Advanced Computer Vision

Elisa Ricci & Nicu Sebe

some slides from S. Hochreiter, I. Laptev, G. Mori, M. Ryoo

Top 10 AI technology trends

Deep learning theory
The information bottleneck principle explains how a deep neural network learns.

Capsule networks
New type of deep neural network that learns with fewer errors and less data, by preserving key hierarchical relationships.

Deep reinforcement learning
This technique combines reinforcement learning with deep neural networks to learn by interacting with the environment.

Generative adversarial networks
A type of unsupervised deep learning system, implemented as two competing neural networks, enabling machine learning with less human intervention.

Lean and augmented data learning
Different techniques that enable a model to learn from less data or synthetic data.

Probabilistic programming
A high-level language that makes it easy for developers to define probability models.

Hybrid learning models
Approach that combines different types of deep neural networks with probabilistic approaches to model uncertainty.

Automated machine learning
Technique for automating the standard workflow of machine learning.

Digital twin
A virtual model used to facilitate detailed analysis and monitoring of physical or psychological systems.

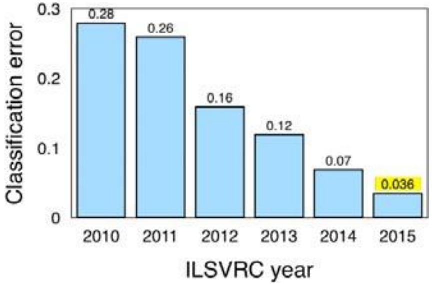
Explainable artificial intelligence
Machine learning techniques that produce more explainable models while maintaining high performance.

Successes of Deep Learning in AI

The New York Times

A Learning Advance in Artificial Intelligence Rivals Human Abilities

FLOWER
FIELD
SKY
CLOUDS



IMAGENET

Deep Learning for self-driving cars

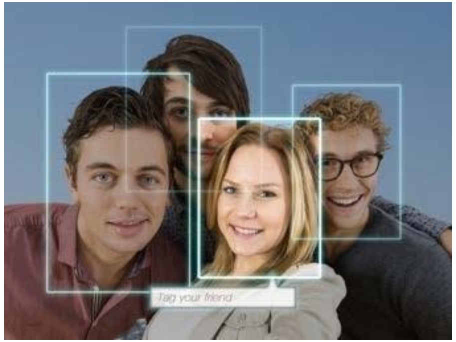
Google's DeepMind Masters Atari Games

Google Translate

English | Russian | Chinese (Simplified)

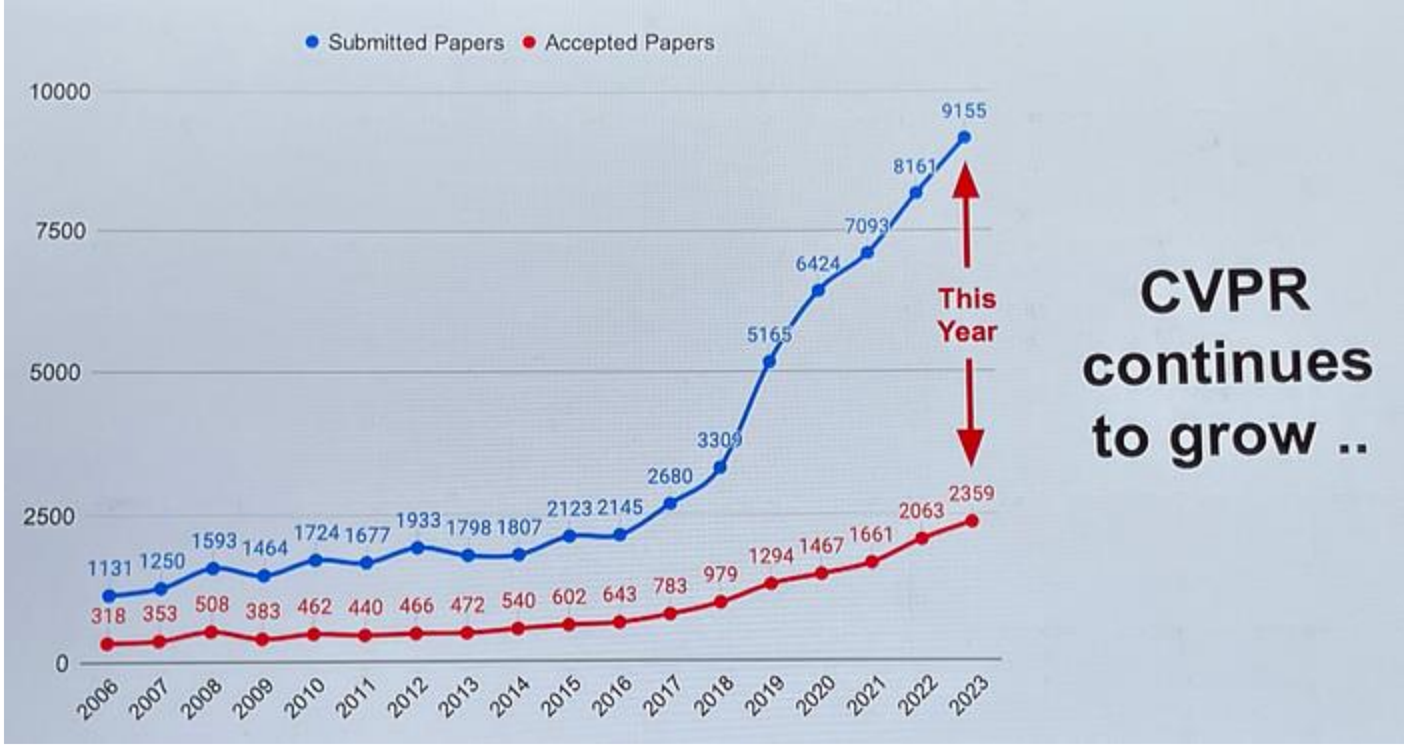
Time flies like an arrow

时间过得很快像箭



Face Recognition

CVPR 2023 Statistics



CVPR 2023 Statistics

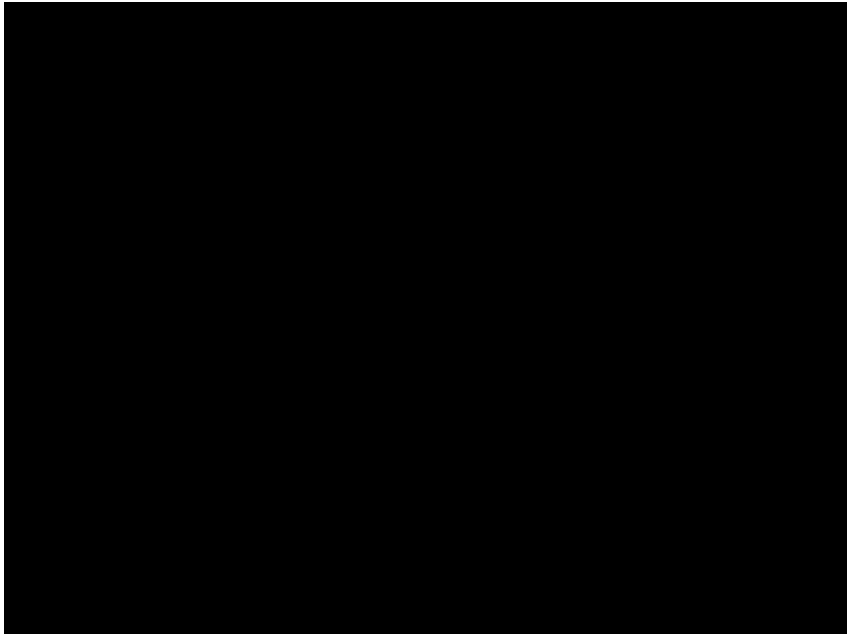
1	3D from multi-view and sensors	1,090	246
2	Image and video synthesis and generation	889	185
3	Humans: Face, body, pose, gesture, movement	813	166
4	Transfer, meta, low-shot, continual, or long-tail learning	688	153
5	Recognition: Categorization, detection, retrieval	673	139
6	Vision, language, and reasoning	631	118
7	Low-level vision	553	126
8	Segmentation, grouping and shape analysis	524	113
9	Deep learning architectures and techniques	485	92
10	Multi-modal learning	450	89

CVPR 2023 Statistics

11	3D from single images	431	91
12	Medical and biological vision, cell microscopy	420	53
13	Video: Action and event understanding	373	83
14	Autonomous driving	359	69
15	Self-supervised or unsupervised representation learning	349	71
16	Datasets and evaluation	344	54
17	Scene analysis and understanding	276	54
18	Adversarial attack and defense	274	61
19	Efficient and scalable vision	252	48
20	Computational imaging	226	53
21	Video: Low-level analysis, motion, and tracking	215	46
22	Vision applications and systems	171	35
23	Vision + graphics	155	32
24	Robotics	141	23
25	Transparency, fairness, accountability, privacy, ethics in vision	129	30
26	Explainable computer vision	107	24
27	Embodied vision: Active agents, simulation	80	14
28	Document analysis and understanding	72	12
29	Machine learning (other than deep learning)	65	14
30	Physics-based vision and shape-from-X	55	12
31	Biometrics	51	11
32	Others	47	12
33	Optimization methods (other than deep learning)	46	12
34	Photogrammetry and remote sensing	38	8
35	Computer vision theory	33	5
36	Computer vision for social good	25	5

So is AI solved?

pedestrian detection FAIL (Volvo S60 Pedestrian Detection System Test)



https://www.youtube.com/watch?v=w2pwxv8rFkU

Pedestrian detection is only active at <22mph and >2mph. Once the brakes have been engaged, they do release afterwards assuming the driver who wasn't initially paying attention is now fully aware of the audible alarm and flashing red lights on the windshield. This is why this guy almost got run over: the system was working as designed but he was not behind the wheel to take back!!!

