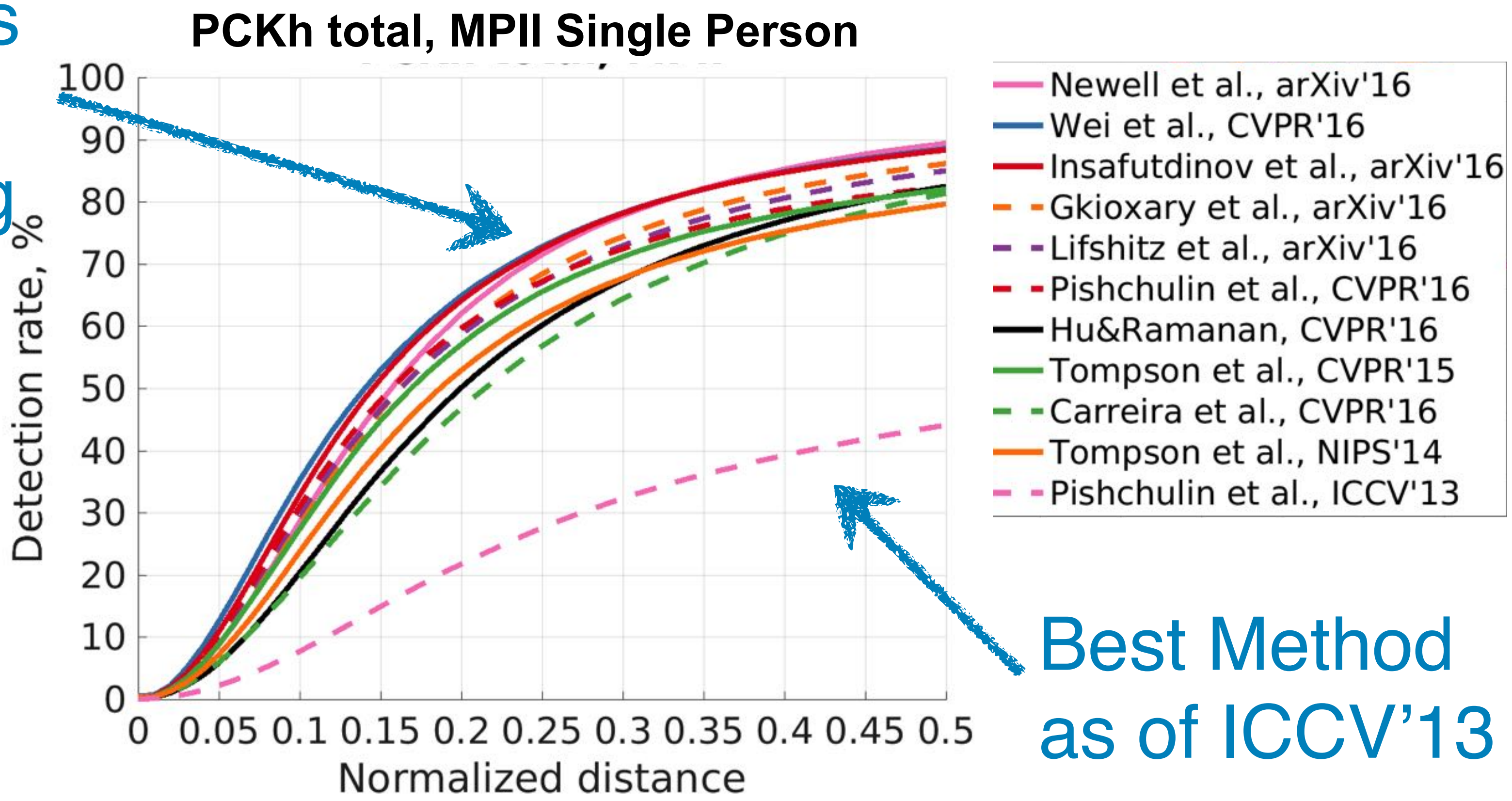
Grouped activities: none

Image ID:

http://human-pose.mpi-inf.mpg.de/

Analysis - overall performance

Best Methods today: deep learning “takes” over

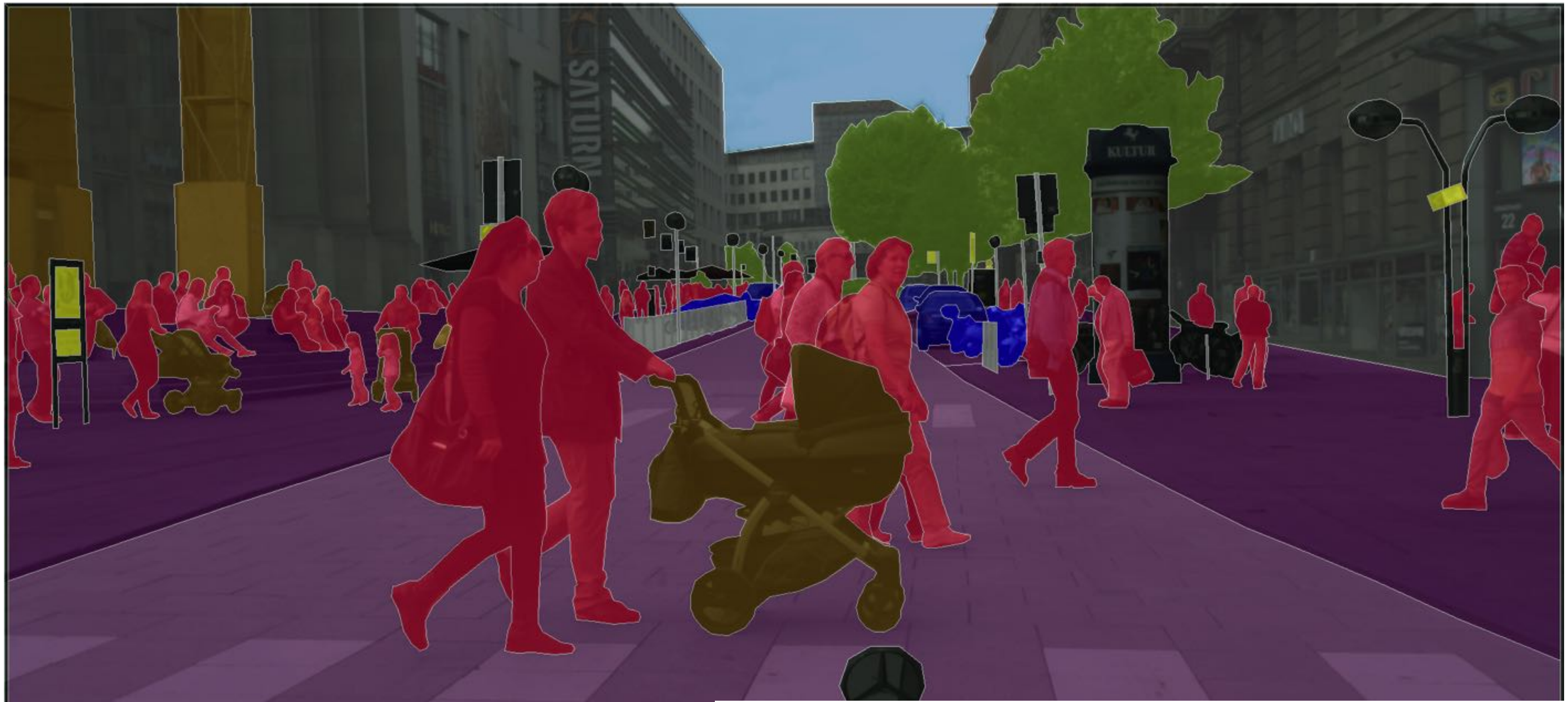


Best Method as of ICCV’13

- ✓ since CVPR’14, dataset has become **de-facto standard benchmark**
- ✓ **large training set** facilitated development of **deep learning methods**

[Cordts,Omran,Ramos,Rehfeld,Enzweiler,Benenson,Franke,Roth,Schiele@cvpr16]

Cityscapes: Large-Scale Datasets for Semantic Labeling of Street Scenes

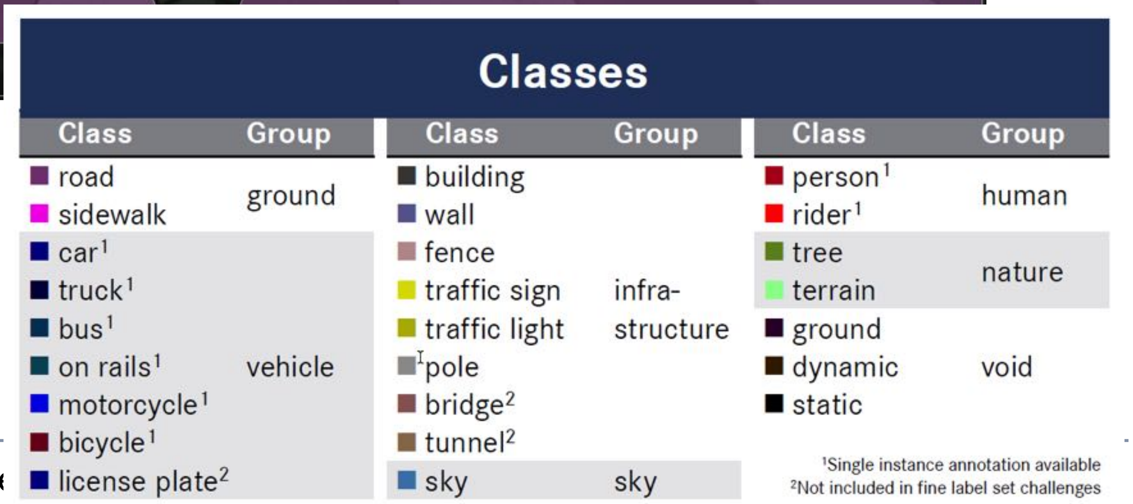


- Joint effort of:

DAIMLER

max planck institut informatik

TECHNISCHE UNIVERSITÄT DARMSTADT



Classes

Class	Group	Class	Group	Class	Group
■ road	ground	■ building	infra-structure	■ person[1]	human
■ sidewalk		■ wall		■ rider[1]	
■ car[1]	vehicle	■ fence		■ tree	nature
■ truck[1]		■ traffic sign		■ terrain	
■ bus[1]		■ traffic light		■ ground	void
■ on rails[1]		■ pole		■ dynamic	
■ motorcycle[1]		■ bridge[2]		■ static	
■ bicycle[1]		■ tunnel[2]			
■ license plate[2]		■ sky	sky		

[1] Single instance annotation available
[2] Not included in fine label set challenges

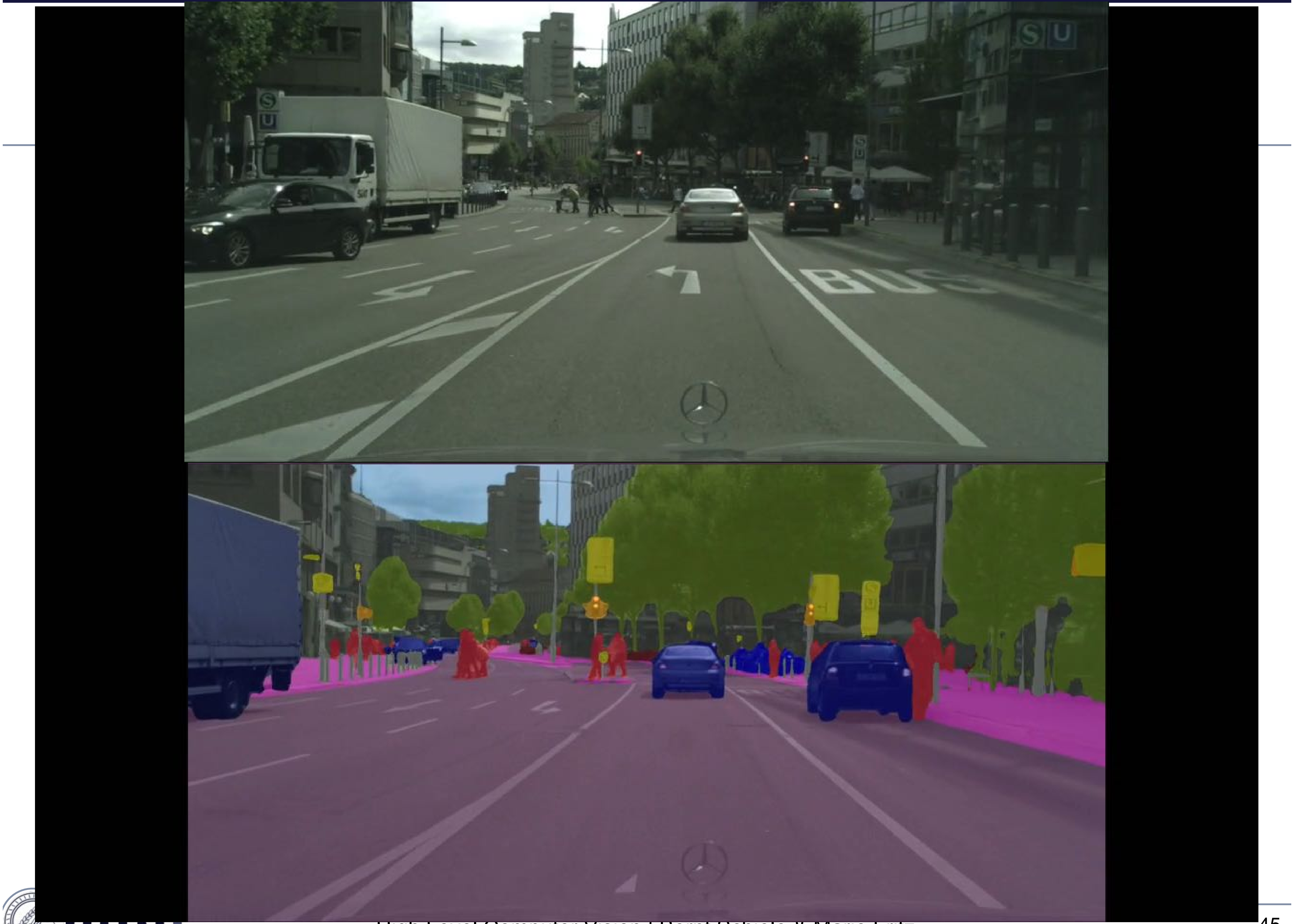

S
U
BUS

Image Description

A female tennis player in action on the court.

A group of young men playing a game of soccer.

A man riding a wave on top of a surfboard.

Image Description

Ours: a person on skis jumping over a ramp

Ours: a skier is making a turn on a course

Ours: a cross country skier makes his way through the snow

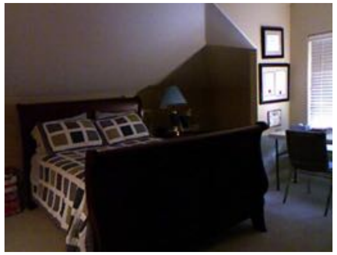
Ours: a skier is headed down a steep slope

Baseline: a man riding skis down a snow covered slope

[Rakshith'17]

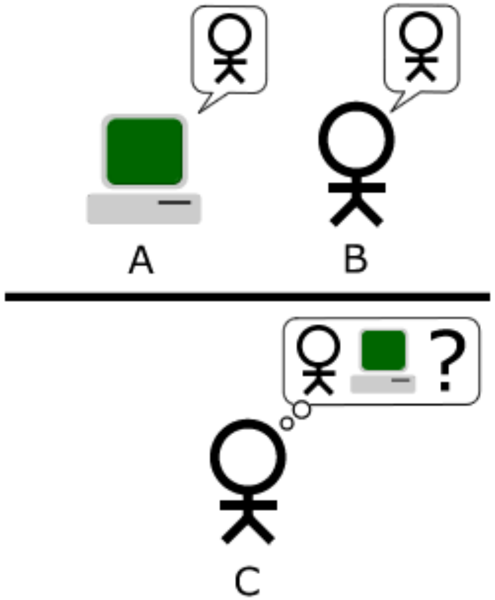
Towards a Visual Turing Challenge

Q: What is the object on the counter in the corner?
A: micro wave



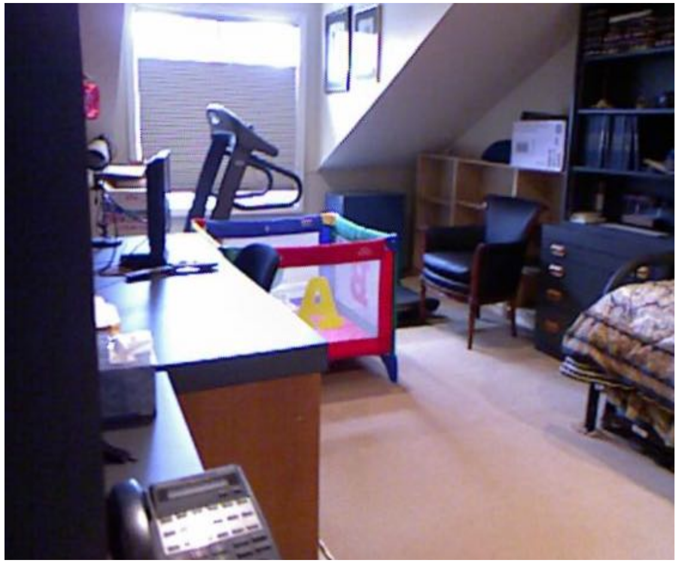
What is the color of the largest object in the scene?
A: brown

Q:How many lights are on?
A: 6



- 1449 RGB-D images (NYU depth dataset)
- 12500 question-answer-pairs
- Publicly available

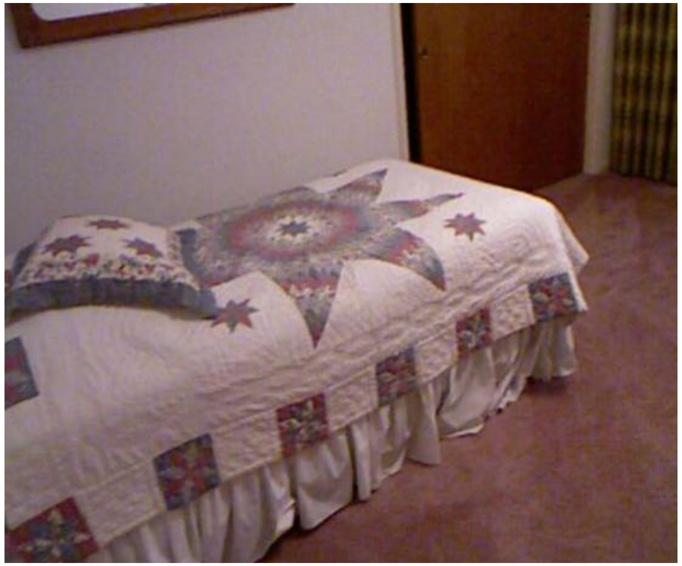
Question Answering Results



What is on the right side of the cabinet?

