Computer Vision

Instructor

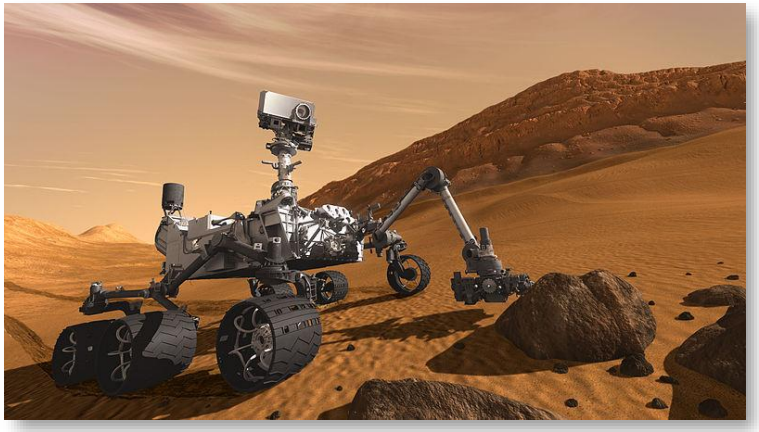
Instructor: Qi Lin
qilin@ouc.edu.cn
https://qilin512.github.io/

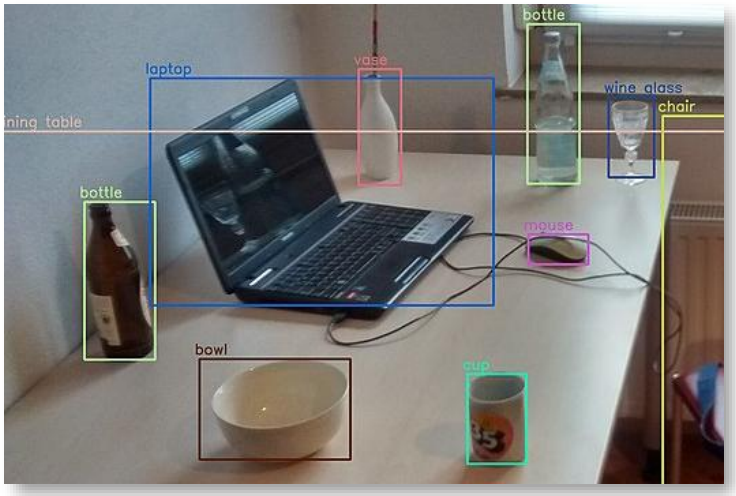
TA:
***, ***@st.ouc.edu.cn

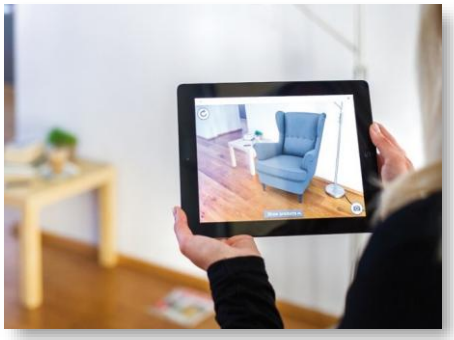
Today

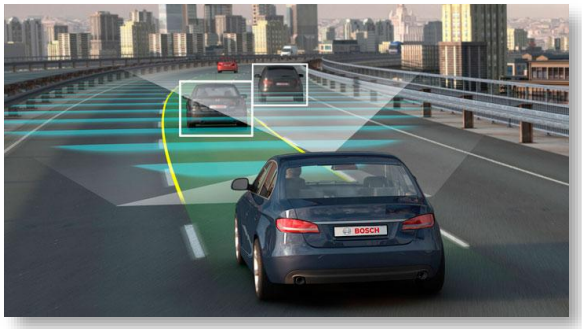
Logistics

Introduction









Slides adopted from S. Lazebnik, D Forsyth, W. Freeman, N. Snavely.

About Me

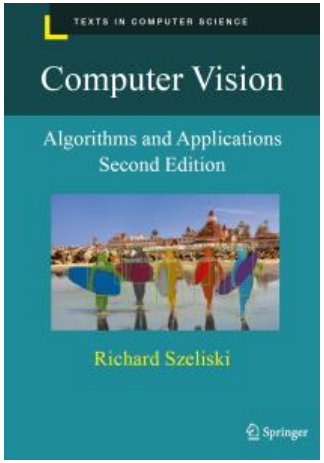
- Research
  - Computer Vision
  - Oceanography Data Analysis
- Teaching
  - UG
    - Object Oriented Programming (2013-2022)
    - Mathematical Logic (2013-2017)
    - Computer Graphics (2016-2020)
    - Principles of Communication (Signals and Systems) (2014)
    - Computer Vision (2021-)
  - PG
    - Computer Vision (2021-)
    - Design and Development of Application System (2014-2017)
    - Image Processing and Pattern Recognition (2017-2018)

Prerequisites

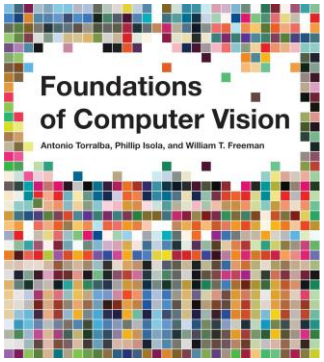
- Prerequisites—*these are essential*! (or willingness/time to pick it up quickly!)
  - Data structures & Algorithms
  - A good working knowledge of python programming
    - If you know Matlab / Python, assignments will be easier
  - Linear algebra
  - Calculus
- Course does ***not*** assume prior imaging experience
  - no image processing, graphics, etc.

Important information

- Textbook:



Rick Szeliski, *Computer Vision: Algorithms and Applications, 2E*
Online at: http://szeliski.org/Book/



Antonio Torralba, Phillip Isola, and William Freeman, *Foundations of Computer Vision*
Online at: https://visionbook.mit.edu/

Topics

- Elementary image representations
- Image Features
- Learned Image Representations
- 3D Vision
- Image Generation

Grading

- **Attendance**
  - 10%
- **Assignments**
  - 20%, about 3 to 5
- **Performance**
  - 5% extra
- **Exam**
  - 70%

Assignments

- **Experiment Platform**:
  - Jupyter Notebook
    - Local Python
    - Baidu AI Studio (aistudio.baidu.com).
    - Huawei Model Art
    - Google Colab
  - Coding
    - Python
    - OpenCV
    - PyTorch / TensorFlow / PaddlePaddle / MindSpore

- **课堂派**:
  - Notice
  - Homework

Academic Integrity

- Assignments will be done solo or in groups (we'll let you know for each assignment)
- Please do not leave any code public on GitHub (or the like) at the end of the semester!
- We will follow the OUC Code of Academic Integrity (https://aco.ouc.edu.cn/2018/0731/c12999a207452/page.psp)
- If you use AI agent (Claude Code, Codex, CoPilot or similar) on coding assignments, you must disclose that with your submission
  - **BUT**: We advise you to do all coding yourself, unassisted. You will learn less, and become less capable experts in vision, if you rely on LLMs.

Outline

- Logistics, requirements
- Key tasks
- Why it is hard
- History of computer vision
- Current state of the art
- Topics covered in class

Today

- Readings
  - Szeliski, Chapter 1 (Introduction)

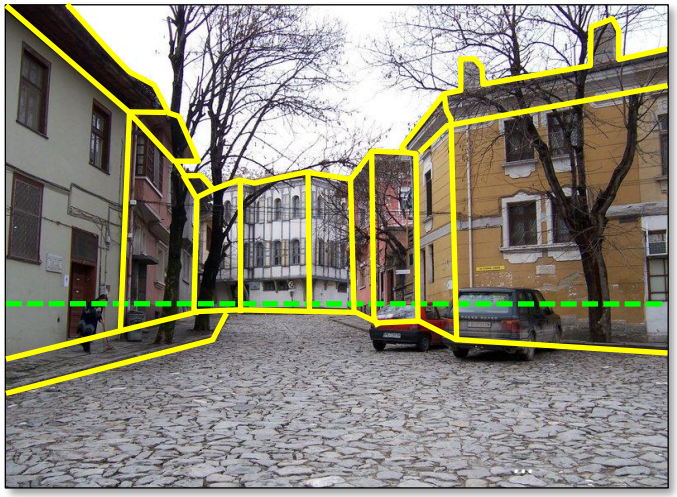
A little story

In 1966, Marvin Minsky asked his undergraduate student Gerald Jay Sussman to “spend the summer linking a camera to a computer and getting the computer to describe what it saw” (Boden 2006, p. 781). We now know that the problem is slightly more difficult than that.

(Szeliski 2010, Computer Vision)

What kind of information can be extracted from an image?

What kind of information can be extracted from an image?



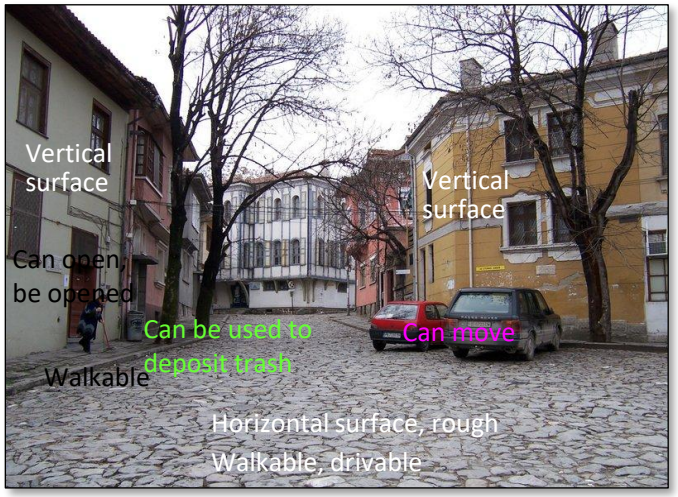
Geometric information

What kind of information can be extracted from an image?

Geometric information

Semantic information

What kind of information can be extracted from an image?



Geometric information

Semantic (?) information – *affordances*

What kind of information can be extracted from an image?

Geometric information

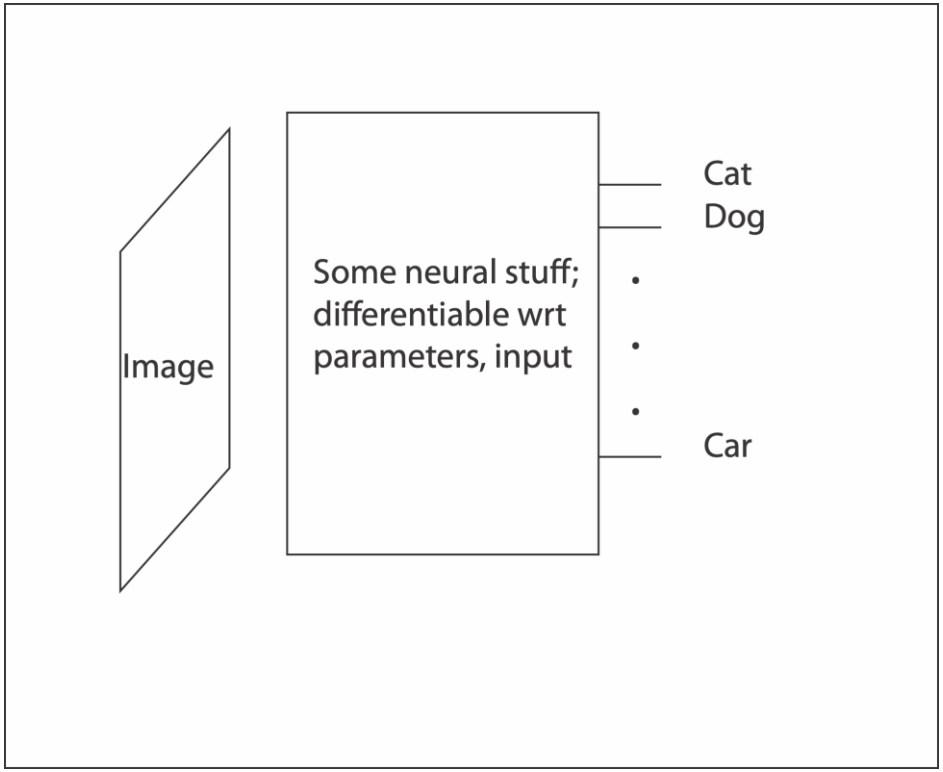
Semantic information

Vision for action

What does vision do?