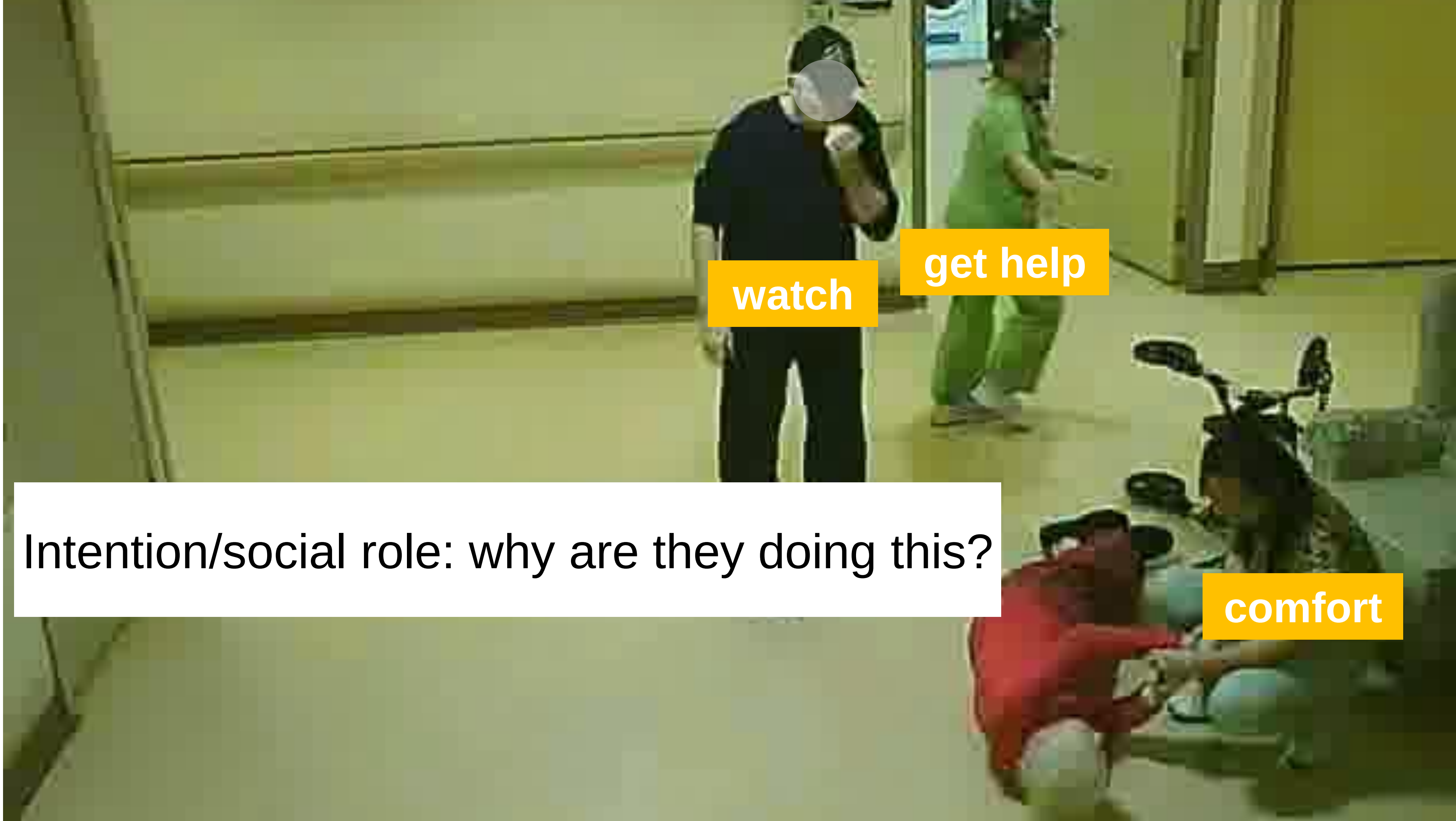
Challenges

What does computer vision involve?


Detection: are there people?

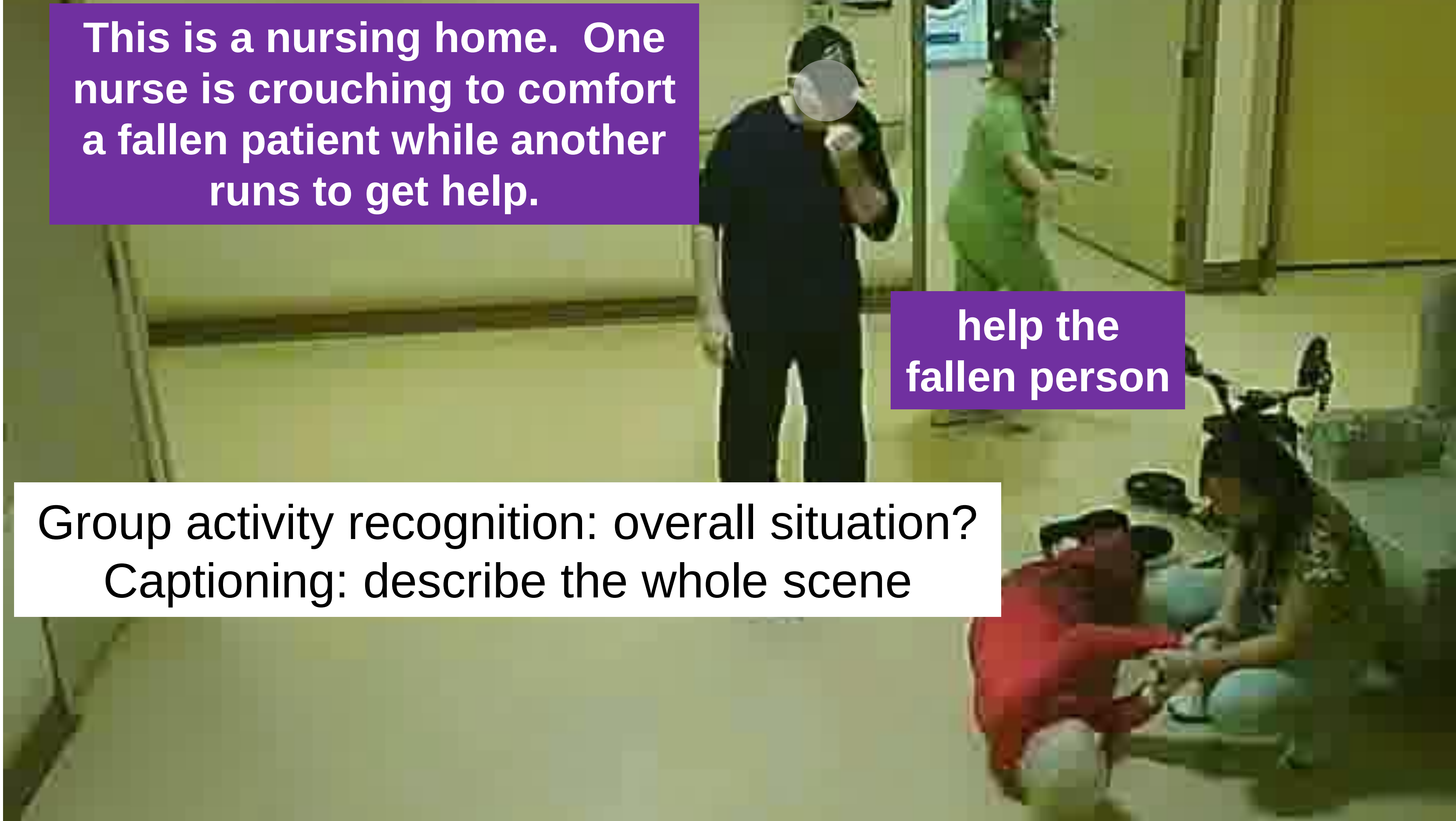
Objects and scenes: where are they?

Action recognition: what are they doing?



Intention/social role: why are they doing this?

This is a nursing home. One nurse is crouching to comfort a fallen patient while another runs to get help.



help the fallen person

Group activity recognition: overall situation?
Captioning: describe the whole scene


indoor scene
This is a nursing home. One nurse is crouching to comfort a fallen patient while another
long term care facility
watch
stand
get help
run
help the fallen person
walker
chair
“AI-complete”: full semantic understanding necessary for success
comfort
squat
fall
floor

Why is computer vision important?

User videos

~300 hours of videos per minute

- Video indexing and retrieval

Monitoring cameras

Streaming videos 24/7

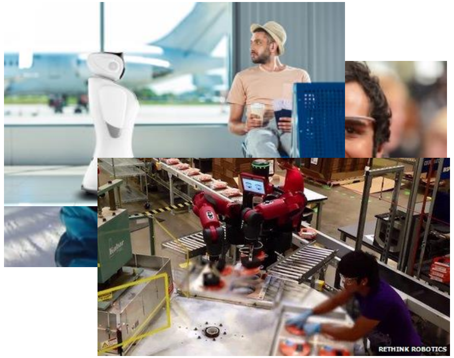
- Surveillance
- Patient/elderly monitoring

Media

Content analysis, experience enrichment

- Recommendation systems
- Advertising
- Sports analytics

Wearables/robots



Streaming videos to be analyzed in real-time

- Lifelogging
- Robot operations and actions

Why is it difficult?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing, etc.