Vision + Language: **bed**

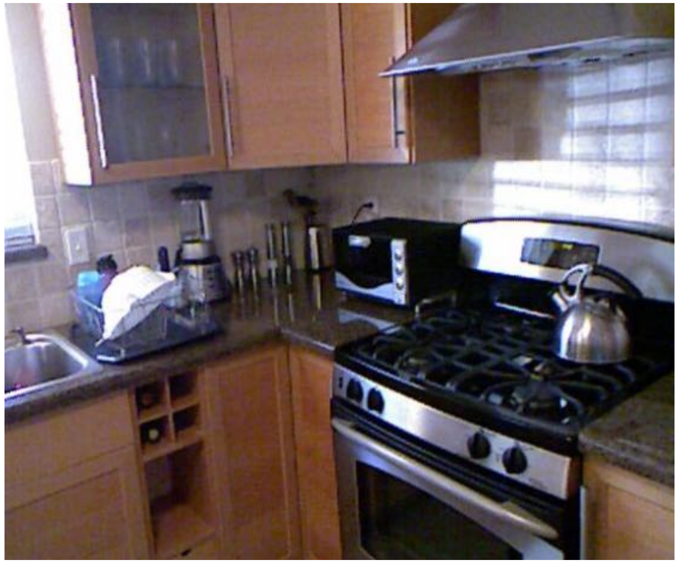
Language Only: **bed**



What objects are found on the bed?

Vision + Language: **bed sheets, pillow**

Language Only: **doll, pillow**



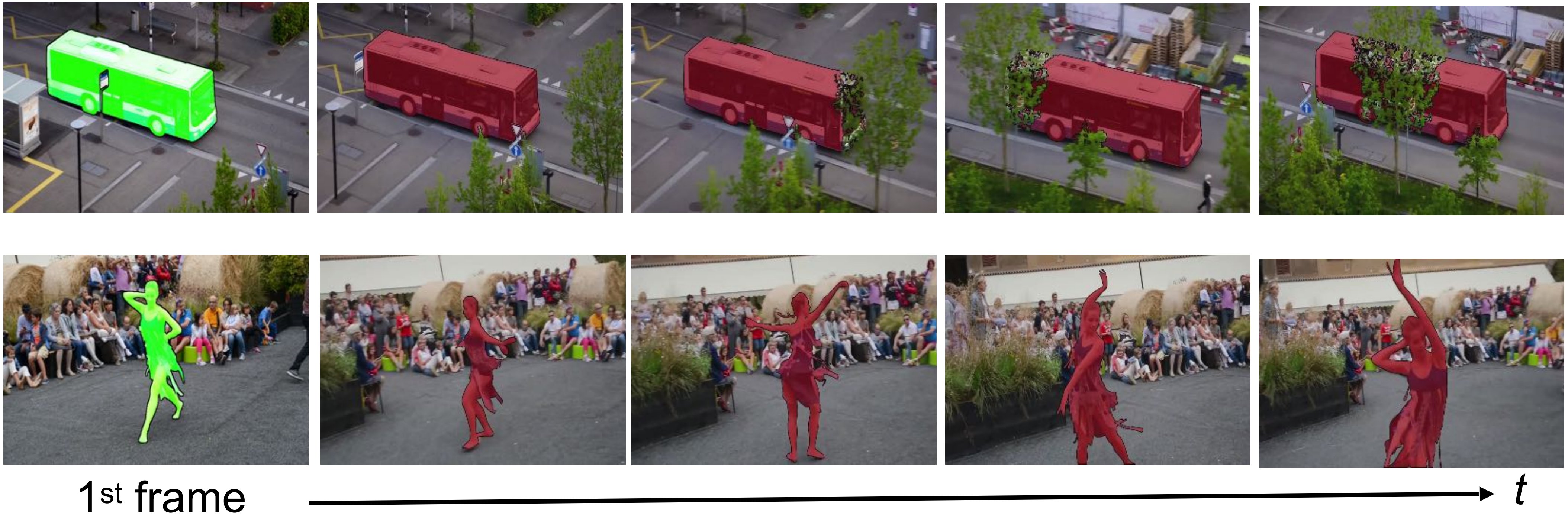
How many burner knobs are there?

Vision + Language: **4**

Language Only: **6**

Video Object Segmentation

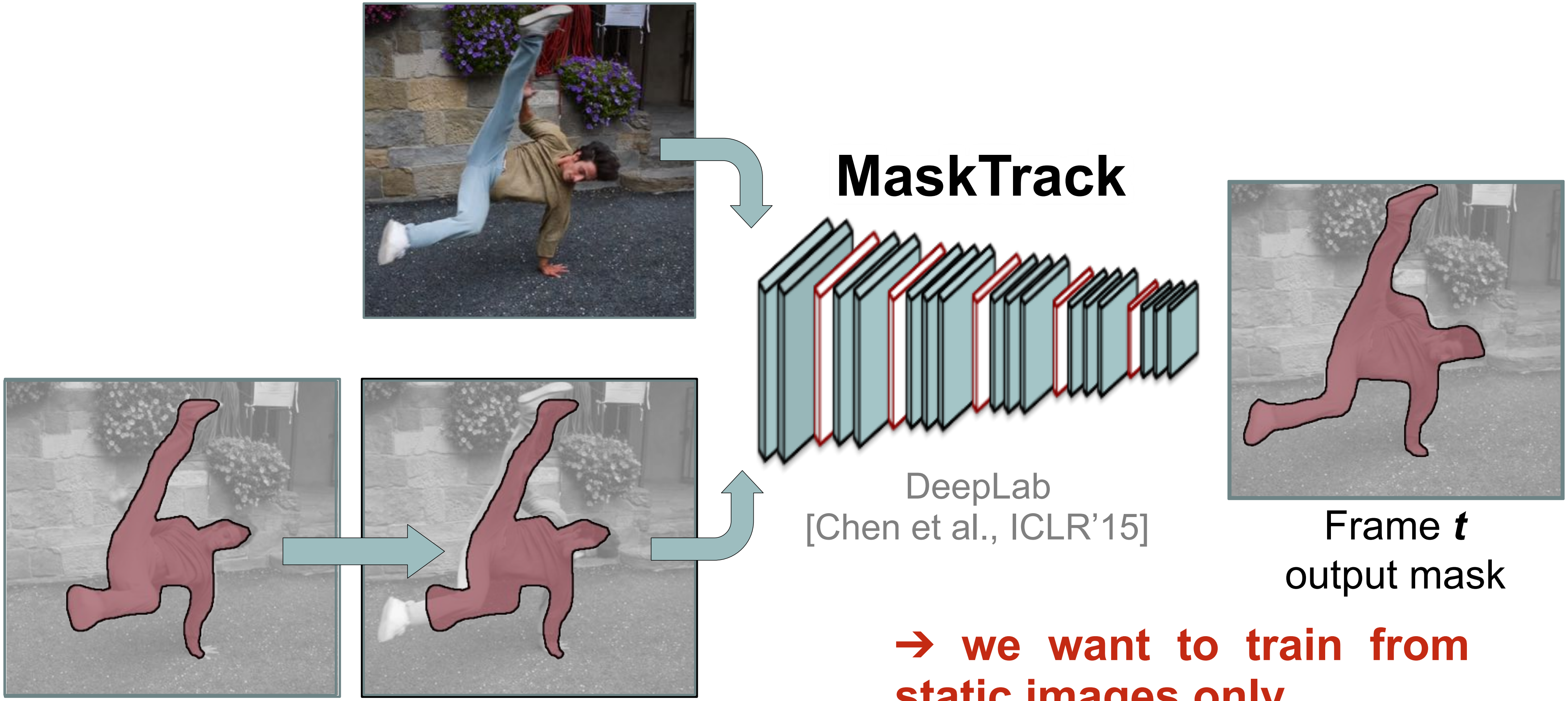
Goal: Separating a specific **foreground object** from **background** in a video given its **1st frame mask annotation**.



DAVIS 2016
[Perazzi et al.'16]

MaskTrack - Proposed Approach

➔ we process video per-frame, using guidance from previous frame



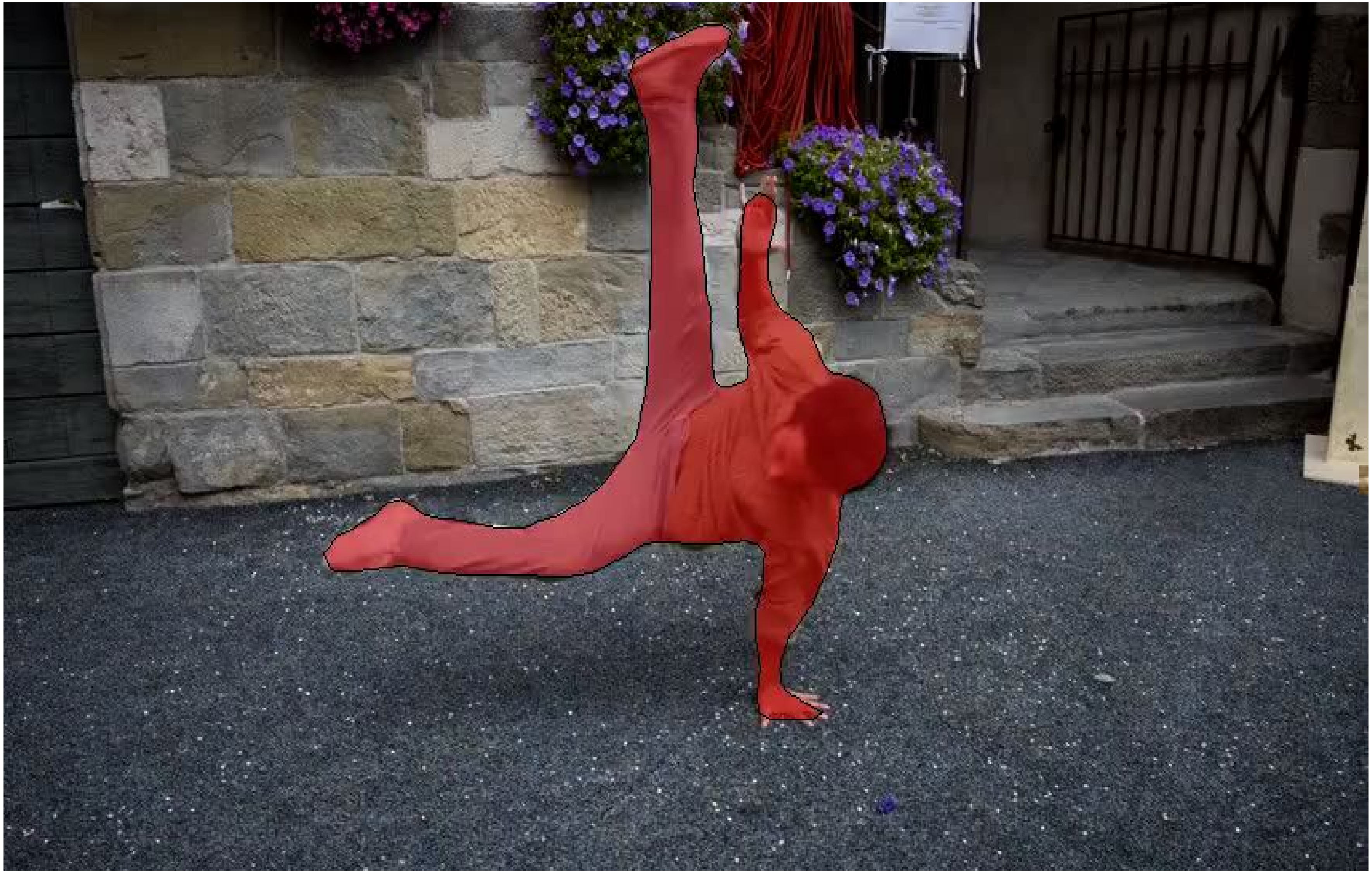
Frame *t-1*
output mask

Frame *t*
input

Frame *t*
output mask

➔ we want to train from static images only

Qualitative Results



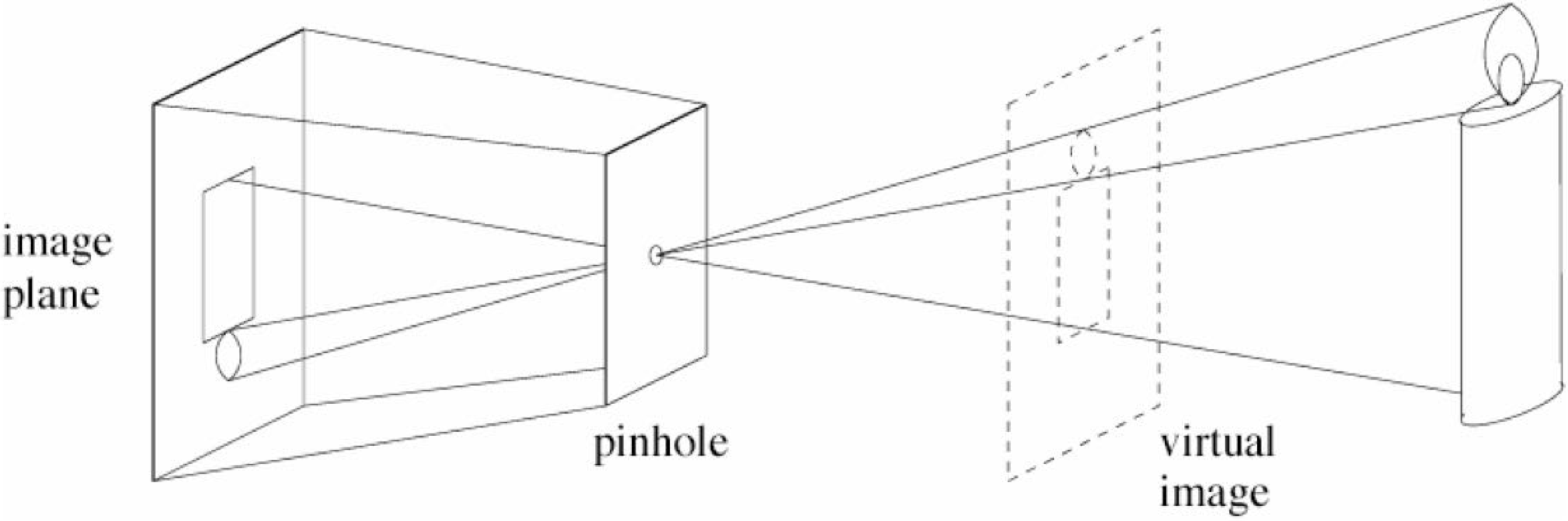
https://www.mpi-inf.mpg.de/masktrack

Basic Concepts and Terminology

Computer Vision vs. Computer Graphics

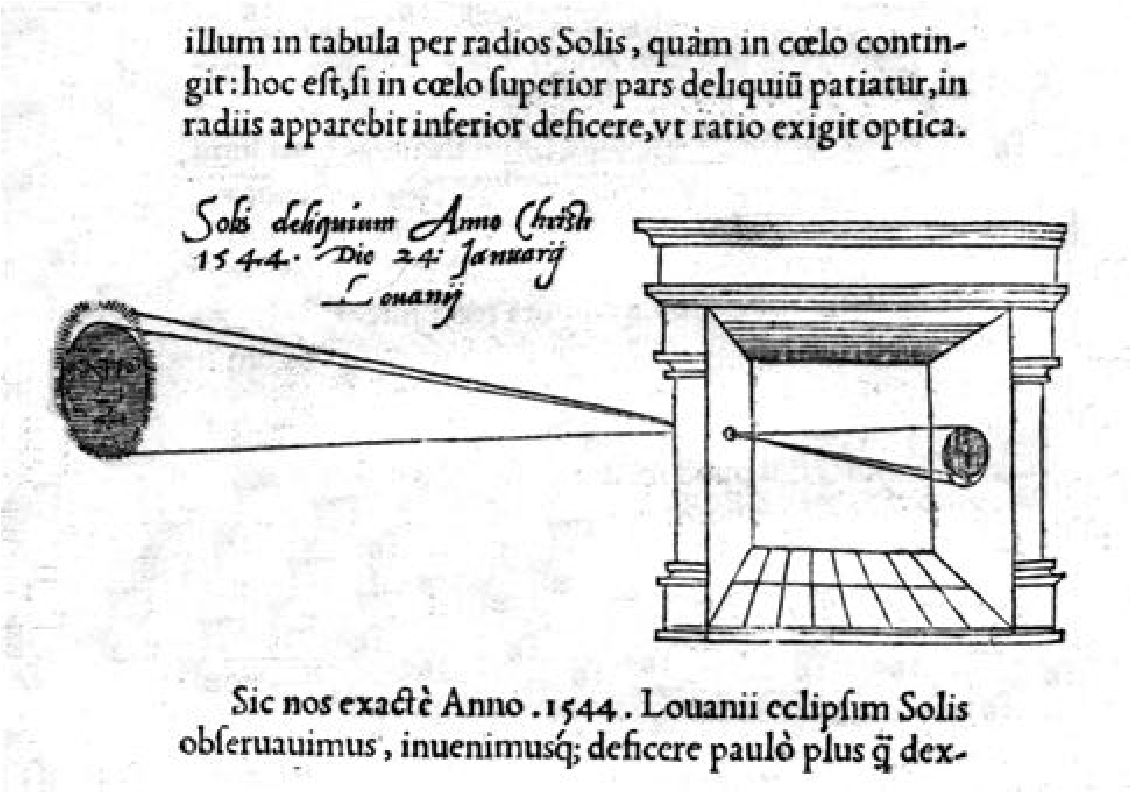
Pinhole Camera (Model)

