



mp

max planck institut
informatik

SIC Saarland Informatics
Campus

Lecture will start at 10h15

**High Level Computer Vision - Introduction
@ April 12, 2023**

Bernt Schiele

<https://cms.sic.saarland/hlcvss23/>

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



mp

max planck institut
informatik

SIC Saarland Informatics
Campus

Computer Vision and Machine Learning Group @ Max-Planck-Institute for Informatics



Bernt Schiele
Computer Vision &
Machine Learning



Jan Eric Lenssen
Geometric Representation Learning

Zeynep Akata
Multimodal Deep Learning
U Tübingen



Margret Keuper
Robust Visual Learning
U Siegen



Gerard Pons-Moll
Real Virtual Humans
U Tübingen



High Level Computer Vision

- Lecturer:
 - ▶ Bernt Schiele (schiele@mpi-inf.mpg.de)
- Assistants:
 - ▶ Sukurt Rao (sukrut.rao@mpi-inf.mpg.de)
 - ▶ Amin Parchami-Aragi (mparcham@mpi-inf.mpg.de)
 - ▶ Paul Soumava (spaul@mpi-inf.mpg.de)
- Language:
 - ▶ English
- mailing list for announcements etc.
 - ▶ register online: <https://cms.sic.saarland/hlcvss22/>

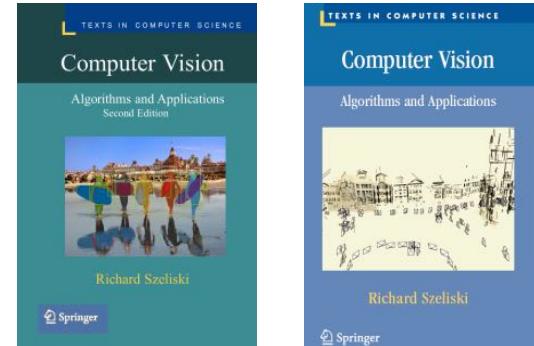


Lecture & Tutorial

- Officially: 2V (lecture) + 2Ü (exercise/tutorial)
 - ▶ Lecture: Wed: 10:15am - 12pm
 - ▶ Tutorial: Mon: 10:15am - 12pm
- typically 1 exercise sheet every 1-2 weeks
 - ▶ part of the final grade - working in **groups of 2-3 students**
 - ▶ some pencil and paper, mostly practical including a larger project
 - ▶ larger project: we/you propose projects, mentoring, final presentation
- 1. tutorial is Python tutorial (Mon, April 17th)
- Exam
 - ▶ written or oral exam (grading 50% exam and 50% exercises)

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://deeplearningbook.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, ...)
 - ▶ ...



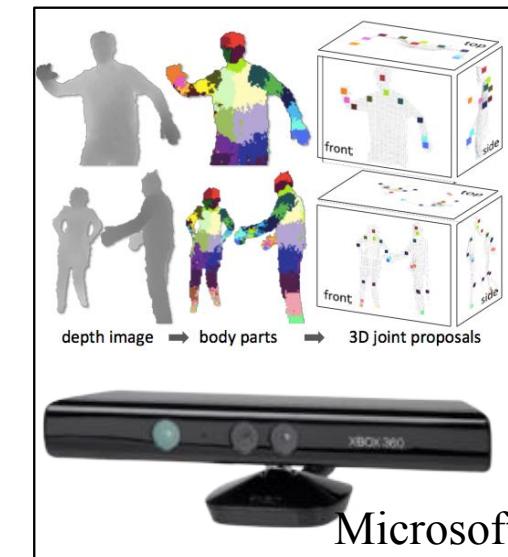
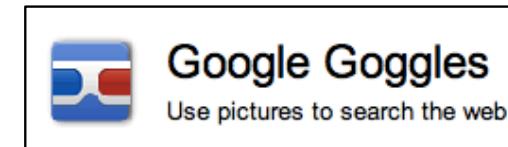
Some Applications

- License Plate Recognition
 - ▶ London Congestion Charge
 - ▶ http://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, Lens, ...
- Vision for Interfaces:
 - ▶ e.g. Microsoft Kinect
- Reconstruction



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
- Case Study
 - ▶ Object Recognition — intuition from human vision...



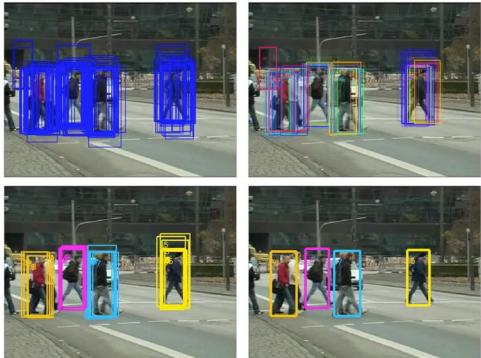


Applications & Appetizers

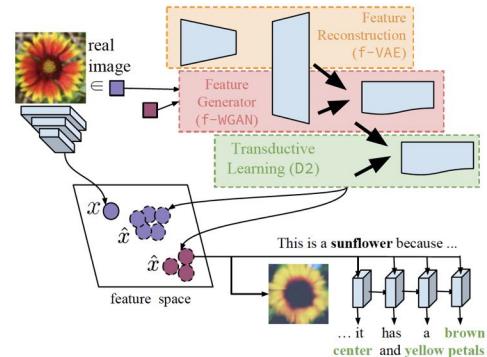
... work from our group

Some Research Highlights from the Past

Multi-Person Tracking



Zero- and Few-Shot Learning



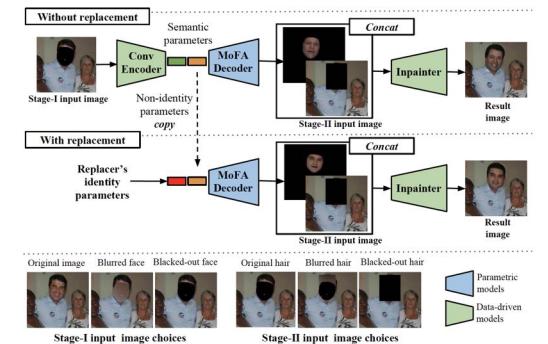
Multi-Person Human Pose Estimation & Tracking



Video Segmentation



Visual Privacy

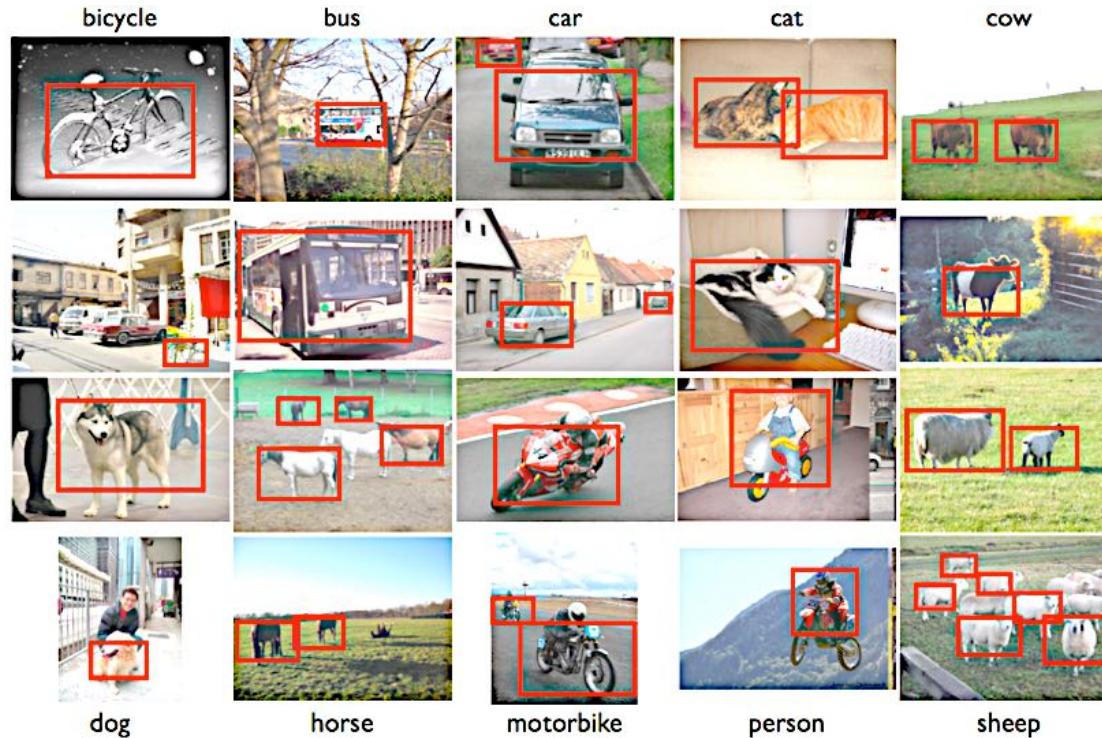


Datasets...

Caltech Pedestrians



Detection & Recognition of Visual Categories



- Challenges:
- multi-scale
 - multi-view
 - multi-class
 - varying illumination
 - occlusion
 - cluttered background
 - articulation
 - high intraclass variance
 - low interclass variance