- *Classification: What is it?*
- *Localization: Where is it?*
- *Detection: Where and what?*
- *Tracking: Where is it going?*
- *Odometry: How have I moved?*
- *Navigation: Where am I?*
- *Modelling: What is the world like?*
- *Control: What should I do?*
- *Speculation: What will it be like if?*

Classification



Detection

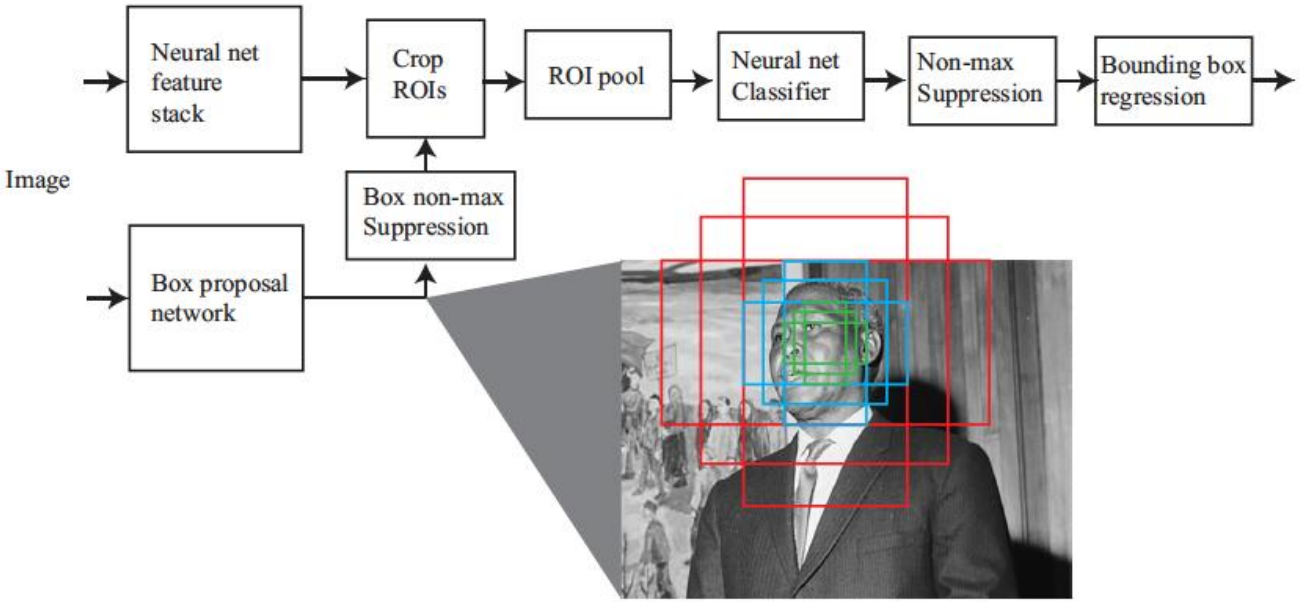
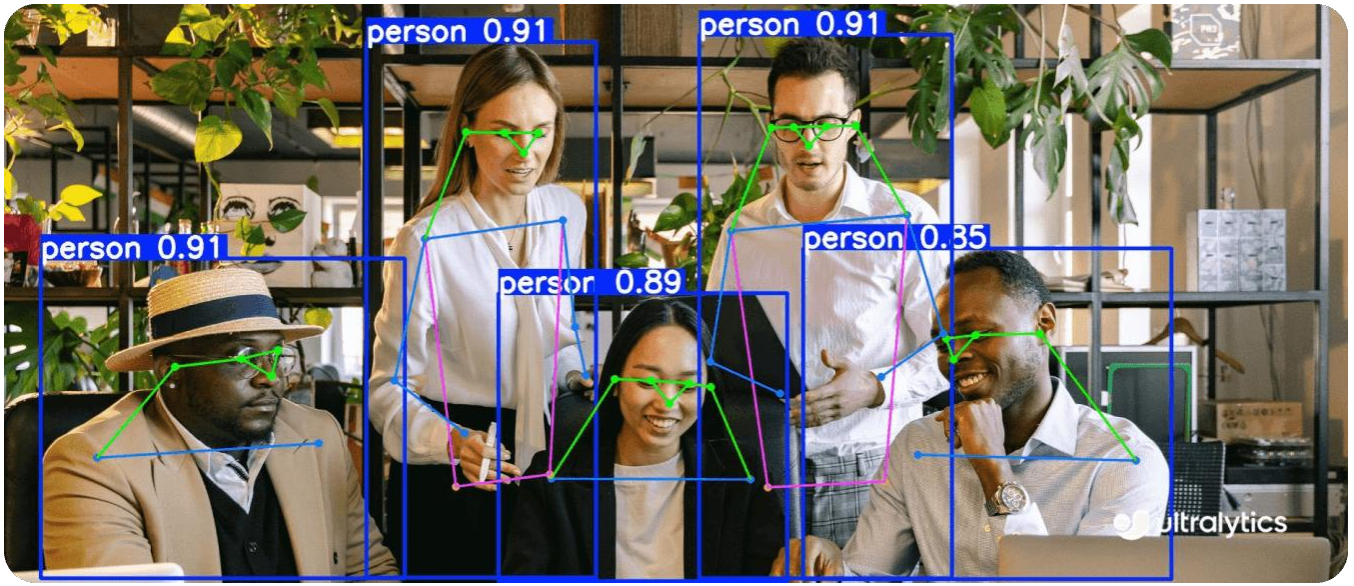


FIGURE 18.8: *Faster RCNN uses two networks. One uses the image to compute "objectness" scores for a sampling of possible image boxes. The samples (called "anchor boxes") are each centered at a grid point. At each grid point, there are nine boxes (three scales, three aspect ratios). The second is a feature stack that computes a representation of the image suitable for classification. The boxes with highest objectness score are then cut from the feature map, standardized with ROI pooling, then passed to a classifier. Bounding box regression means that the relatively coarse sampling of locations, scales and aspect ratios does not weaken accuracy.*