- (simple) standard and abstract model today
  - box with a small hole in it



Camera Obscura

- around 1519, Leonardo da Vinci (1452 - 1519)
  - http://www.acmi.net.au/AIC/CAMERA_OBSCURA.html
  - “when images of illuminated objects … penetrate through a small hole into a very dark room … you will see [on the opposite wall] these objects in their proper form and color, reduced in size … in a reversed position owing to the intersection of the rays”



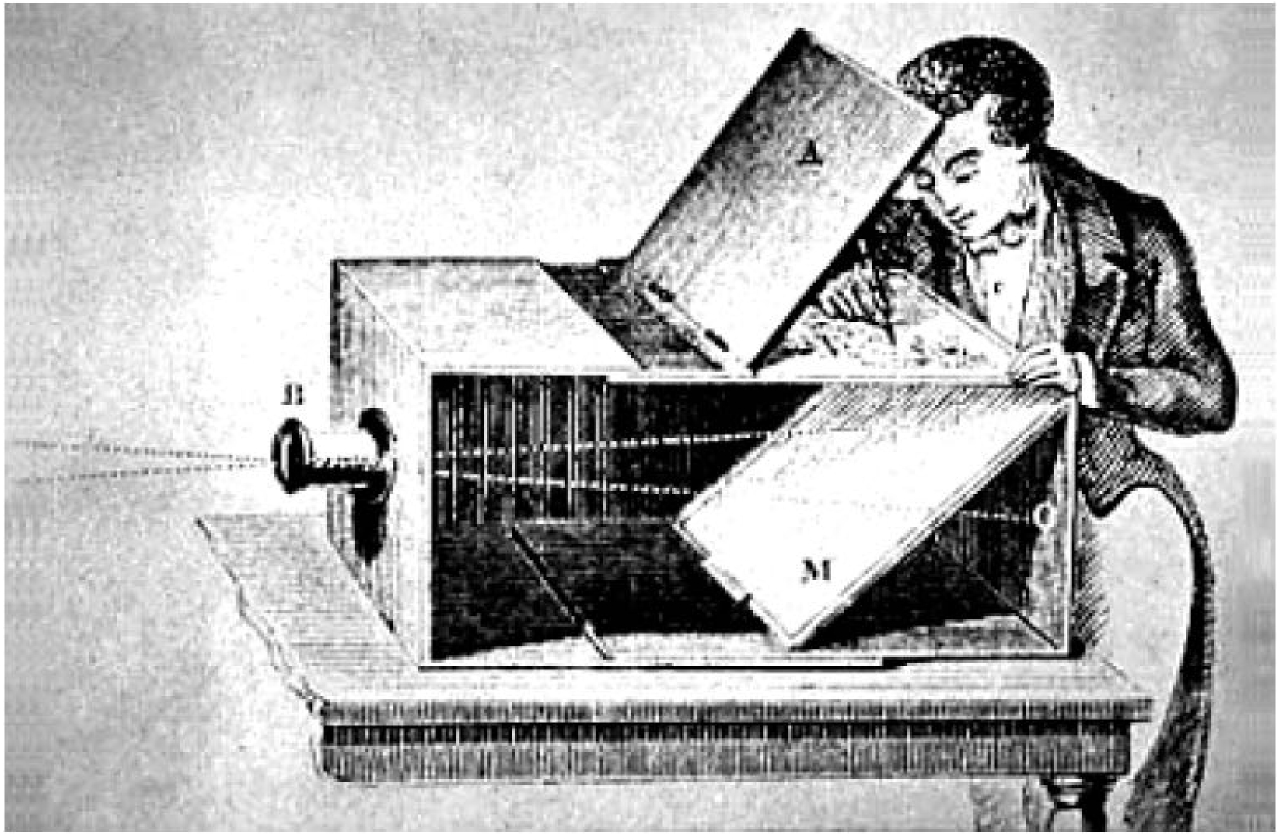
illum in tabula per radios Solis, quàm in cœlo contingit: hoc eſt, ſi in cœlo ſuperior pars deliquiũ patiatur, in radiis apparebit inferior deficere, vt ratio exigit optica.

Solis deliquium Anno Christi
1544. Die 24 Januarij
Louanij

Sic nos exactè Anno .1544. Louanii eclipſim Solis obſeruauimus, inuenimusq; deficere paulò plus q̃ dex-

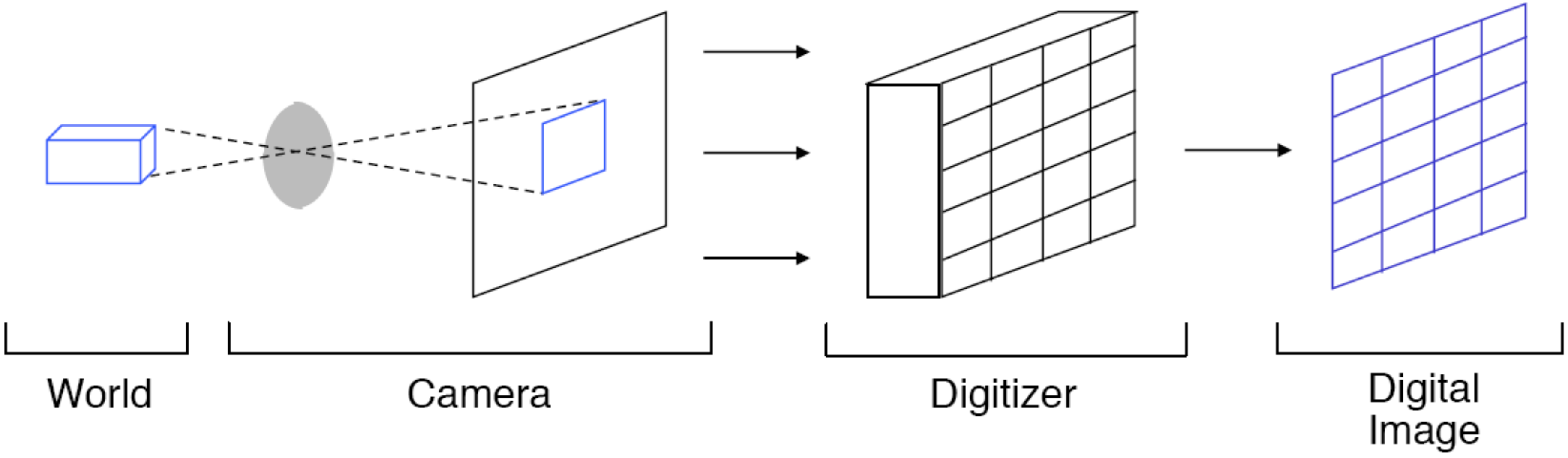
Principle of pinhole....

- ...used by artists
  - (e.g. Vermeer 17th century, dutch)
- and scientists



Digital Images

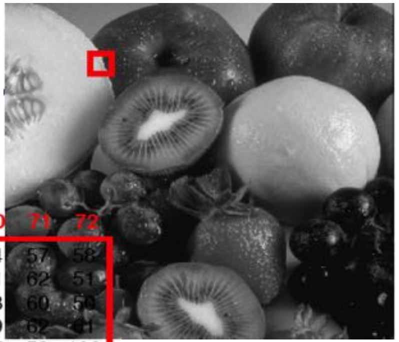
- Imaging Process:
  - (pinhole) camera model
  - digitizer to obtain digital image



(Grayscale) Image

- ‘Goals’ of Computer Vision
  - how can we recognize fruits from an array of (gray-scale) numbers?
  - how can we perceive depth from an array of (gray-scale) numbers?
  - …

- ‘Goals’ of Graphics
  - how can we generate an array of (gray-scale) numbers that looks like fruits?
  - how can we generate an array of (gray-scale) numbers so that the human observer perceives depth?
  - …



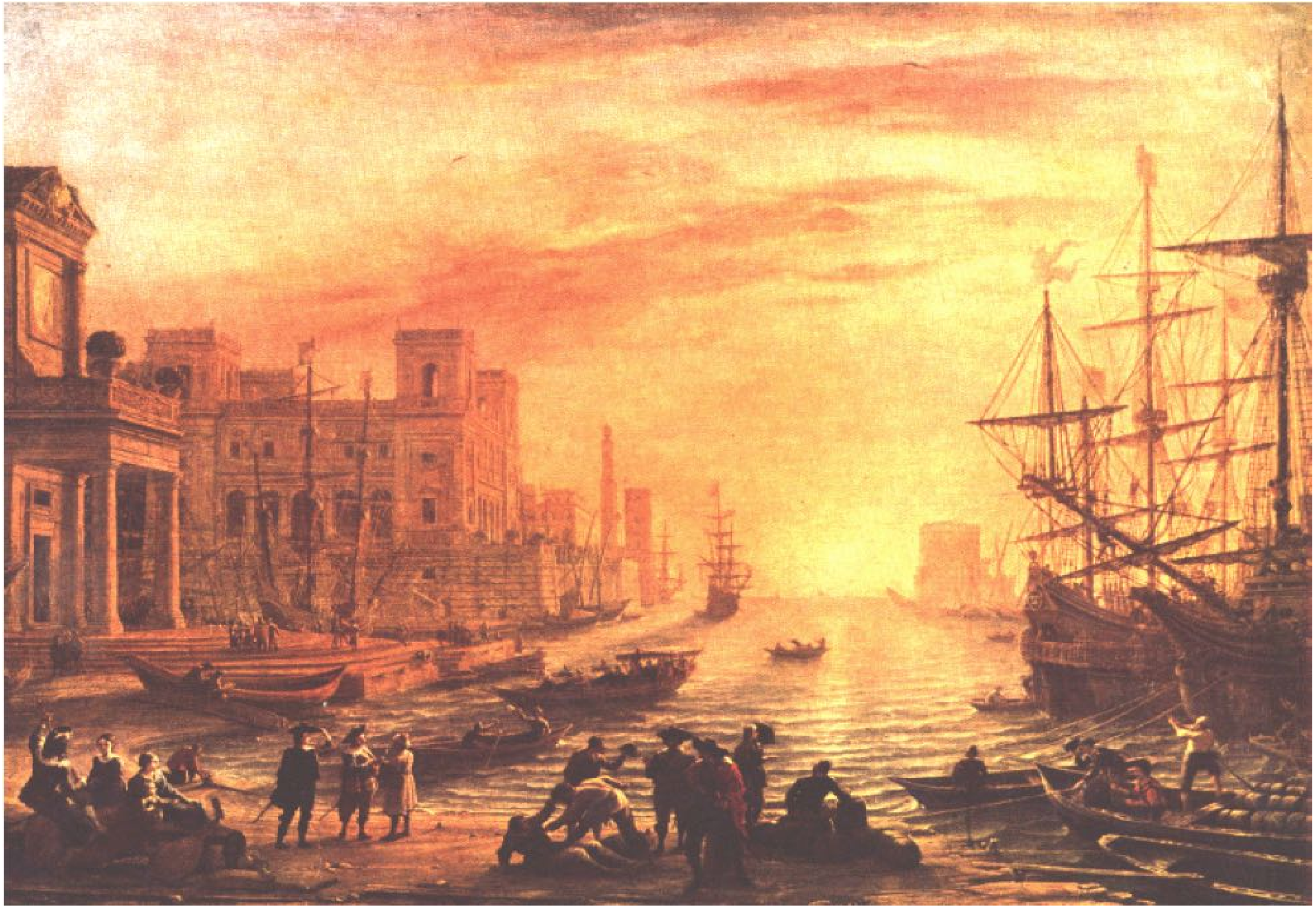
- computer vision = the problem of ‘inverse graphics’ …?

x =	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y = 41	210	209	204	202	197	247	143	71	64	80	84	54	54	57	58
42	206	196	203	197	195	210	207	56	63	58	53	53	61	62	51
43	201	207	192	201	198	213	156	69	65	57	55	52	53	60	58
44	216	206	211	193	202	207	208	57	69	60	55	77	49	62	61
45	221	206	211	194	196	197	220	56	63	60	55	46	97	58	106
46	209	214	224	199	194	193	204	173	64	60	59	51	62	56	48
47	204	212	213	208	191	190	191	214	60	62	66	76	51	49	55
48	214	215	215	207	208	180	172	188	69	72	55	49	56	52	56
49	209	205	214	205	204	196	187	196	86	62	66	87	57	60	48
50	208	209	205	203	202	186	174	185	149	71	63	55	55	45	56
51	207	210	211	199	217	194	183	177	209	90	62	64	52	93	52
52	208	205	209	209	197	194	183	187	187	239	58	68	61	51	56
53	204	206	203	209	195	203	188	185	183	221	75	61	58	60	60
54	200	203	199	236	188	197	183	190	183	196	122	63	58	64	66
55	205	210	202	203	199	197	196	181	173	186	105	62	57	64	63