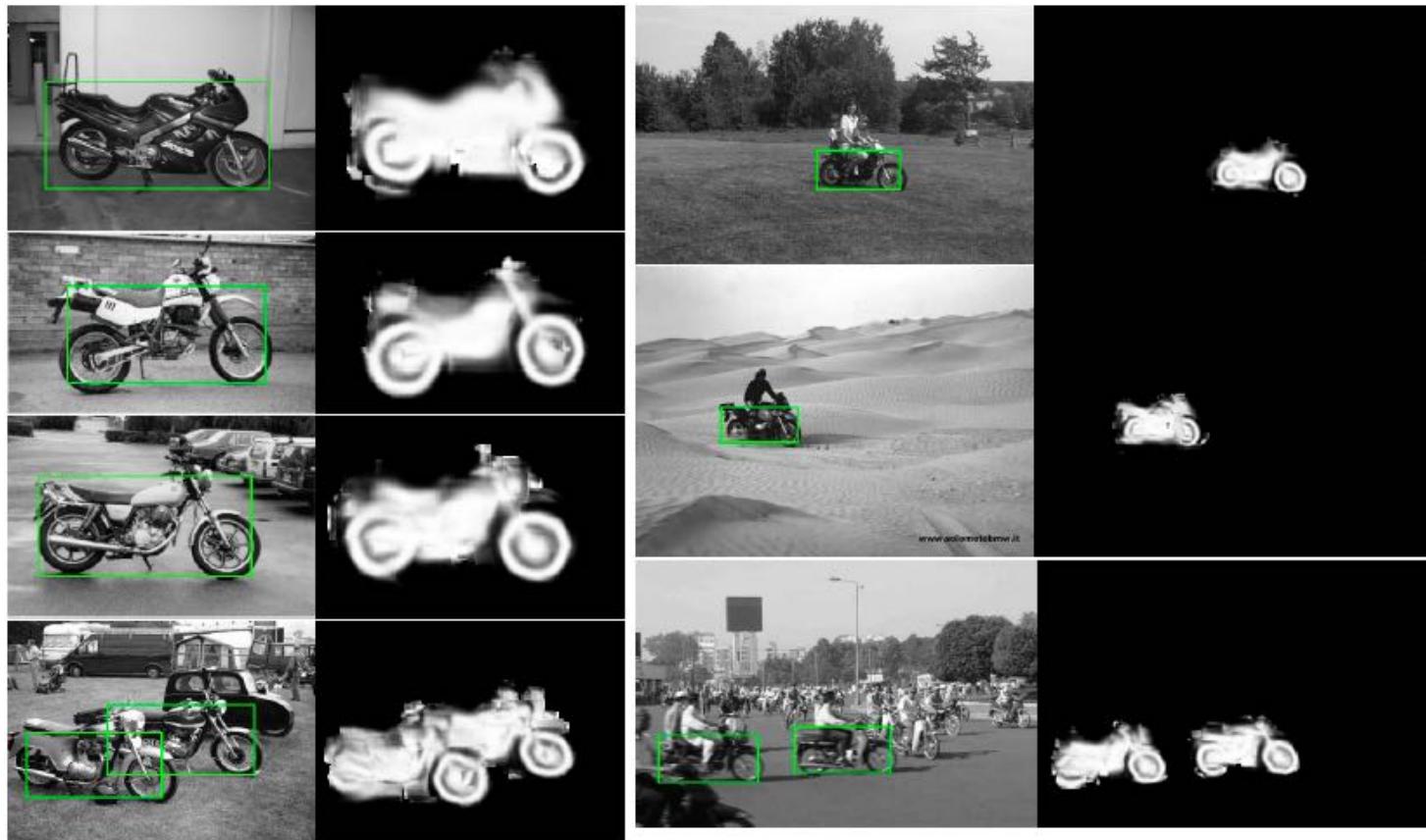
Challenges of Visual Categorization

- high intra-class variation
- low inter-class variation

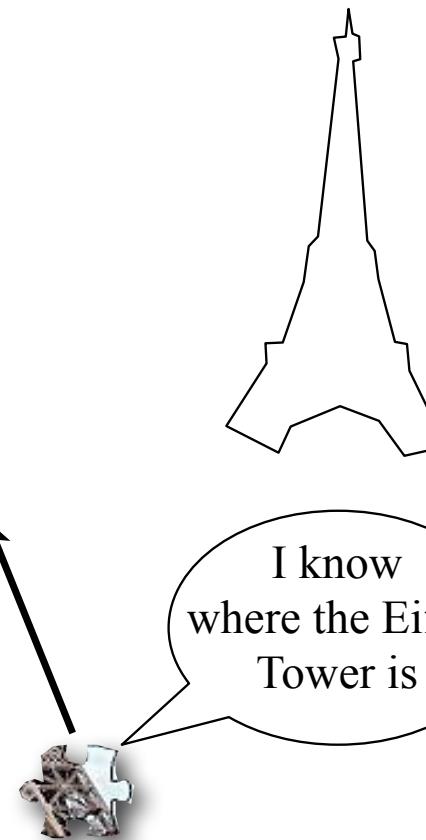
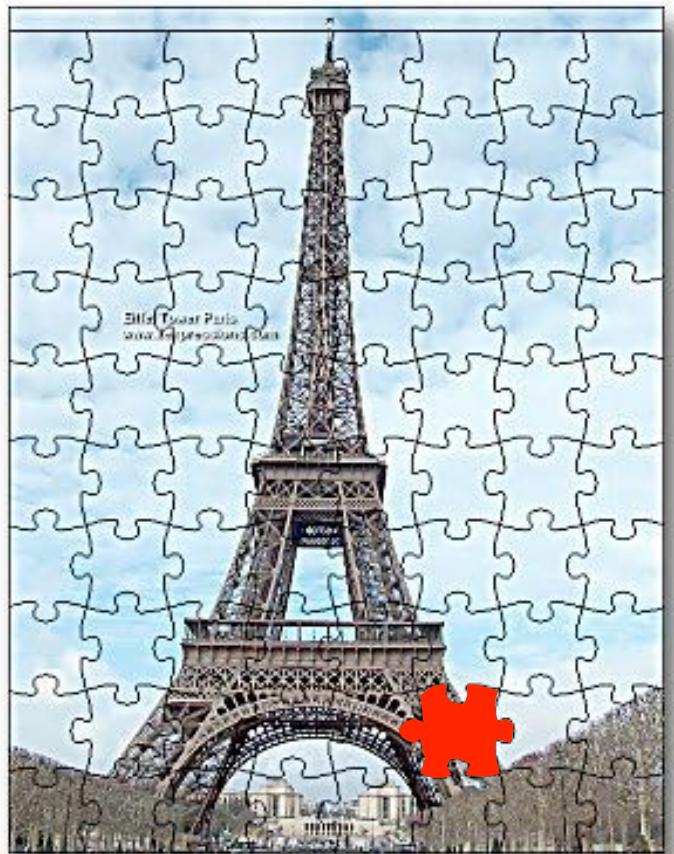


- high intra-class variation

Sample Category: Motorbikes

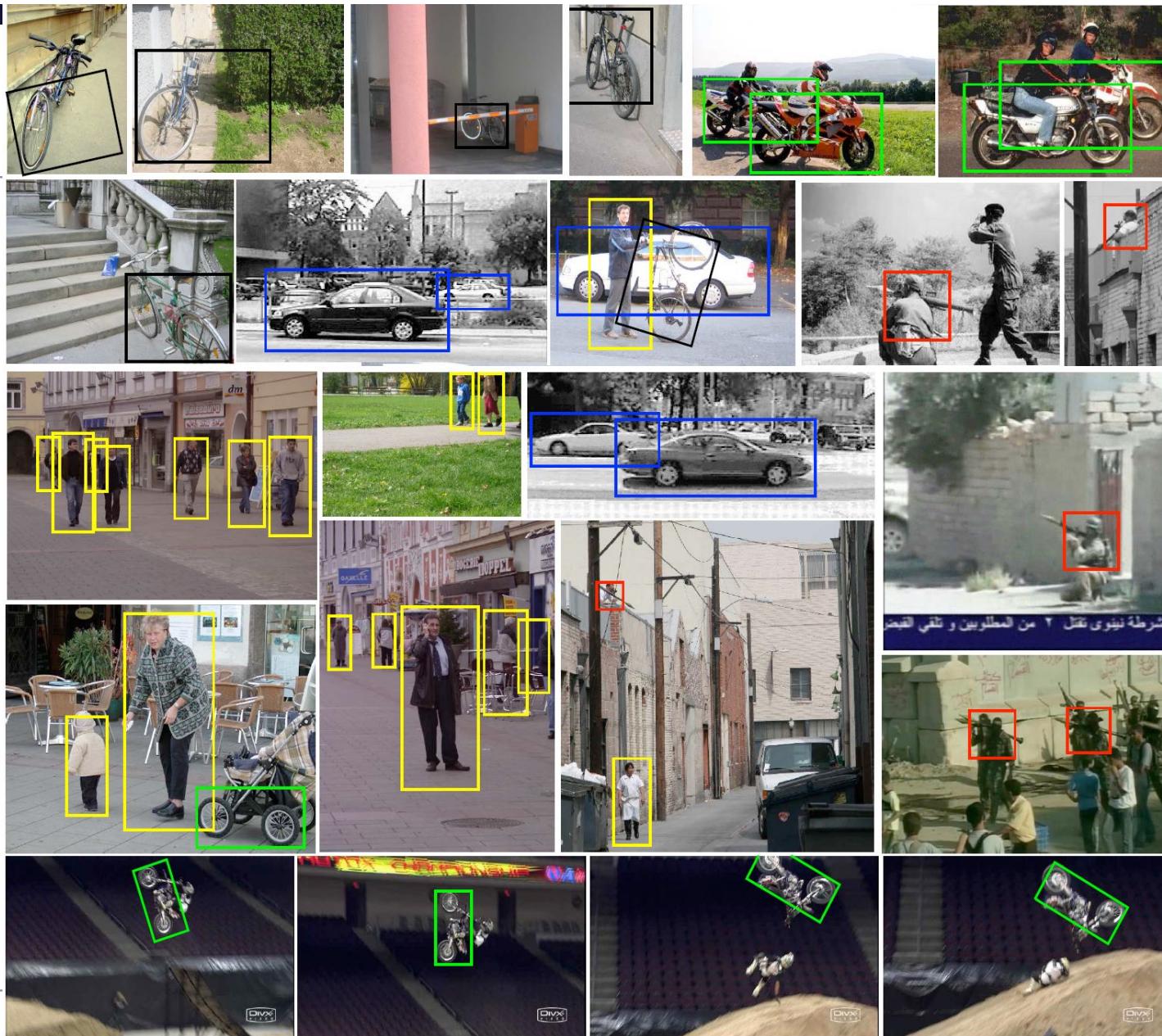


Basic Idea



global

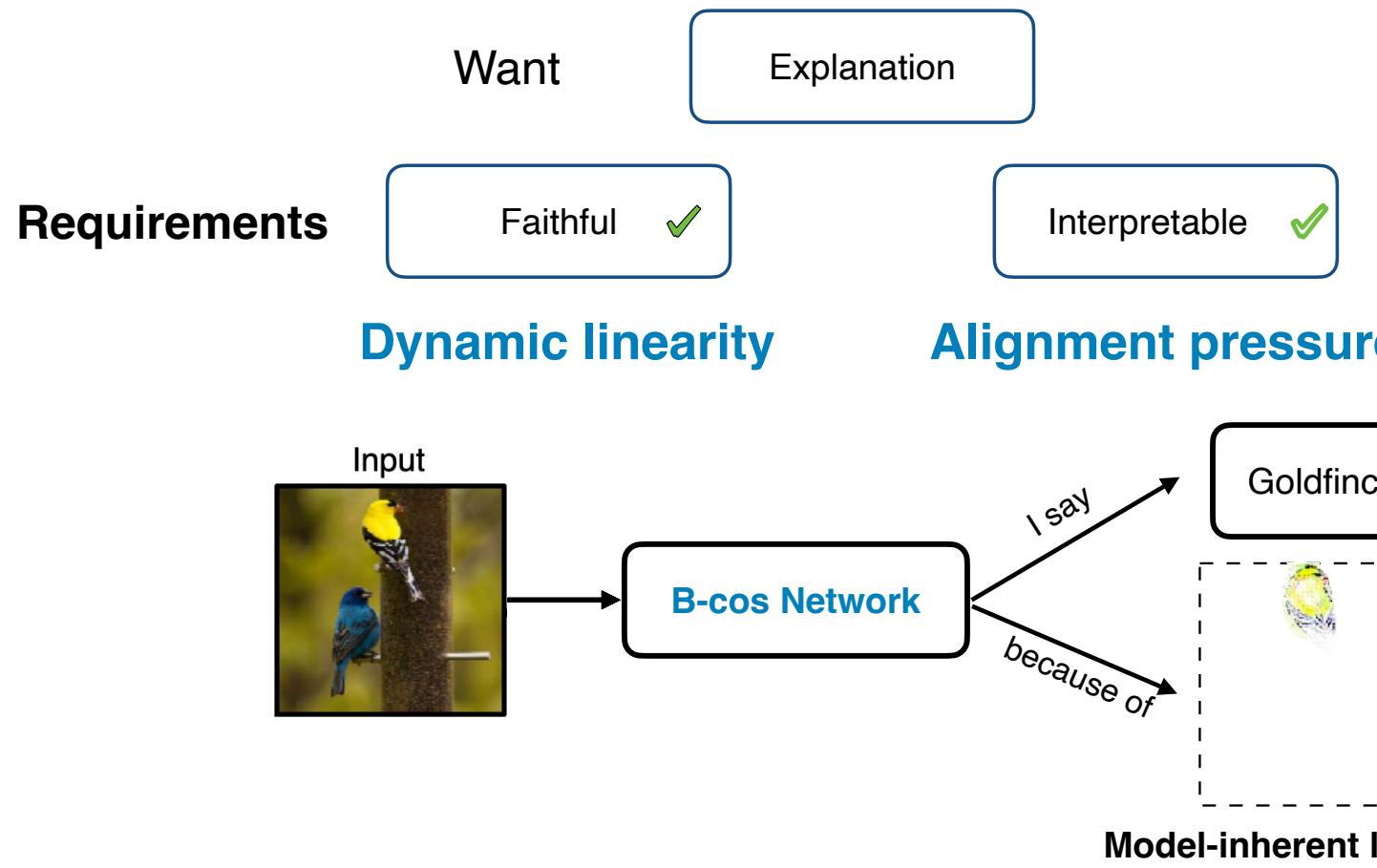
local



Computer Vision & Machine Learning: Current Research Themes

- **Interpretability, Robustness** and **Security** of Deep Learning in Computer Vision
 - ▶ **Inherently Interpretable** Deep Neural networks — CVPR'21, CVPR'22
 - ▶ **Robustness** of Deep Models:
Bright and Dark Side of Scene Context — NeurIPS'18, CVPR'19, ECCV'20
 - ▶ **Security** of Deep Models
Reverse Engineering and **Stealing** of Deep Models — ICLR'18, CVPR'19, ICLR'20

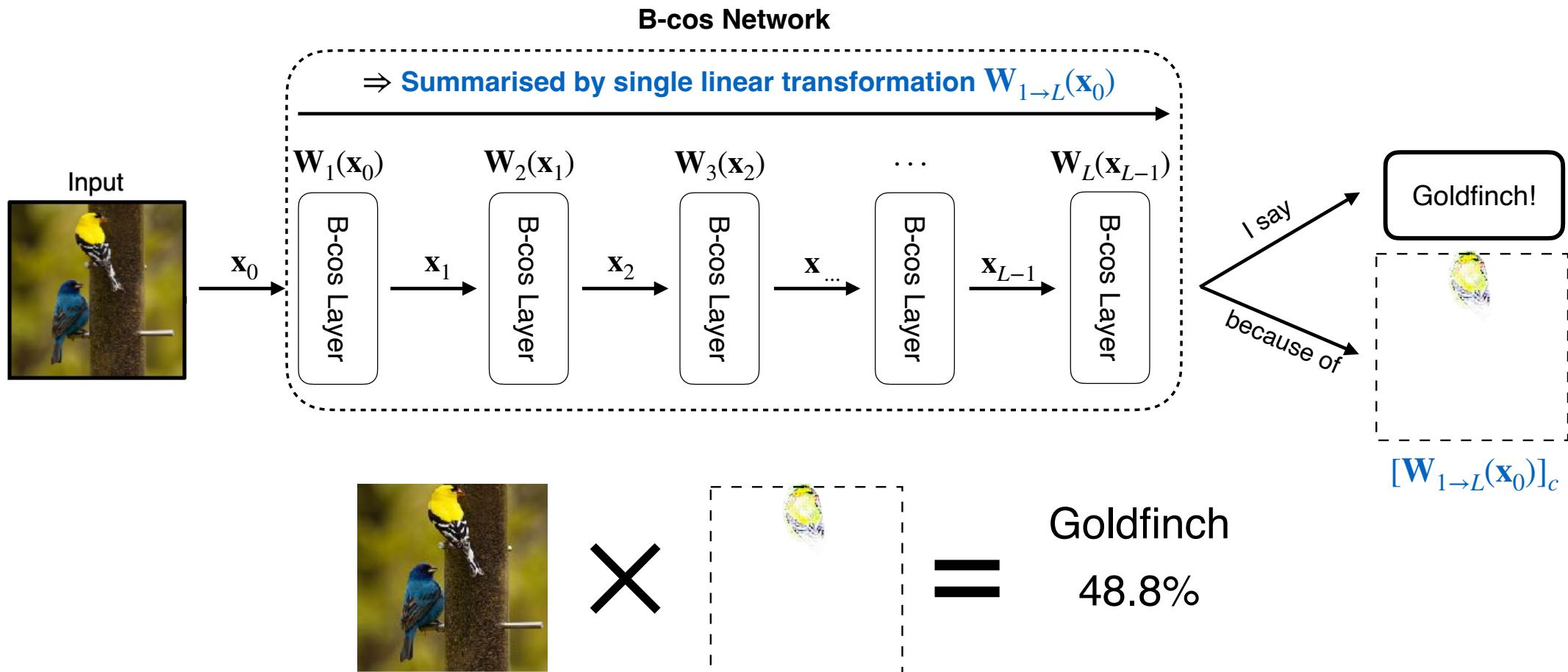
Motivation: we aim for Inherent Interpretability