Action *Hugging*:


…

- **Manual collection of training samples is prohibitive:** many classes, rare occurrence




…

- **Vocabulary is not well-defined**

Action *Open*:



…

[slide credit: Ivan Laptev]

Why is it difficult?

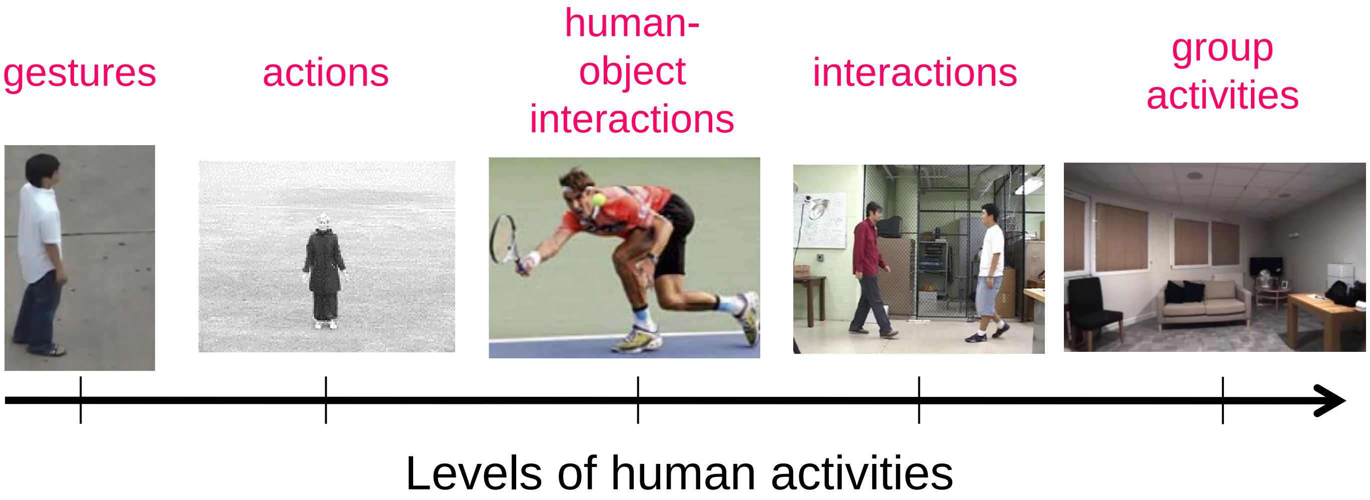
Humans Outperform Computers

- Humans generalize better that deep learning systems
- Humans can handle out of distribution data
- Humans are more robust to domain shifts
- Humans can handle new situations, objects, environments
- Humans are better in building abstractions

[slide credit: Sepp Hochreiter]

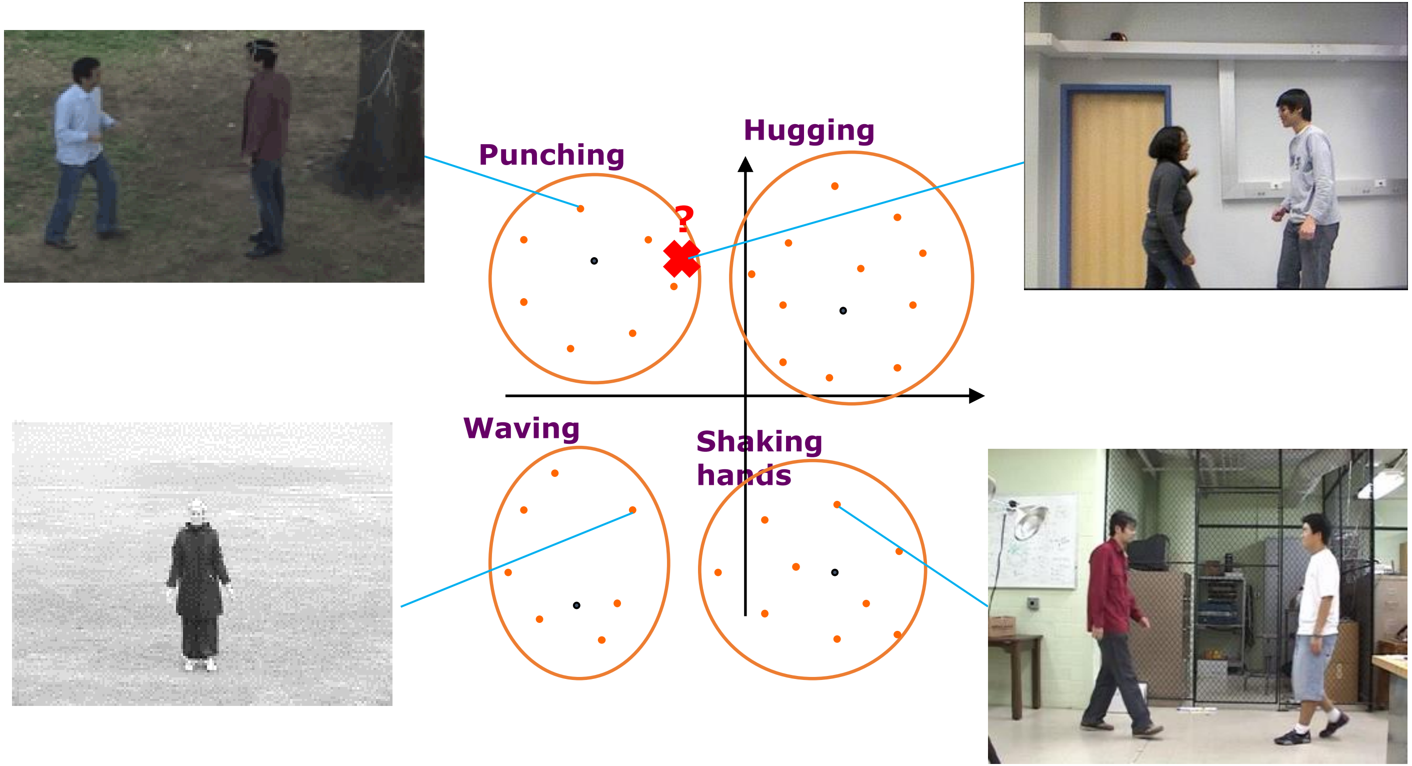
Example: Human activity recognition

- There are various types/levels of activities
  - The ultimate goal is to make computers recognize all of them reliably



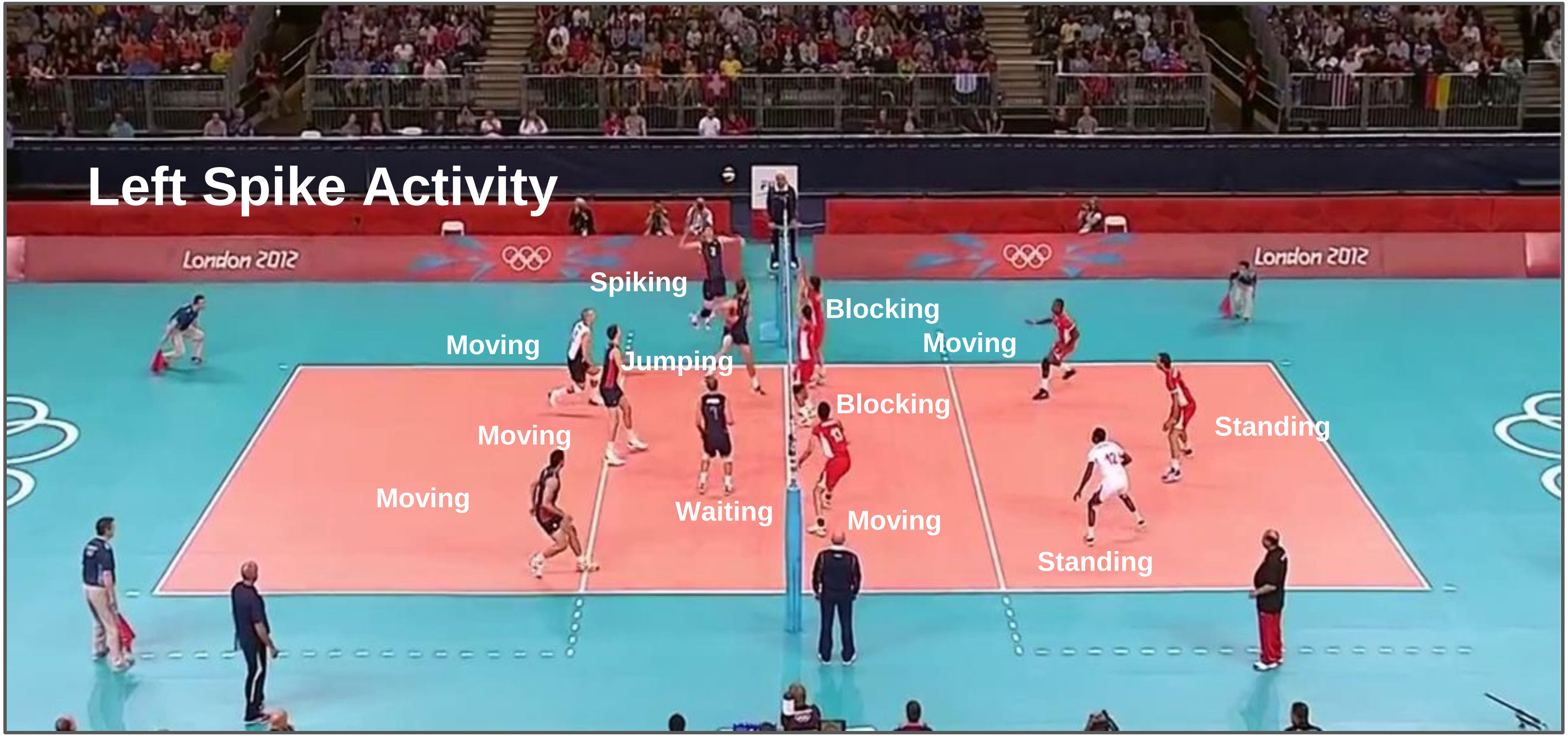
Example: Activity classification

- Categorization of segmented videos
  - Input = a video segment containing only one activity



Example: Action Recognition

Group activity recognition



Relationships?