Visual Cues for Image Analysis

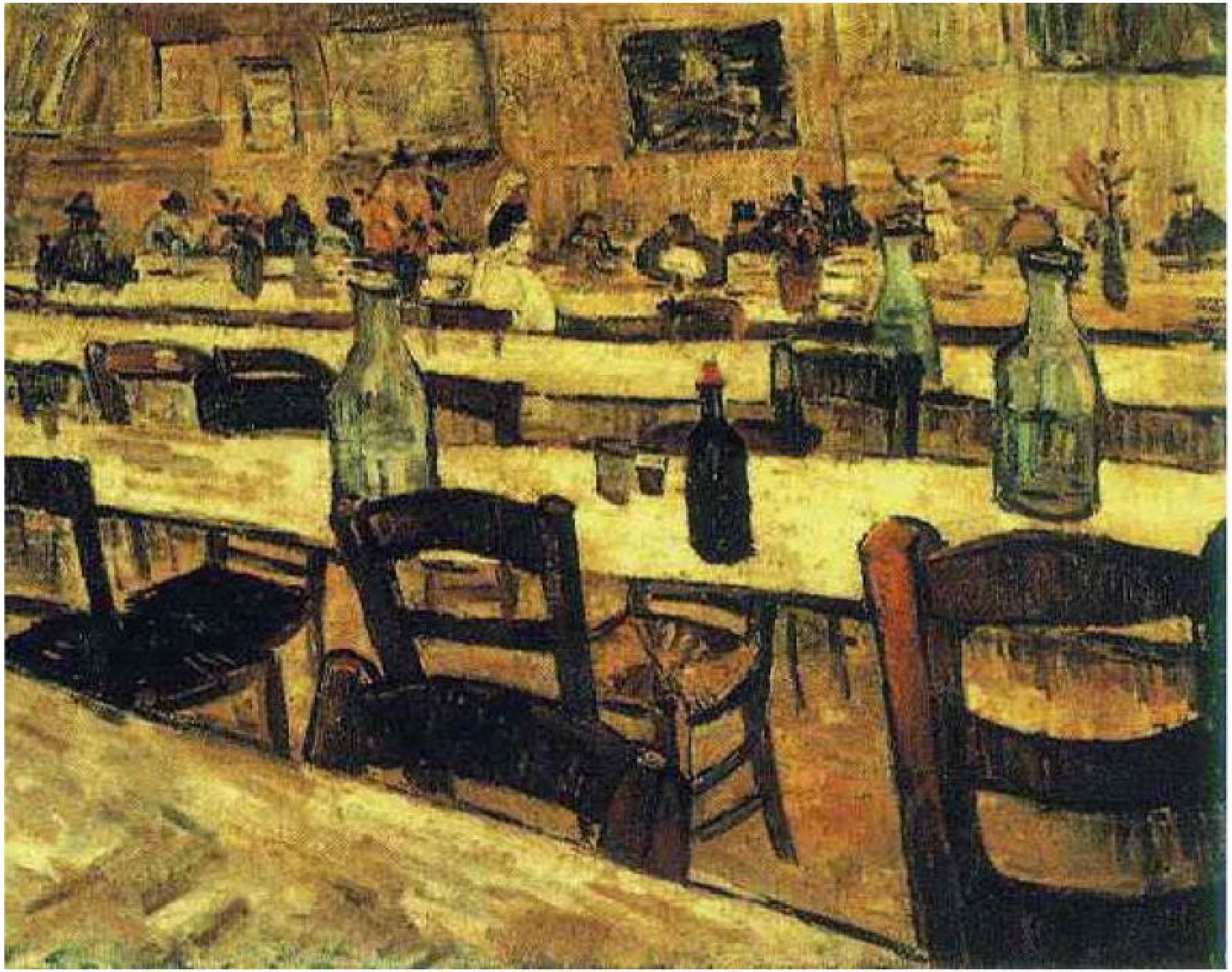
... in art and visual illusions

1. Case Study:
Human & Art - Recovery of 3D Structure



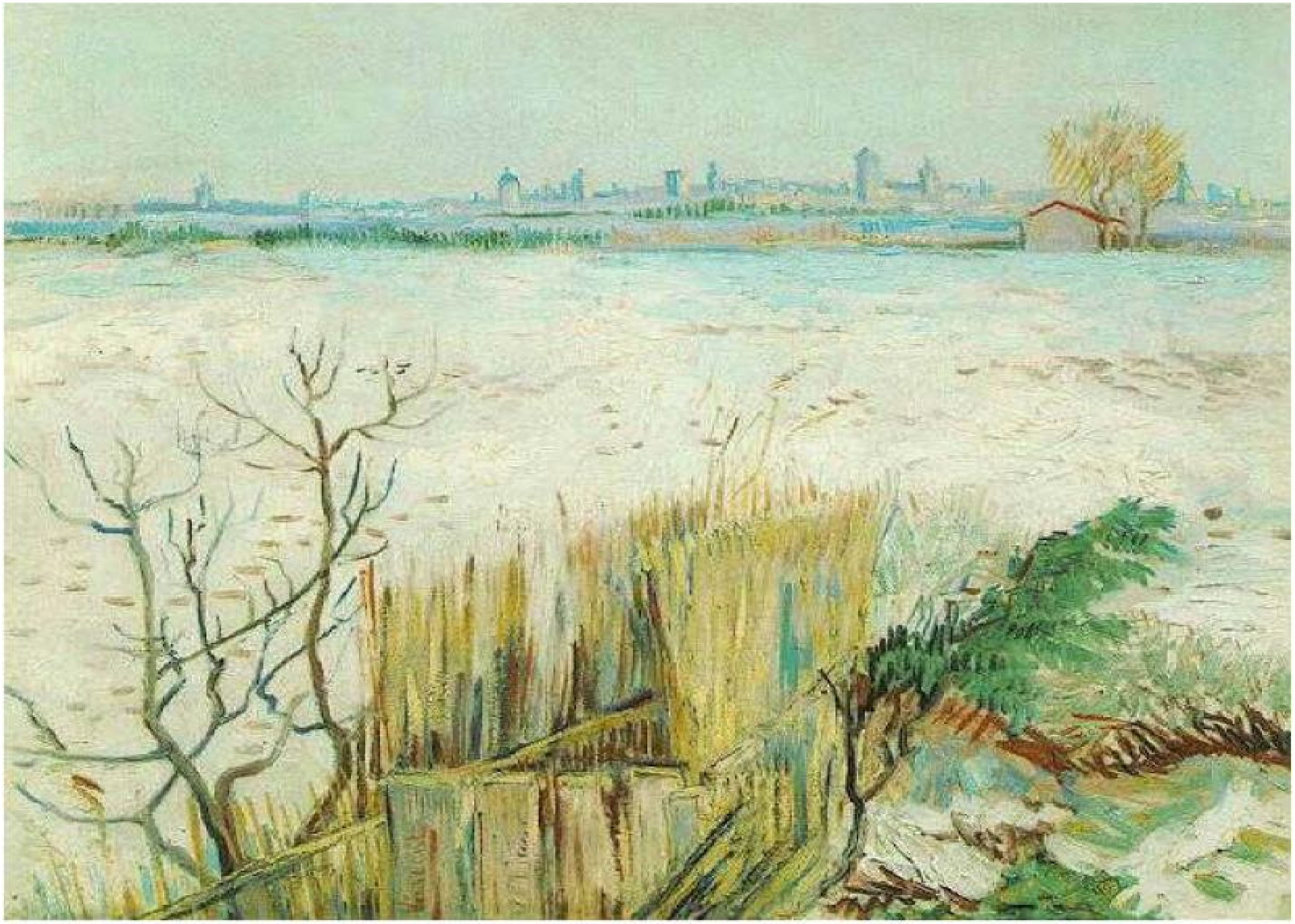
1. Case Study:
Human & Art - Recovery of 3D Structure

1. Case Study: Human & Art - Recovery of 3D Structure



Vincent van Gogh *Interior of a Restaurant at Arles* 1888

1. Case Study: Human & Art - Recovery of 3D Structure

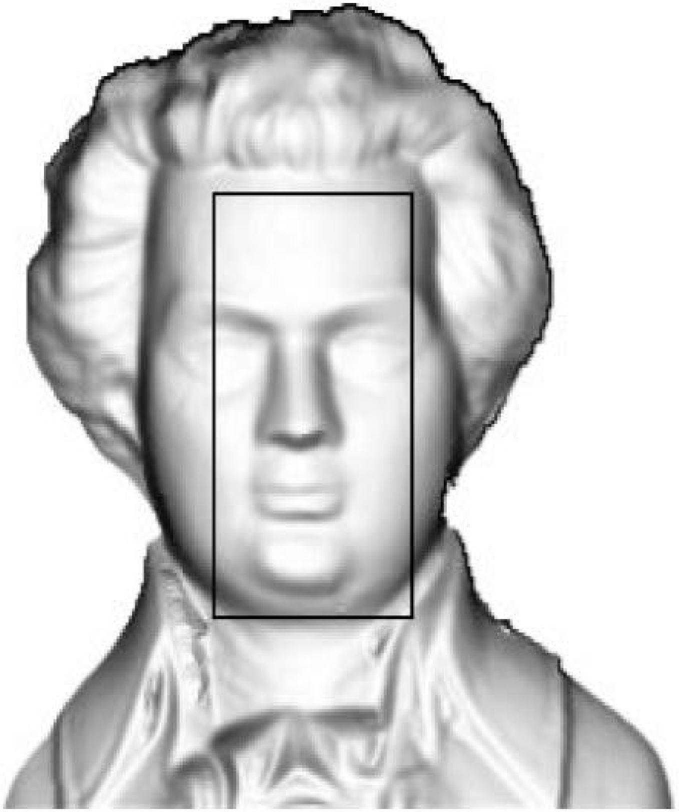


Vincent van Gogh *Snowy Landscape with Arles in the Background* 1888

1. Case Study
Computer Vision - Recovery of 3D Structure

- take all the cues of artists and 'turn them around'
  - exploit these cues to **infer** the structure of the world
  - need **mathematical** and **computational models** of these cues

- sometimes called 'inverse graphics'



http://www.vrvis.at/ar2/adm/shading/

A ‘trompe l’oeil’

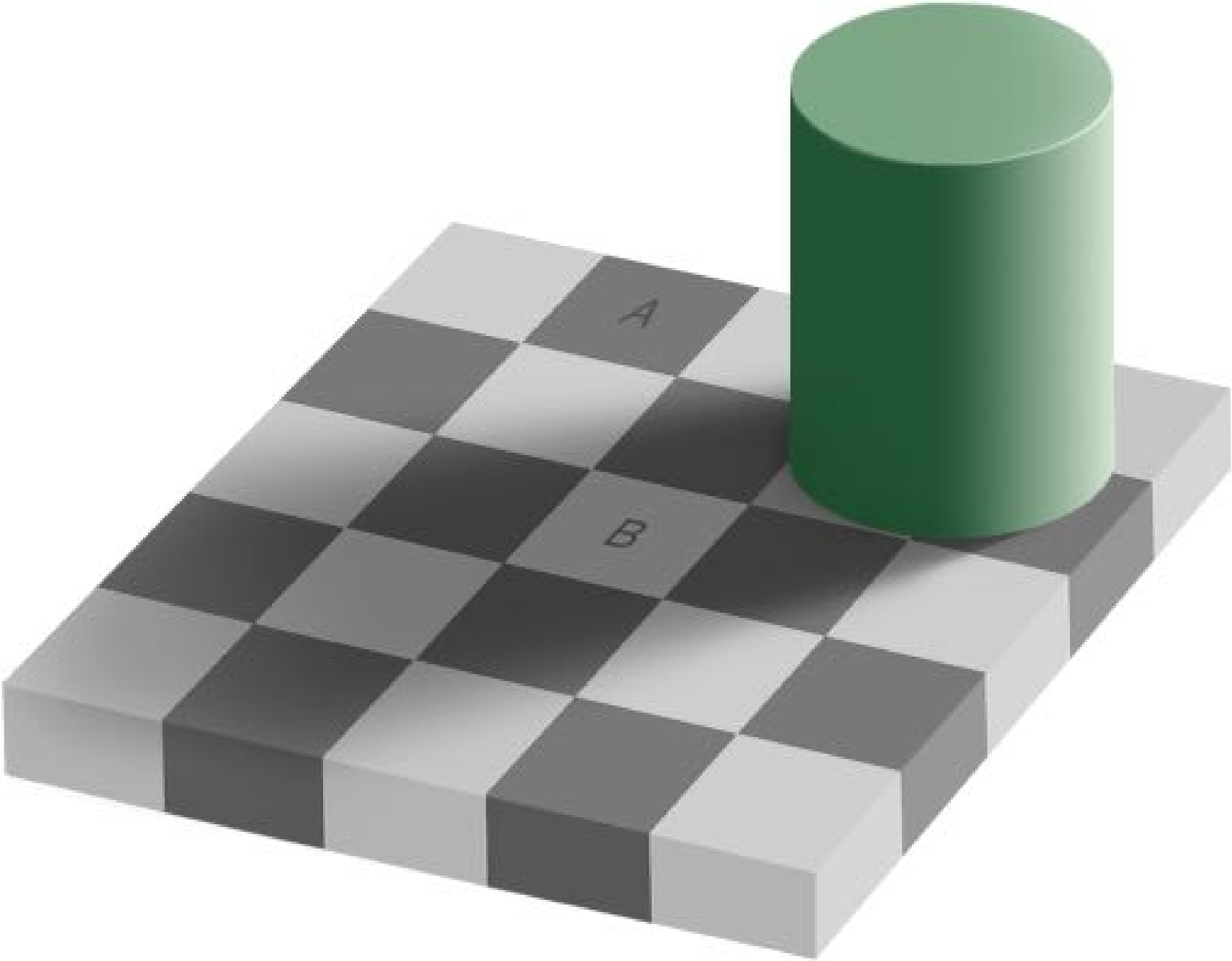
- depth-perception
  - movement of ball stays the same
  - location/trace of shadow changes

Motion in Depth

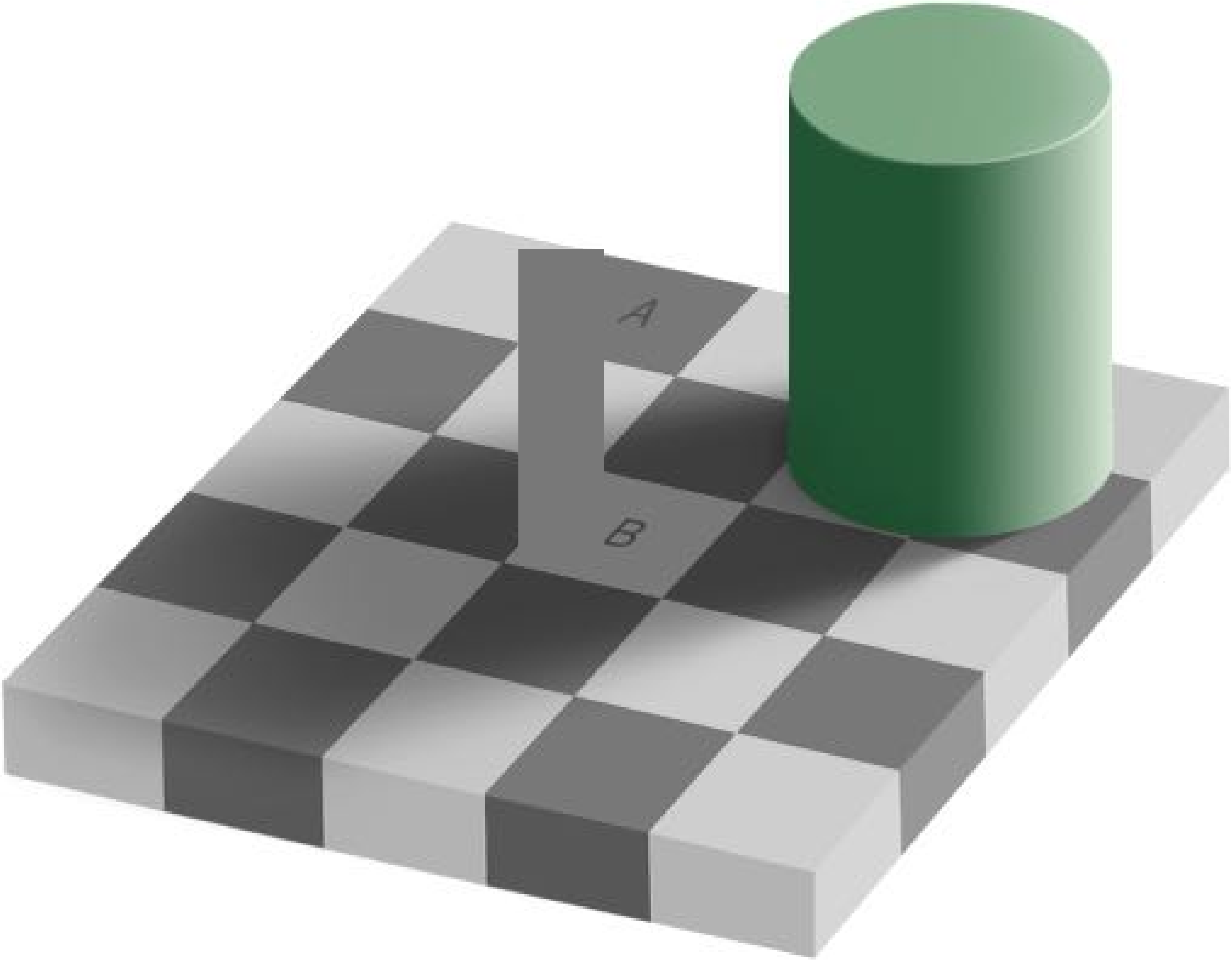
Another ‘trompe l’oeil’

- illusory motion
  - only shadows changes
  - square is stationary

Color & Shading



Color & Shading



2. Case Study: Computer Vision & Object Recognition

- is it more than inverse graphics?
- how do you recognize
  - the banana?
  - the glas?
  - the towel?
- how can we make computers to do this?
- ill posed problem:
  - missing data
  - ambiguities
  - multiple possible explanations