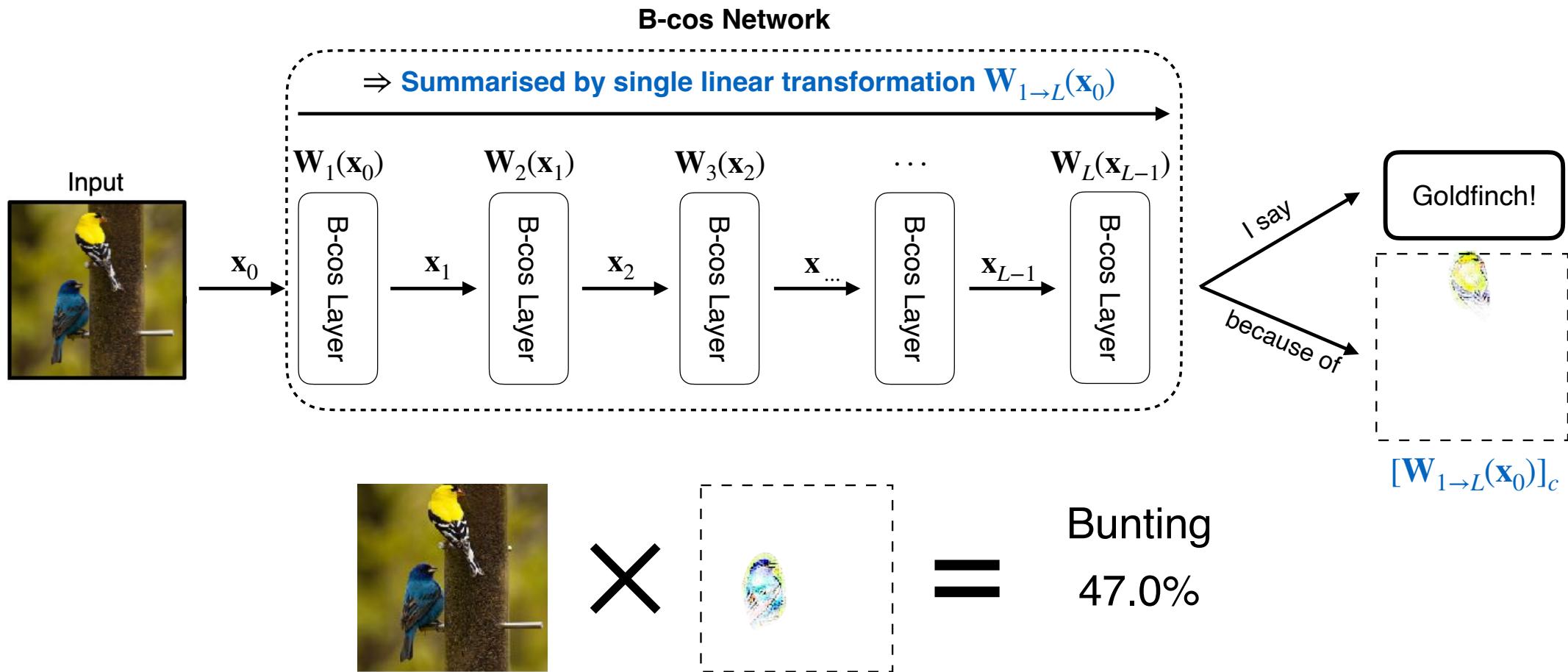
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)



Dynamic linearity



Dynamic linearity



Visualisations of $W_{1 \rightarrow L}(x)$

Input image



Input image



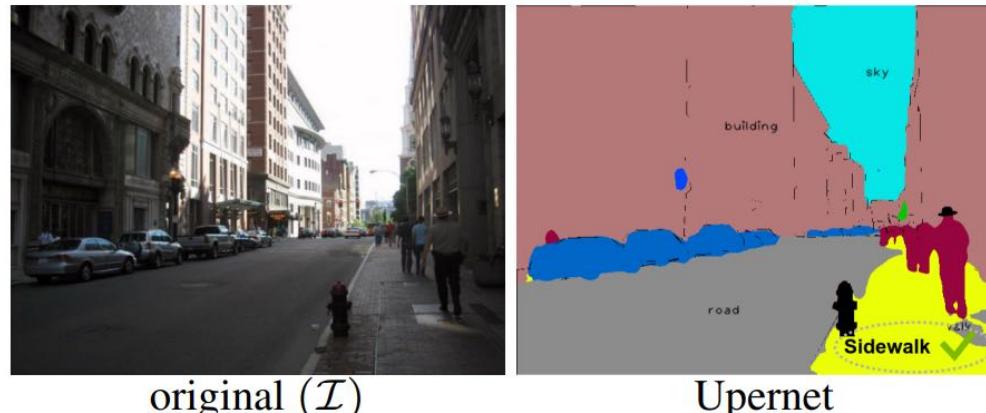
Computer Vision & Machine Learning: Current Research Themes

- **Interpretability, Robustness** and **Security** of Deep Learning in Computer Vision
 - ▶ **Inherently Interpretable** Deep Neural networks — CVPR'21, CVPR'22
 - ▶ **Robustness** of Deep Models:
Bright and Dark Side of Scene Context — NeurIPS'18, CVPR'19, ECCV'20
 - ▶ **Security** of Deep Models
Reverse Engineering and **Stealing** of Deep Models — ICLR'18, CVPR'19, ICLR'20

The Bright and the Dark Side of Scene Context

- Current models heavily rely on scene context:

- ▶ Original image with cars on the left side:



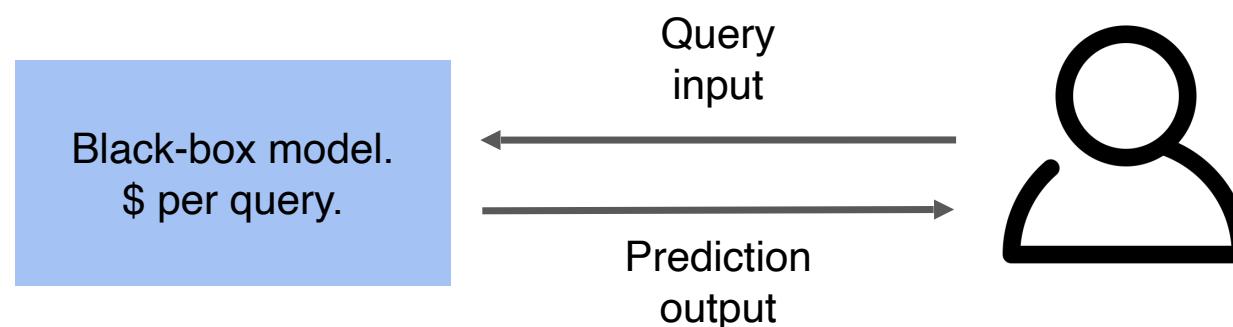
- ▶ Same image without those cars:

Computer Vision & Machine Learning: Current Research Themes

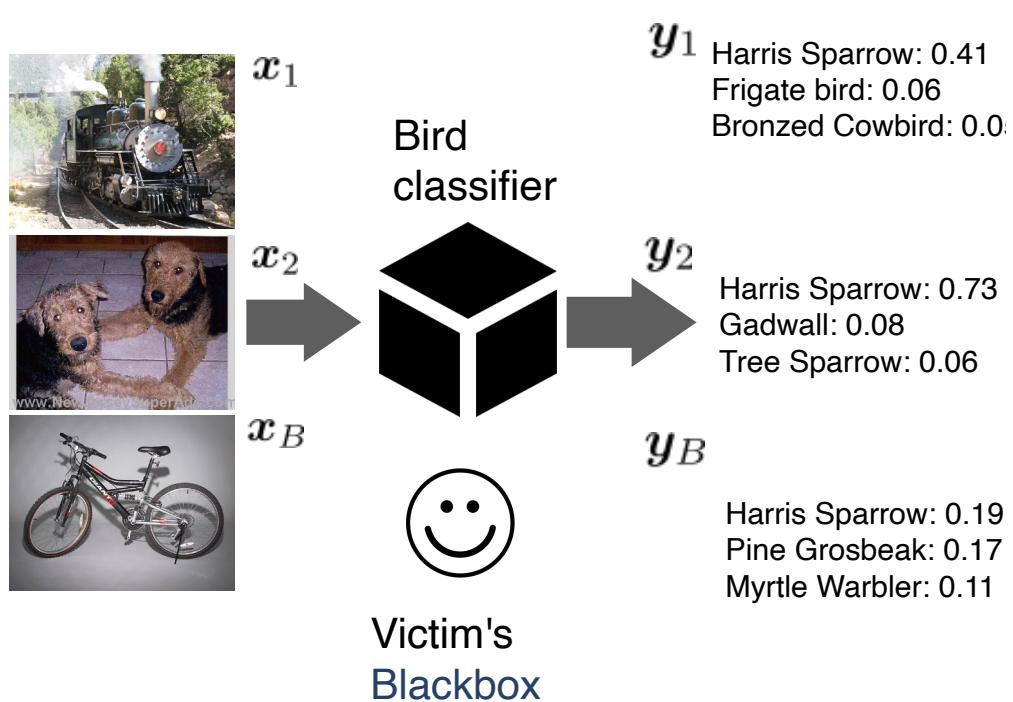
- **Interpretability, Robustness** and **Security** of Deep Learning in Computer Vision
 - ▶ **Inherently Interpretable** Deep Neural networks — CVPR'21, CVPR'22
 - ▶ **Robustness** of Deep Models:
Bright and Dark Side of Scene Context — NeurIPS'18, CVPR'19, ECCV'20
 - ▶ **Security** of Deep Models
Reverse Engineering and **Stealing** of Deep Models — ICLR'18, CVPR'19, ICLR'20

Providing ML Models is a Business Model

- **Input in, prediction out.** Ask \$ per query.
 - ▶ ML models are **black boxes** !
 - ▶ not shared: **architecture, parameters, hyperparameter** details (IPs)
- **Research question:**
 - ▶ can an adversary **steal the functionality of the model** ?

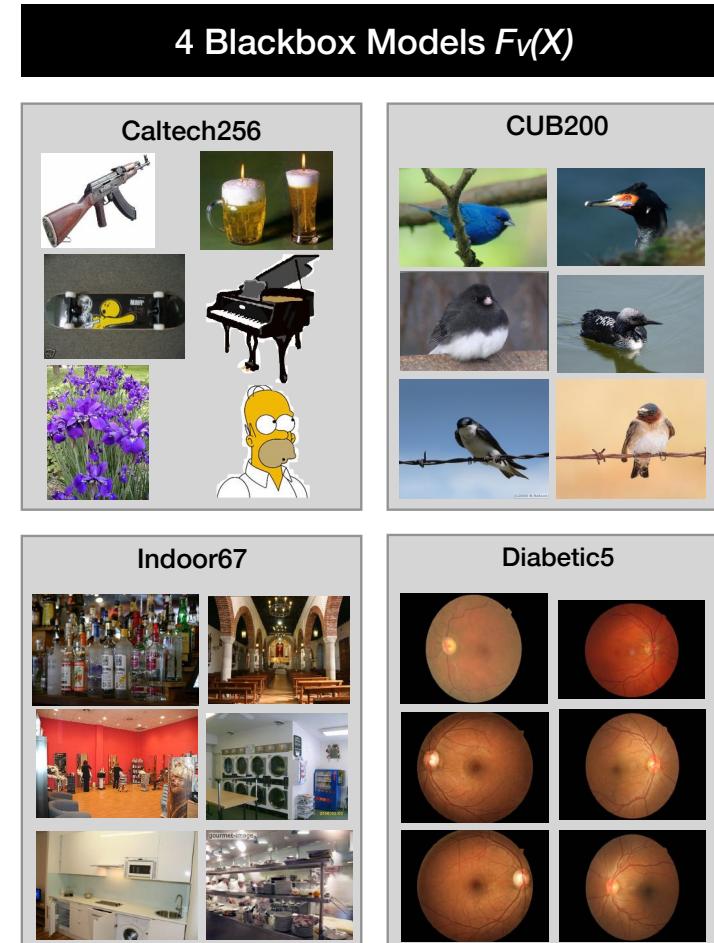


Functionality Stealing: Knock-Off Nets



Transfer Set Construction: $x_i \stackrel{\pi}{\sim} P_A(X)$

- Simple method: $\pi = \text{random}$
 - ▶ sample images randomly (without replacement)
 - ▶ prone to querying irrelevant images



Can we Learn with $\pi = \text{Random?}$ Yes!