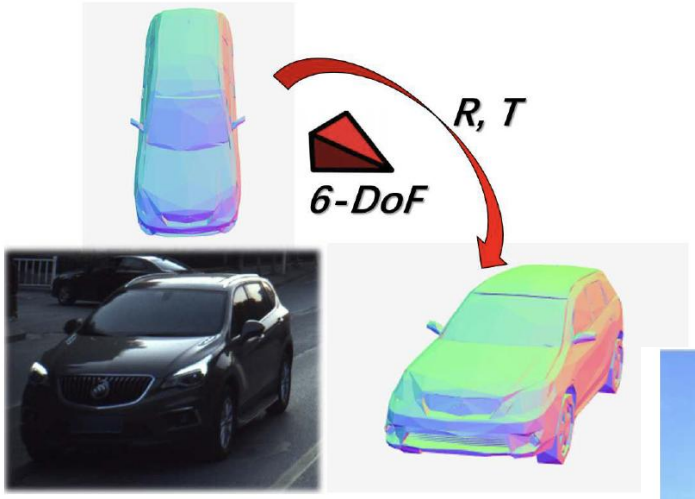
Detection and localization in 2D

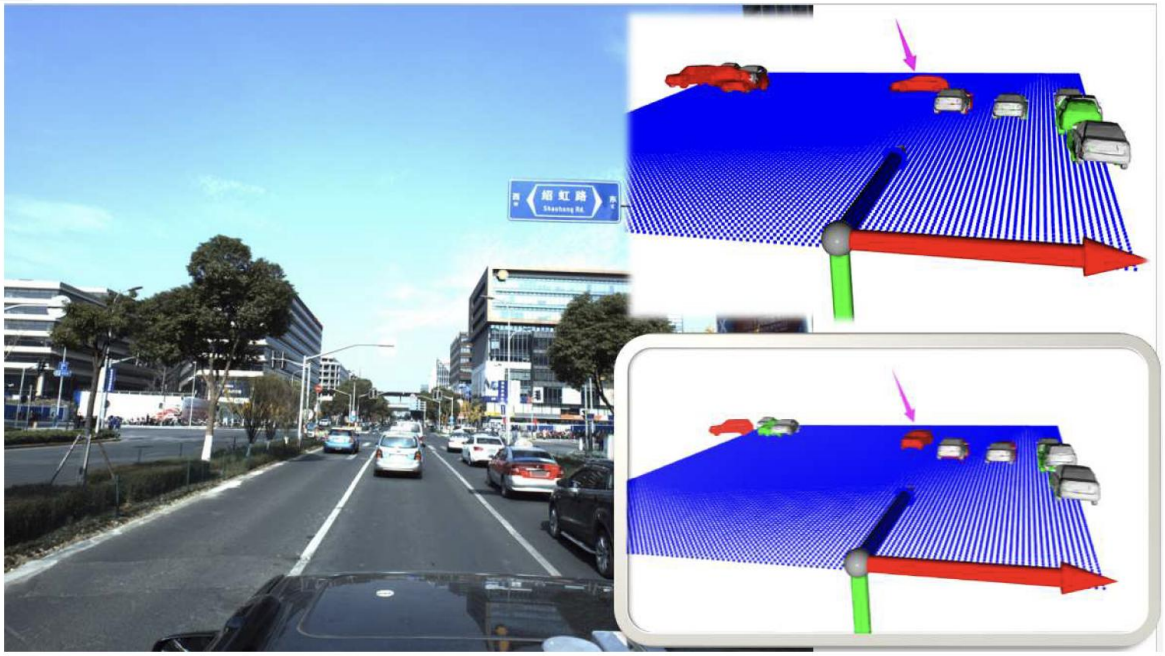


YOLOv11 documentation by Ultralytics

Localization in 3D

from detection





al, *6D-VNet: End-to-end 6DoF Vehicle Pose Estimation from Monocular RGB Images*

Lane detection

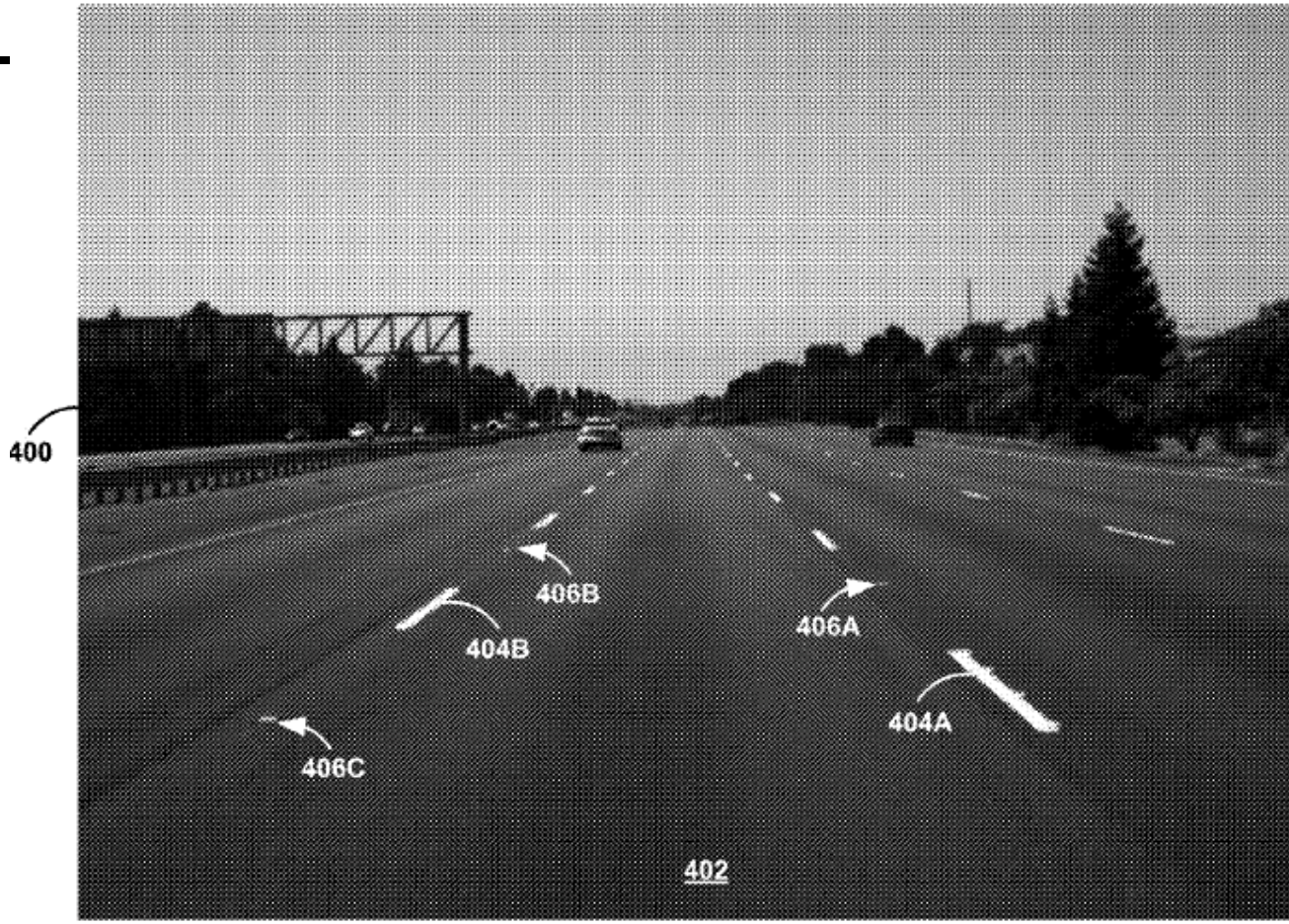


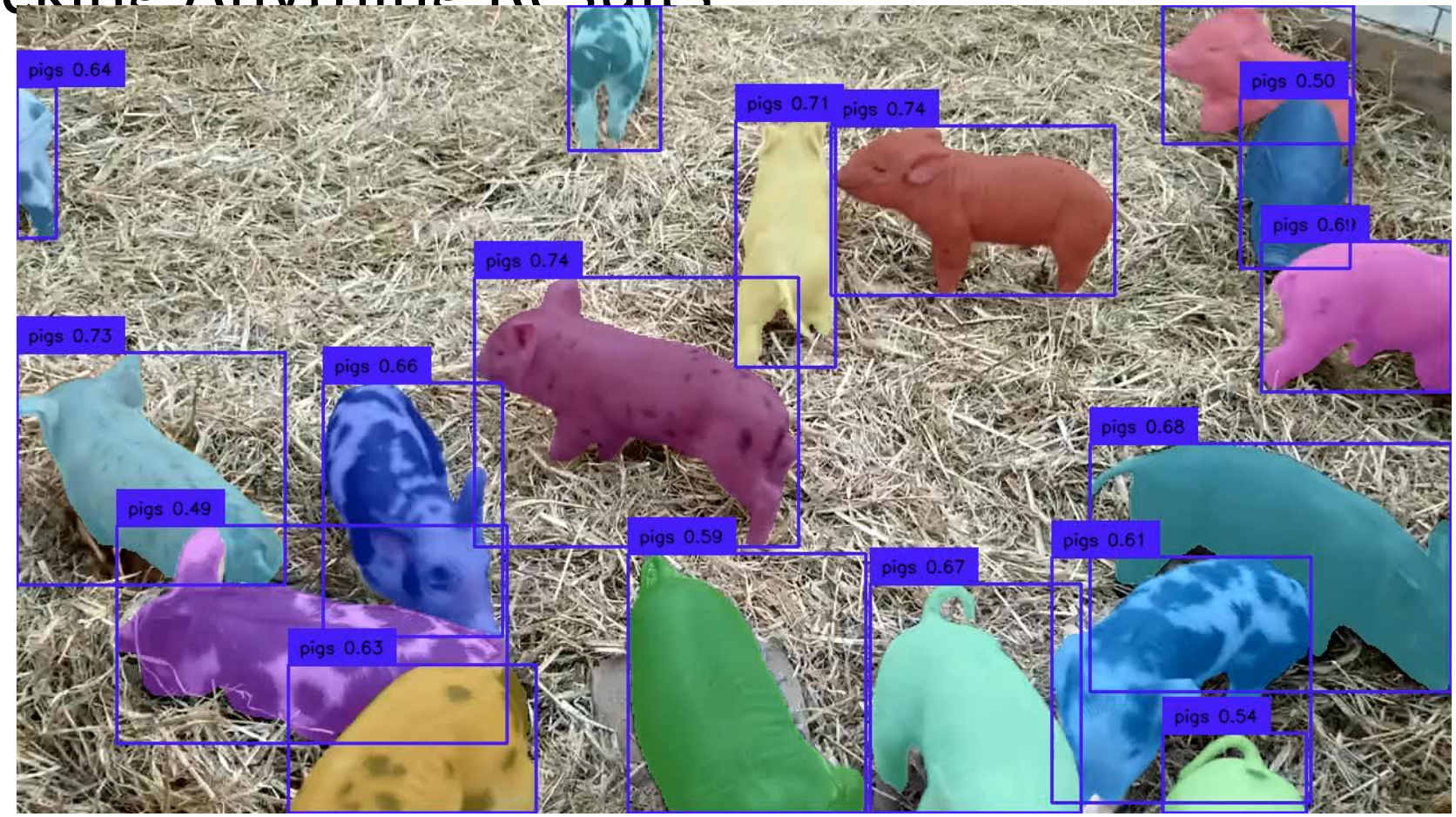
FIGURE 4A

US 9081385

Waymo and Google 2012

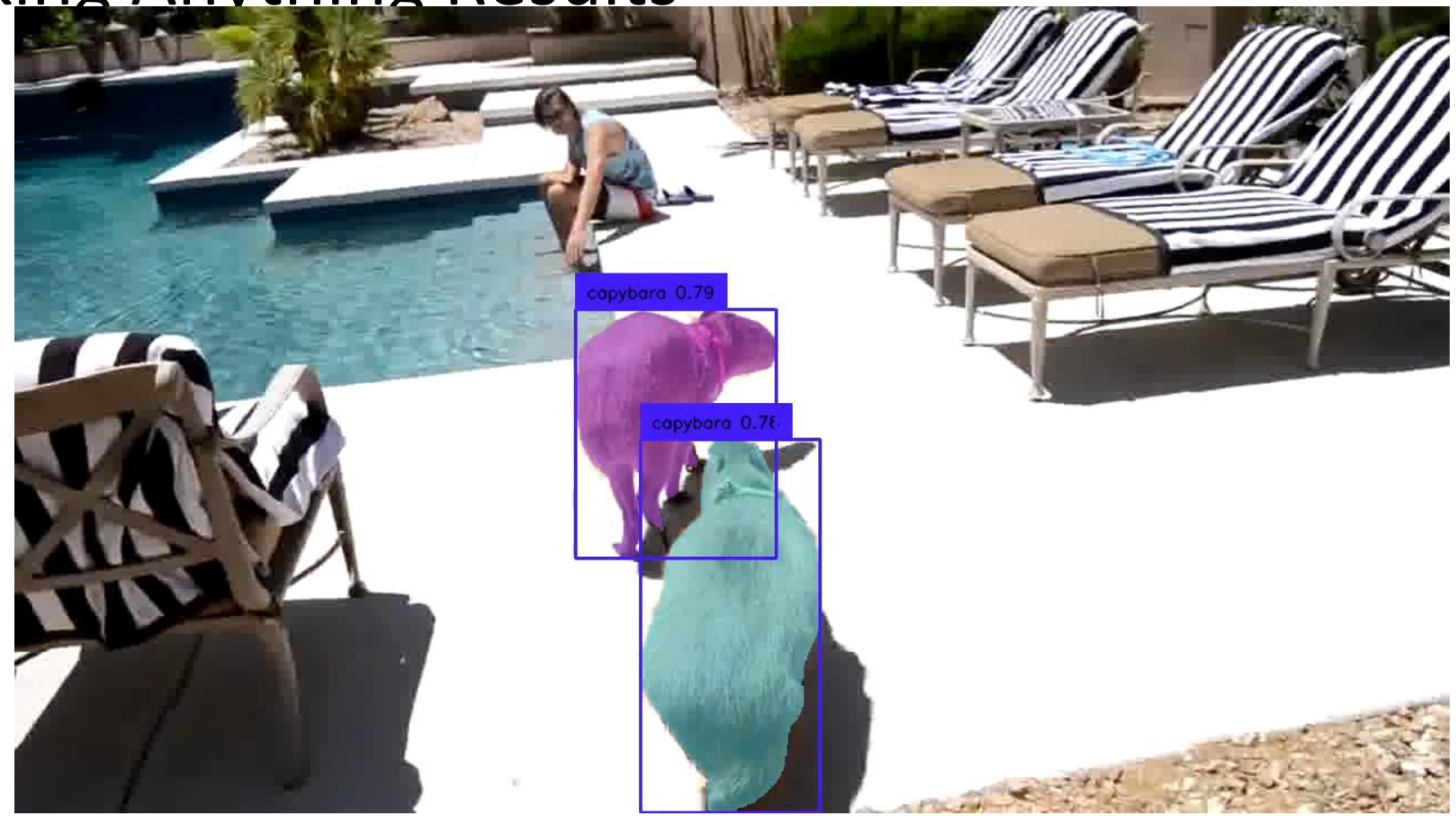
Strategy: detect markers (reflective paint), join up
exercise in robust fitting of curves

Tracking Anything Results



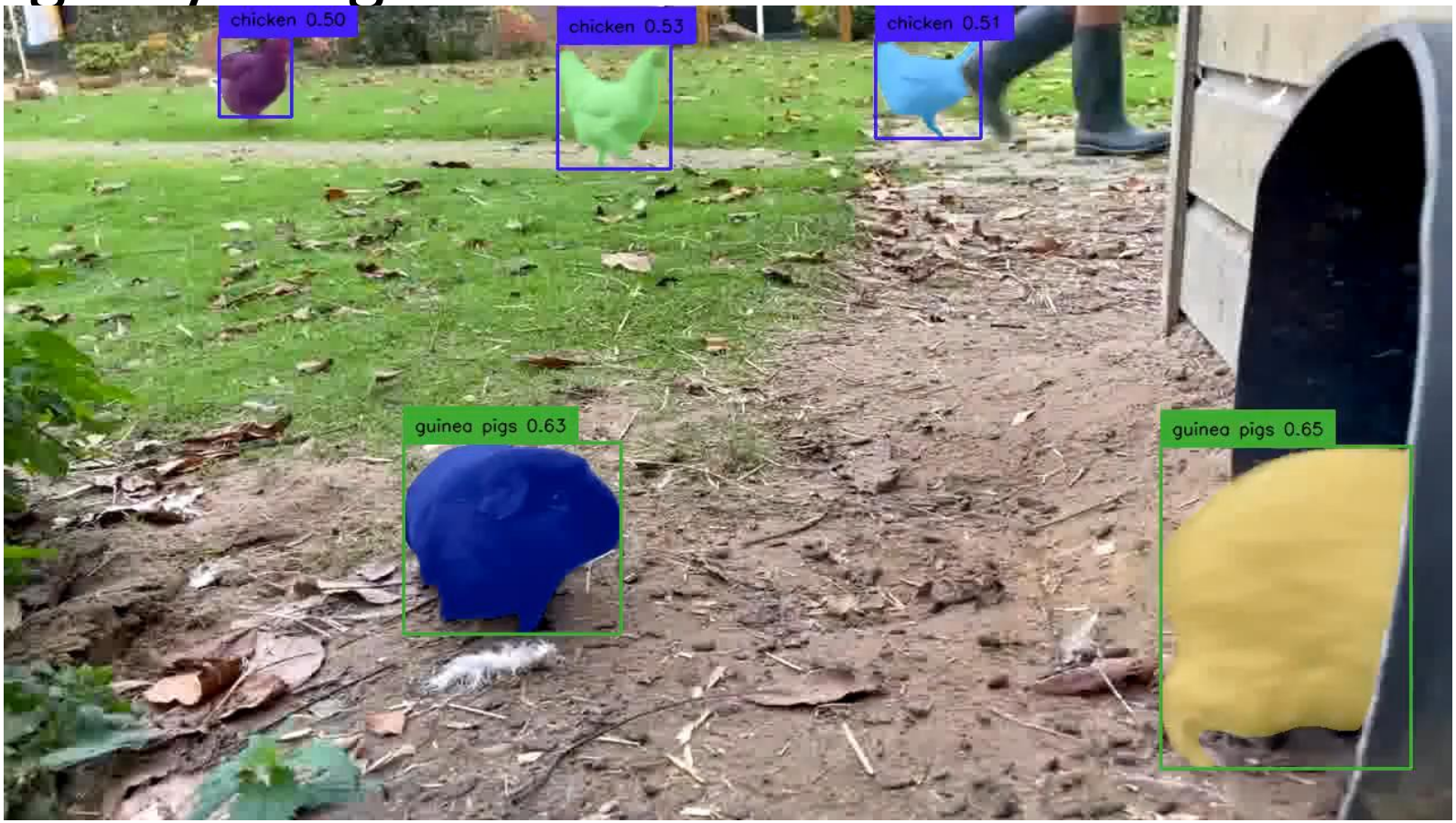
Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

Tracking Anything Results



Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

Tracking Anything Results



Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

Visual odometry

https://github.com/MAC-VO/MAC-VO/blob/main/asset/ICRAvideo.gif

Extreme odometry

- https://www.youtube.com/watch?v=fBiataDpGIo

Every image tells a story

- Goal of computer vision: perceive the “story” behind the picture
- Compute properties of the world
  - 3D shape
  - Names of people or objects
  - What happened?

The goal of computer vision

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

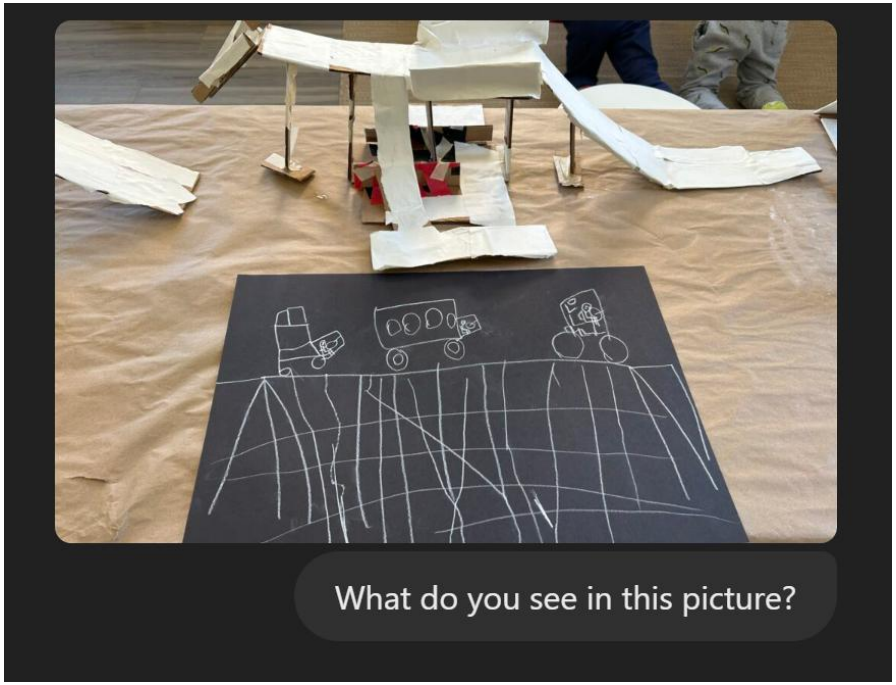
Can computers match human perception?

- Yes and no
  - humans are better at “hard” things, are more robust, and make inferences quickly and cheaply
- But huge progress
  - Accelerating in the last 10 years due to deep learning, large vision-language models
  - What is considered “hard” keeps changing

What do you see in this picture? What do you think happened before this picture was taken?

In the picture, we see the famous Montparnasse derailment from 1895. It shows a steam locomotive that has crashed through the station wall and is hanging precariously over the street below. The photograph captures the aftermath of a dramatic accident where the train failed to stop while approaching the buffer stop at the Gare Montparnasse terminal in Paris.

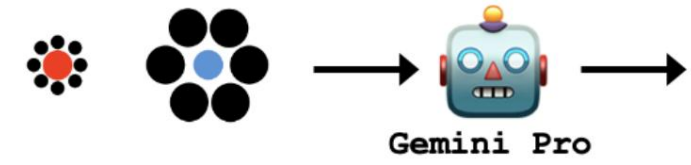
Before this picture was taken, the train, arriving from Granville to Paris, was running late. The driver, trying to make up for lost time, was likely going too fast and unable to stop the train in time upon reaching the station. The train overran the buffer stop, crossed the station concourse, crashed through a 60 cm thick wall, and ended up nosediving onto the Place de Rennes below. Remarkably, only one person, a woman selling newspapers on the street, was killed by falling masonry, and several others were injured. This image has become a symbol of dramatic accidents and is frequently used to illustrate the need for careful management of technology and speed.


What do you see in this picture?