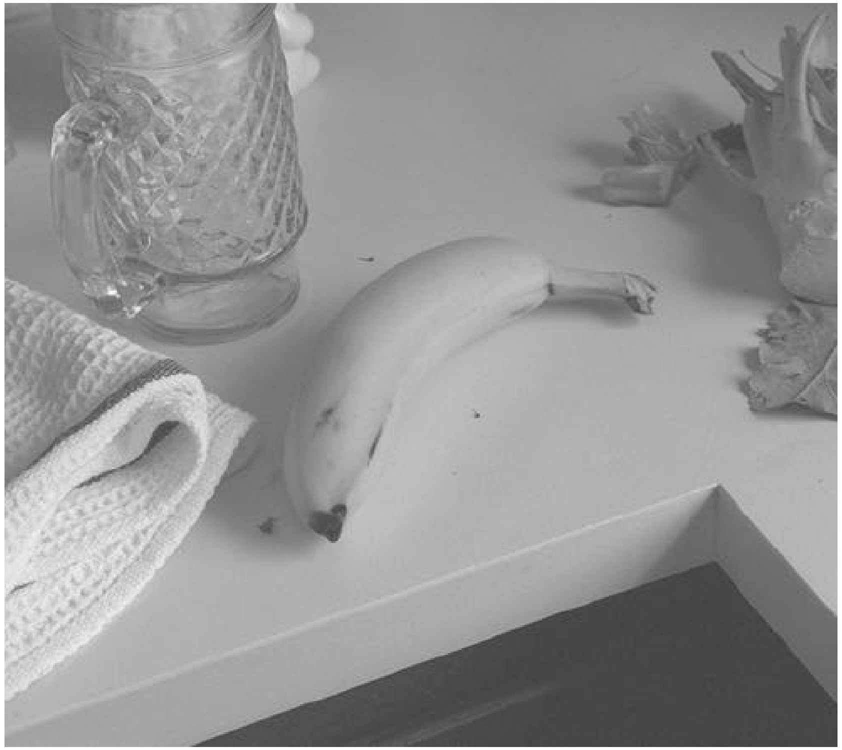
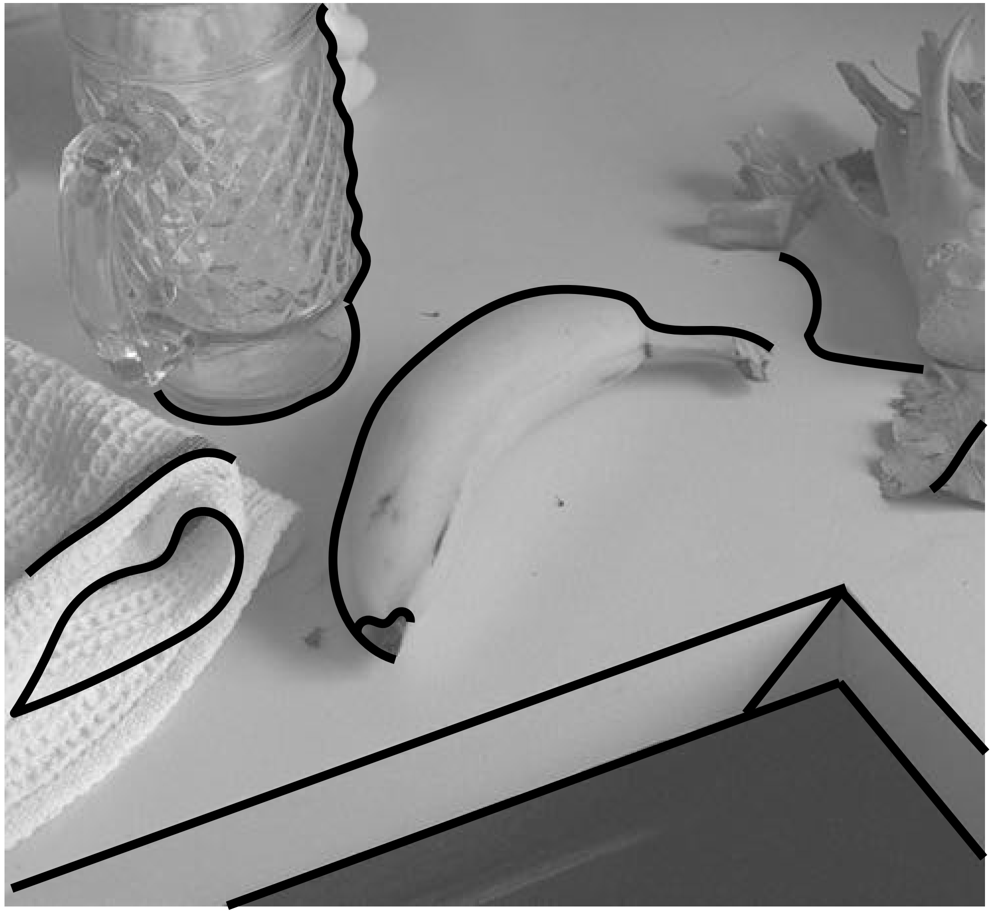
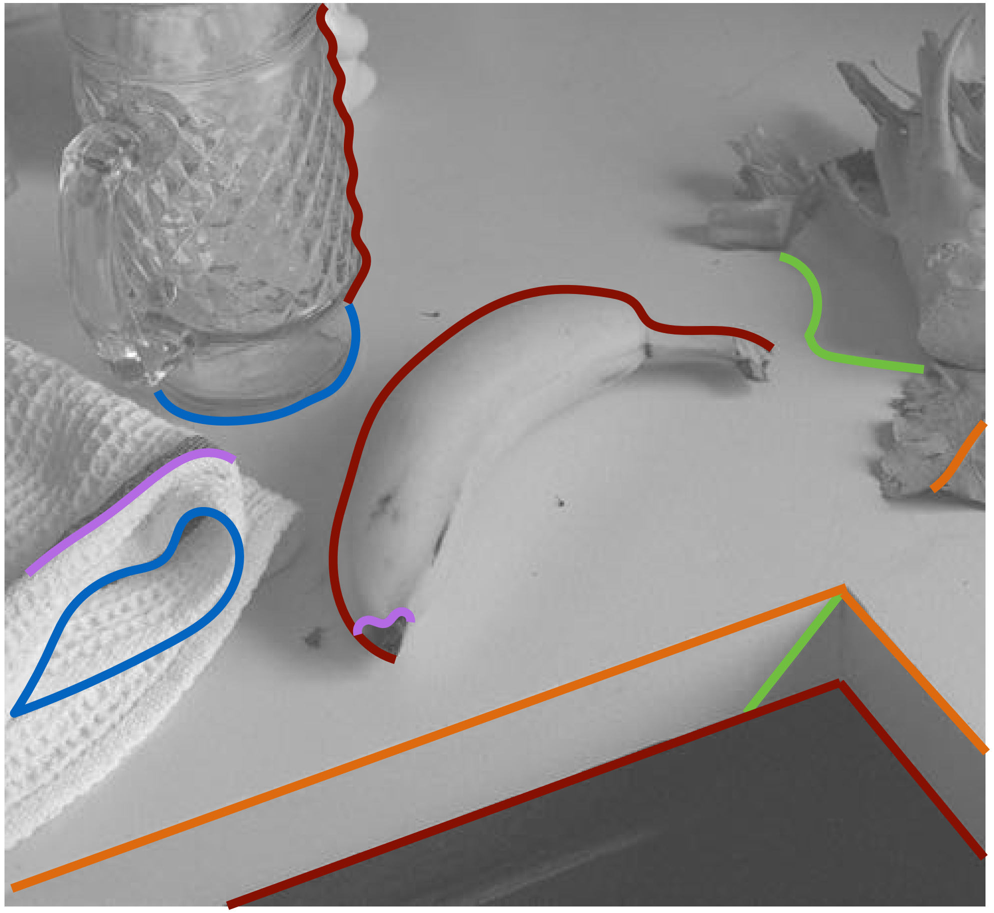


Image Edges:
What are edges? Where do they come from?



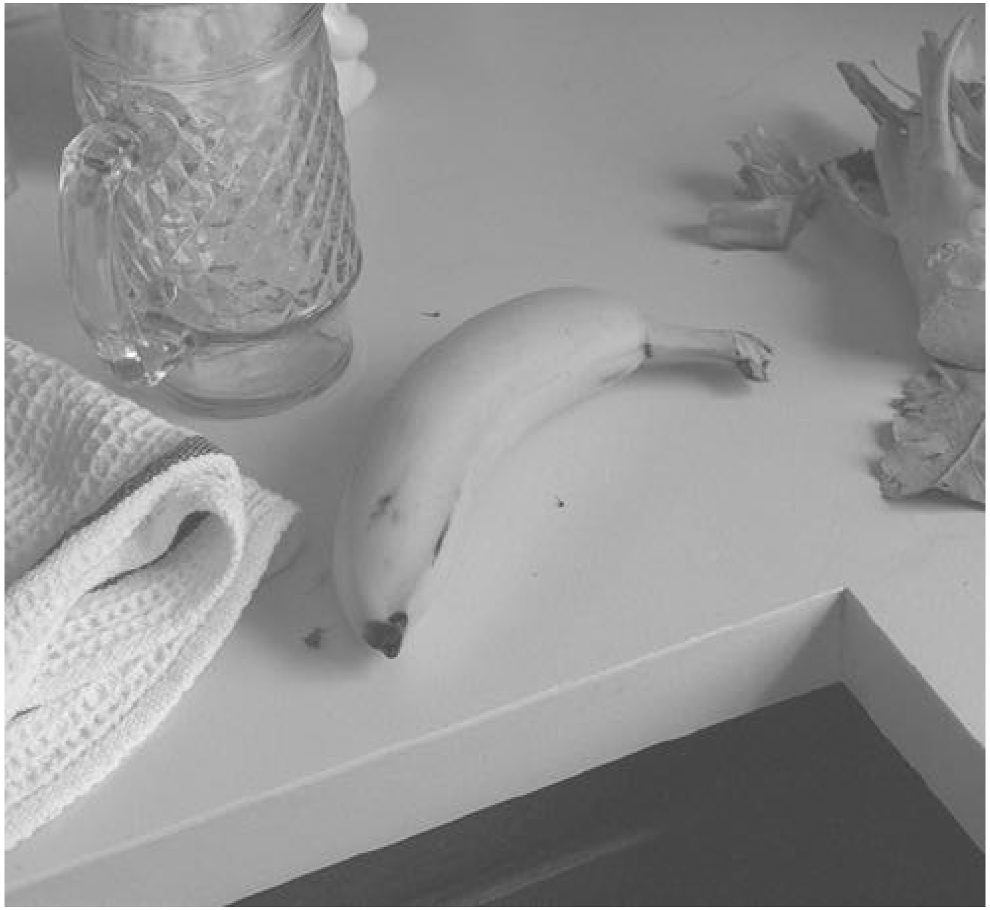
- Edges are changes in pixel brightness

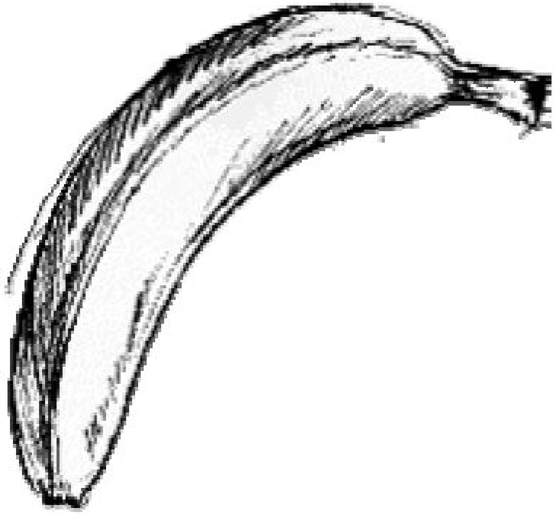
Image Edges:
What are edges? Where do they come from?

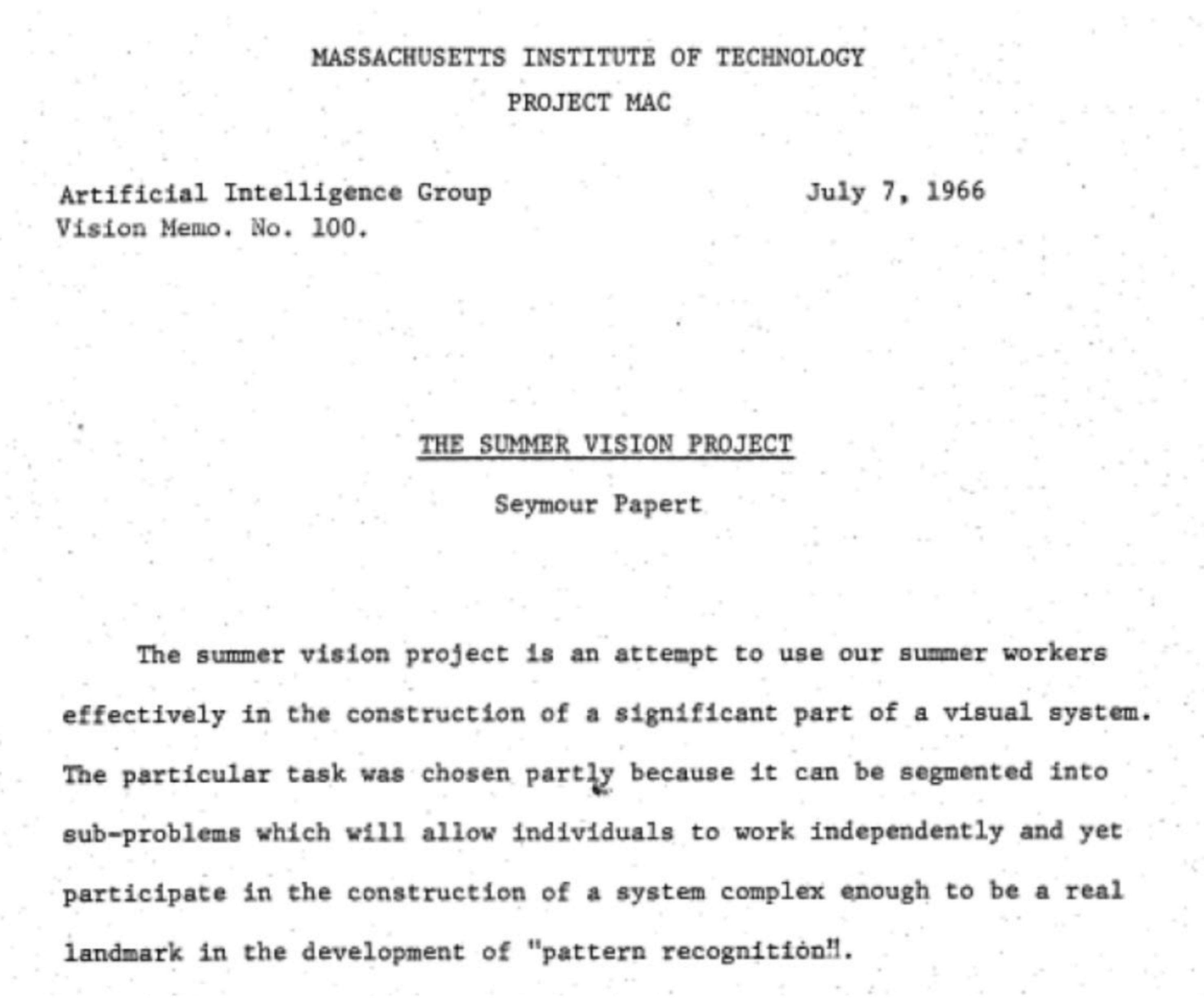


- Edges are changes in pixel brightness
  - **Foreground/Background Boundaries**
  - **Object-Object-Boundaries**
  - **Shadow Edges**
  - **Changes in Albedo or Texture**
  - **Changes in Surface Normals**

Line Drawings: Good Starting Point for Recognition?







MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

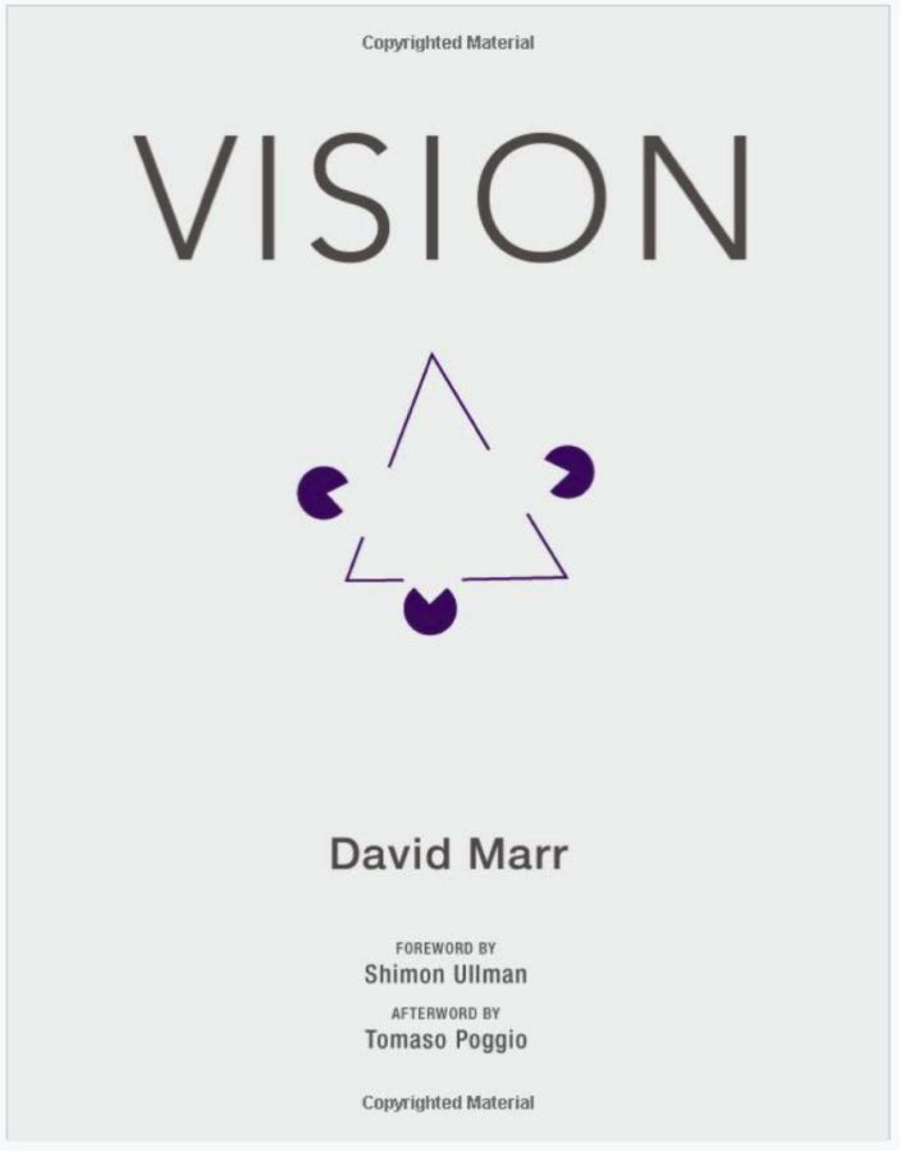
July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