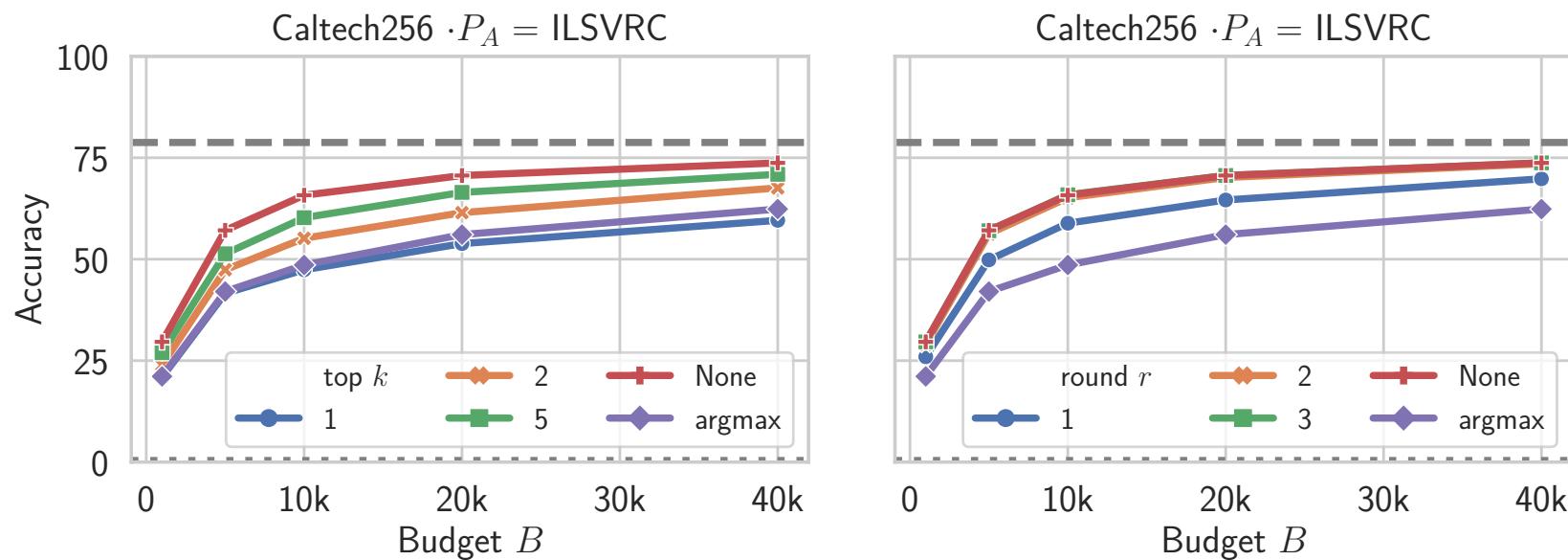
		random			
P_A		Caltech256	CUBS200	Indoor67	Diabetic5
Closed	$P_V(F_V)$	78.8 (1×)	76.5 (1×)	74.9 (1×)	58.1 (1×)
	$P_V(\text{KD})$	82.6 (1.05×)	70.3 (0.92×)	74.4 (0.99×)	54.3 (0.93×)
Closed	D^2	76.6 (0.97×)	68.3 (0.89×)	68.3 (0.91×)	48.9 (0.84×)
Open	ILSVRC	75.4 (0.96×)	68.0 (0.89×)	66.5 (0.89×)	47.7 (0.82×)
	OpenImg	73.6 (0.93×)	65.6 (0.86×)	69.9 (0.93×)	47.0 (0.81×)

accuracy(victim blackbox)

accuracy(knockoff)

$\Rightarrow > 0.81 \times$ accuracy of blackbox recovered

Learning with Less Information? Yes!



⇒ Robust to various **passive defense mechanisms**:
e.g. argmax, top-k, rounding, ...



mp

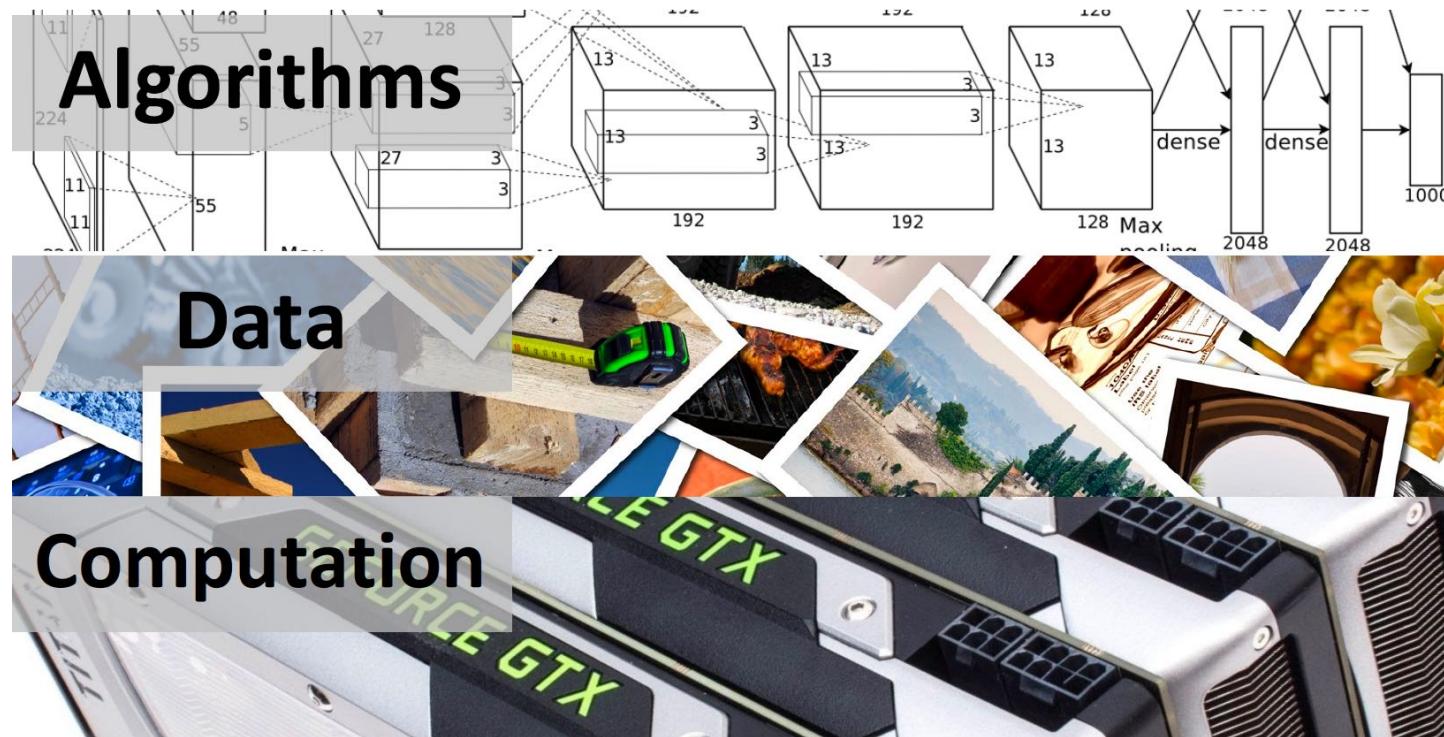
max planck institut
informatik

SIC Saarland Informatics
Campus

**Deep Learning
has become an important tool
for essentially all computer vision tasks**

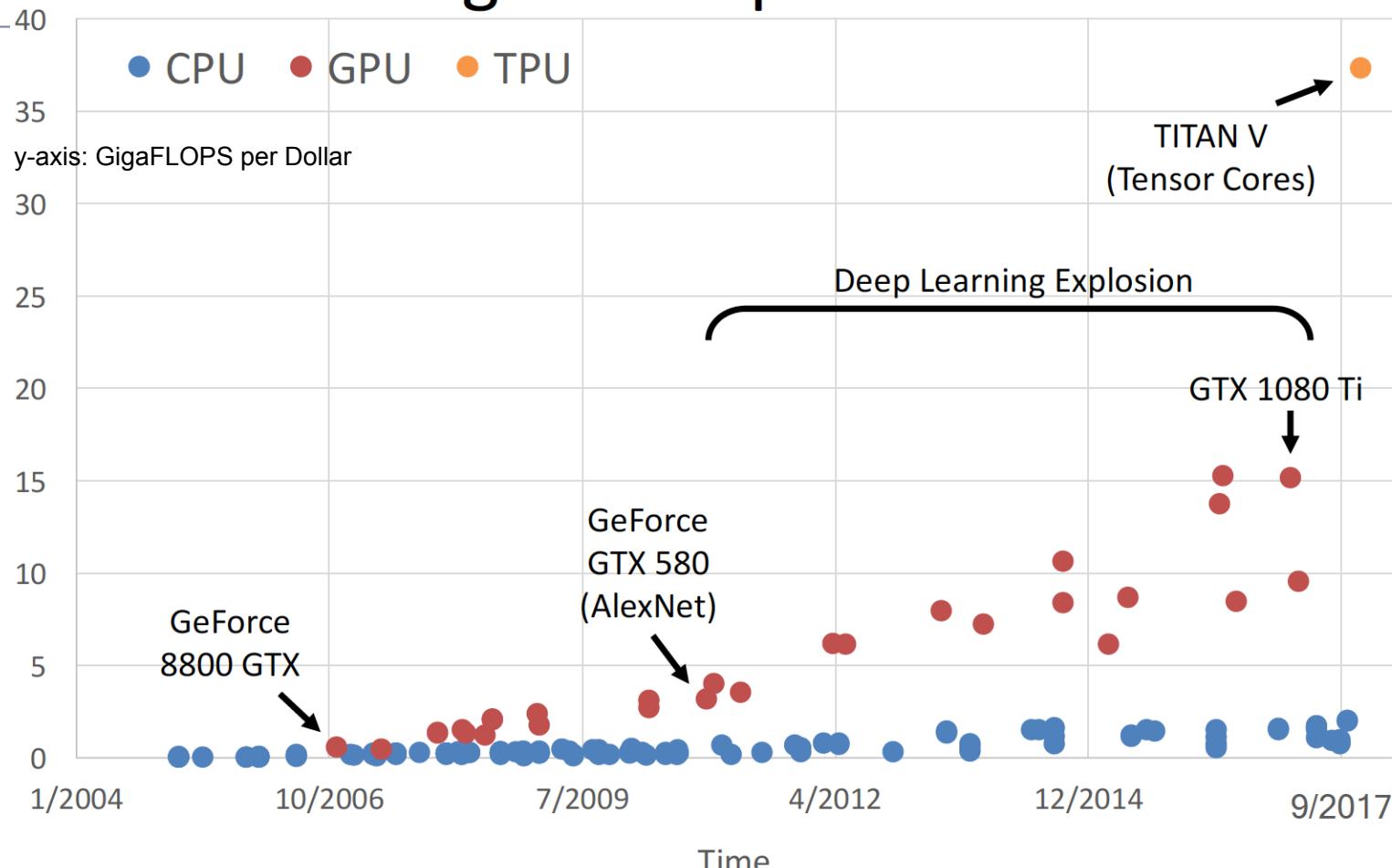
Let's briefly discuss CNNs
(Convolutional Neural Networks)

Ingredients for Deep Learning



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

GigaFLOPs per Dollar



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



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



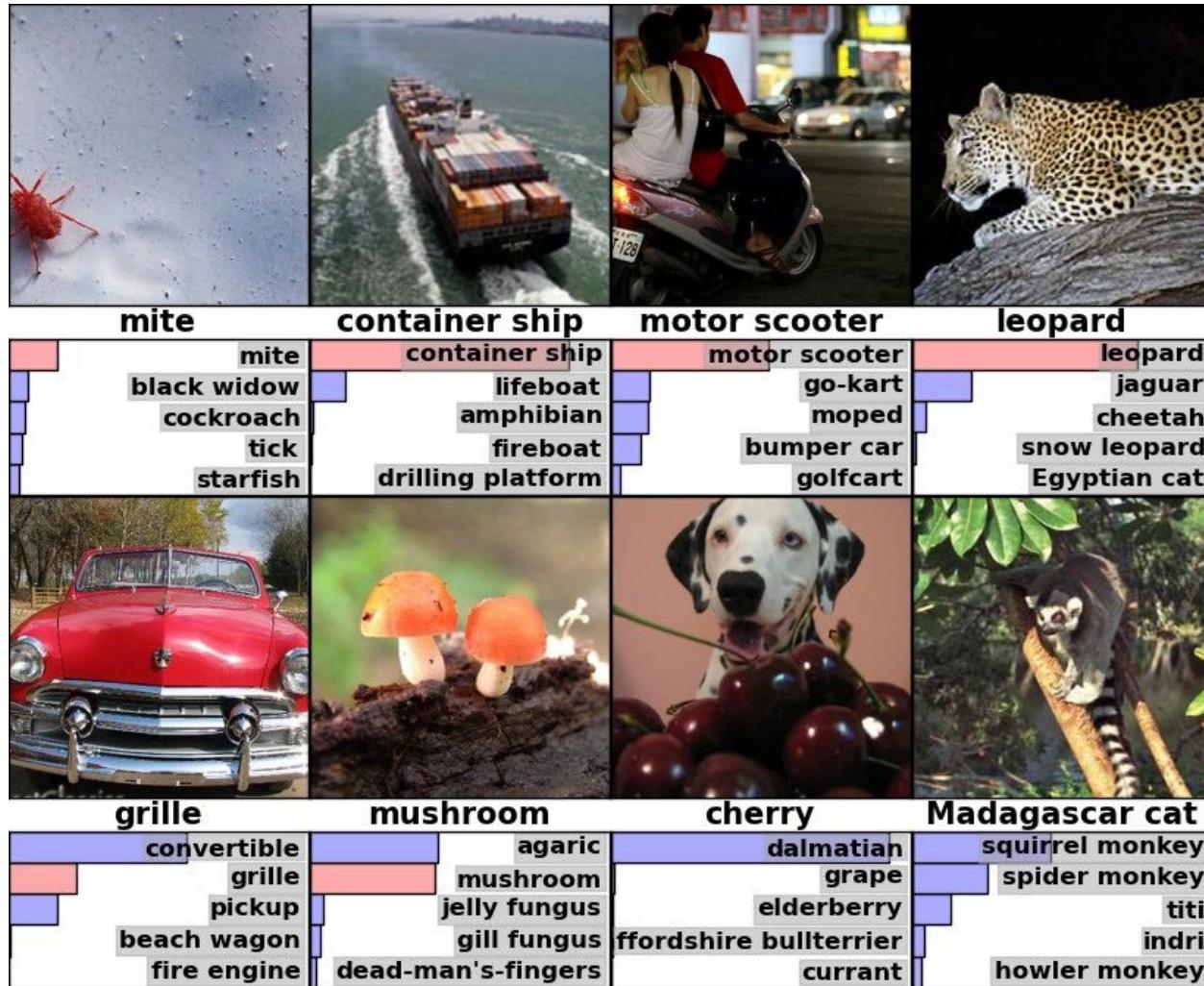
Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