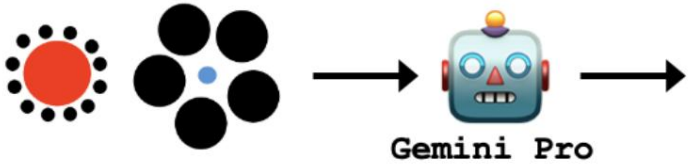
In this image, you can see a creative and playful representation of a bridge construction made from paper materials. There are several sections of the bridge, including ramps and supports, all crafted from white and red paper, arranged over a brown paper surface. Beside this bridge model, there is a black paper with white chalk drawings that depict vehicles like trucks and motorcycles, perhaps to simulate a traffic scenario on the bridge. It appears to be a hands-on educational activity, possibly designed to teach children about engineering, construction, and physics through interactive play. The scene is set in a room where children are present, indicating that this might be a classroom or workshop setting.

Current models still make very silly mistakes

Prompt: "Which is bigger, the blue circle or the red circle?"



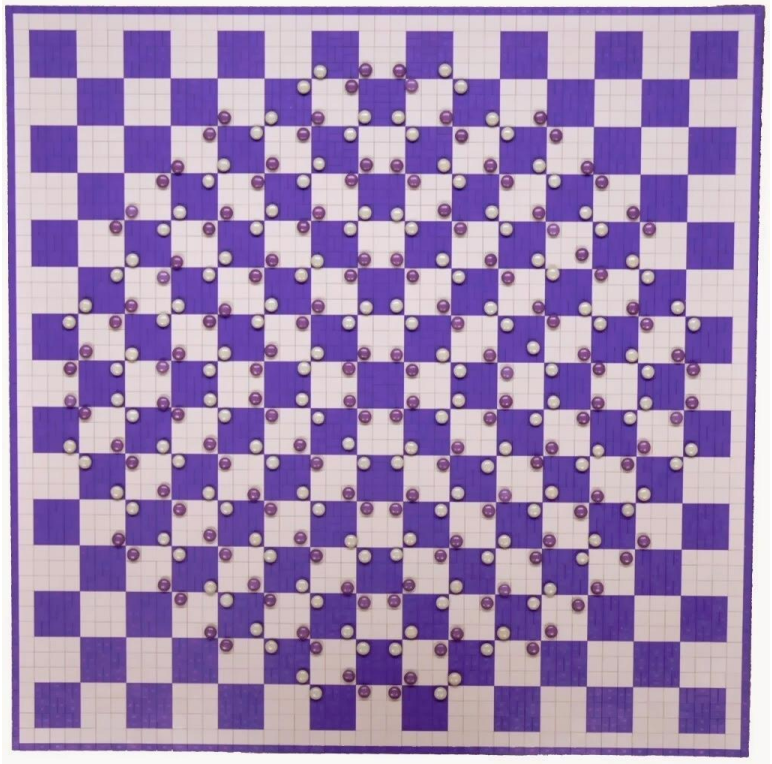
"This is a classic optical illusion! The red and blue circles are actually the same size. Our brains tend to perceive the red circle as smaller because it's surrounded by larger black circles. The blue circle, surrounded by smaller black circles, appears larger in comparison."



"This is a classic optical illusion! The red and blue circles are actually the same size. Our brains tend to perceive the red circle as larger because it's surrounded by smaller circles, while the blue circle is surrounded by larger circles. This is an example of the Ebbinghaus illusion."

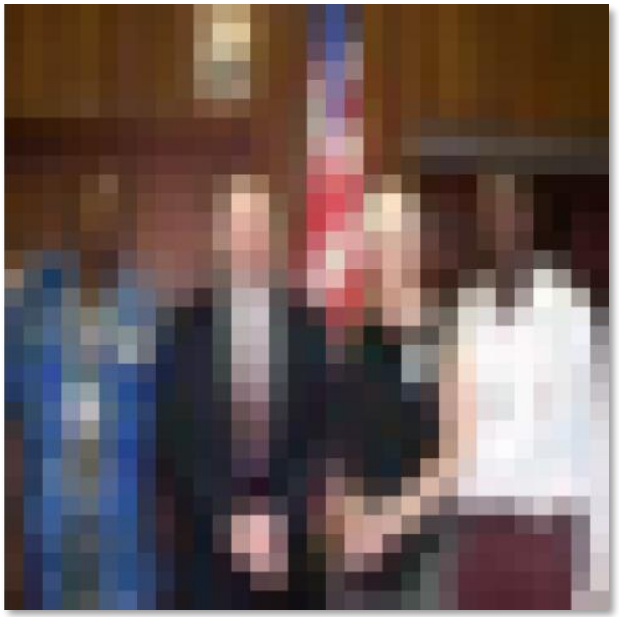
[Tomer Ullmann, *The Illusion-Illusion: Vision Language Models See Illusions Where There are None*, arXiv 2024]

Human perception has its shortcomings



https://twitter.com/pickover/status/1460275132958662657/

But humans can tell a lot about a scene from a little information…



Source: “80 million tiny images” by Torralba, et al.

…still, vision is hard even for humans

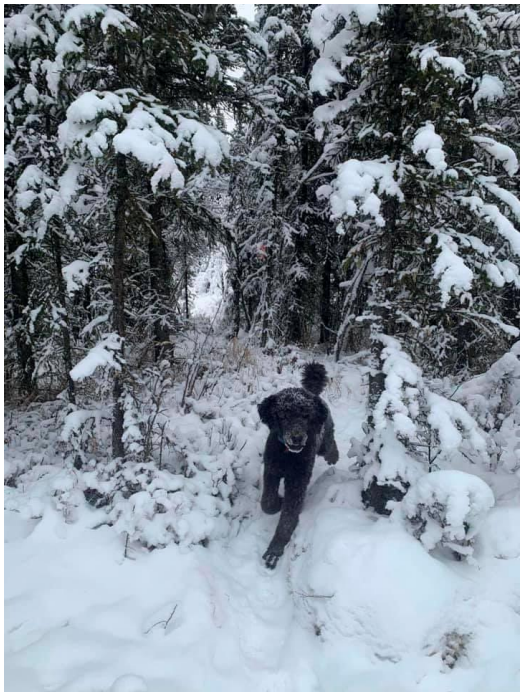


Image source

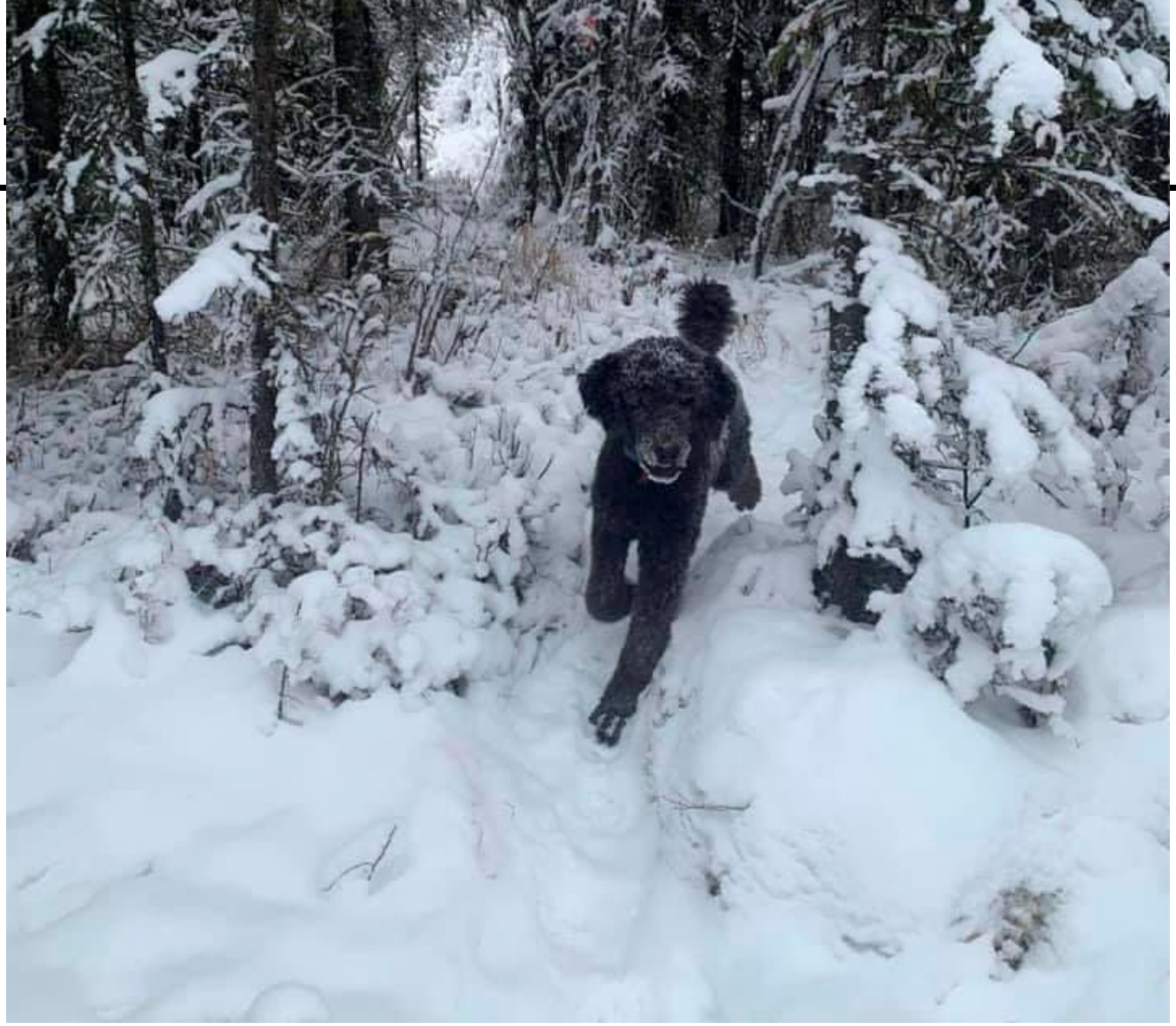
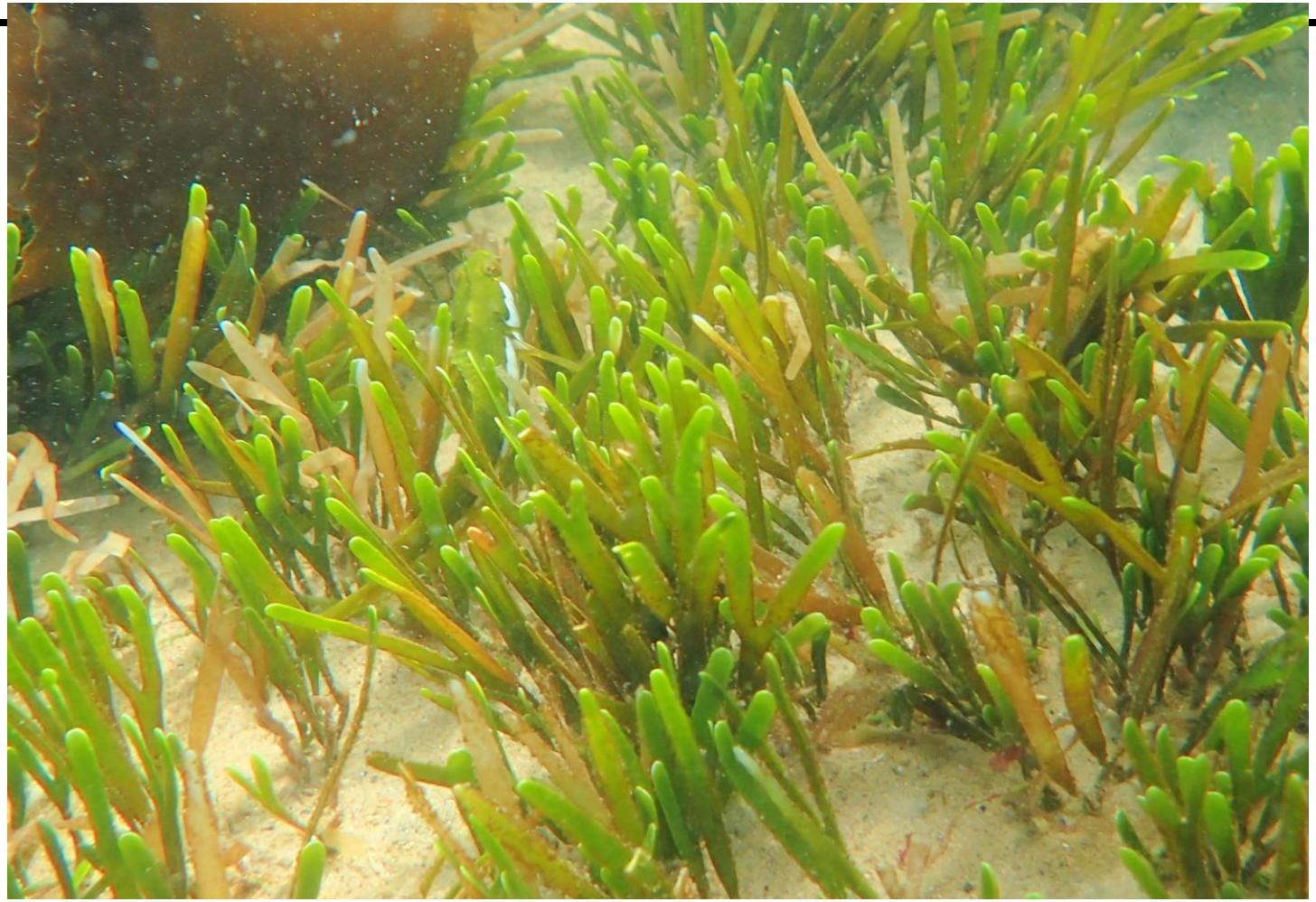