Copyrighted Material

VISION

David Marr

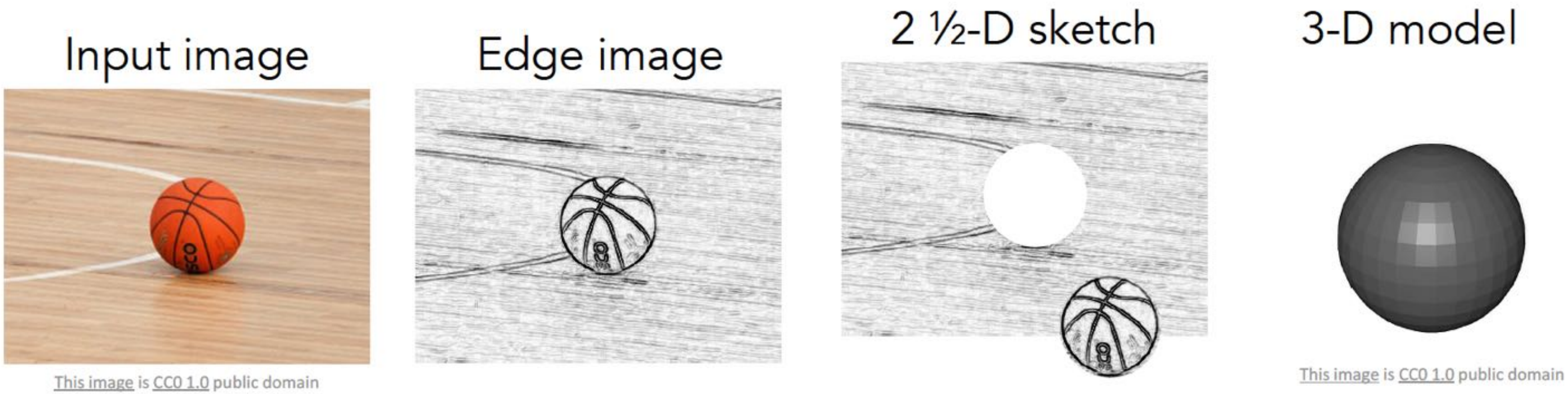
FOREWORD BY
Shimon Ullman

AFTERWORD BY
Tomaso Poggio

Copyrighted Material

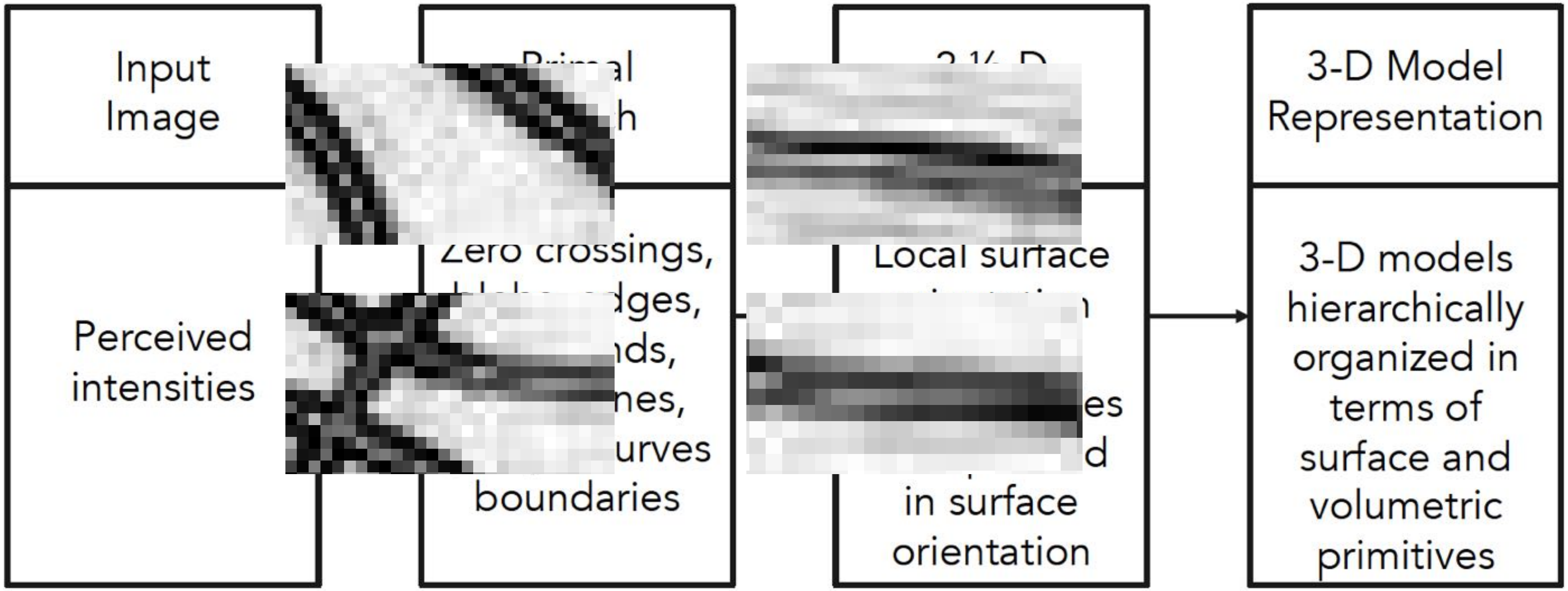
David Marr, 1970s

slide credit: Fei-Fei, Justin Johnson, Serena Yeung



This image is CC0 1.0 public domain

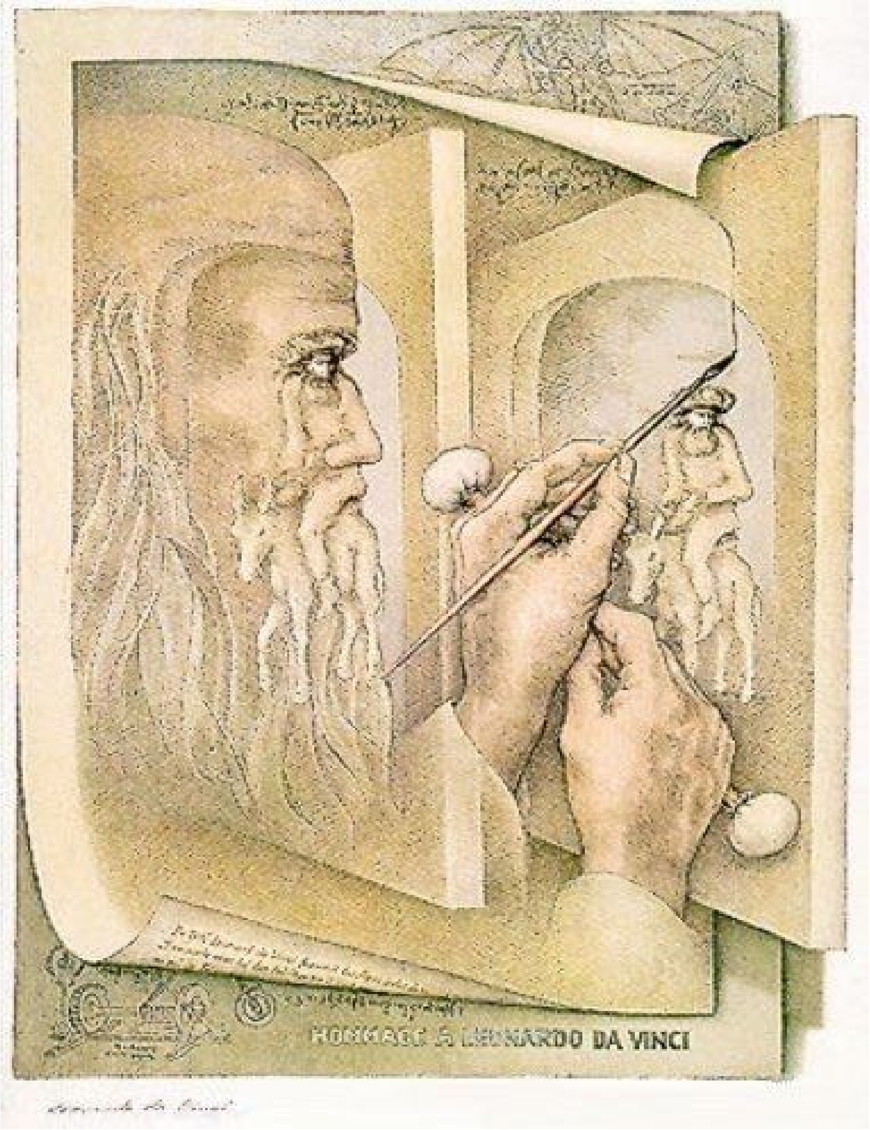
This image is CC0 1.0 public domain



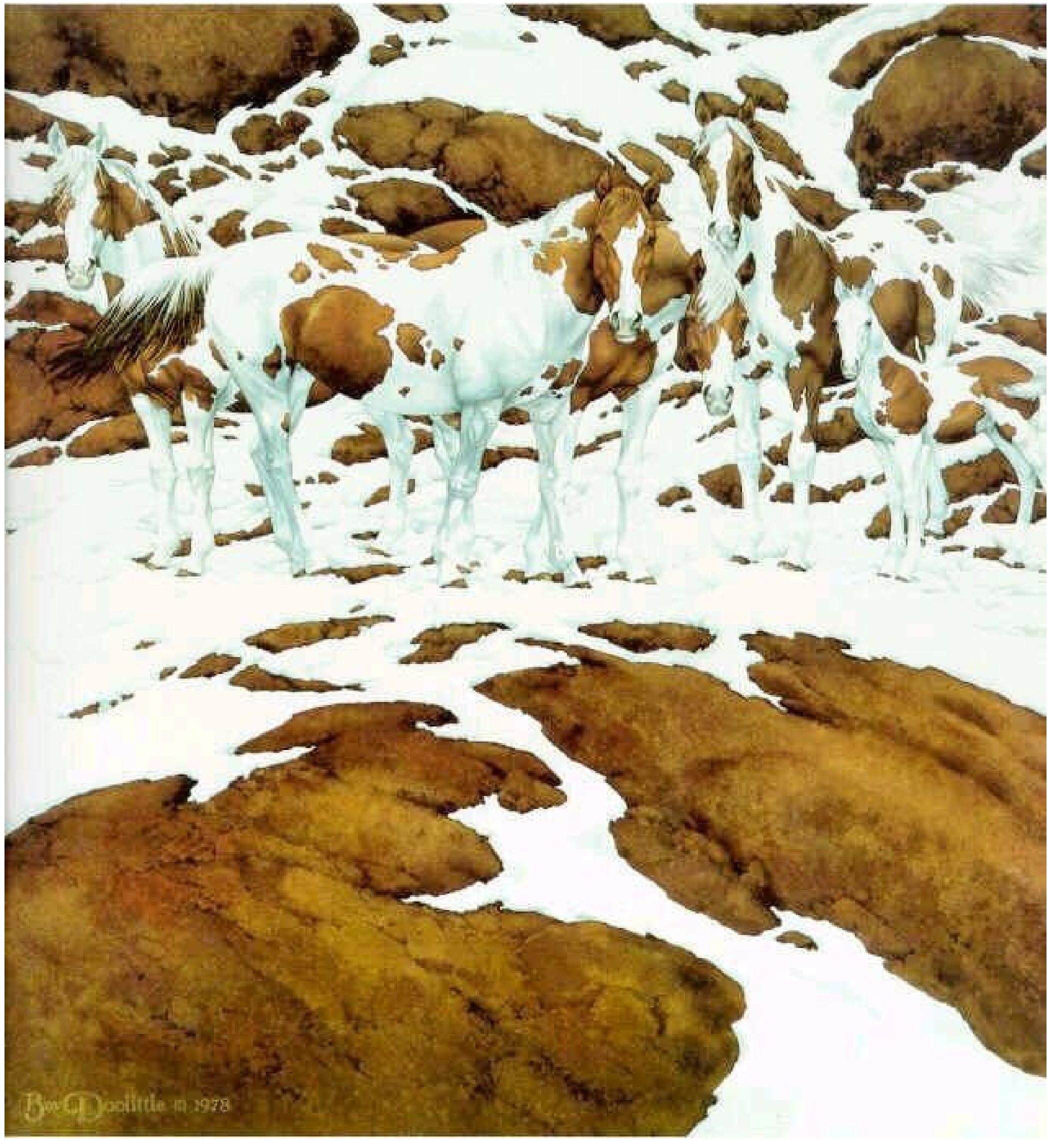
Stages of Visual Representation, David Marr, 1970s