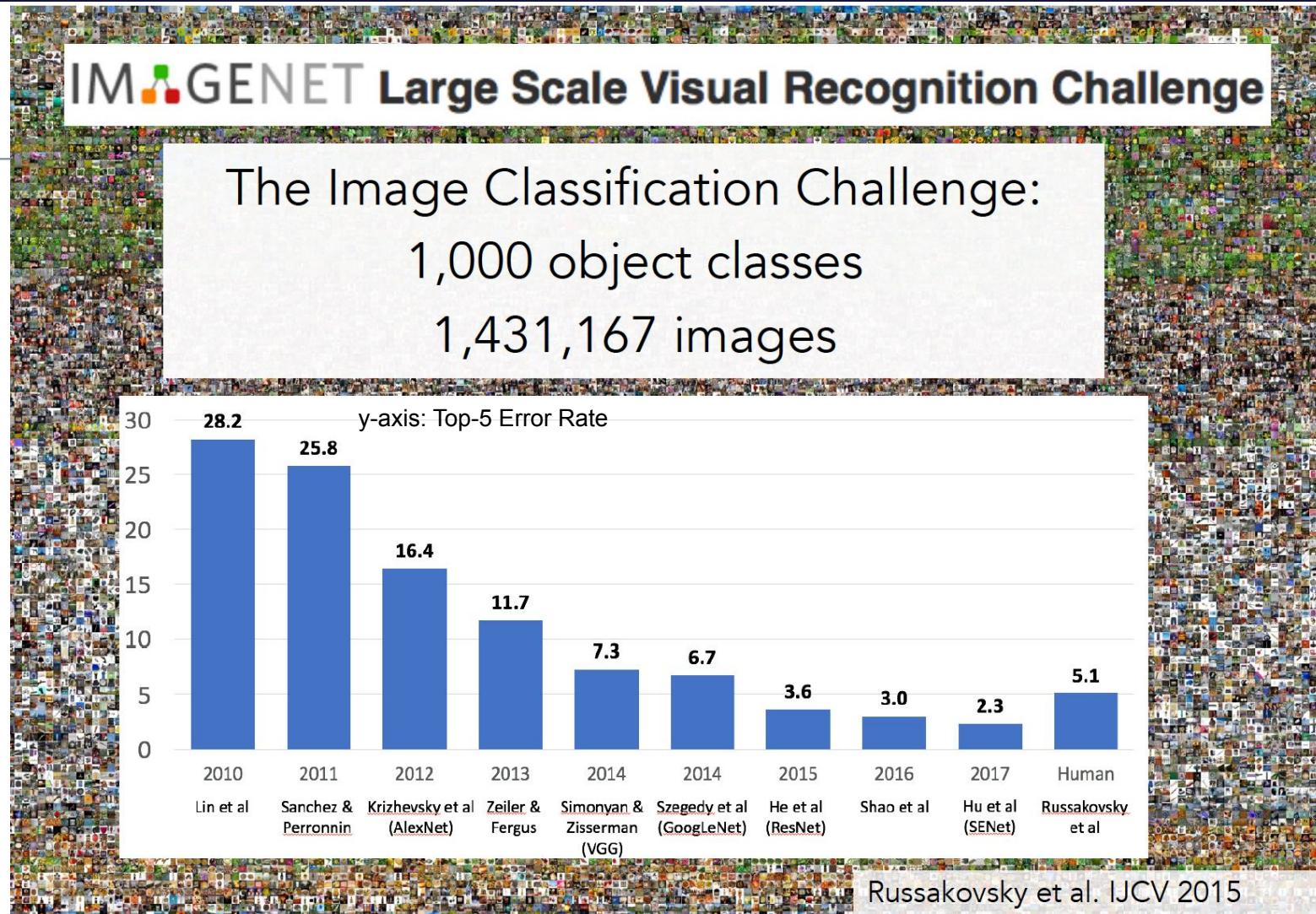


Russakovsky et al. IJCV 2015

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

Validation classification



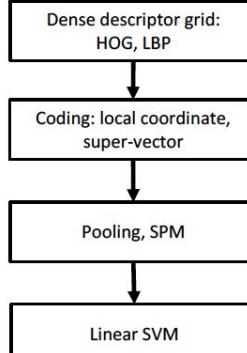


slide credit: Fei-Fei, Justin Johnson, Serena Yeung

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

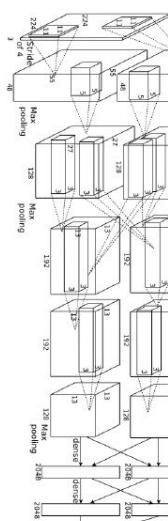


[Lin CVPR 2011]

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

Pooling
 Convolution
 n
 Softmax
 Other



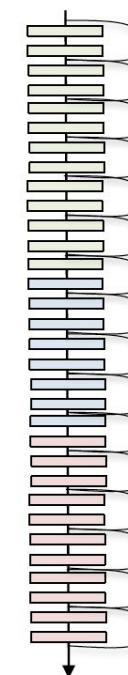
[Szegedy arxiv 2014]

VGG



Year 2015

MSRA



[He ICCV 2015]

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

How deep is enough?

11

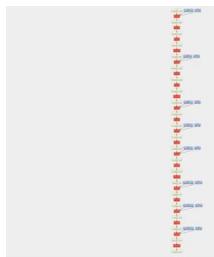
AlexNet (2012)



How deep is enough?

13

AlexNet (2012)



VGG-M (2013)



VGG-VD-16 (2014)

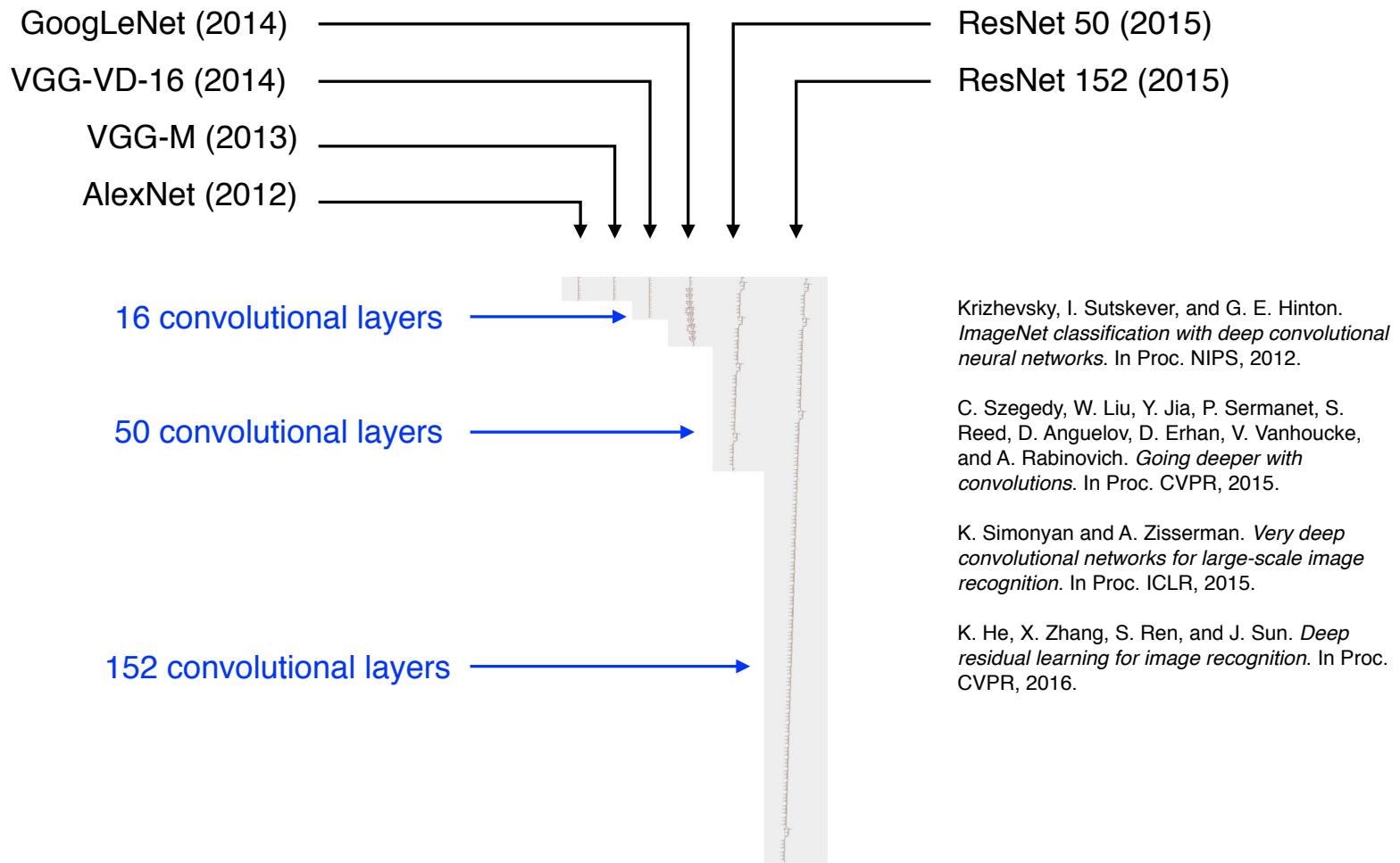


GoogLeNet (2014)



How deep is enough?

15





■ ■ ■ p

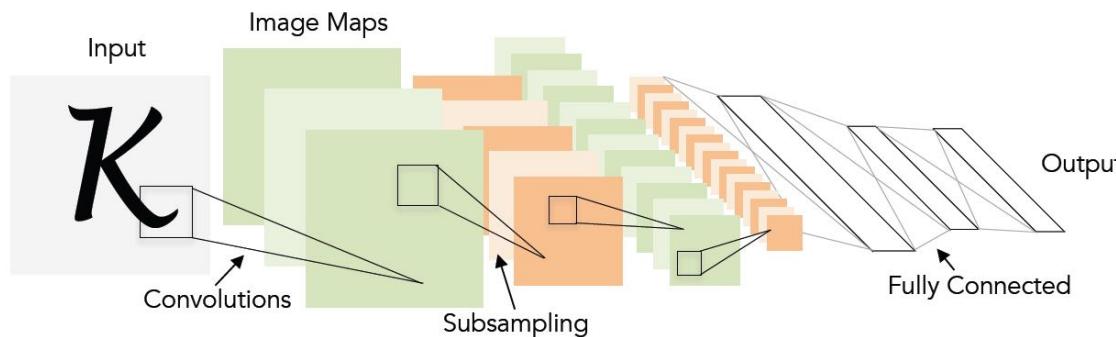
max planck institut
informatik

SIC Saarland Informatics
Campus

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

1998

LeCun et al.



of transistors



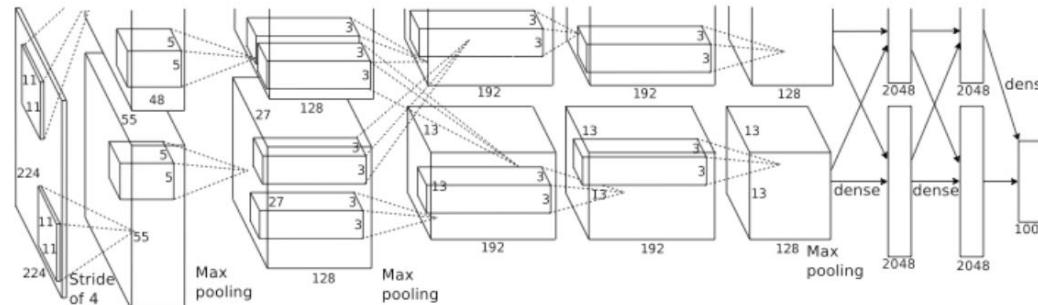
10^6

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



■ ■ ■ p

max planck institut
informatik

SIC Saarland Informatics
Campus

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

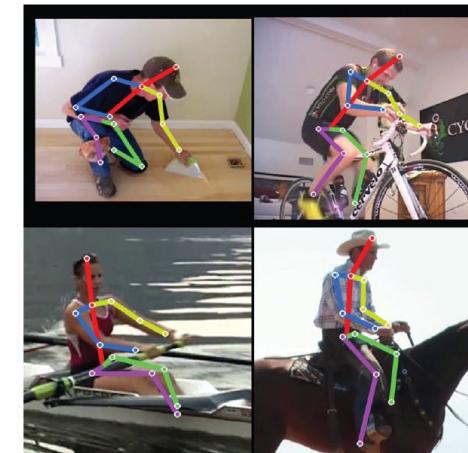
but there exist many other computer vision tasks
where Deep Learning is also an essential ingredient

a few examples...