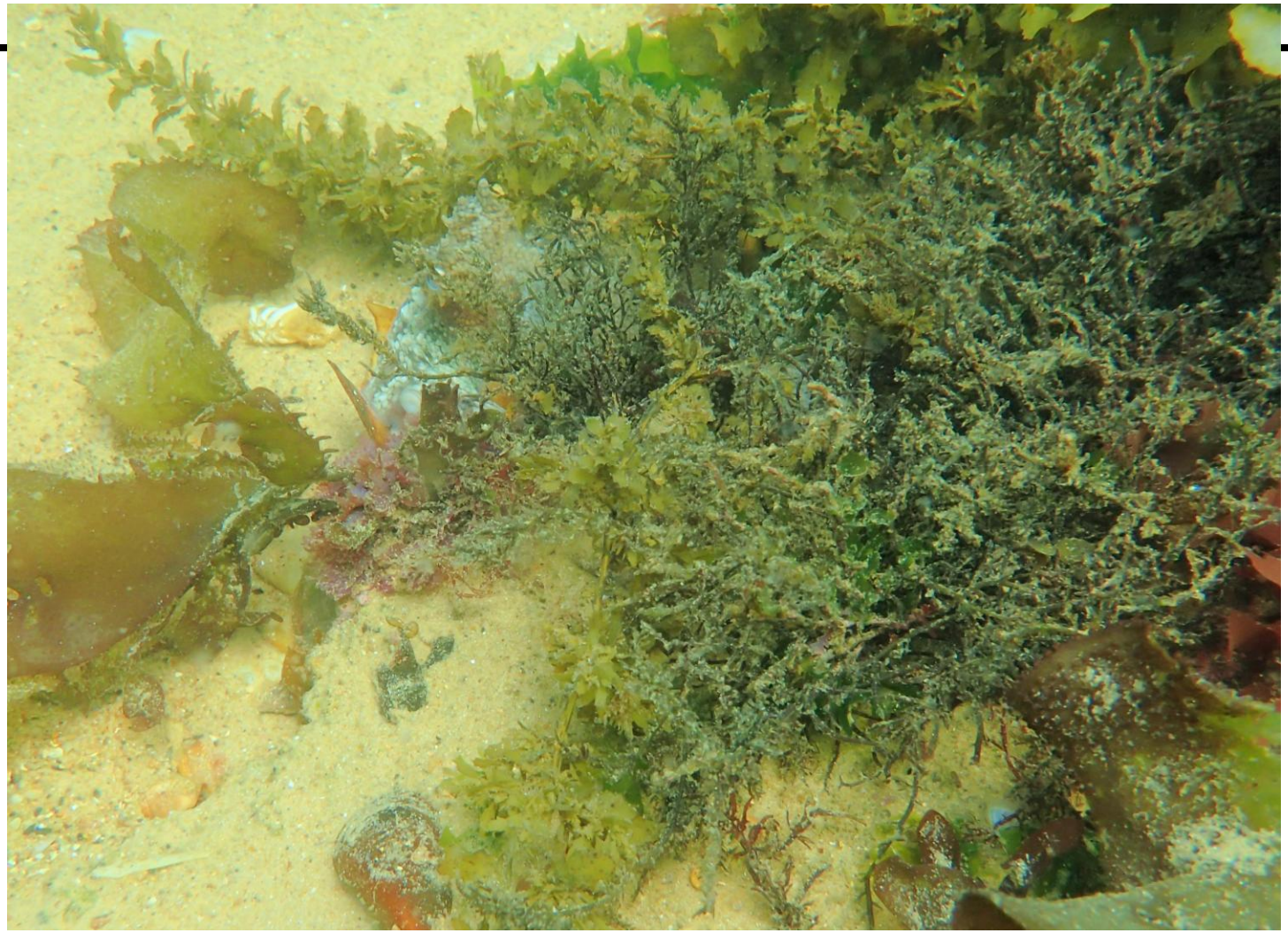
Image source

…still, vision is hard even for humans

Figure from Marr (1982), att

Is t





Why is computer vision difficult?

Viewpoint variation

Illumination

Scale

Why is computer vision difficult?

Intra-class variation

Motion (Source: S. Lazebnik)

Background clutter

Occlusion

Challenges: local ambiguity

slide credit: Fei-Fei, Fergus & To

But there are lots of visual cues we can use…

Source: S. Lazebnik

Bottom line

- Perception is an inherently ambiguous problem
  - Many different 3D scenes could have given rise to a given 2D image

Artist Julian Beever with his anamorphic Coke bottle

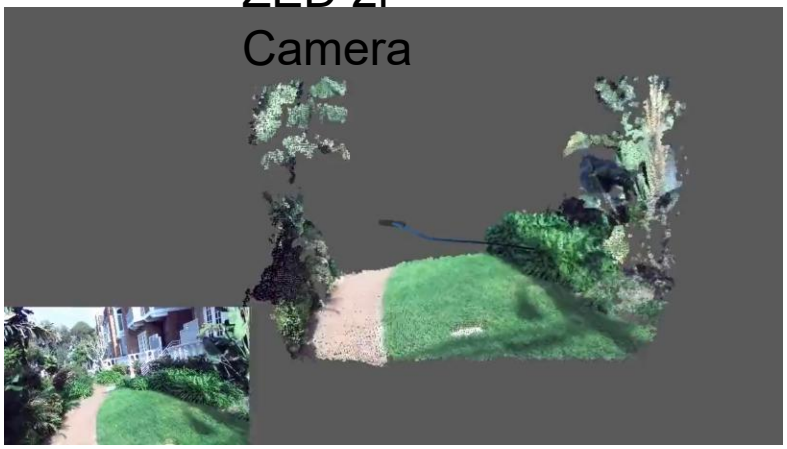
  - We often must use prior knowledge about the world's structure

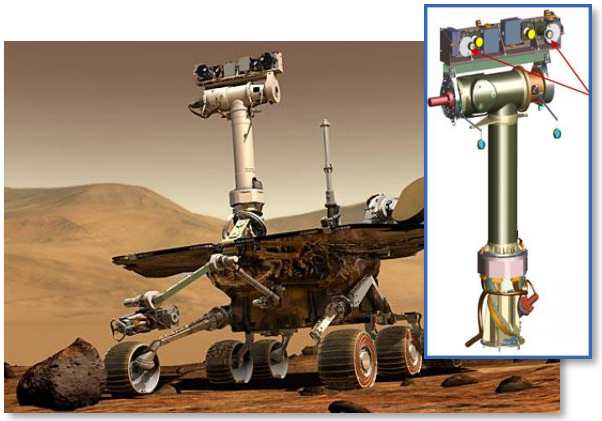
The goal of computer vision

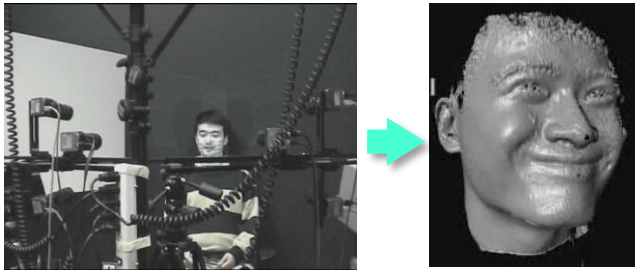
The goal of computer vision

- Compute the 3D shape of the world

ZED 2i
Camera

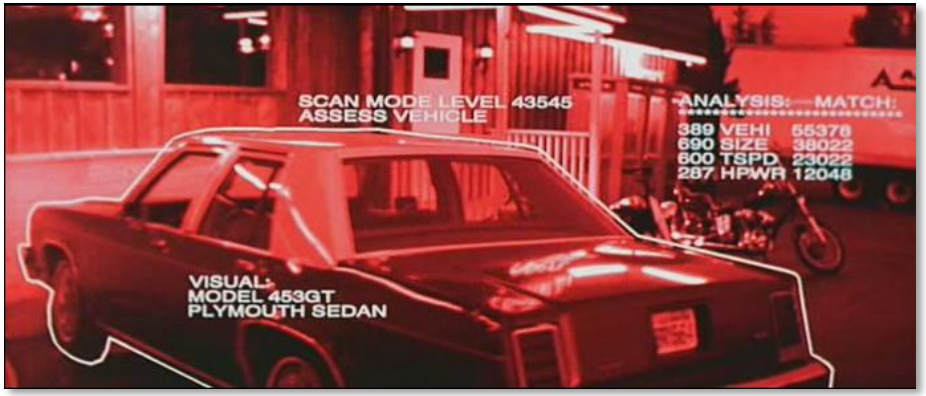






The goal of computer vision

- Recognize objects and people



Terminator 2, 1991

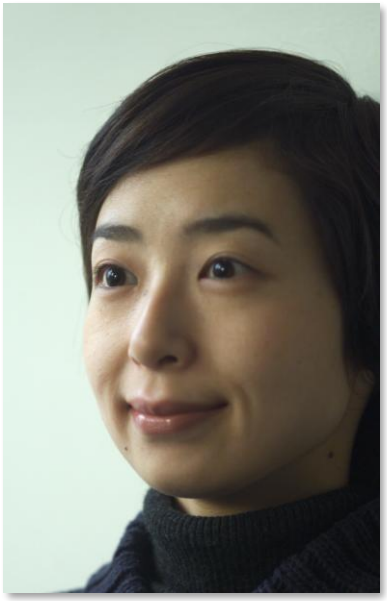
The goal of computer vision

- “Enhance” images


GSUENY

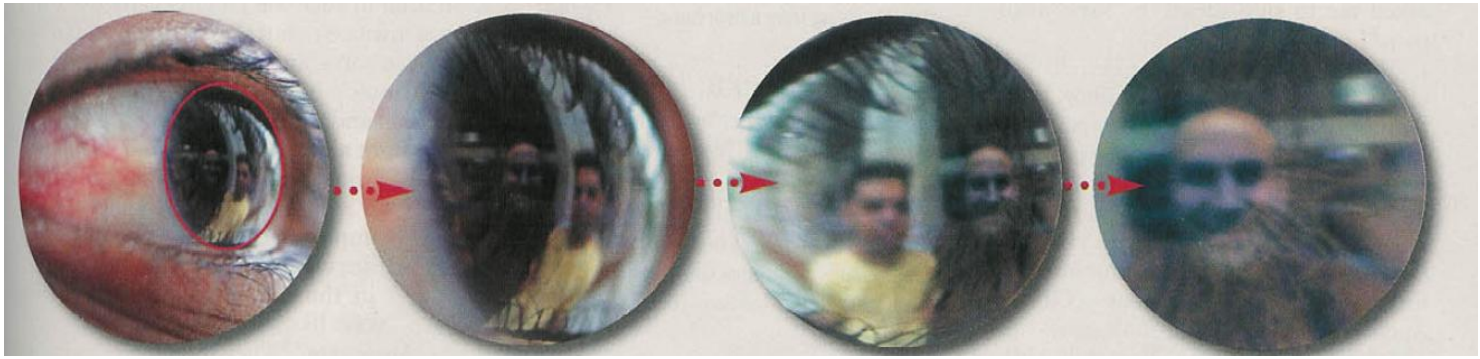
The goal of computer vision

- Forensics



Source: Nayar and Nishino, “Eyes for Rel

Source: Nayar and Nishino, “Eyes for Rel



Source: Nayar and Nishino, “Eyes for Rel

The goal of computer vision

- Improve photos (“Computational Photography”)

Super-resolution (source: 2d3)

Low-light photography
(credit: Hasinoff et al., SIGGRAPH ASIA 2016)

Depth of field on cell phone camera
(source: Google Research Blog)

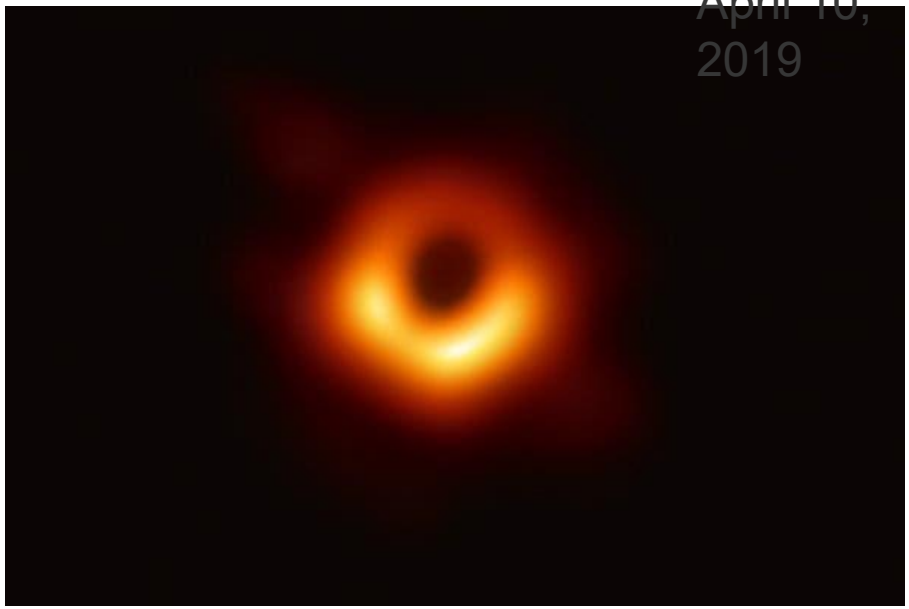
Removing objects
(Google Magic Eraser)