Left team:

- **Spike** by a player
- **Fake jump** by another player

Right team:

- 2 players jump to **block**

Learning Vision: Human vs. Computer

Computer Vision: Static Images (or Videos)

- Directly learn to detect objects
- Learn to segment

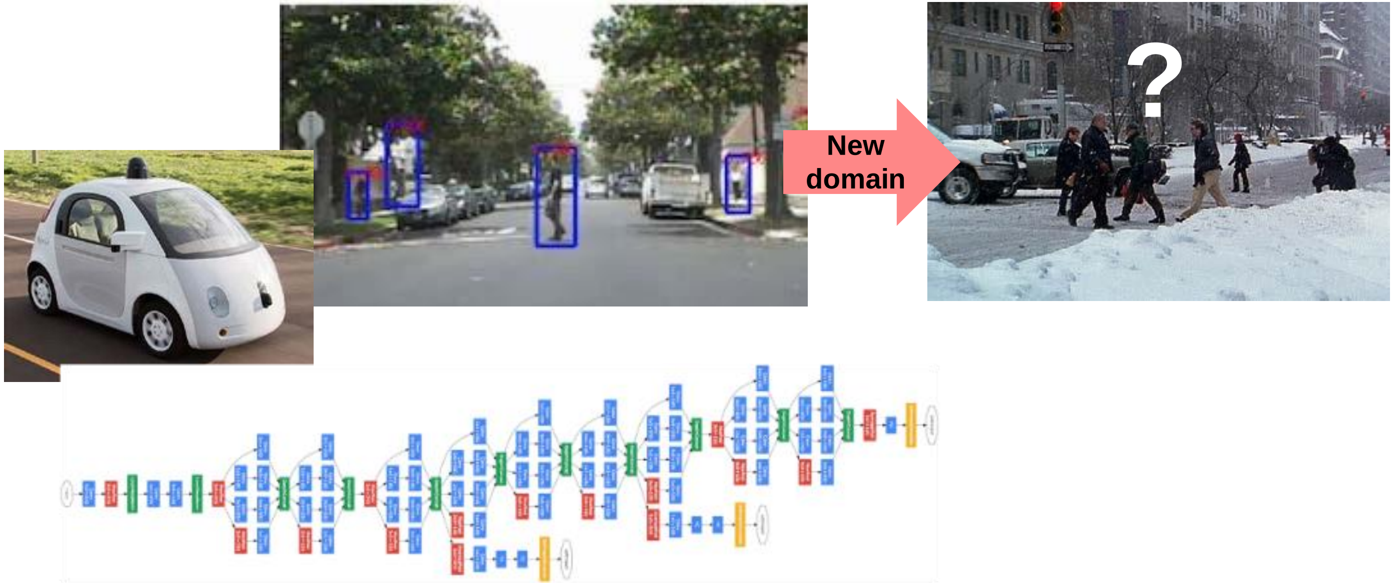
Human children: Sequence of related images

- Learn unity of objects
- Learn object permanence
- Learn to solve the binding problem

[slide credit: Sepp Hochreiter]

Major limitation of deep learning

Not data efficient: Learning requires **millions** of labeled examples, models do not generalize well to new domains; not like humans!



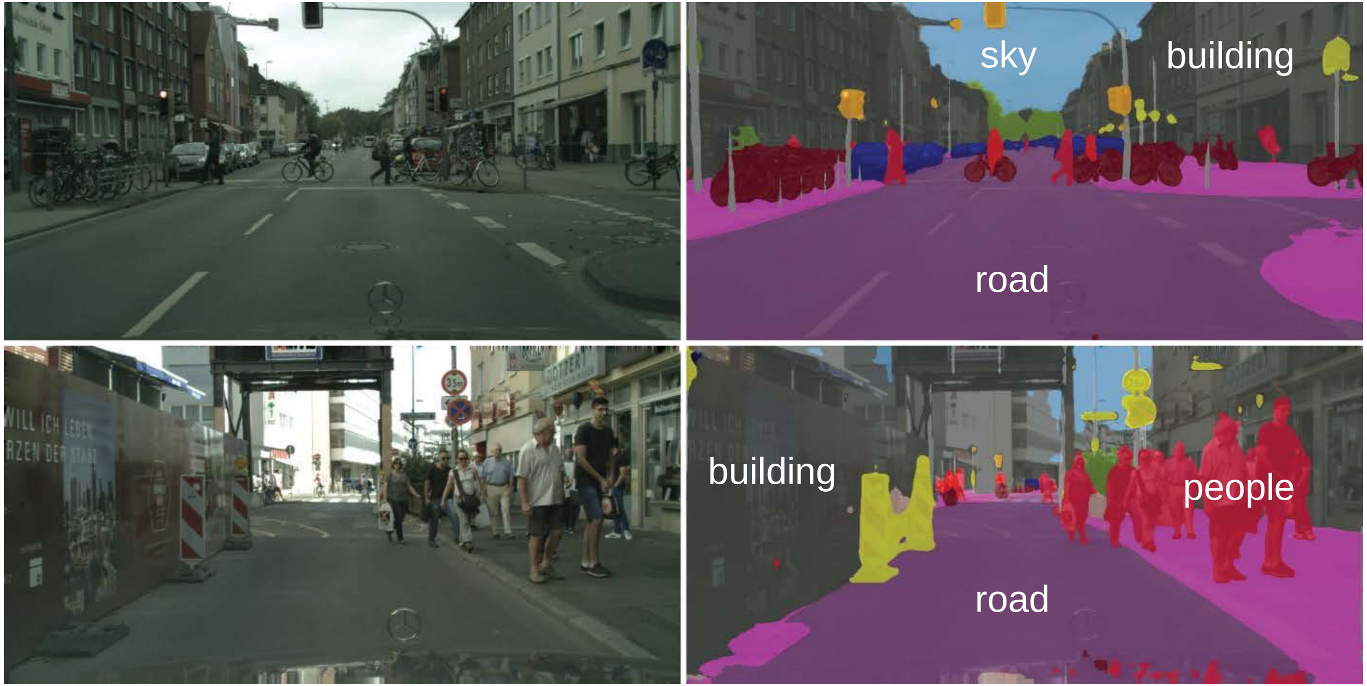
“What you saw is not what you get”

What your net is trained on

What it is asked to label

“Dataset Bias”
“Domain Shift”
“Domain Adaptation”
“Domain Transfer”