Human Pose Estimation

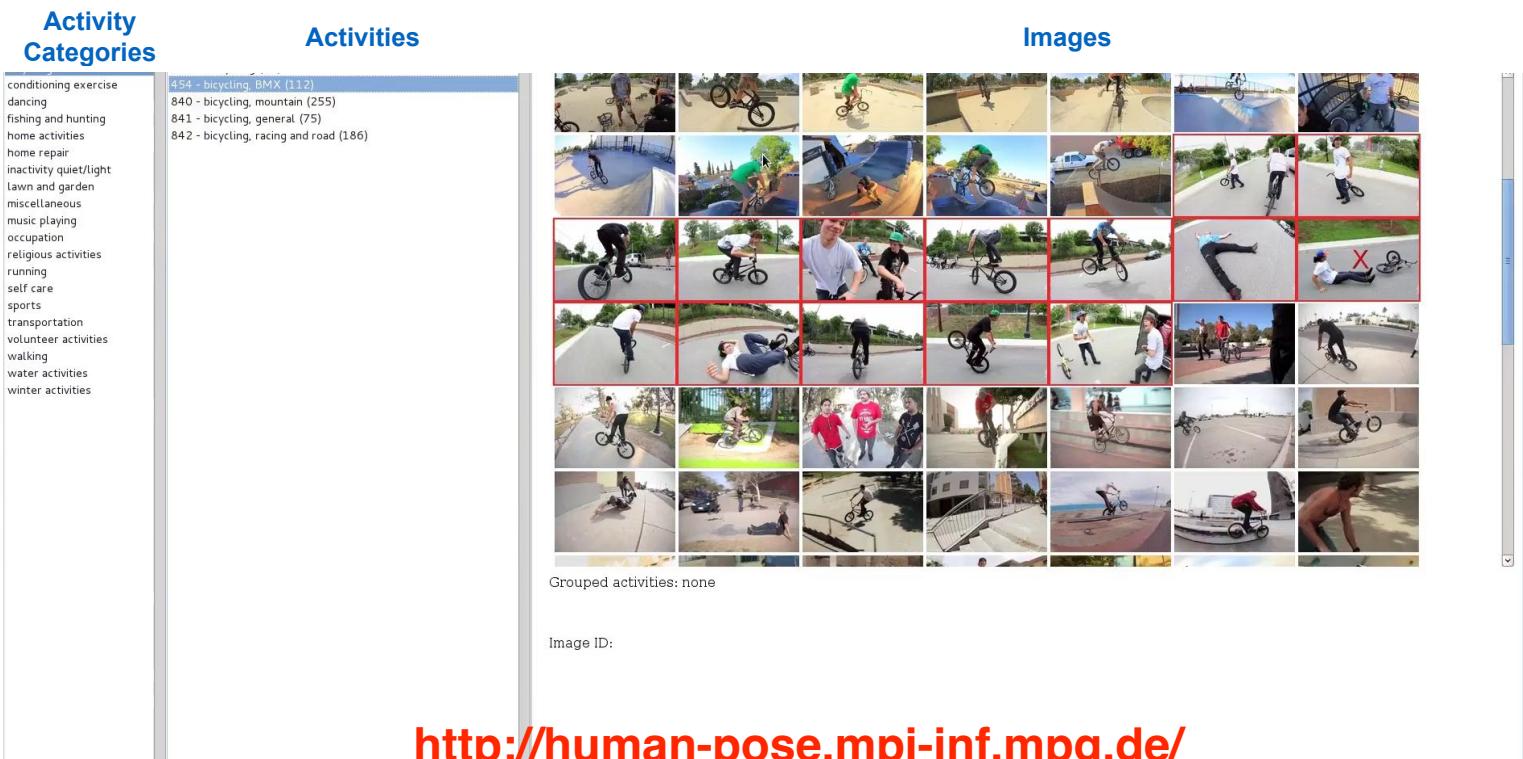
- **Single Person Pose Estimation** - two “phases”

- ▶ **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], ...



MPII Human Pose Dataset: Dataset demo

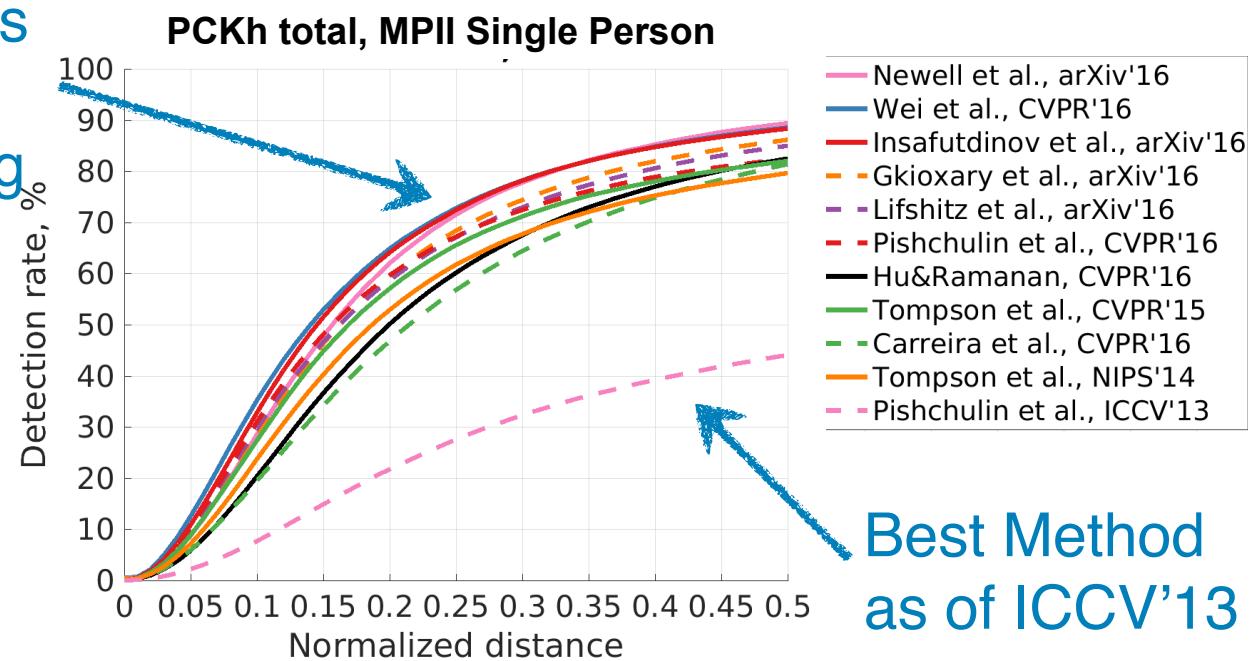
- 410 human activities (after merging similar activities)
- over 40,000 annotated poses
- over 1.



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

Analysis - overall performance

Best Methods
today:
deep learning
“takes” over



- ✓ 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



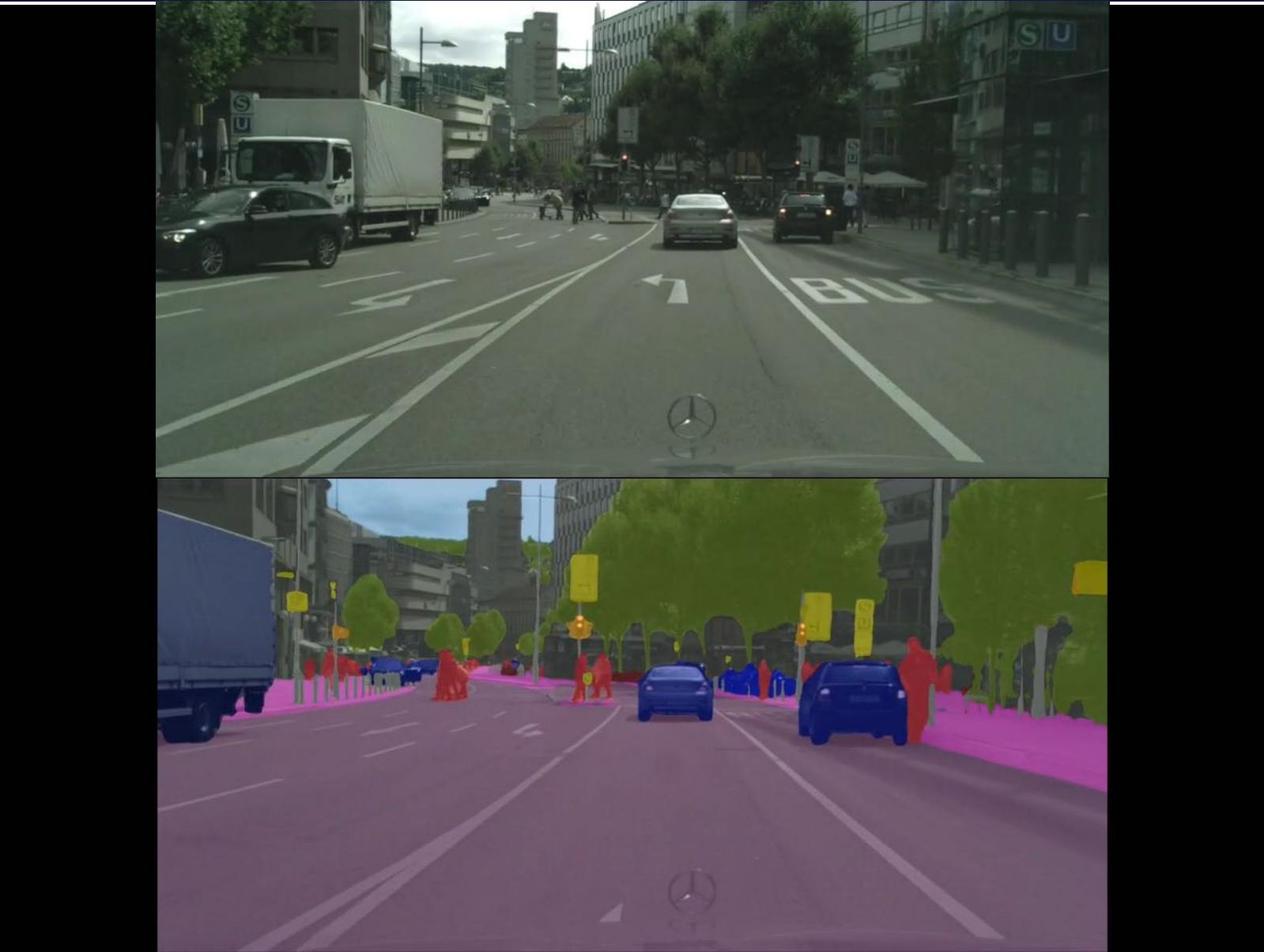
- Joint effort of:



High Level Com-

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

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



High Level Computer Vision | Bernt Schiele

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]

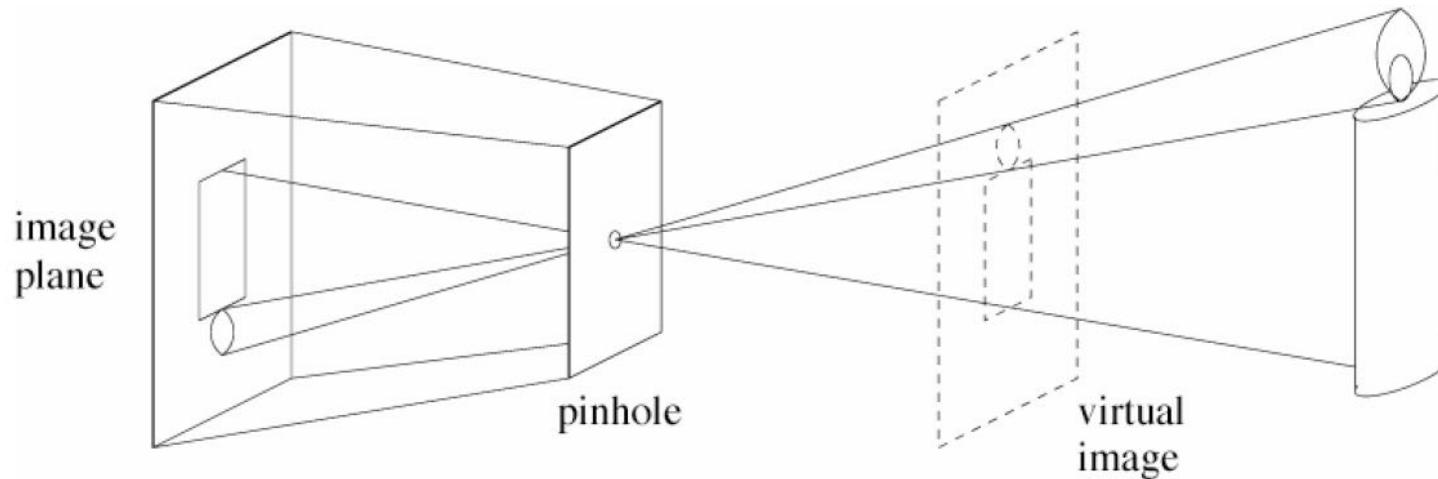


Basic Concepts and Terminology

Computer Vision vs. Computer Graphics

Pinhole Camera (Model)