The New York Times

Darkness Visible, Finally: Astronomers Capture First Ever Image of a Black Hole

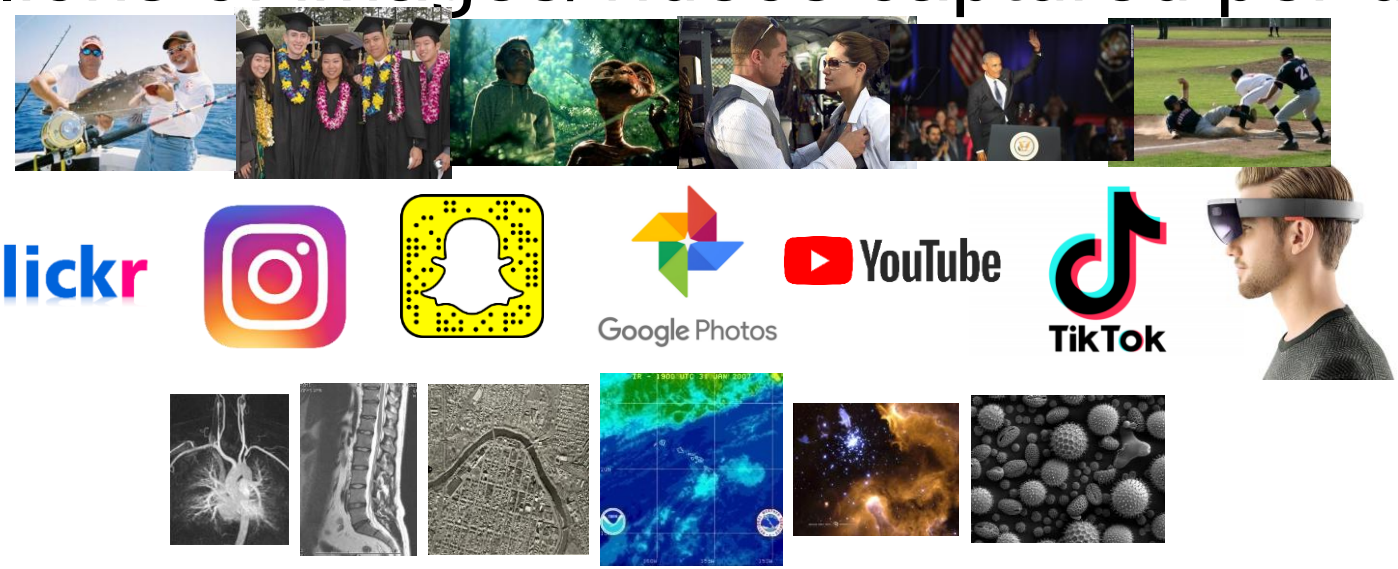
Astronomers at last have captured a picture of one of the most secretive entities in the cosmos.

April 10, 2019



Why study computer vision?

- Billions of images/videos captured per day



- Huge number of potential applications
- The next slides show the current state of the art

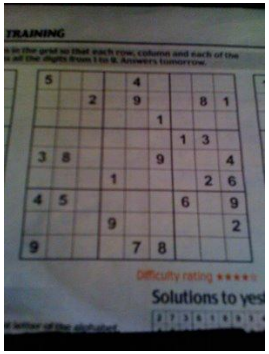
Optical character recognition (OCR)

- If you have a scanner, it probably came with OCR software

Digit recognition, AT&T labs (1990's)
http://yann.lecun.com/exdb/lenet/

License plate readers
http://en.wikipedia.org/wiki/Automatic_numbe

Automatic check processing

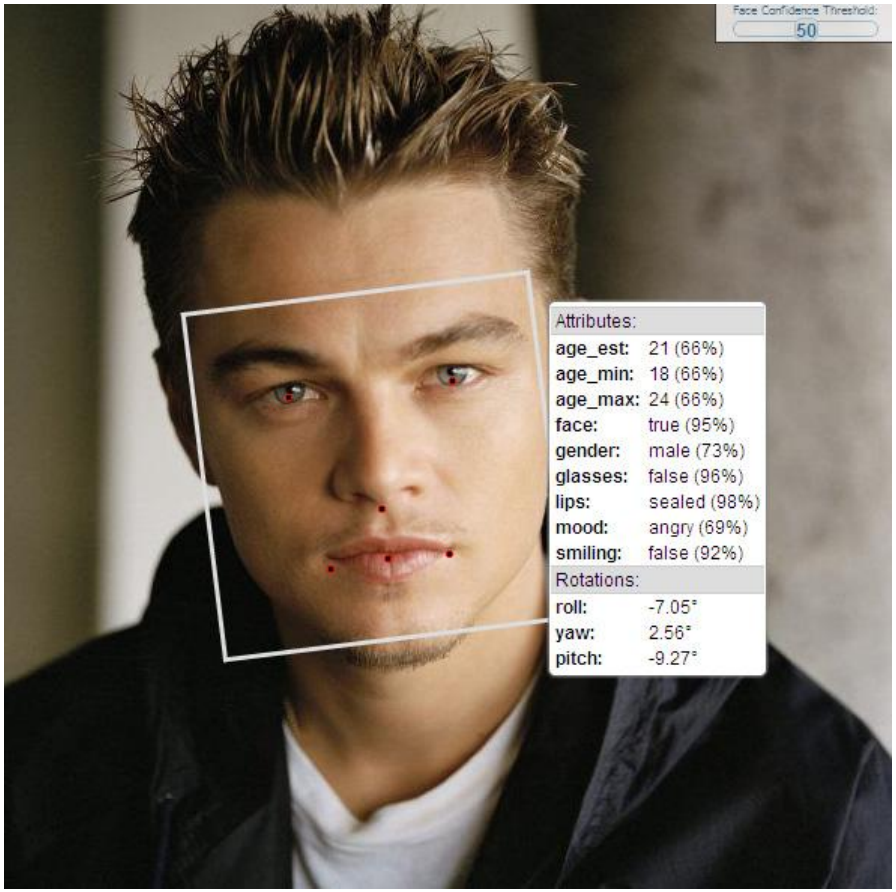


Sudoku grabber
http://sudokugrab.blogspot.co

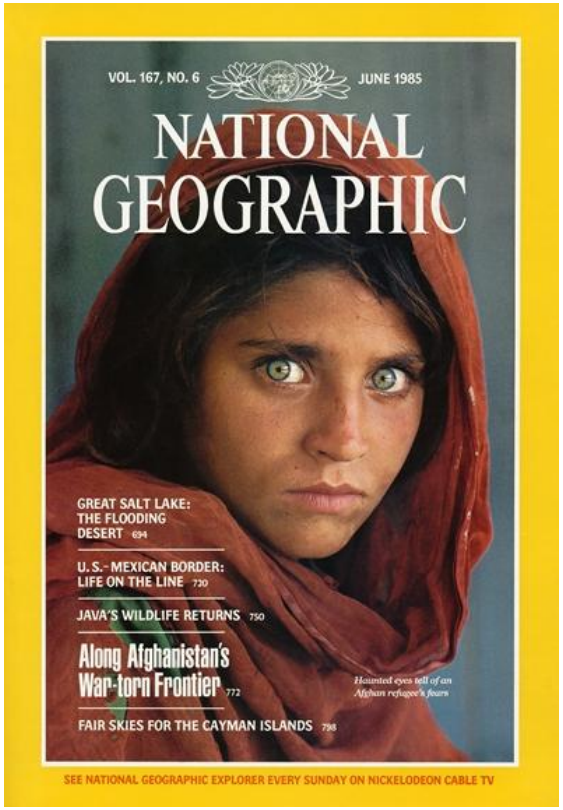
Face detection

- Nearly all cameras detect faces in real time
  - (Why?)

Face analysis and recognition



Vision-based biometrics

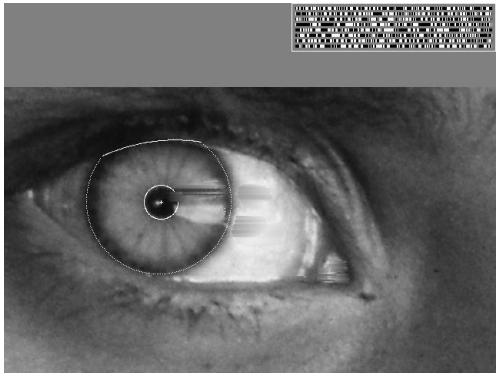


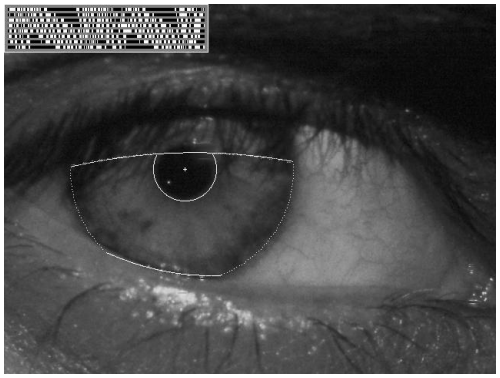
Who is she?

Source: S. Seitz

Vision-based biometrics

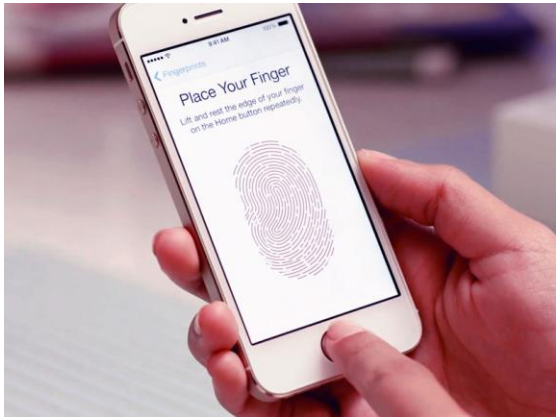
"*How the Afghan Girl was Identified by Her Iris Patterns*" Read the story





Source: S. Seitz

Login without a password



Fingerprint scanners on many new smartphones and other devices

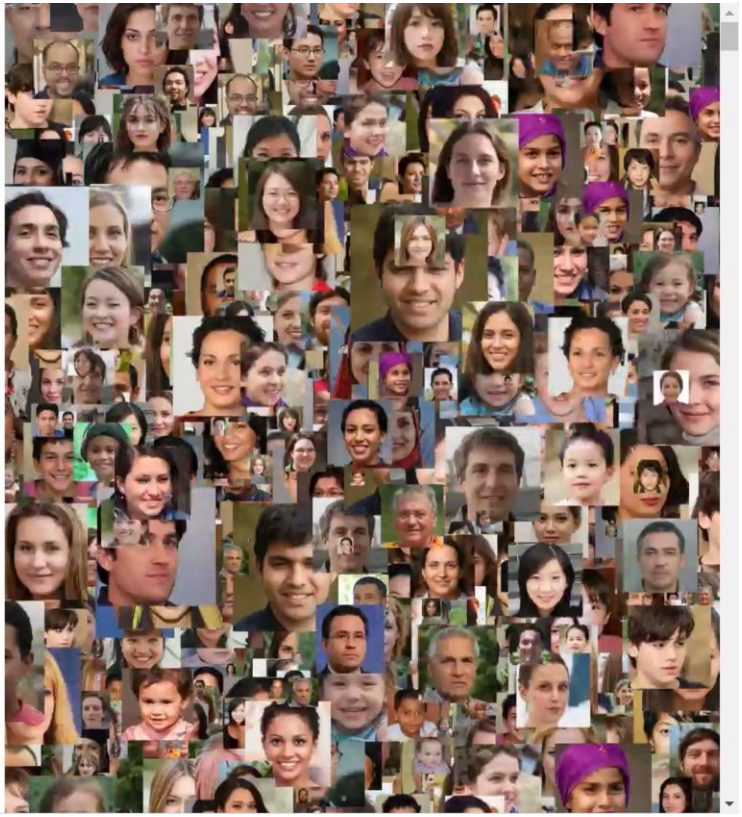
Face unlock on Apple iPhone X
See also http://www.sensiblevision.com/

The New York Times

Account

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.



New York Times, Jan. 18, 2020
by Kashmir Hill

Researchers warn peace sign photos could expose fingerprints

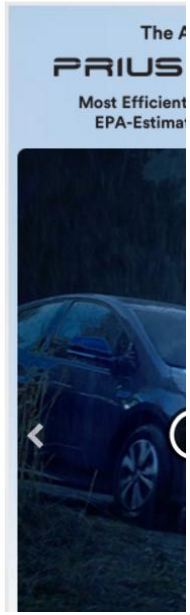
But the likelihood of anyone actually using images to recreate prints is pretty slim.

Jamie Rigg, @jmerigg
01.13.17 in Security

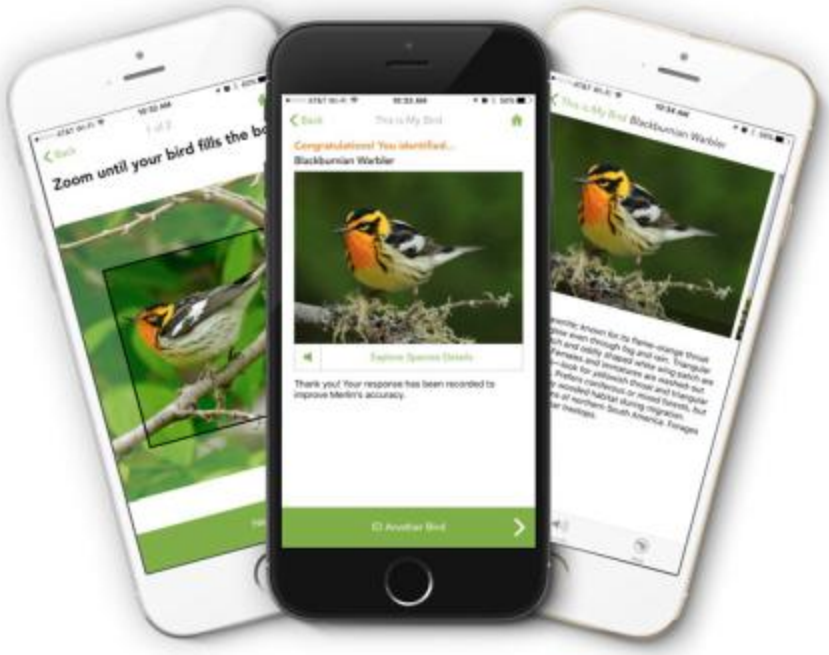
Comments

1721
Shares

Getty



Bird identification



Merlin Bird ID (based on Cornell Tech technology!)