slide credit: Fei-Fei, Justin Johnson, Serena Yeung

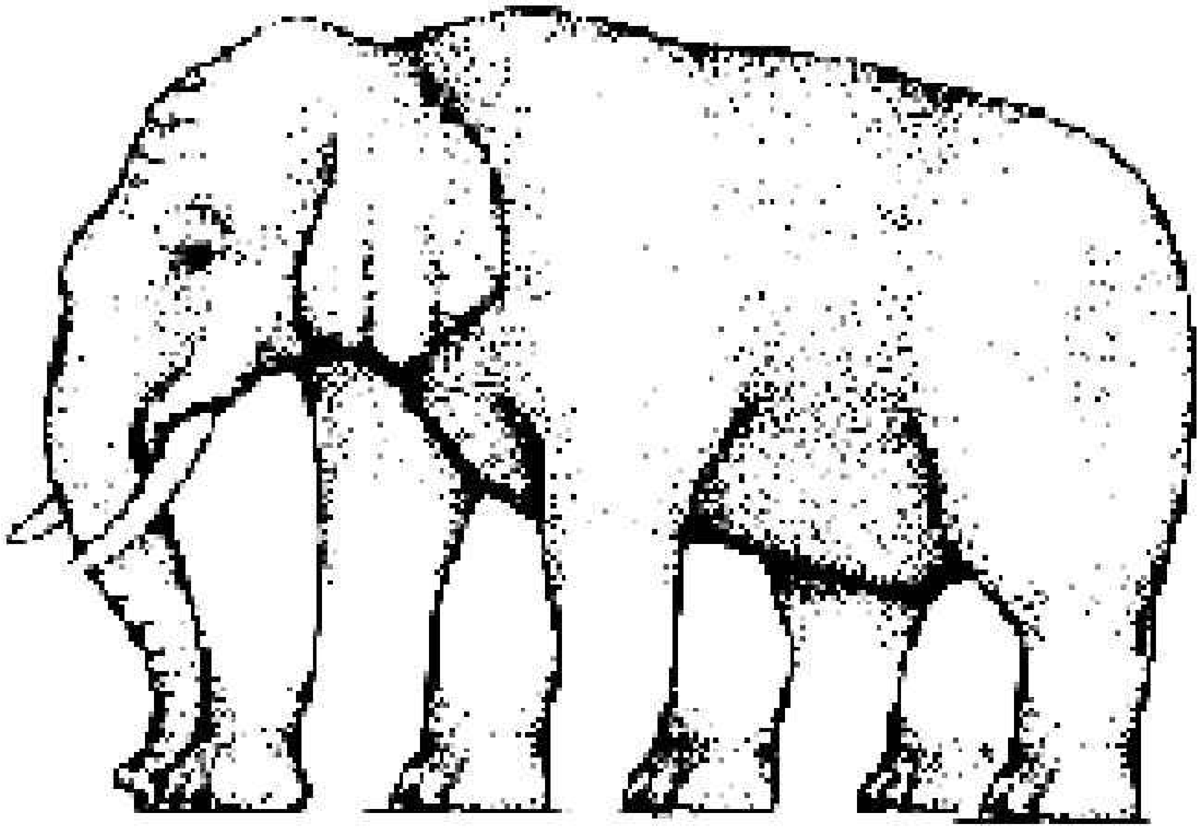
Complexity of Recognition



Complexity of Recognition

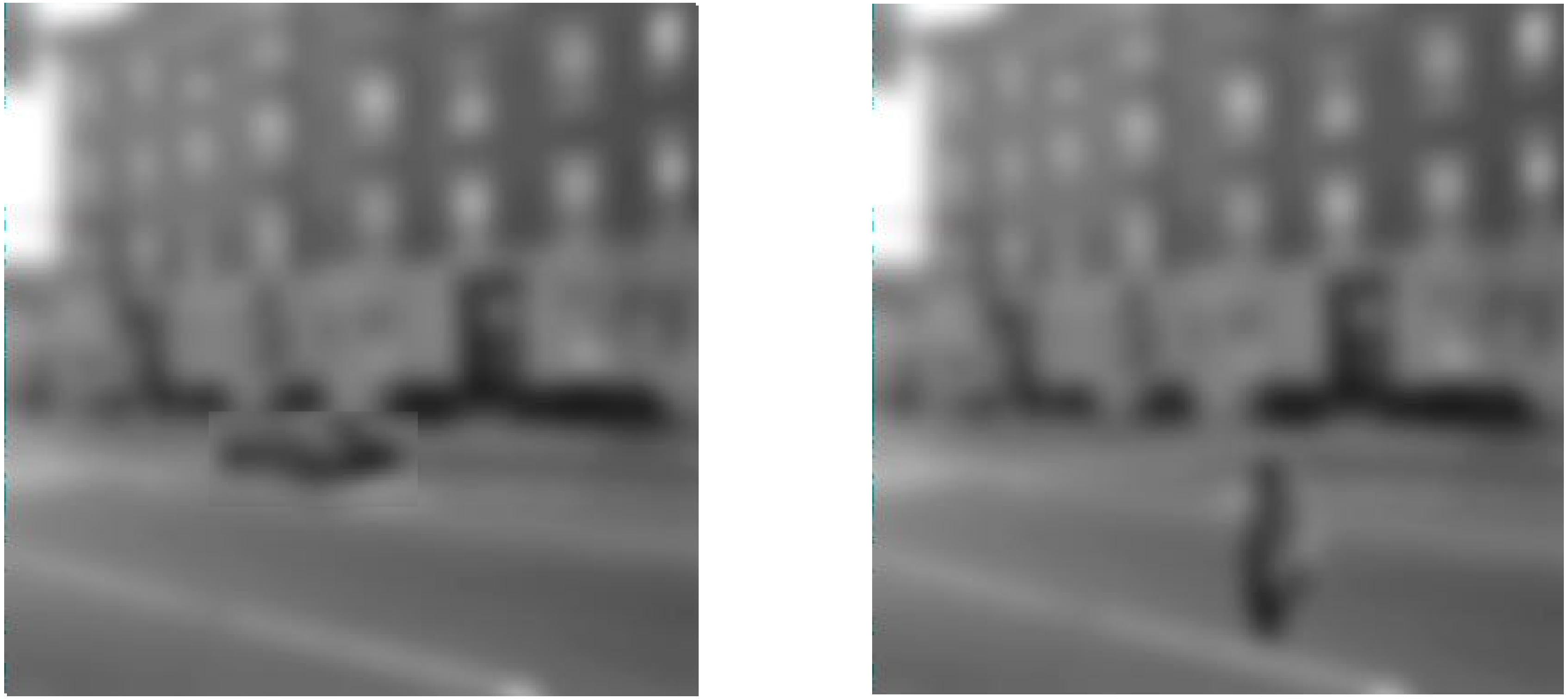


Complexity of Recognition



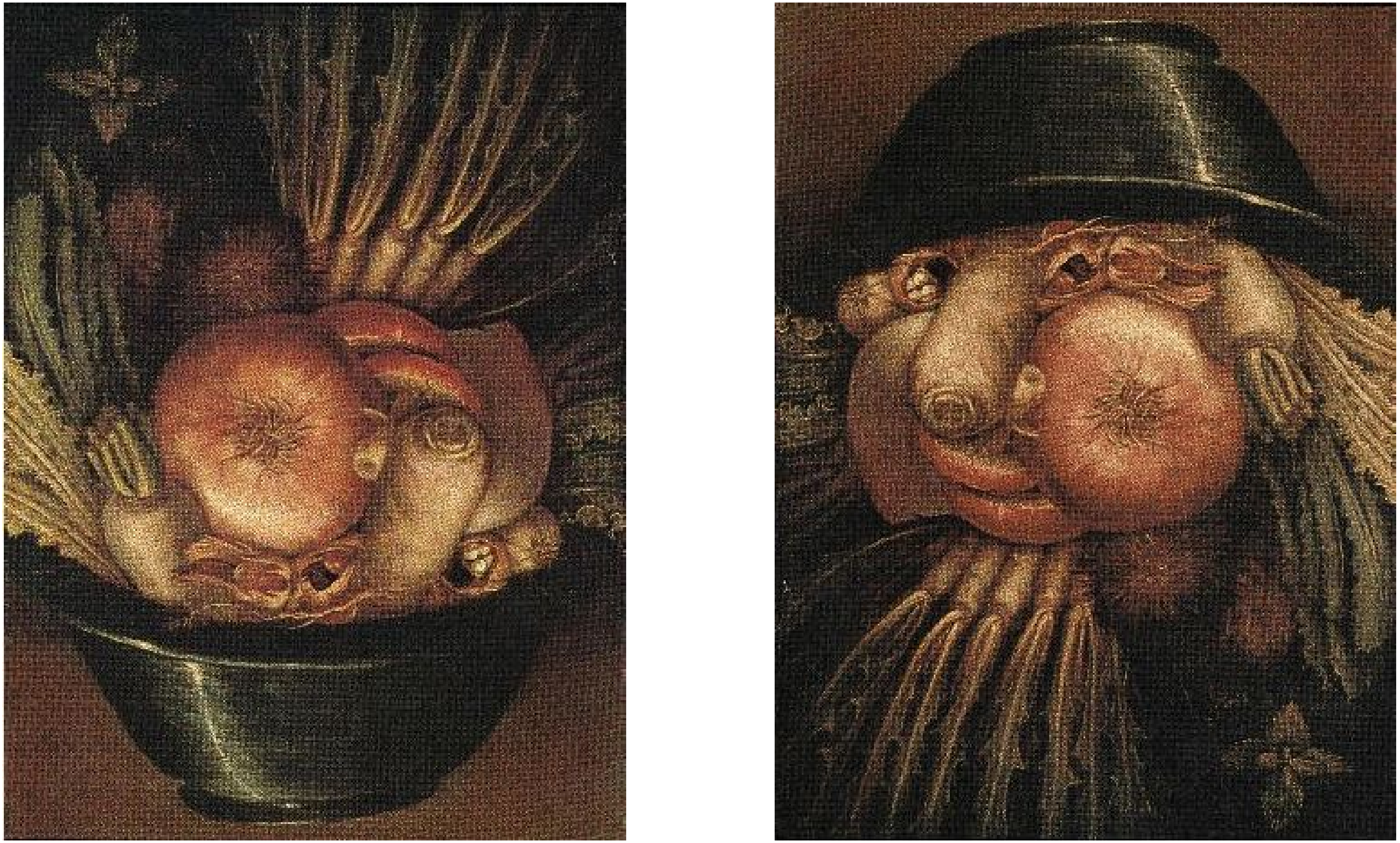
Recognition: the Role of Context

- Antonio Torralba

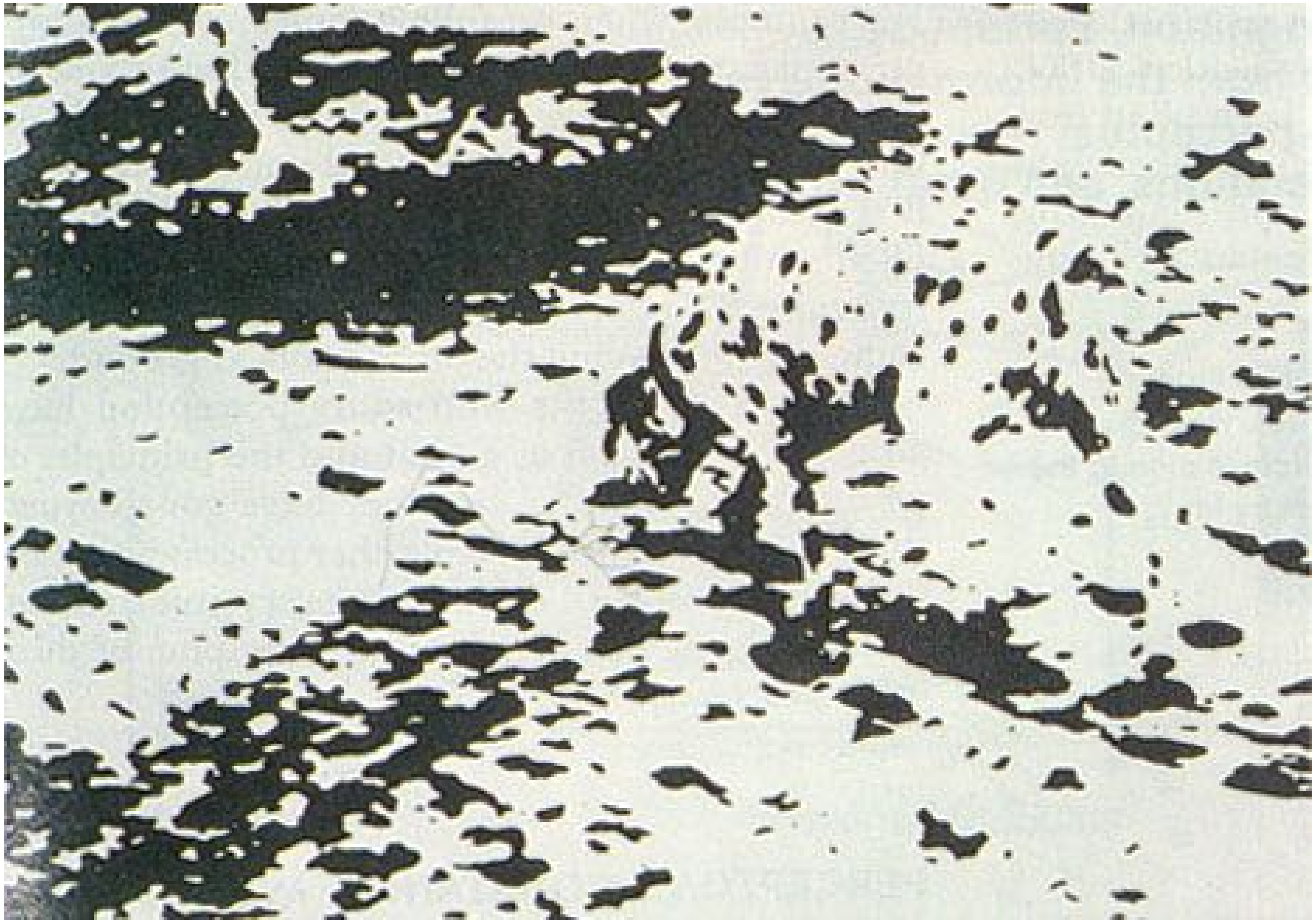


Recognition: the role of Prior Expectation

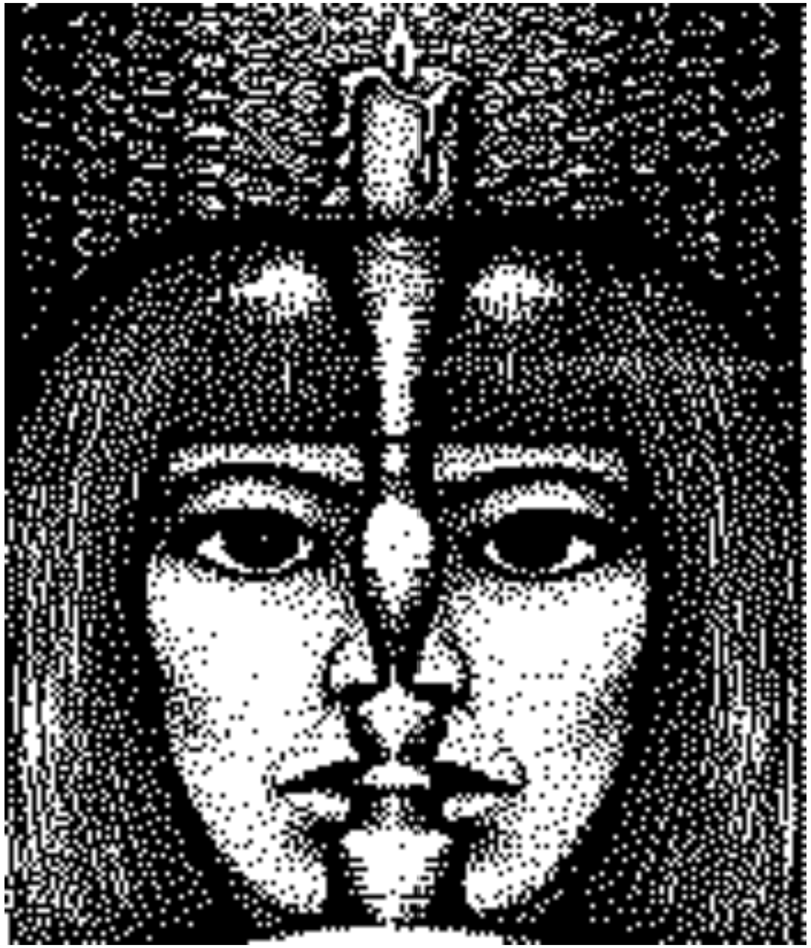
- Guiseppe Arcimboldo



Complexity of Recognition

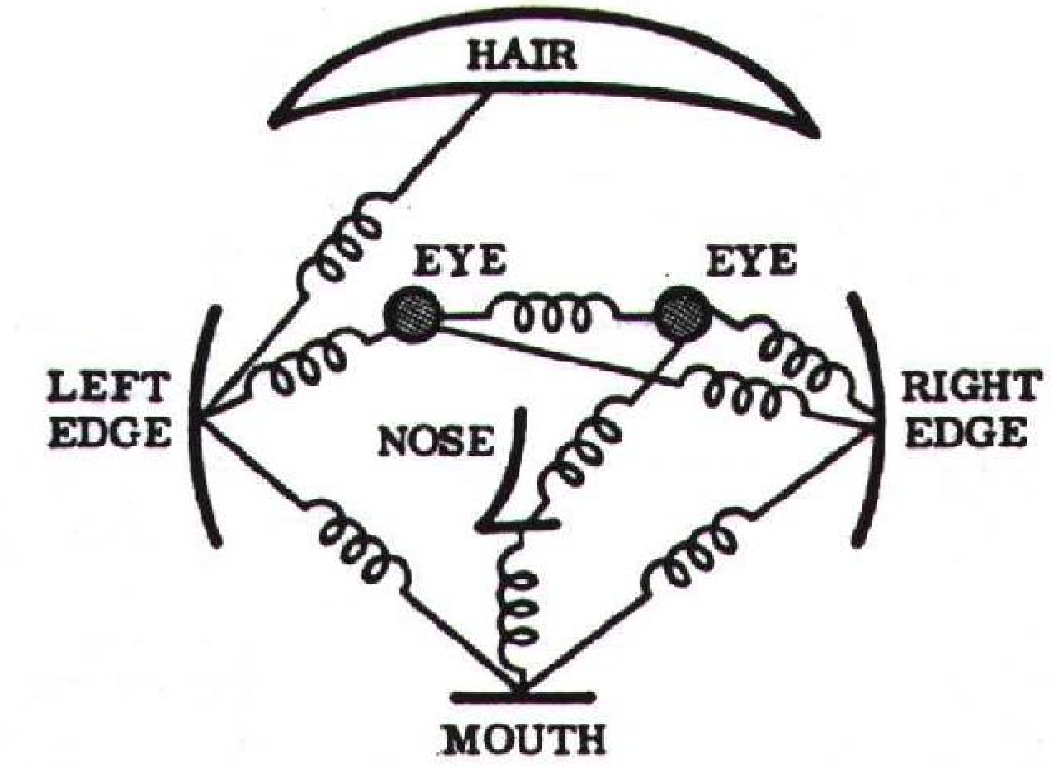


One or Two Faces ?

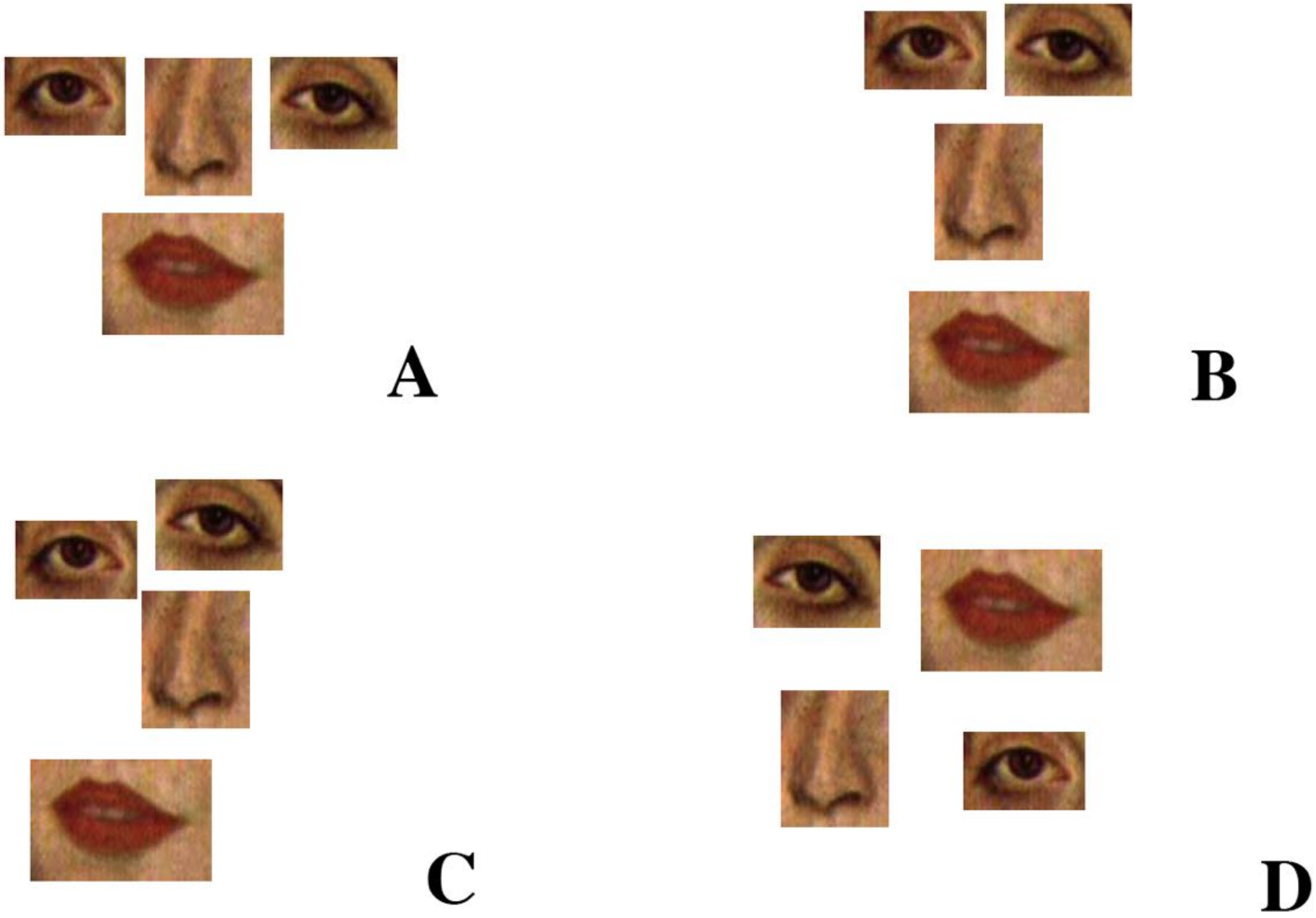


Class of Models: Pictorial Structure

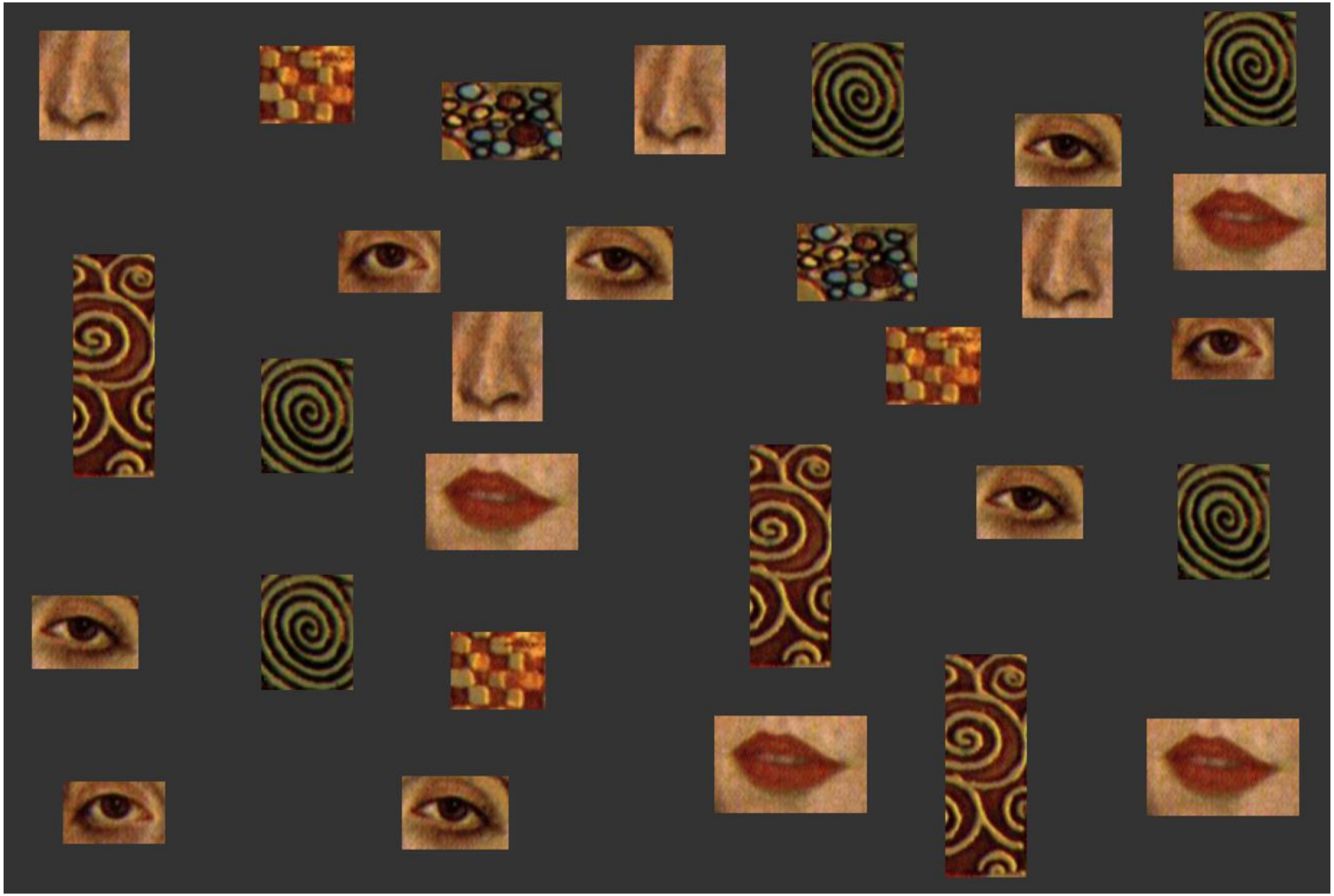
- Fischler & Elschlager 1973
- Model has two components
  - parts
    (2D image fragments)
  - structure
    (configuration of parts)