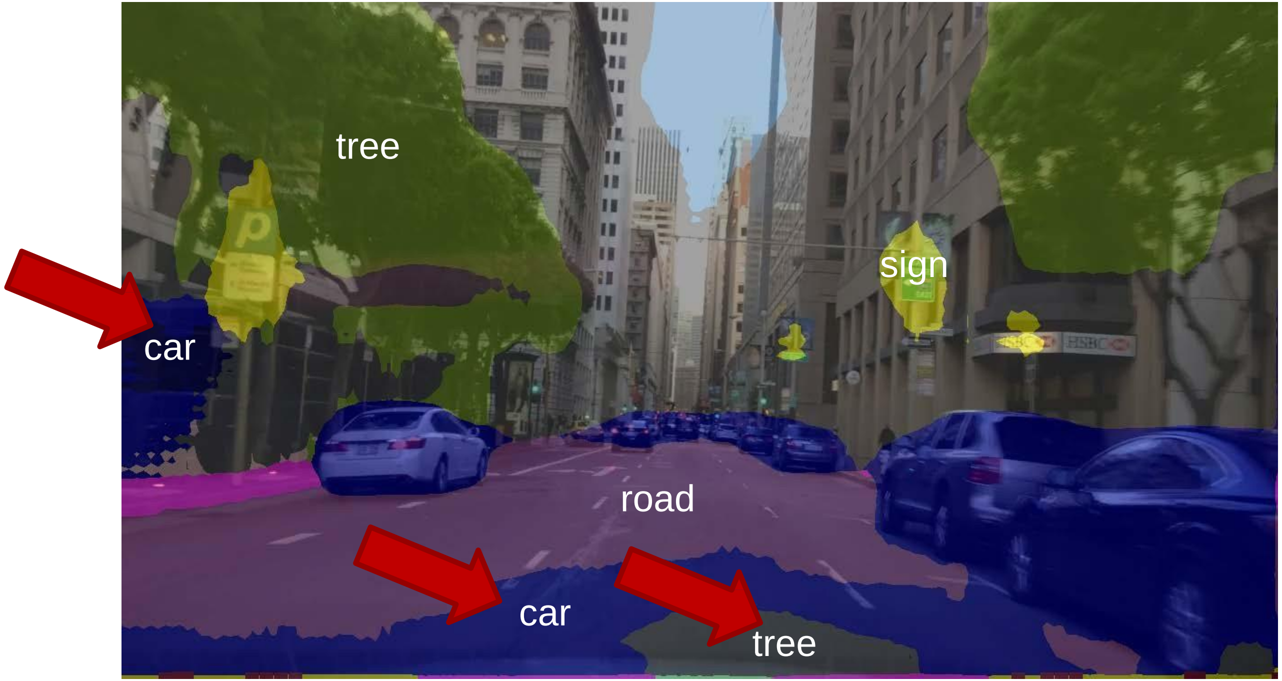
Example: Scene segmentation



Train on Cityscapes, Test on Cityscapes

J. Hoffman, D. Wang, F. Yu, T. Darrell, FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation, Arxiv 2016

Domain shift: Cityscapes to SF



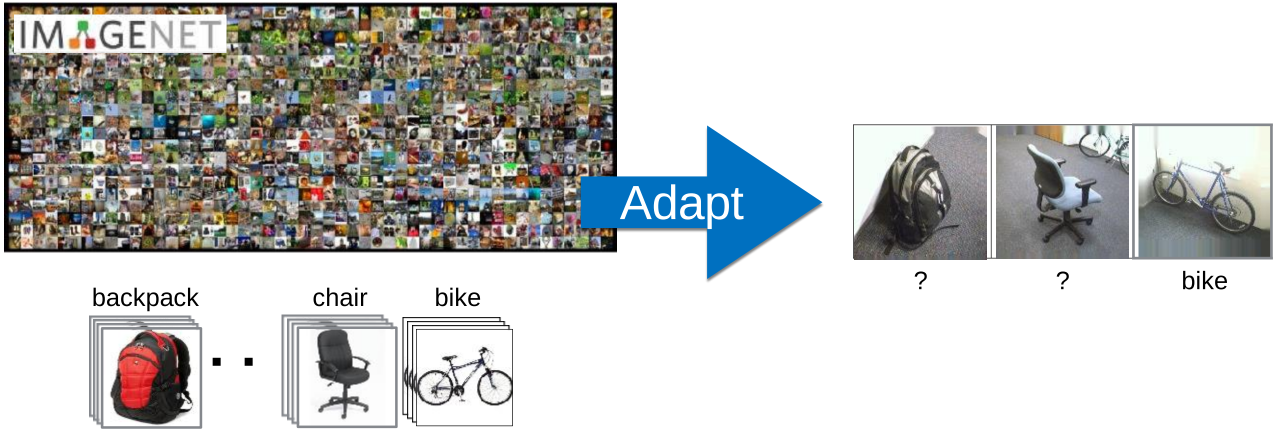
Train on Cityscapes, Test on San Francisco Dashcam

J. Hoffman, D. Wang, F. Yu, T. Darrell, FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation, Arxiv 2016

No tunnels in CityScapes?...

J. Hoffman, D. Wang, F. Yu, T. Darrell, FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation, Arxiv 2016

Domain Adaptation from source to target distribution



Source Domain $\sim P_S(X,Y)$ $\neq$ Target Domain $\sim P_T(Z,H)$

lots of **labeled** data

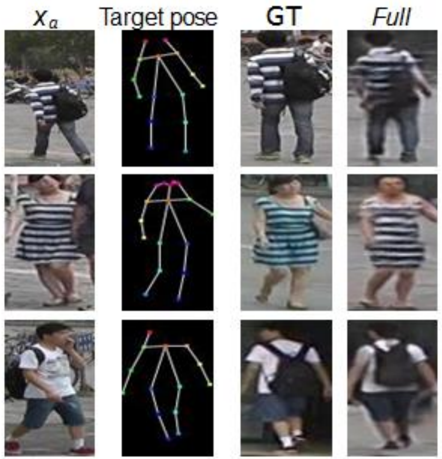
unlabeled or limited labels

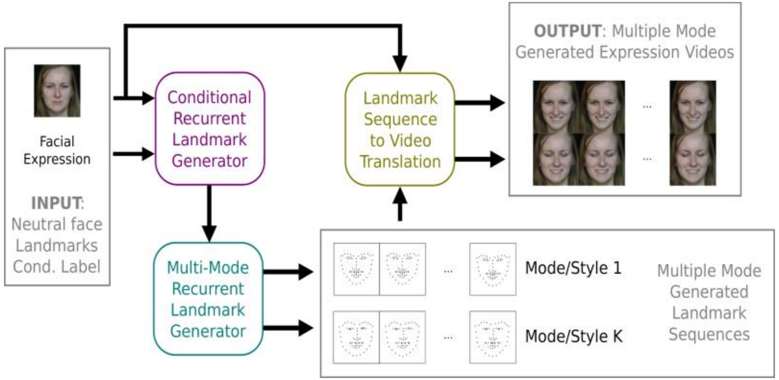
$$D_S = \{(\mathbf{x}_i, y_i), \forall i \in \{1,\dots,N\}\}$$

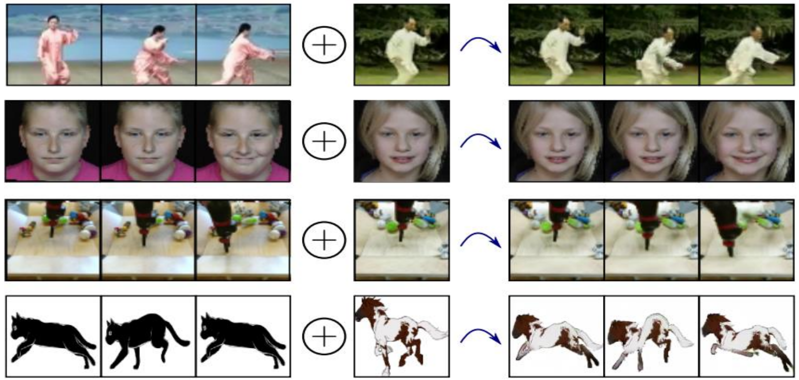
$$D_T = \{(\mathbf{z}_j, ?), \forall j \in \{1,\dots,M\}\}$$