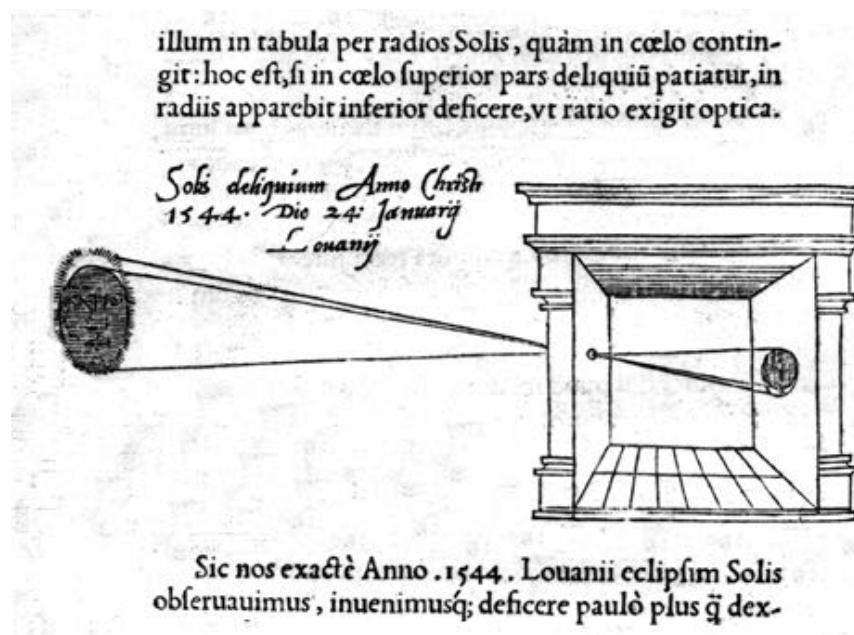
- (simple) standard and abstract model today
 - ▶ box with a small hole in it



Camera Obscura

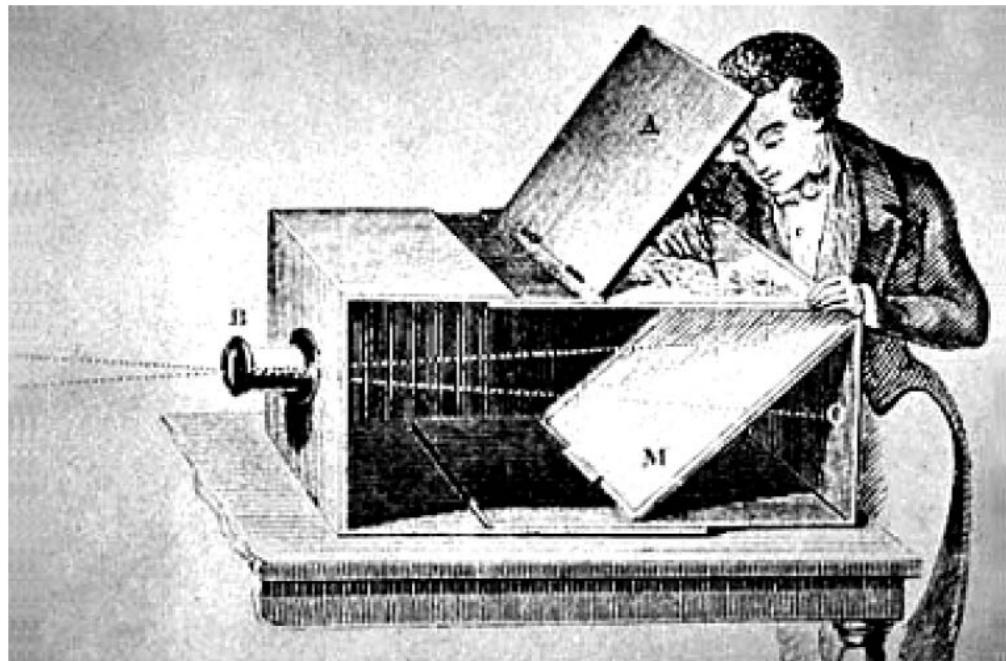
- around 1519, Leonardo da Vinci (1452 - 1519)
 - ▶ http://www.acmi.net.au/AIC/CAMERA_OBSCURA.html

▶ “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”



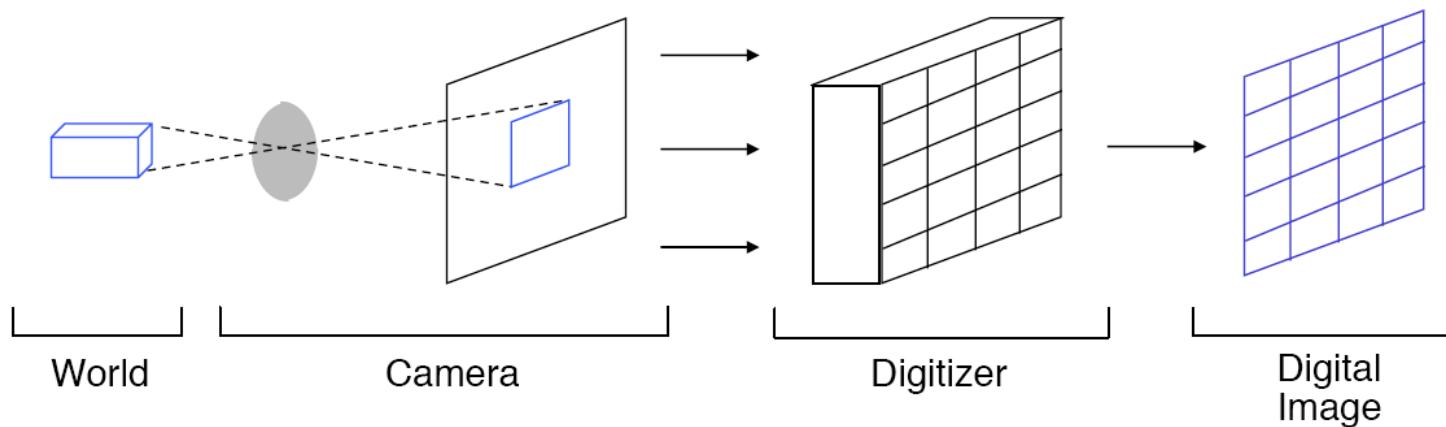
Principle of pinhole....

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



Digital Images

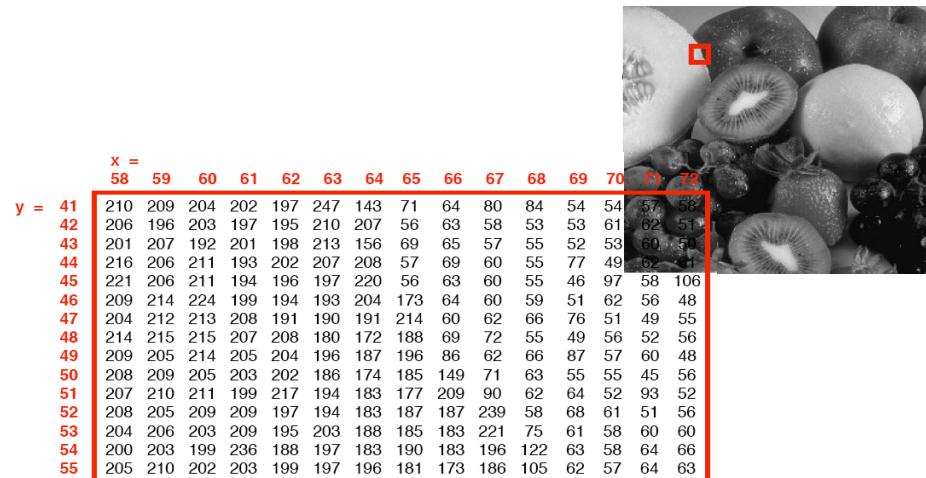
- Imaging Process:
 - ▶ (pinhole) camera model
 - ▶ digitizer to obtain digital image



(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?
 - ▶ ...
- computer vision = the problem of ‘inverse graphics’ ...?

- ‘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?
 - ▶ ...





Case Study

... object recognition

Case Study: Computer Vision & Object Recognition

- how do you recognize
 - ▶ the banana?
 - ▶ the glass?
 - ▶ 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

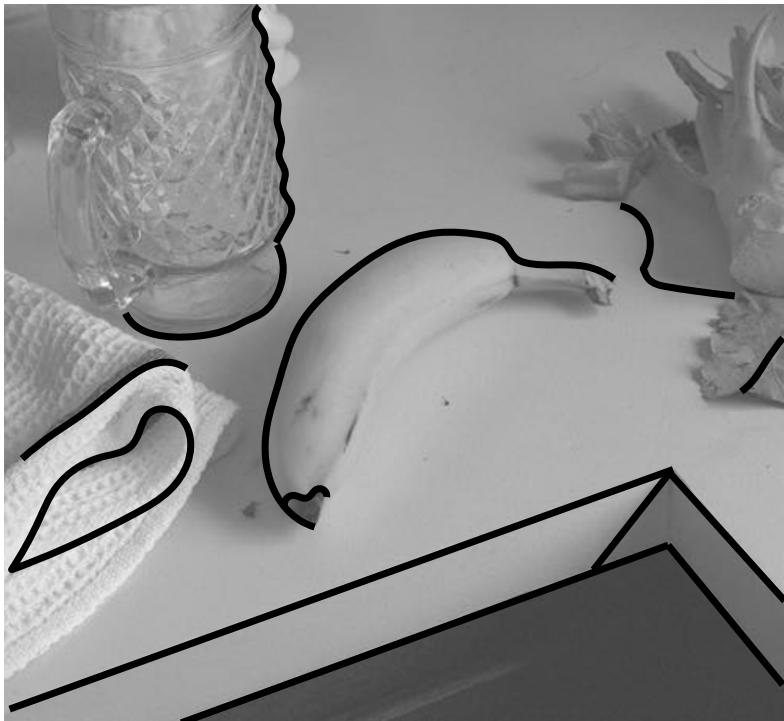
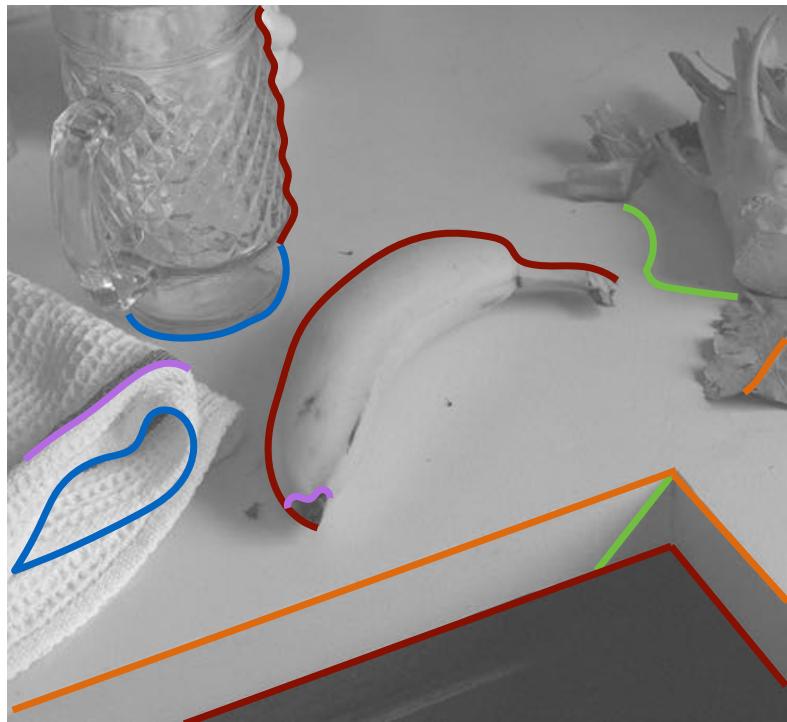


Image Edges: What are edges? Where do they come from?



- 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

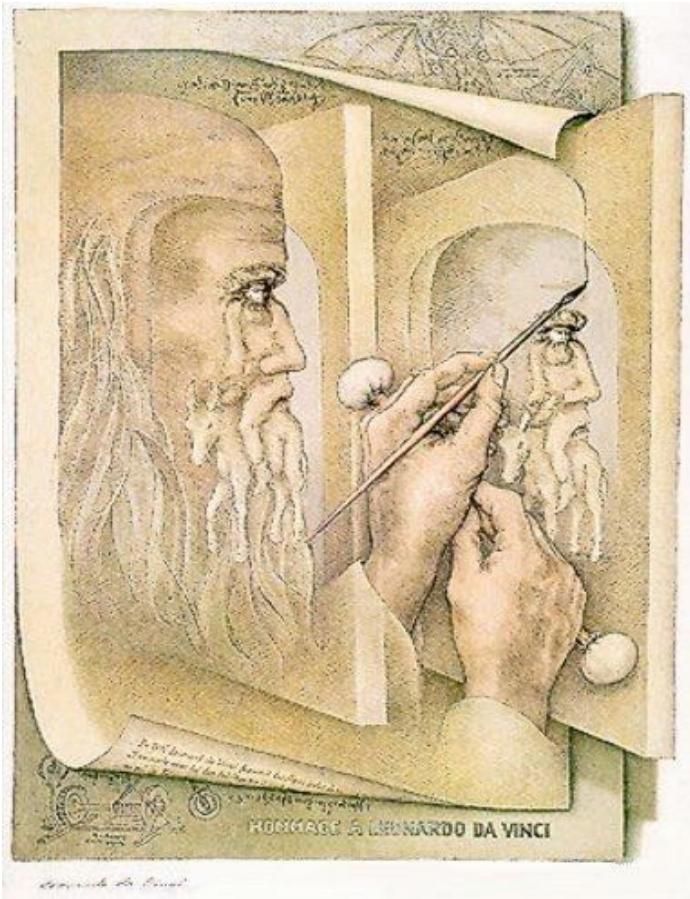
Seymour Papert

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".