Special effects: shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Source: S. Seitz

Special effects: motion capture

Pirates of the Carribean, Industrial Light and Magic

Source: S. Seitz

Los Angeles Times

MOVIES

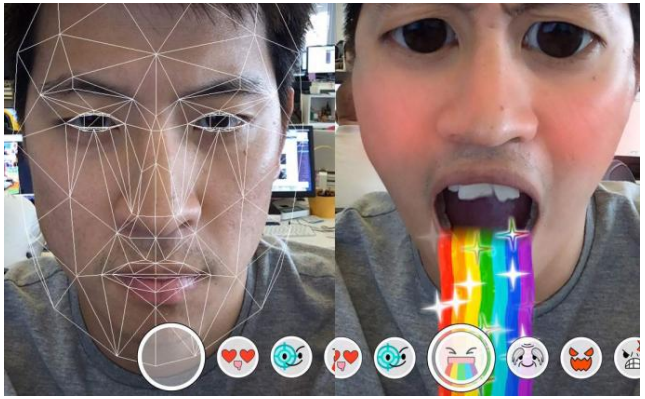
Robert De Niro said no green screen. No face dots. How 'The Irishman's' de-aging changes Hollywood

Makeup and wig work got Robert De Niro partway to his character, Frank Sheeran, at 41, left. It took a specially built camera and visual artists to get all the way there, as before-and-after images show. (Netflix)

By JOSH ROTTENBERG | STAFF WRITER JAN. 2, 2020

Los Angeles Times

3D face tracking w/ consumer cameras



Snapchat Lenses

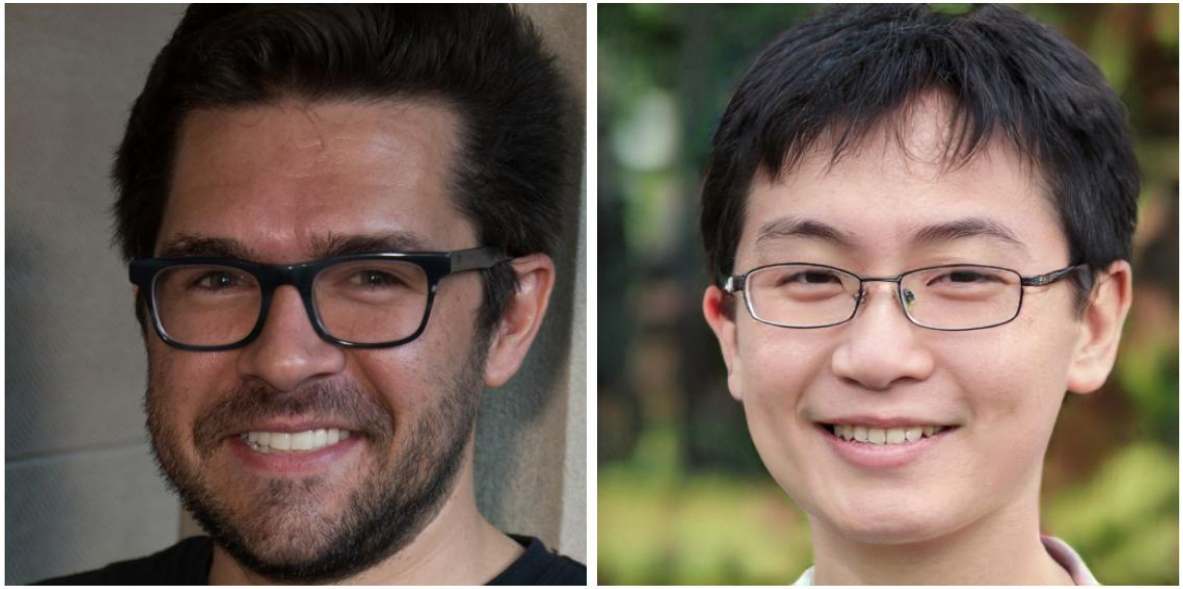
Face2Face system (Thies et al.)

Image synthesis

Karras, et al., *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, ICLR 2018

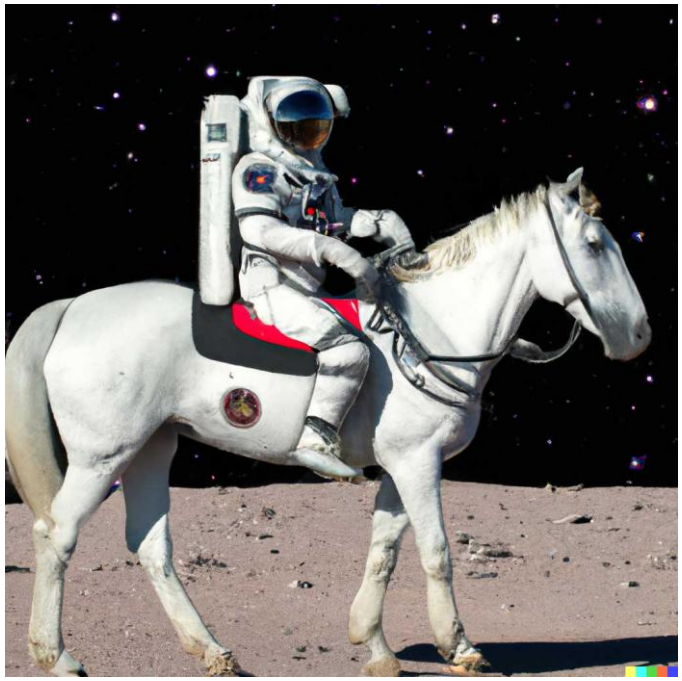
Which face is real?

Click on the person who is real.



https://www.whichfaceisreal.com/

Image synthesis



“An astronaut riding a horse in a photorealistic style” – DALL-E 2

“A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat” – Imagen

Sports

Sportvision first down line
Explanation on www.howstuffworks.com

Source: S. Seitz

Smart cars

- Mobileye
- Tesla Autopilot
- Safety features in many cars

Self-driving cars

Waymo

Robotics

NASA's Mars Curiosity Rover

https://en.wikipedia.org/wiki/Curiosity_(rover)

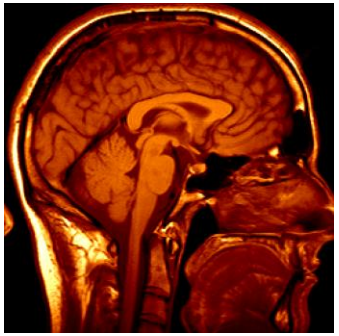
Amazon Picking Challenge

http://www.robocup2016.org/en/events/amazon-picking-challenge/

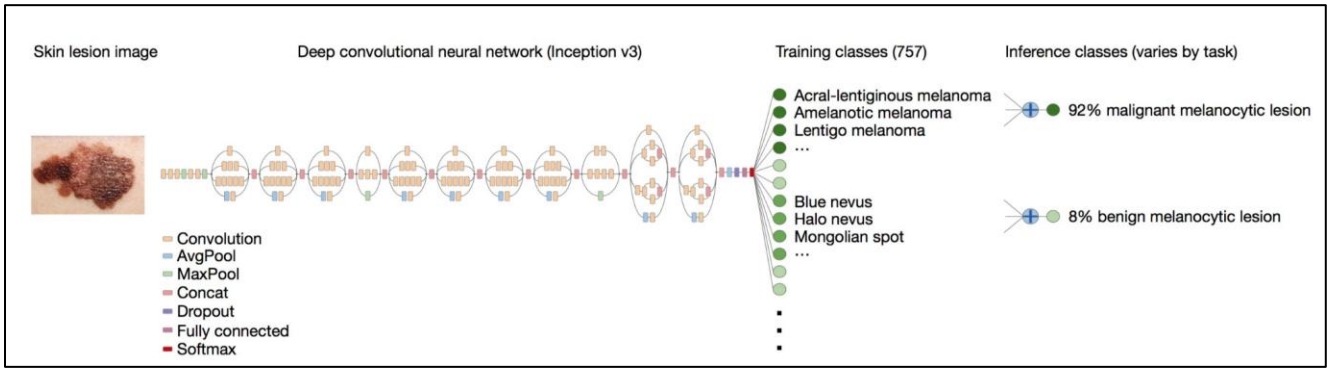
Amazon Prime Air

Amazon Scout

Medical imaging



3D imaging
(MRI, CT)



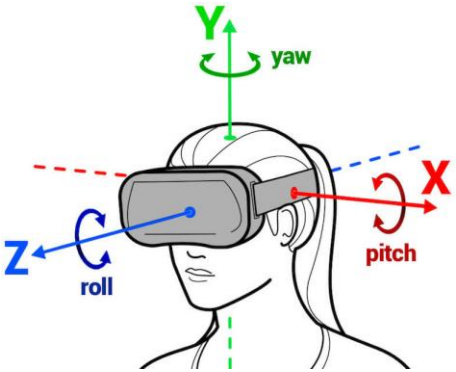
Skin cancer classification with deep learning
https://cs.stanford.edu/people/esteva/nature/

INVESTING 3/25/2014 @ 5:43PM | 70,399 views

Facebook Buys Oculus, Virtual Reality Gaming Startup, For $2 Billion

+ Comment Now + Follow Comments

Virtual & Augmented Reality



6DoF head tracking

Hand & body tracking

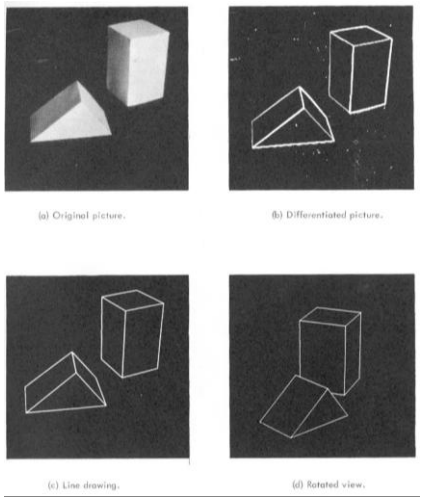
3D scene understanding

3D-360 video capture

Outline

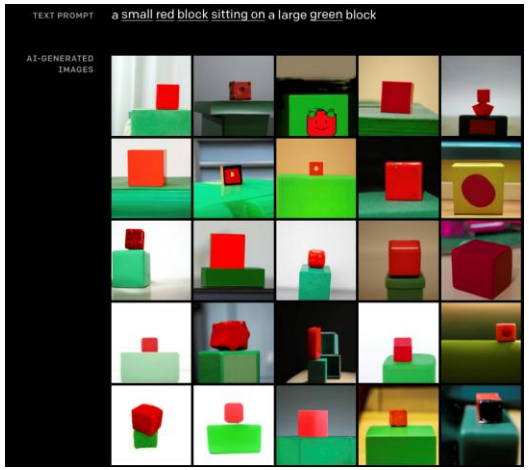
- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision

How it started



L. G. Roberts, 1963

How it's going



OpenAI DALL-E, 2020

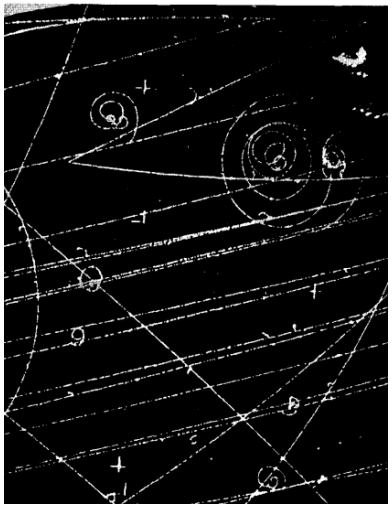
Decade by decade

- **1960s**: Blocks world, image processing and pattern recognition
- **1970s**: Key recovery problems defined: structure from motion, stereo, shape from shading, color constancy. Attempts at knowledge-based recognition
- **1980s**: Fundamental and essential matrix, multi-scale analysis, corner and edge detection, optical flow, geometric recognition as alignment
- **1990s**: Multi-view geometry, statistical and appearance-based models for recognition, first approaches for (class-specific) object detection
- **2000s**: Local features, generic object recognition and detection
- **2010s**: Deep learning, big data

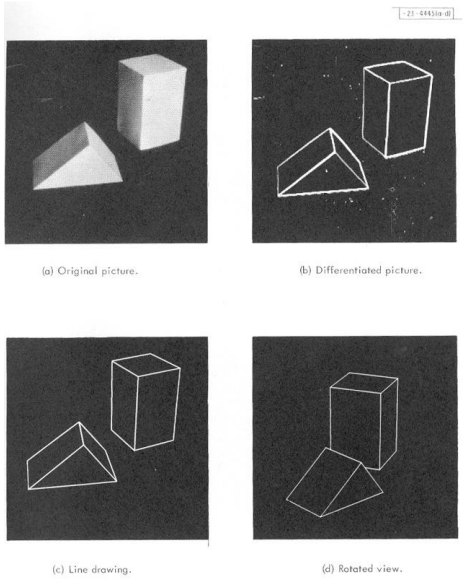
- For much more detail: see Prof Lazebnik's historical overview