Domain Shift: Image and Video Generation

Deep Generative Models for Image/Video Generation & Animation

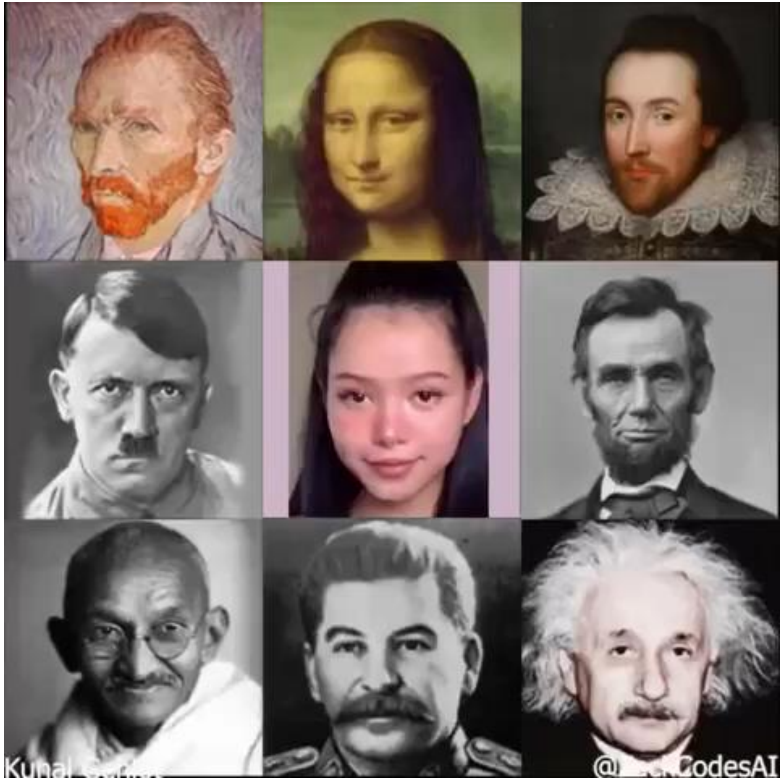






Arbitrary Object Animation without 3D modeling

Learned Keypoints

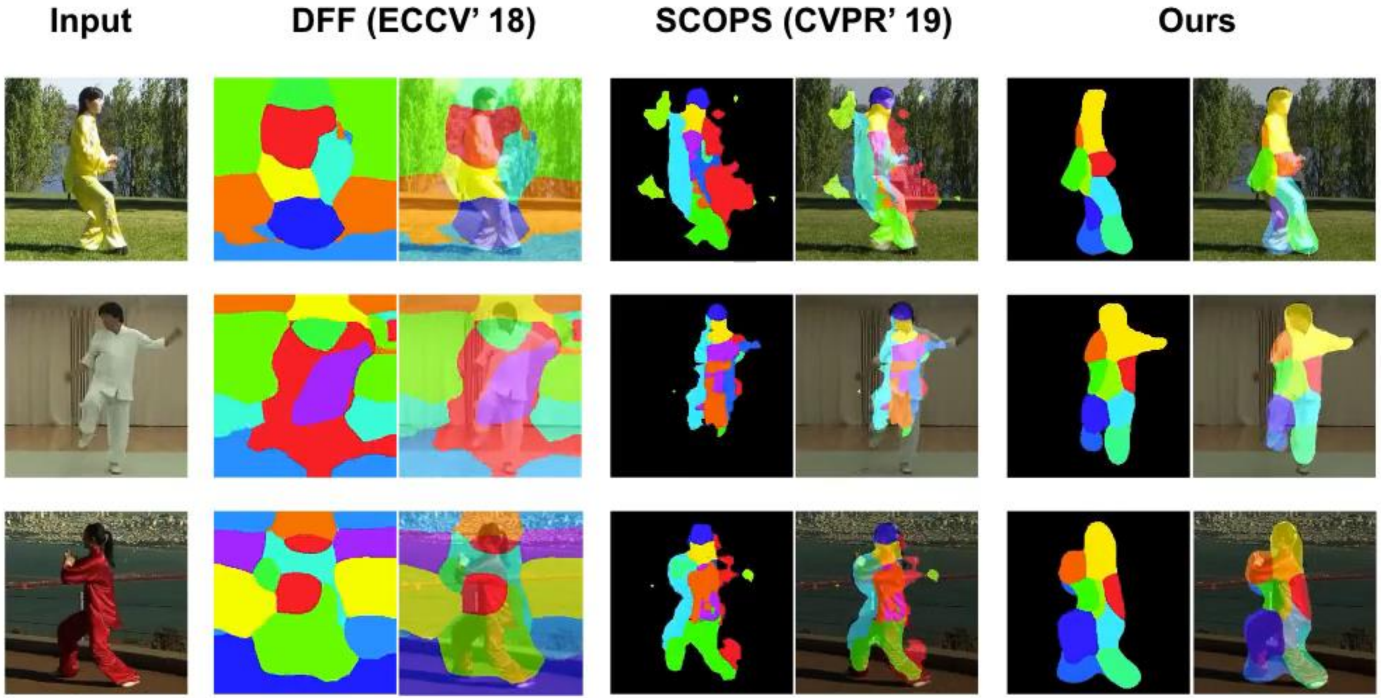


Deep Fakes …


BANCA NAȚIONALĂ A ROMÂNIEI
BNR
GUVERNATOR
CASIER CENTRAL
5
5 LEI CINCI
CINCI LEI

Deep Fakes …

Tai-Chi-HD



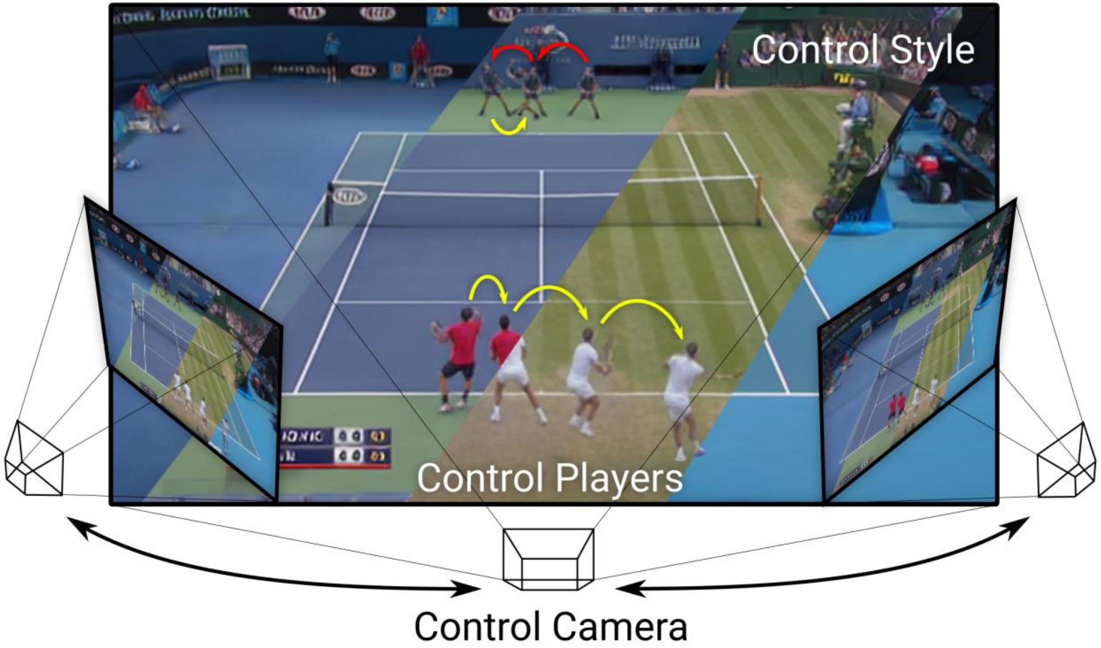

humans
SCALE
talkens.ai
TALKENS.AI

Playable Video Generation

Menapace, et al., “Playable Video Generation”, in CVPR 2021

https://github.com/willi-menapace/PlayableVideoGeneration

Playable Environments



- Learn a model that represents the observed environment
- Allow the user to input actions to the model through a controller at test time

Playable Environments

Menapace et al, “Playable Environments: Video Manipulation in Space and Time”, in CVPR 2022

https://github.com/willi-menapace/PlayableEnvironments

Playable Environments

Menapace et al, "Playable Environments: Video Manipulation in Space and Time", in CVPR 2022

https://github.com/willi-menapace/PlayableEnvironments

Playable Environments

Menapace et al, “Playable Environments: Video Manipulation in Space and Time”, in CVPR 2022

https://github.com/willi-menapace/PlayableEnvironments

Topics

- Scene recognition and understanding
- Object detection/localization/recognition
- Pixel-level prediction
  - Semantic segmentation
  - Depth estimation
- Motion analysis and activity recognition
- Biometrics: face, gesture, body pose, emotions
- Social signals processing

Topics

- Image and video generation
- Tiny ML for computer vision
- Transfer learning and deep domain adaptation
- Explainable AI
- Privacy, transparency and ethics in vision

Course organization

- Alternate oral presentations with practical sessions:
  - We will present in detail some of the basic code implementation to provide details on how things are done in practice
- Final evaluation
  - Students (groups) will be asked to choose from a list of recent important articles
  - Briefly present the main ideas
  - Propose extensions that could improve/generalize the approach

Course timeline

- Classes on:
  - September 11, 12, 19, 25, 26
  - October 2, 3 (possible pause for ICCV in this week), 9, 10, 16, 17, 30, 31
  - November 6, 7, 13, 14, 20, 21, 27, 28

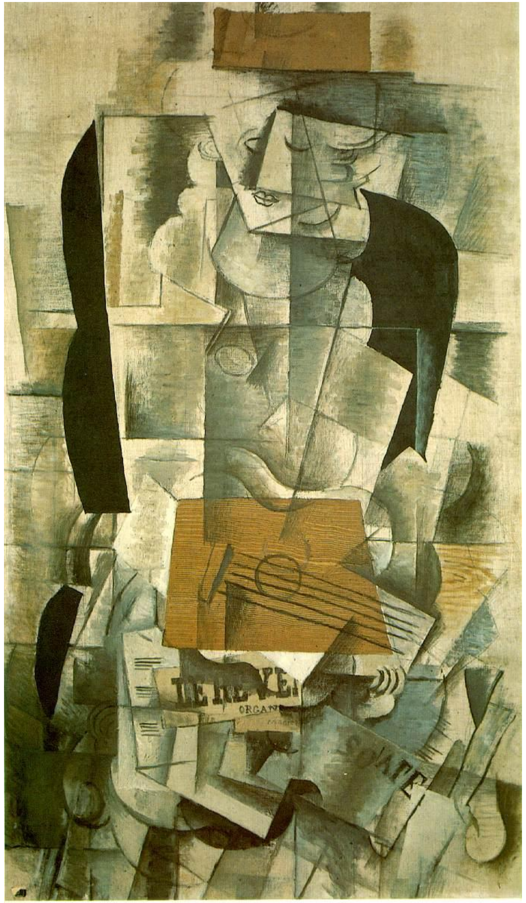
- Student presentations on:
  - December 4, 5, 11

Scene Understanding and Object Recognition

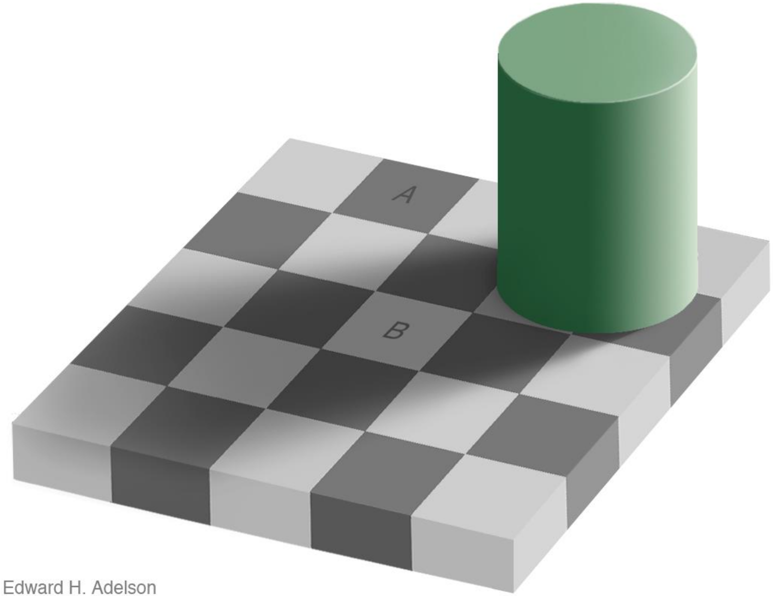
Why is Vision Interesting?