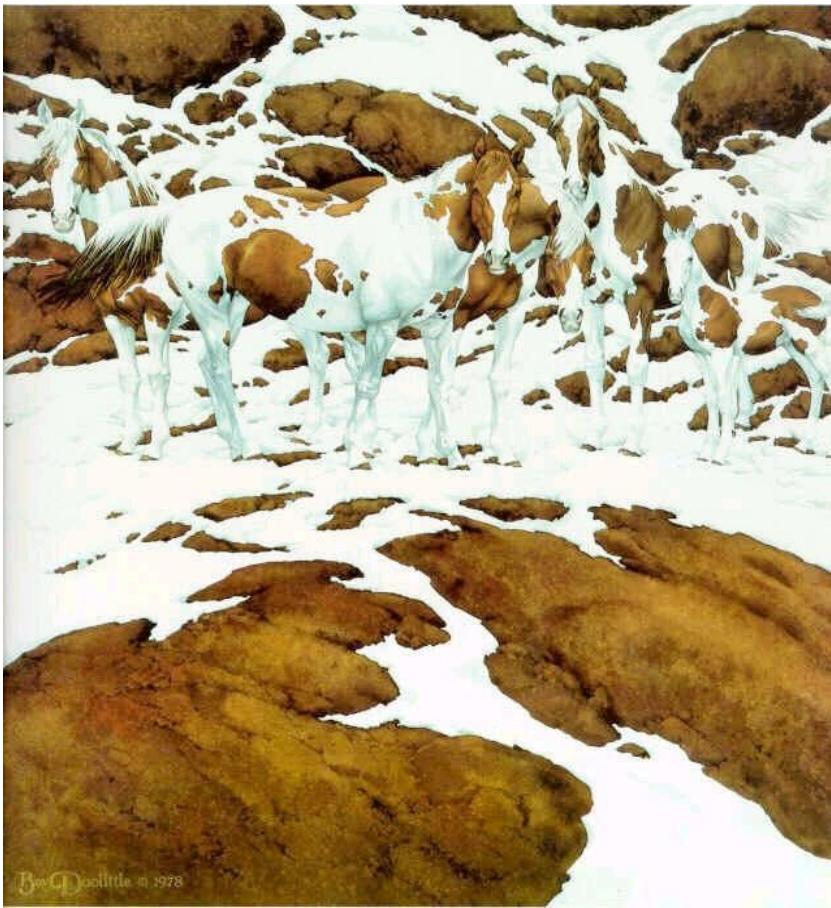
slide credit: Fei-Fei, Justin Johnson, Serena Yeung



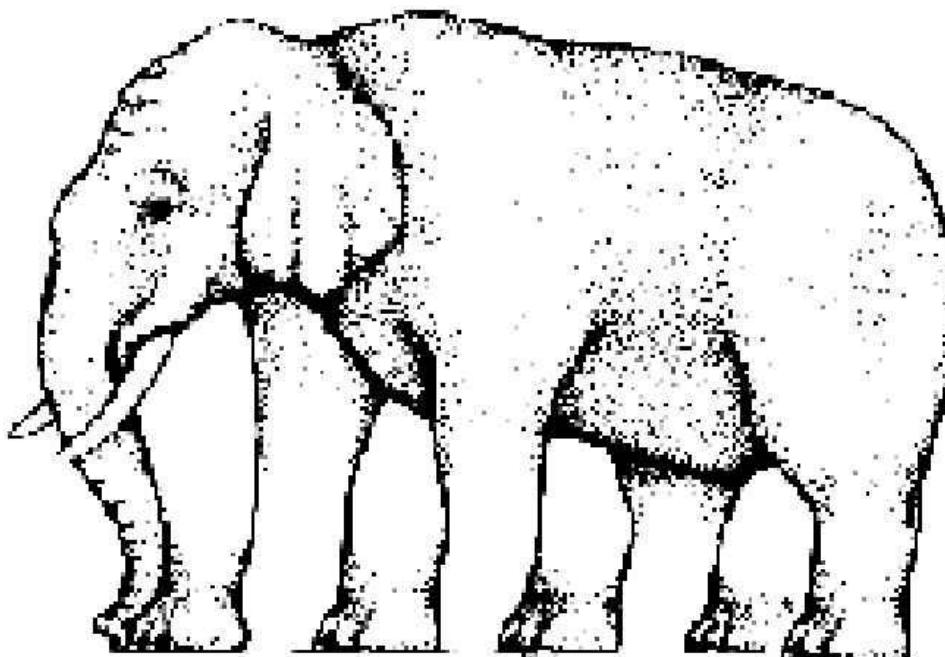
Complexity of Recognition



Complexity of Recognition



Complexity of Recognition



Recognition: the Role of Context

- Antonio Torralba

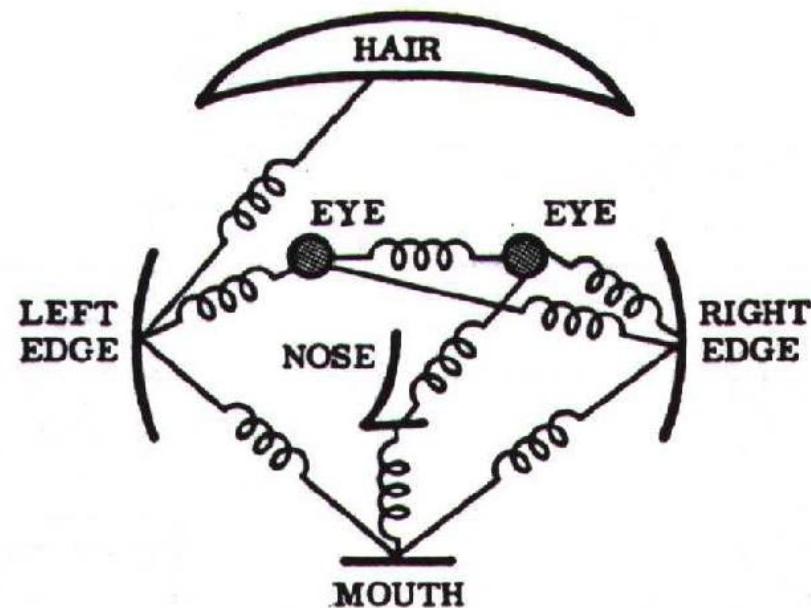


Complexity of Recognition



Class of Models: Pictorial Structure

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



Deformations



A



B



C

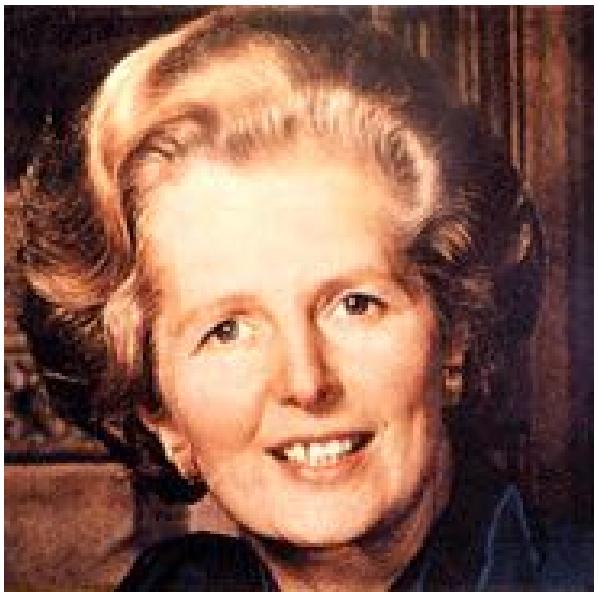


D

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'



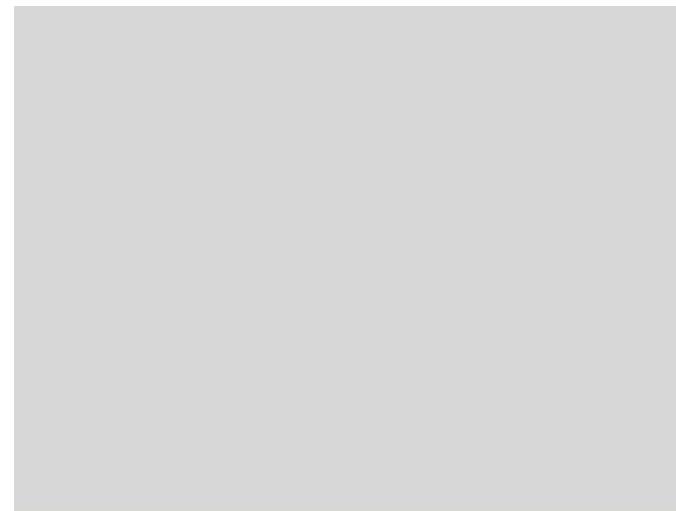
Which Level is right for Object Classes?

- 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 :-)

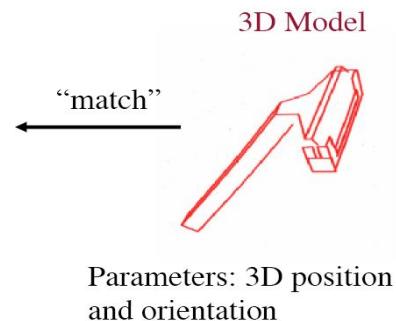
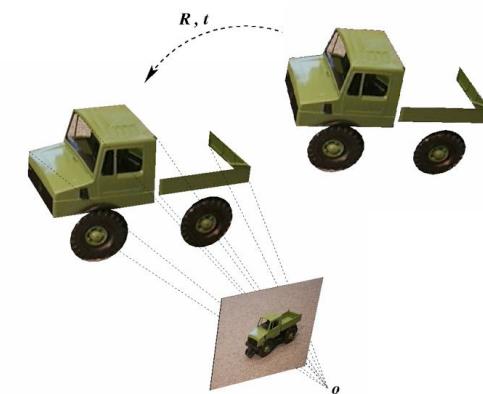
Object Recognition & Segmentation

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

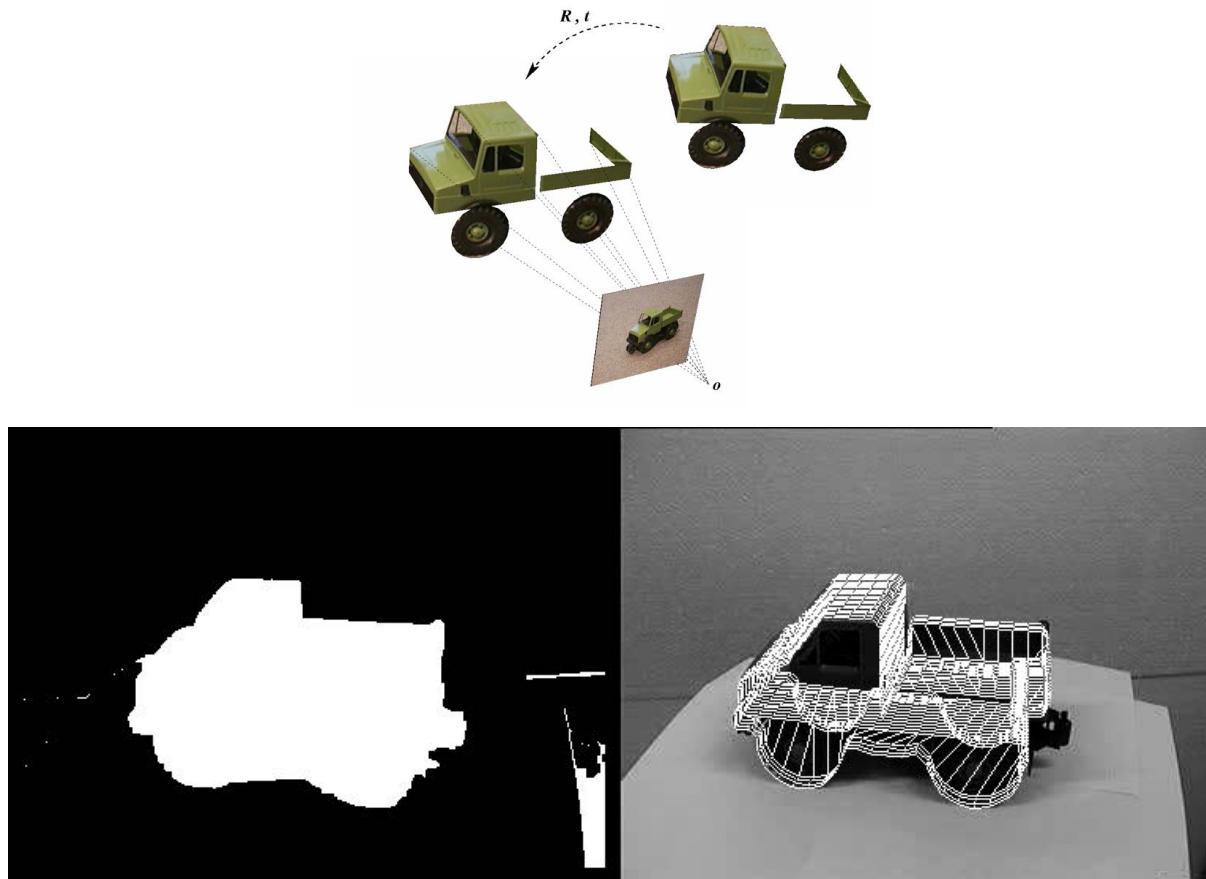


Object Recognition & Localization

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



Localization: Example Video



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
- Case Study
 - ▶ Object Recognition — intuition from human vision...