Adapted from J. Malik

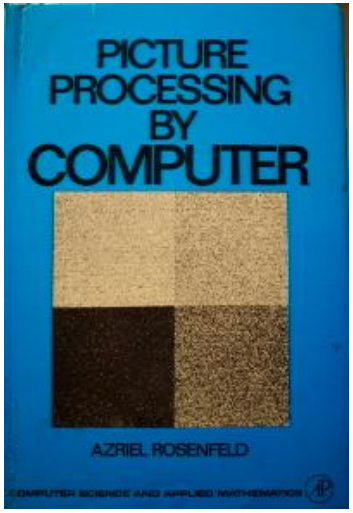
Origins



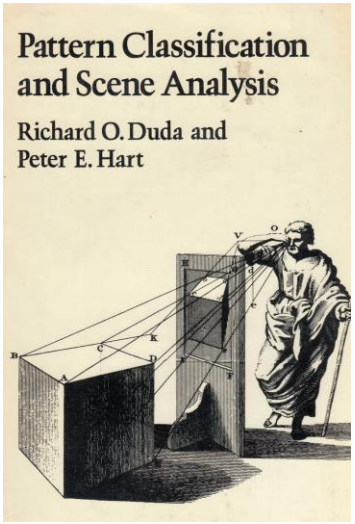
Hough, 1959



Roberts, 1963



Rosenfeld, 1969

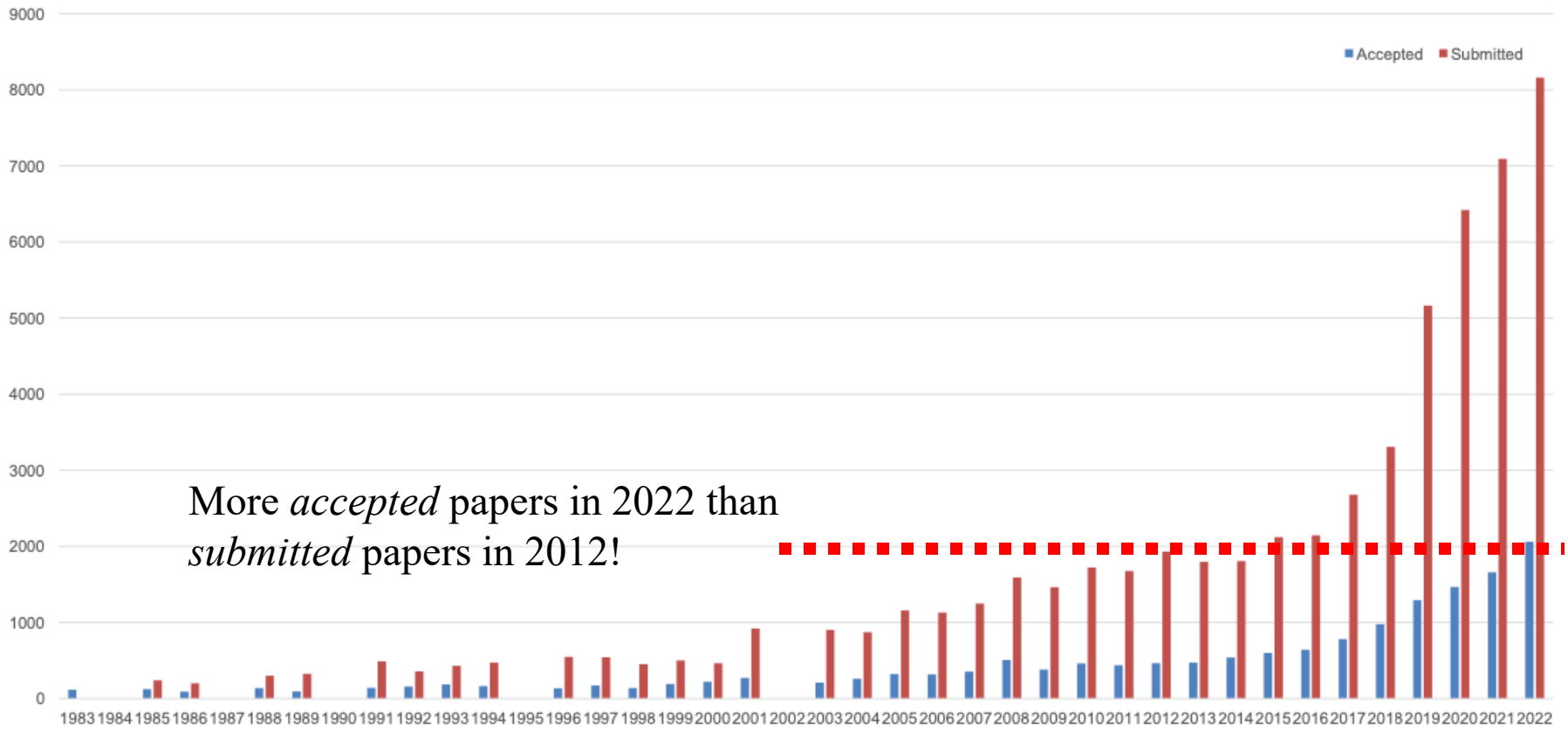


Duda & Hart, 1972

Current state of the art

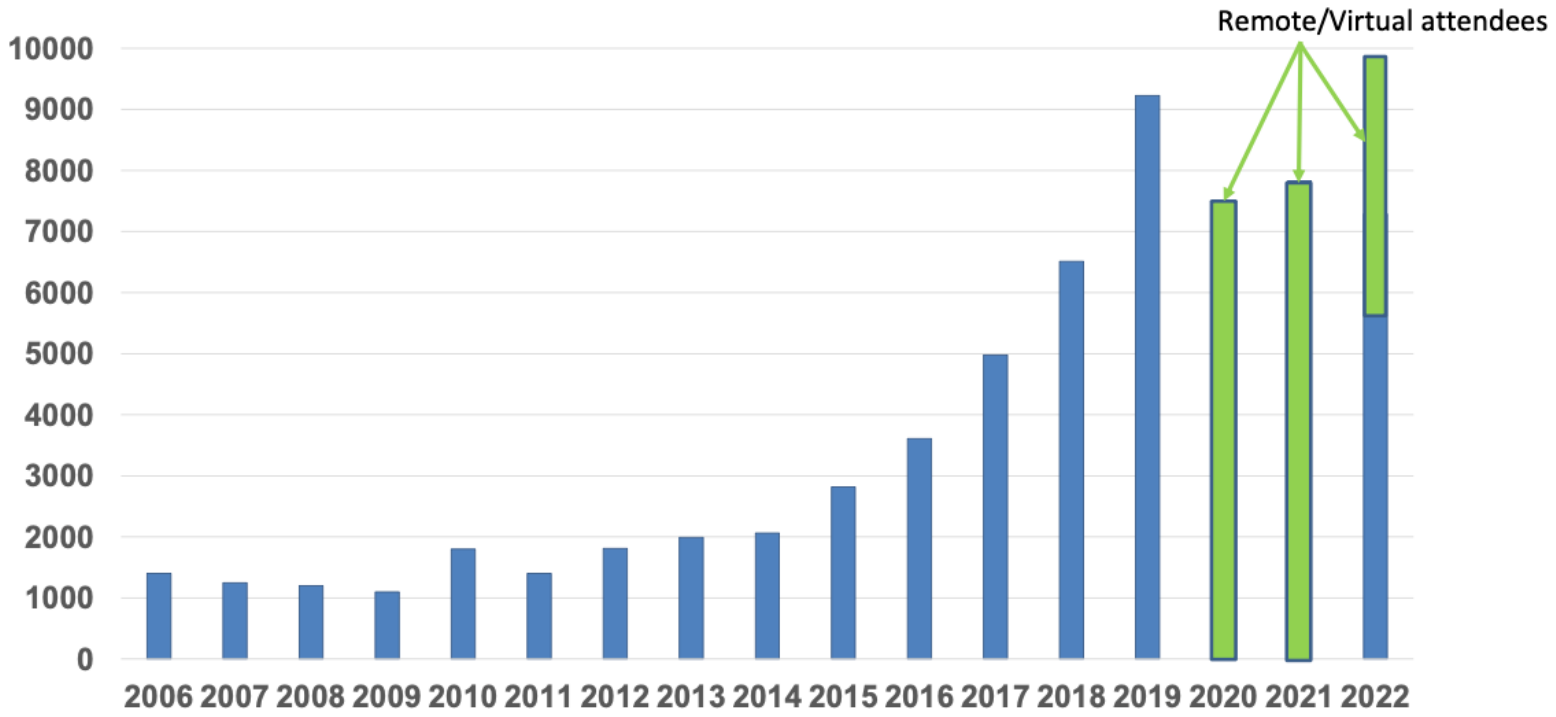
- You just saw many examples of current systems.
  - Many of these are less than 10 years old
- Computer vision is an active research area, and rapidly changing
  - Many new apps in the next 5 years
  - Deep learning and generative methods powering many modern applications
- Many startups across a dizzying array of areas
  - Generative AI, robotics, autonomous vehicles, medical imaging, construction, inspection, VR/AR, …

Growth of the field: CVPR papers



Source: CVPR 2022 opening sides

Growth of the field: CVPR attendance



Source: CVPR 2022 opening sides

Google Scholar

Top publications

Categories English

	Publication	h5-index	h5-median
1.	Nature	376	552
2.	The New England Journal of Medicine	365	639
3.	Science	356	526
4.	The Lancet	301	493
5.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	299	509
6.	Advanced Materials	273	369
7.	Nature Communications	273	366
8.	Cell	269	417
9.	Chemical Reviews	267	438
10.	Chemical Society reviews	240	368

Vision

Vision

Vision

Guide2Research

Top Computer Science Conferences

*Ranking is based on **Conference H5-index>=12** provided by Google Scholar Metrics*

Show Due only | All Categories | All Countries | Search by keyword

Rank	Publisher	Conference Details	H5-index	Impact Score
1	IEEE	**CVPR : IEEE/CVF Conference on Computer Vision and Pattern Recognition** Jun 21, 2021 - Jun 24, 2021 - Nashville , United States http://cvpr2021.thecvf.com/	299	51.98
2		**NeurIPS : Neural Information Processing Systems (NIPS)** Dec 6, 2021 - Dec 14, 2021 - Online , Online https://nips.cc/	198	33.49
3	IEEE	**ICCV : IEEE/CVF International Conference on Computer Vision** Oct 11, 2021 - Oct 17, 2021 - Montreal , Canada http://iccv2021.thecvf.com/home	176	32.51
4	Springer	**ECCV : European Conference on Computer Vision** Oct 11, 2021 - Oct 17, 2021 - Montreal , Canada http://iccv2021.thecvf.com/	144	25.91
5		**AAAI : AAAI Conference on Artificial Intelligence** Feb 2, 2021 - Feb 9, 2021 - Vancouver , Canada https://aaai.org/Conferences/AAAI-21/	126	25.57

Andrej Karpathy blog

The state of Computer Vision and AI: we are really, really far.

Oct 22, 2012

The picture above is funny.

But for me it is also one of those examples that make me sad about the outlook for AI and for Computer Vision. What would it take for a computer to understand this image as you or I do? I challenge you to think explicitly of all the pieces of knowledge that have to fall in place for it to make sense. Here is my short attempt:

- You recognize it is an image of a bunch of people and you understand they are in a hallway
- You recognize that there are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.
- You recognize Obama from the few pixels that make up his face. It helps that he is in his suit and that he is surrounded by other people with suits.
- You recognize that there's a person standing on a scale, even though the scale occupies only very few white pixels that blend with the background. But, you've used the person's pose and knowledge of how people interact with objects to figure it out.
- You recognize that Obama has his foot positioned just slightly on top of the scale. Notice the language I'm using: It is in terms of the 3D structure of the scene, not the position of the leg in the 2D coordinate system of the image.
- You know how physics works: Obama is leaning in on the scale, which applies a force on it. Scale measures force that is applied on it, that's how it works => it will over-estimate the weight of the person standing on it.
- The person measuring his weight is not aware of Obama doing this. You derive this because you know his pose, you understand that the field of view of a person is finite, and you understand that he is not very likely to sense the slight push of Obama's foot.
- You understand that people are self-conscious about their weight. You also understand that he is reading off the scale measurement, and that shortly the over-estimated weight will confuse him because it will probably be much higher than what he expects. In other words, you reason about implications of the events that are about to unfold seconds after this photo was taken, and especially about the thoughts and how they will develop inside people's heads. You also reason about what pieces of information are available to people.
- There are people in the back who find the person's imminent confusion funny. In other words you are reasoning about state of mind of people, and their view of the state of mind of another person. That's getting frighteningly meta.
- Finally, the fact that the perpetrator here is the president makes it maybe even a little more funnier. You understand what actions are more or less likely to be undertaken by different people based on their status and identity.

The state of Computer Vision and AI: we are really, really far.

Oct 22, 2012

The picture above is funny.

But for me it is also one of those examples that make me sad about the outlook for AI and for Computer Vision. What would it take for a computer to understand this image as you or I do? I challenge you to think explicitly of all the pieces of knowledge that have to fall in place for it to make sense. Here is my short attempt:

- You recognize it is an image of a bunch of people and you understand they are in a hallway
- You recognize that there are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.
- You recognize Obama from the few pixels that make up his face. It helps that he is in his suit and that he is surrounded by other people with suits.
- You recognize that there's a person standing on a scale, even though the scale occupies only very few white pixels that blend with the background. But, you've used the person's pose and knowledge of how people interact with objects to figure it out.
- You recognize that Obama has his foot positioned just slightly on top of the scale. Notice the language I'm using: It is in terms of the 3D structure of the scene, not the position of the leg in the 2D coordinate system of the image.
- You know how physics works: Obama is leaning in on the scale, which applies a force on it. Scale measures force that is applied on it, that's how it works => it will over-estimate the weight of the person standing on it.
- The person measuring his weight is not aware of Obama doing this. You derive this because you know his pose, you understand that the field of view of a person is finite, and you understand that he is not very likely to sense the slight push of Obama's foot.
- You understand that people are self-conscious about their weight. You also understand that he is reading off the scale measurement, and that shortly the over-estimated weight will confuse him because it will probably be much higher than what he expects. In other words, you reason about implications of the events that are about to unfold seconds after this photo was taken, and especially about the thoughts and how they will develop inside people's heads. You also reason about what pieces of information are available to people.
- There are people in the back who find the person's imminent confusion funny. In other words you are reasoning about state of mind of people, and their view of the state of mind of another person. That's getting frighteningly meta.
- Finally, the fact that the perpetrator here is the president makes it maybe even a little more funnier. You understand what actions are more or less likely to be undertaken by different people based on their status and identity.