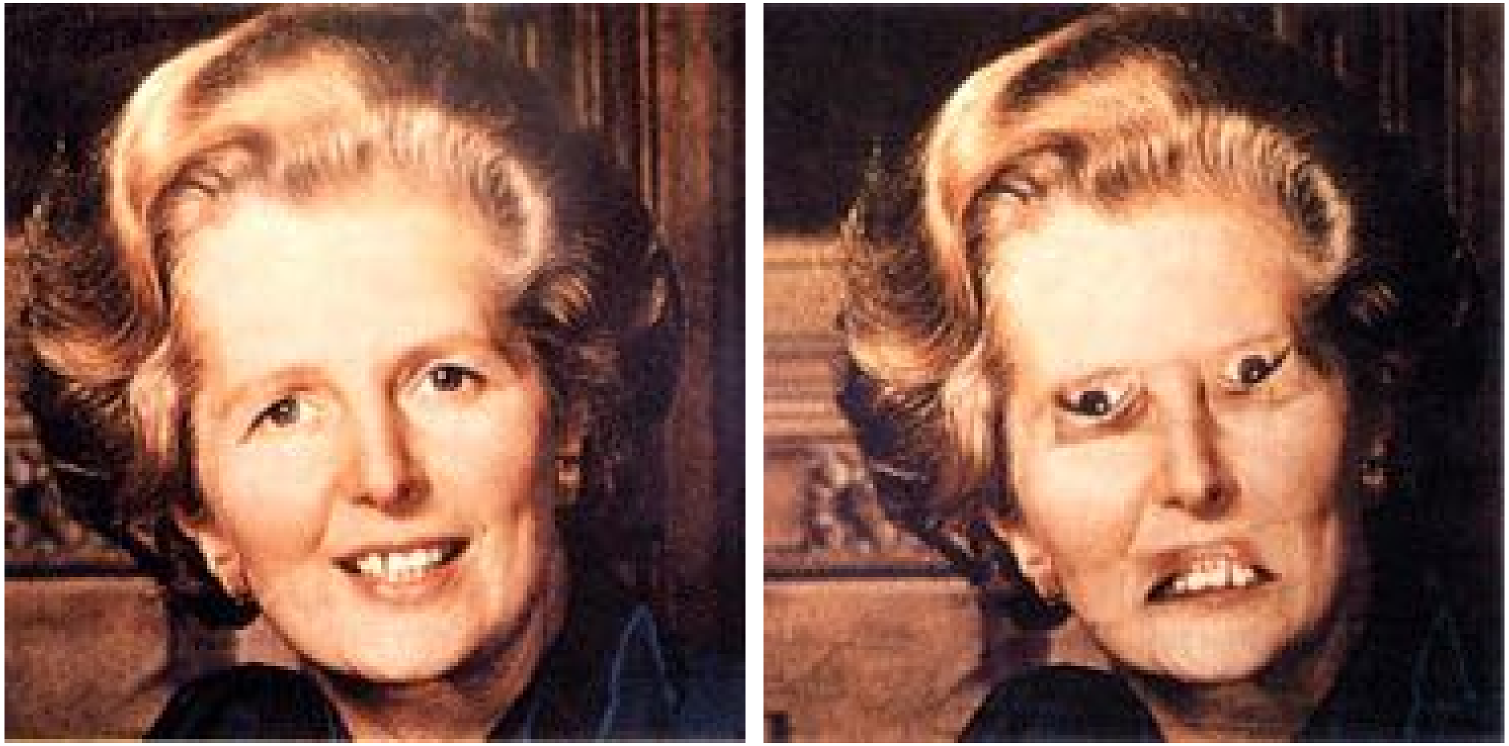
Deformations



Clutter



Example



Recognition, Localization, and Segmentation

a few terms

… let’s briefly define what we mean by that

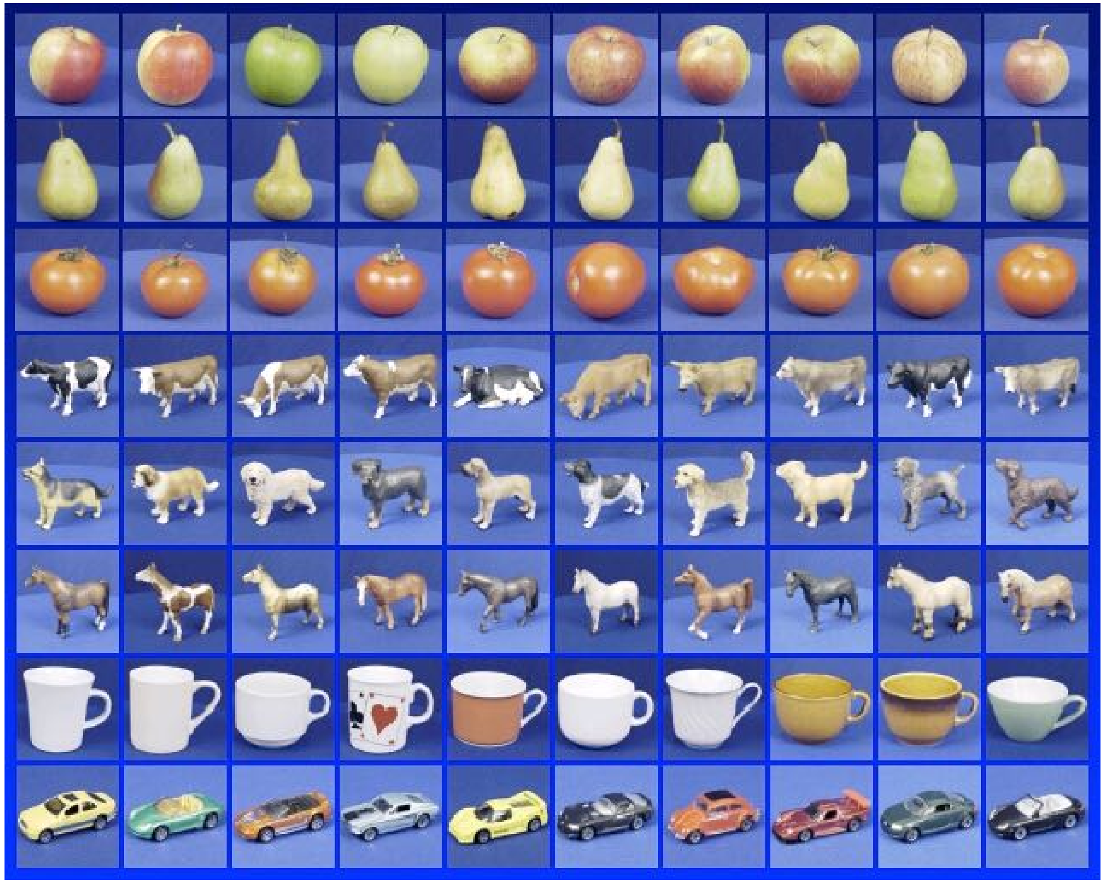
Object Recognition: First part of this Computer Vision class

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your pencil, your dog, your car
  - Object **Classification**
    - recognize any pencil, any dog, any car
    - also called: generic object recognition, object categorization, …

- Recognition and
  - **Segmentation**: separate pixels belonging to the foreground (object) and the background
  - **Localization/Detection**: position of the object in the scene, pose estimate (orientation, size/scale, 3D position)

Object Recognition: First part of this Computer Vision class

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your apple, your cup, your dog
  - Object **Classification**
    - recognize any apple, any cup, any dog
    - also called: generic object recognition, object categorization, …
    - typical definition: ‘basic level category’



Which Level is right for Object Classes?

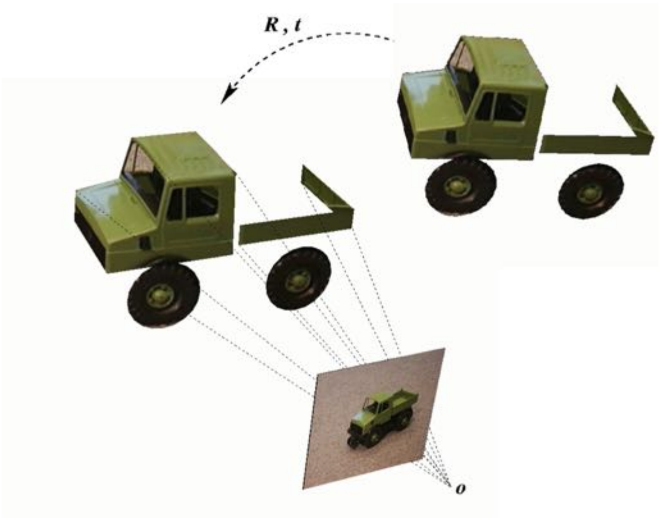
- Basic-Level Categories
  - the highest level at which category members have **similar perceived shape**
  - the highest level at which a **single mental image** can reflect the entire category
  - the highest level at which a person uses similar **motor actions** to interact with category members
  - the level at which human subjects are usually **fastest** at identifying category members
  - the first level named and understood by **children**
  - (while the definition of basic-level categories depends on culture there exist a remarkable consistency across cultures...)
- Most recent work in object recognition has focused on this problem
  - we will discuss several of the most successful methods in the lecture :-)

Object Recognition & Segmentation

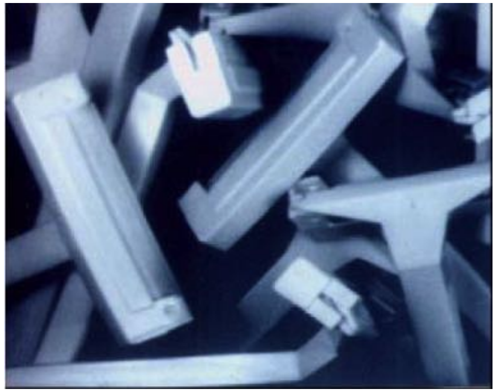
- Recognition and
  - **Segmentation**: separate pixels belonging to the foreground (object) and the background

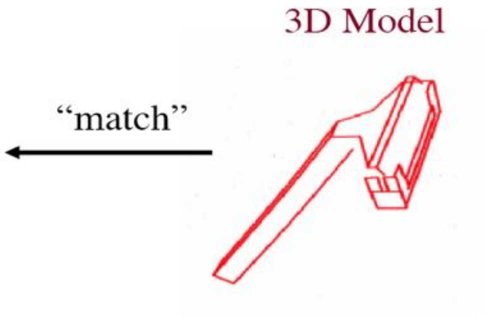
Object Recognition & Localization

- Recognition and
  - **Localization**: to position the object in the scene, estimate the object's pose (orientation, size/scale, 3D position)



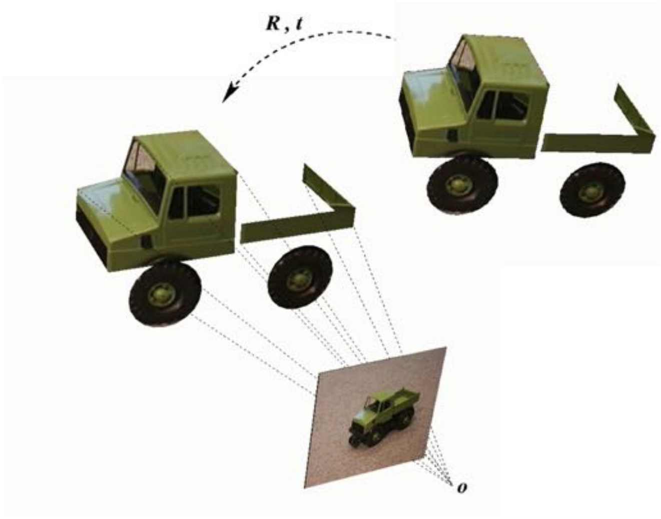
  - Example from David Lowe:

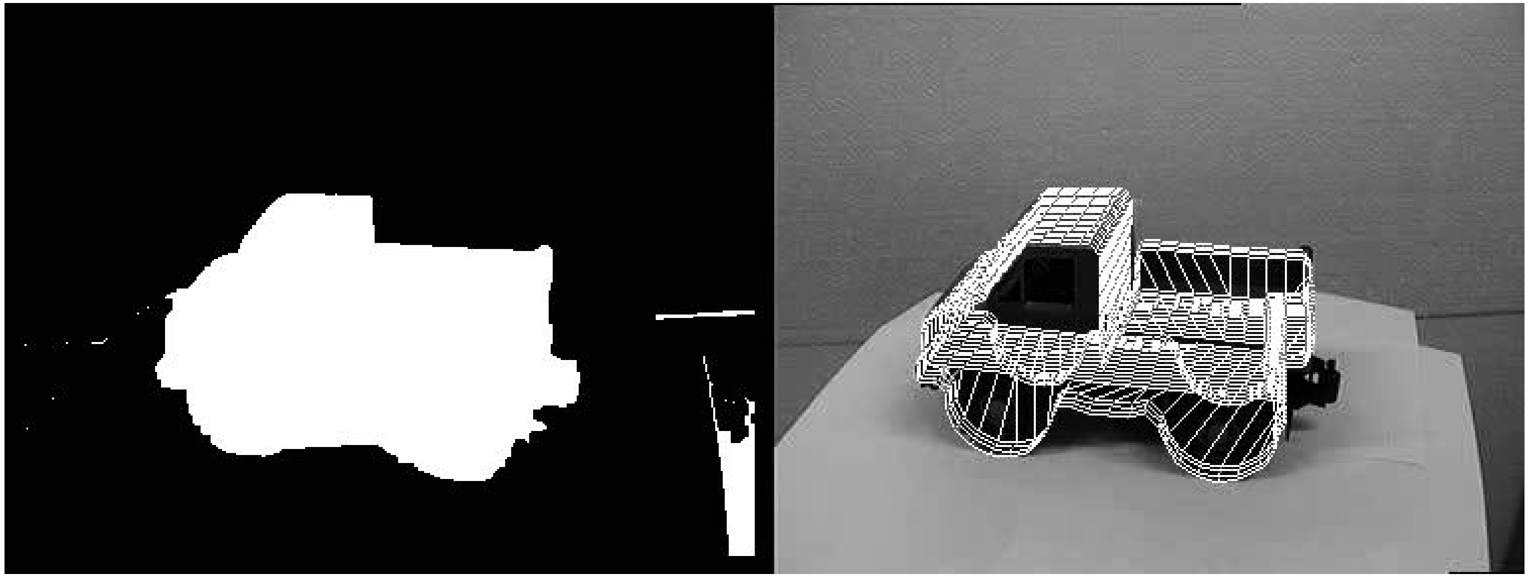




Parameters: 3D position and orientation

Localization: Example Video 1





Localization: Example Video 2

Object Recognition

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your pencil, your dog, your car
  - Object **Classification**
    - recognize any pencil, any dog, any car
    - also called: generic object recognition, object categorization, …

- Recognition and
  - **Segmentation**: separate pixels belonging to the foreground (object) and the background
  - **Localization**: position the object in the scene, estimate pose of the object (orientation, size/scale, 3D position)

Goals of today’s lecture

- First intuitions about
  - What is computer vision?
  - What does it mean to see and how do we (as humans) do it?
  - How can we make this computational?

- Applications & Appetizers

- Role of Deep Learning
    - with several slides taken from Fei-Fei Li, Justin Johnson, Serena Yeung @ Stanford

- 2 case studies:
  - Recovery of 3D structure
    - slides taken from Michael Black @ Brown University / MPI Intelligent Systems
  - Object Recognition
    - intuition from human vision...