- Psychology
  - ~ 35% of cerebral cortex is for vision
  - Vision is how we experience the world
- Engineering
  - Want machines to interact with the world
  - Digital images are everywhere

Measurement vs. Perception

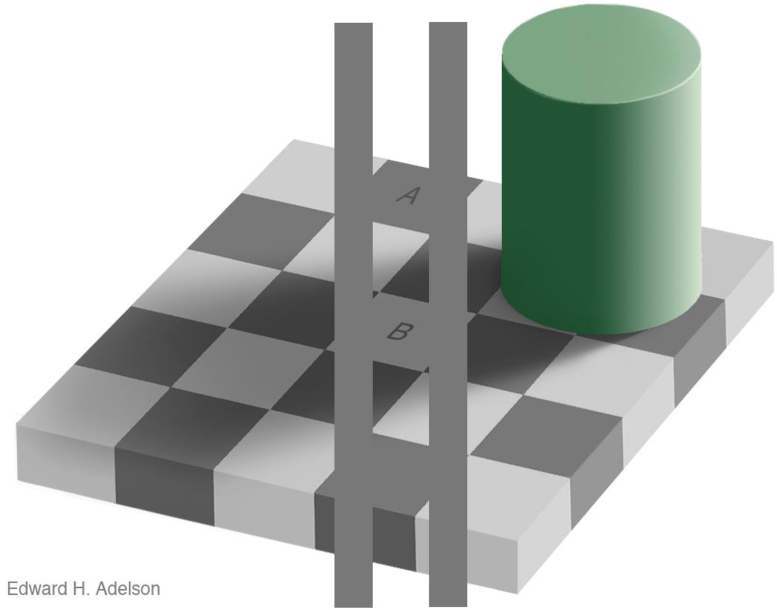


Brightness: Measurement vs. Perception



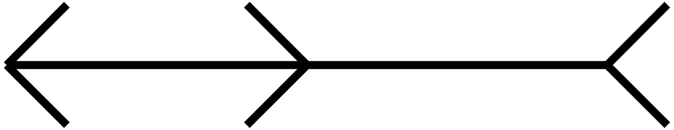
Edward H. Adelson

Brightness: Measurement vs. Perception

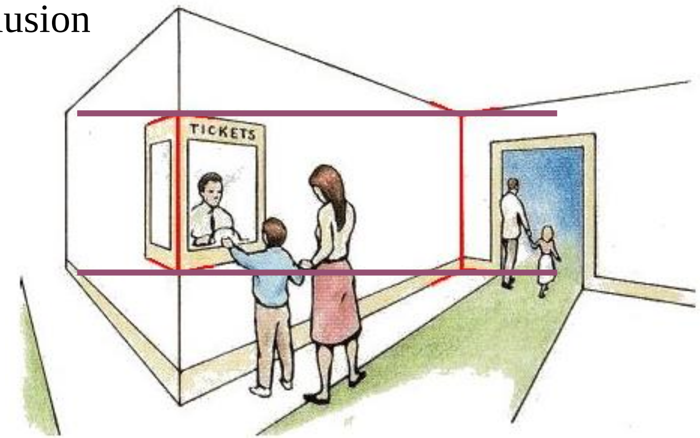


Proof!

Lengths: Measurement vs. Perception

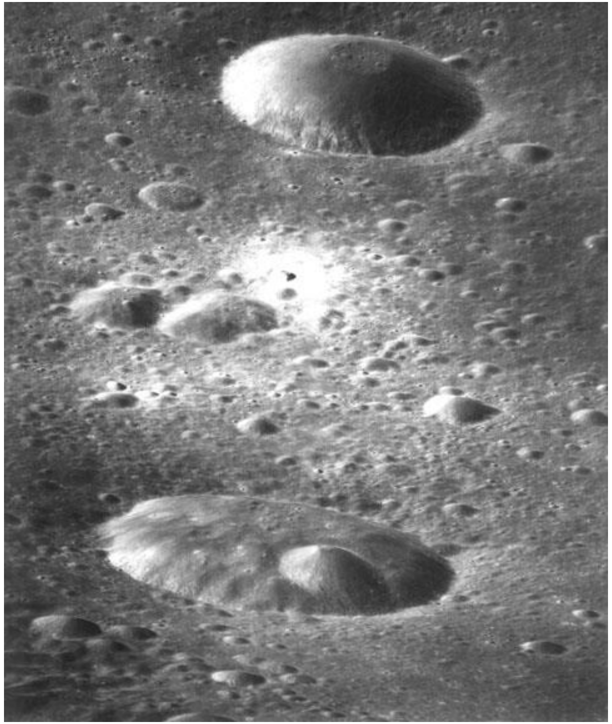


Müller-Lyer Illusion



https://michaelbach.de/ot/sze-muelue/index.html

Vision is Inferential: Prior Knowledge

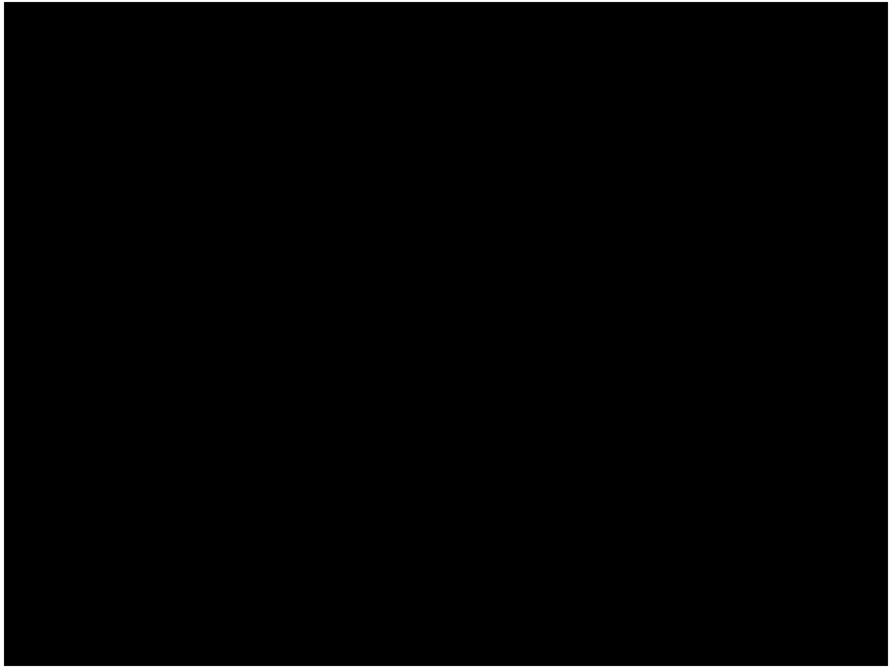


Vision is Inferential: Prior Knowledge

- Our visual system is very sophisticated
- Humans can interpret images successfully under a wide range of conditions – even in the presence of very limited cues

Playing with Perspective

- Perspective gives us very strong depth cues
  ➪ hence we can perceive a 3D scene by viewing its 2D representation (i.e. image)
- An example where perception of 3D scenes is misleading:



“Ames room”

A clip from "The computer that ate Hollywood" documentary. Dr. Vilayanur S. Ramachandran.

- *"What if I don't care about this wishy-washy human perception stuff? I just want to make my robot go!"*
- Small Reason:
  - For measurement, other sensors are often better (in DARPA Grand Challenge, vision was barely used!)
  - For navigation, you still need to learn!
- Big Reason:
- The goals of computer vision (**what + where**) are in terms of what <u>humans</u> care about

So what do humans care about?

Verification: is that a bus?

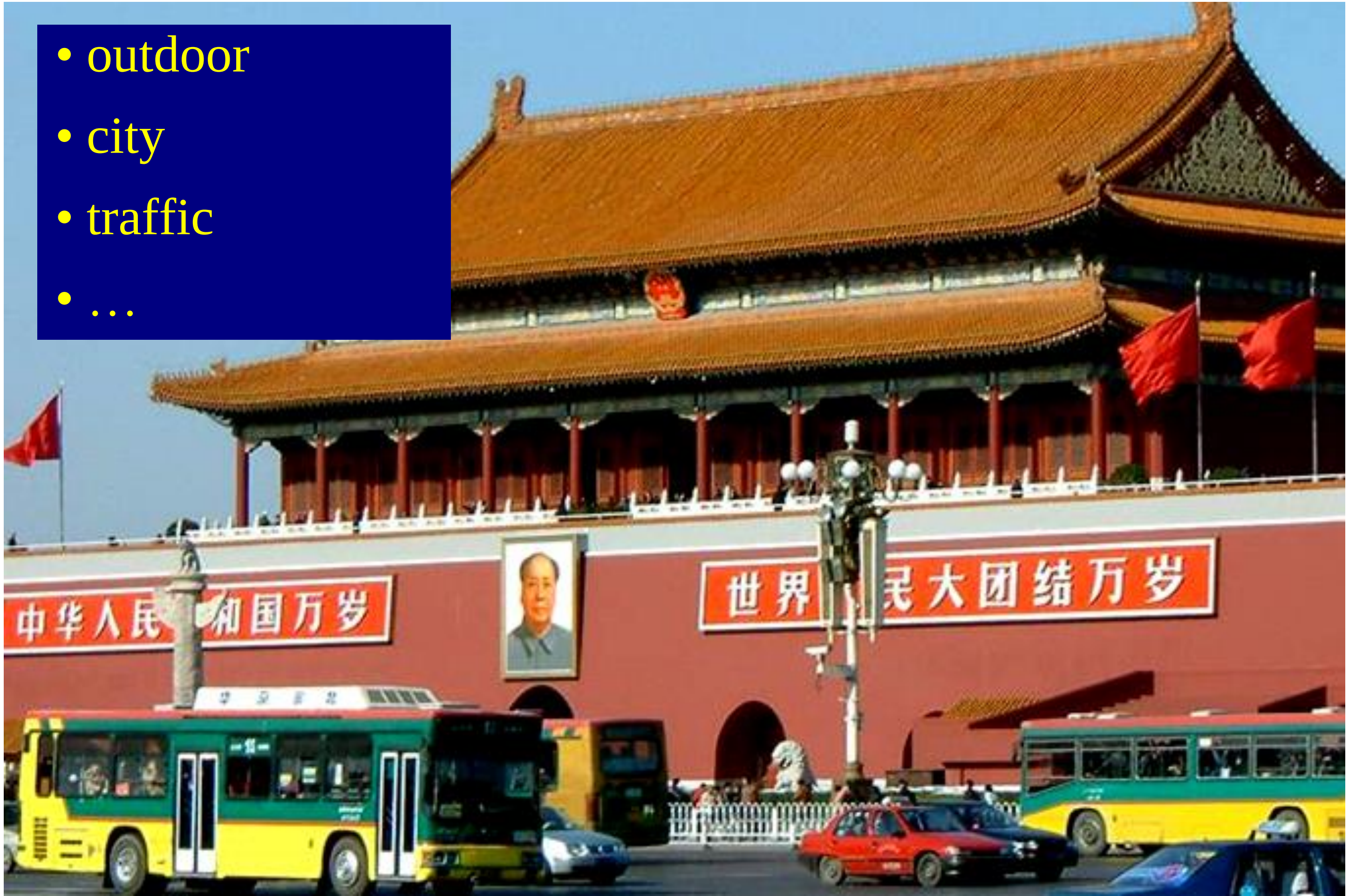
Detection: are there cars?

Identification: is that a picture of Mao?

Object categorization

Scene and context categorization

- outdoor
- city
- traffic
- …



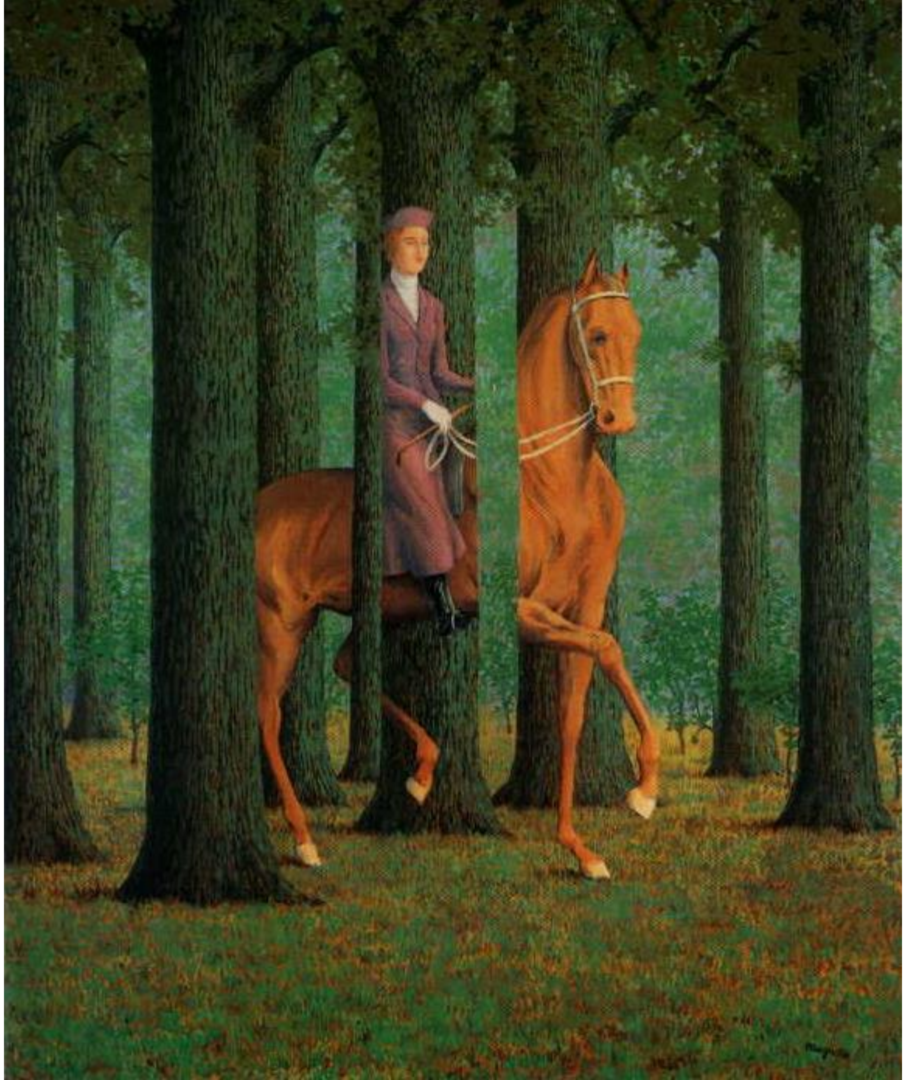
Rough 3D layout, depth ordering

Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

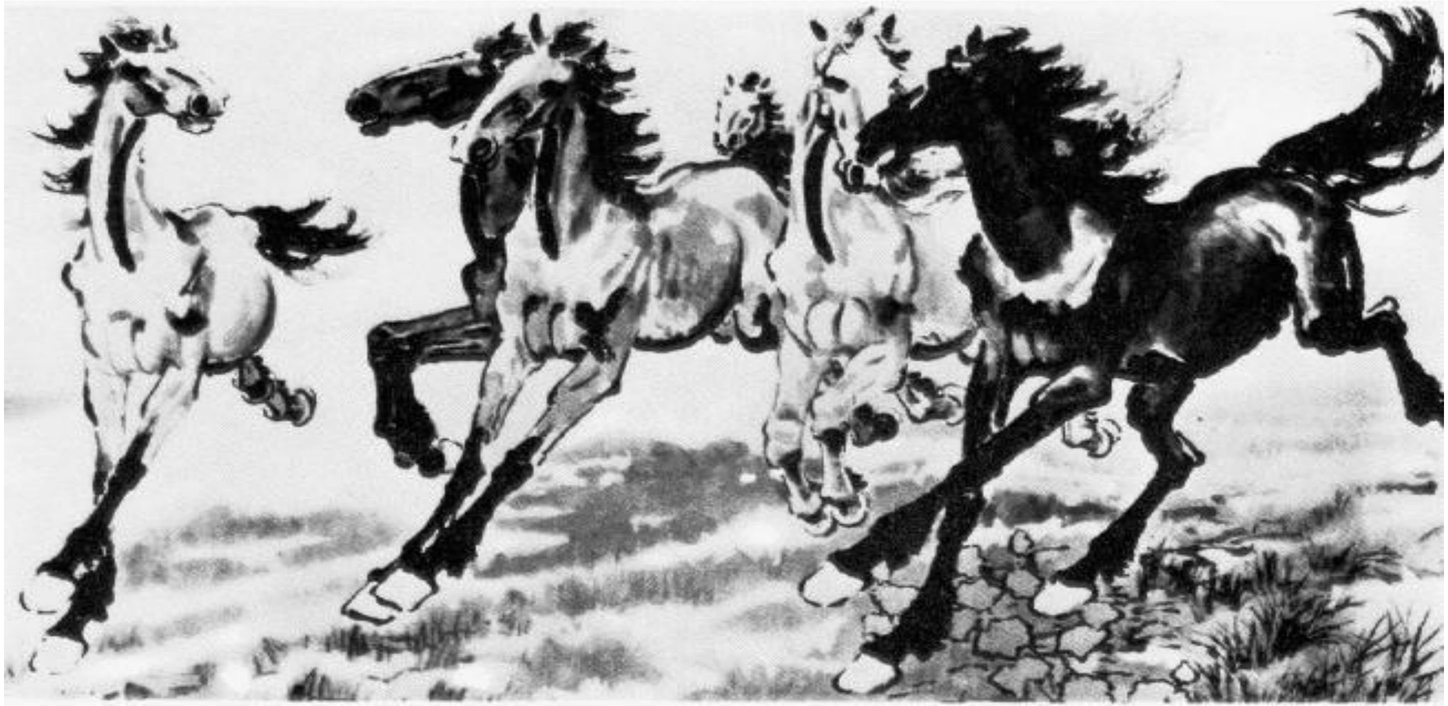
Challenges 3: occlusion



Magritte, 1957

Challenges 4: scale

Challenges 5: deformation



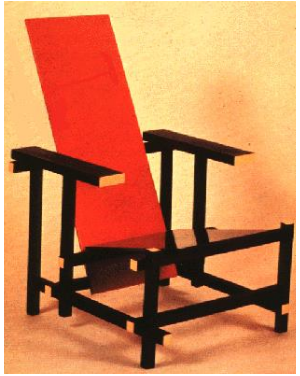
Xu, Beihong 1943

Challenges 6: background clutter

Klimt, 1913

Challenges 7: object intra-class variation



Challenges 8: local ambiguity