

Computer Vision

Course webpage URL: <http://luthuli.cs.uiuc.edu/~daf>

And follow links

Vision group at Illinois



- **David Forsyth**
 - Fulton Watson Copp chair; Marr prize, 1993; 2 ex students with Marr prizes; IEEE Tech. Achievement, Fellow; ACM Fellow; EIC IEEE TPAMI; 4 patents



- **Jim Rehg**
 - HCESC director; multiple famous ex-students, best paper awards; 26 patents



- **Derek Hoiem**
 - best paper, CVPR 2006; ACM Doctoral Dissertation honorable mention; Sloan Fellow; PAMI-TC Young Researcher; major startup



- **Lana Lazebnik**
 - Microsoft Faculty Fellow; Sloan Fellow; Koenderink Prize (2016); EIC IJCV



- **Alex Schwing**
 - Visual learning, segmentation and GAN models



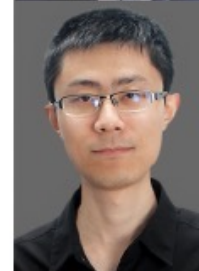
- **Saurabh Gupta**
 - Linking visual sensing to motion



- **Liangyan Gui**
 - Understanding human movement

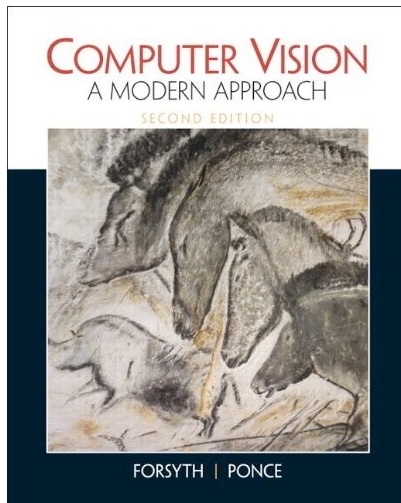


- **Shenlong Wang**
 - Simulation and sensing for autonomous vehicles



- **Yuxiong Wang**
 - Learning to detect and classify with very little data

Vision group



Well-known ex-students:
Lana Lazebnik (UIUC)
Tamara Berg (UNC)
Pinar Duygulu (Hacettepe U.)
Ian Endres
Ali Farhadi (UW)
Varsha Hedau
Nazli Ikizler (Hacettepe U.)
Brett Jones
Kevin Karsch
Zicheng Liao
Deva Ramanan (CMU)
Raj Sodhi
Gang Wang (now Alibaba)
Amin Sadeghi
Zicheng Liao (Zhejiang U.)

The New Computer Vision

D.A. Forsyth

Likely about ~~2024~~ 2026

Cover design opportunity!

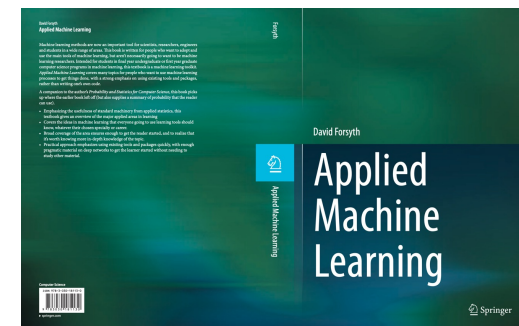
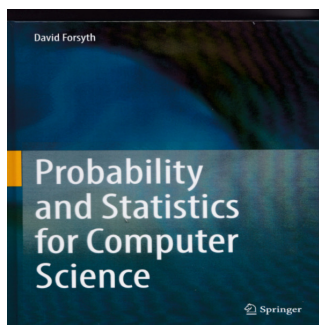
Startups:

Lightform

Revery.ai

Reconstruct

Depix





Top publications

Categories		English	
	Publication	<u>h5-index</u>	<u>h5-median</u>
1.	Nature	<u>488</u>	745
2.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	<u>440</u>	689
3.	The New England Journal of Medicine	<u>434</u>	897
4.	Science	<u>409</u>	633
5.	Nature Communications	<u>375</u>	492
6.	The Lancet	<u>368</u>	678
7.	Neural Information Processing Systems	<u>337</u>	614
8.	Advanced Materials	<u>327</u>	420
9.	Cell	<u>320</u>	482
10.	International Conference on Learning Representations	<u>304</u>	584
11.	JAMA	<u>298</u>	498
12.	Science of The Total Environment	<u>297</u>	436
13.	IEEE/CVF International Conference on Computer Vision	<u>291</u>	484
14.	Angewandte Chemie International Edition	<u>281</u>	361
15.	Nature Medicine	<u>274</u>	474
16.	Journal of Cleaner Production	<u>272</u>	359
17.	International Conference on Machine Learning	<u>268</u>	424
18.	Chemical Reviews	<u>267</u>	461
19.	Proceedings of the National Academy of Sciences	<u>267</u>	405
20.	IEEE Access	<u>266</u>	364

.com

Home / Best Conferences - Computer Science

Best Computer Science Conferences

The ranking of leading conferences for Computer Science was prepared by Research.com, one of the primary websites for Computer Science research offering accurate data on scientific output since 2014.

The spot on the list is based on Impact Score data collected on 21-11-2023. It was based on a careful analysis of as much as 3,040 conference profiles and websites. [Show more](#)

All resear All publi All countr Paper submission open

Search by name

Rank	Conference Details	Impact Score
1 Rank	 Computer Vision and Pattern Recognition 18-06-2023 - 22-06-2023 - Vancouver	62.60
2 Rank	 International Conference on Computer Vision 11-10-2021 - 11-10-2021 - Montreal	43.80
3 Rank	 Neural Information Processing Systems 12-12-2023 - 14-12-2023 - New Orleans	40.60

4 Rank	 International Conference on Learning Representations 01-05-2023 - 05-05-2023 - Kigali	38.10
5 Rank	 AAAI Conference on Artificial Intelligence 07-02-2023 - 14-02-2023 - Washington DC	35.00
6 Rank	 International Conference on Machine Learning 17-07-2022 - 23-07-2022 - Baltimore	30.10
7 Rank	 European Conference on Computer Vision 29-09-2024 - 04-10-2024 - Milan	29.60
8 Rank	 Meeting of the Association for Computational Linguistics 22-05-2022 - 27-05-2022 - Dublin	28.90
9 Rank	 Empirical Methods in Natural Language Processing 06-12-2023 - 10-12-2023 - Singapore City	21.90
10 Rank	 International Joint Conference on Artificial Intelligence 23-07-2022 - 29-07-2022 - Vienna	21.10

Class mechanics

Boeing has a website that mirrors my class website

Mine is

<http://luthuli.cs.uiuc.edu/~daf/ForBoeing/classwebpage.html>

A usual week (current planning):

- lecture movies come out at least a week before
(19, 22, 29 Jan are there already)
- you watch lecture movies, read notes
- at meeting, we discuss:
 - issues
 - quiz questions
 - assignments

Outline

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

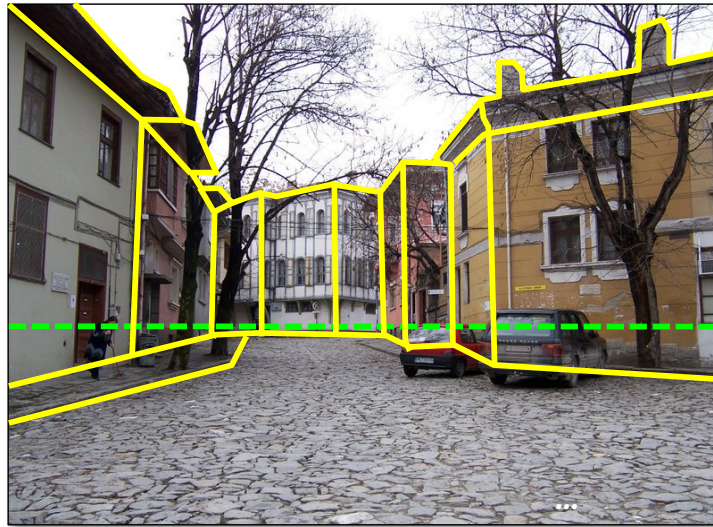
What does vision do?

- *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?*

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?

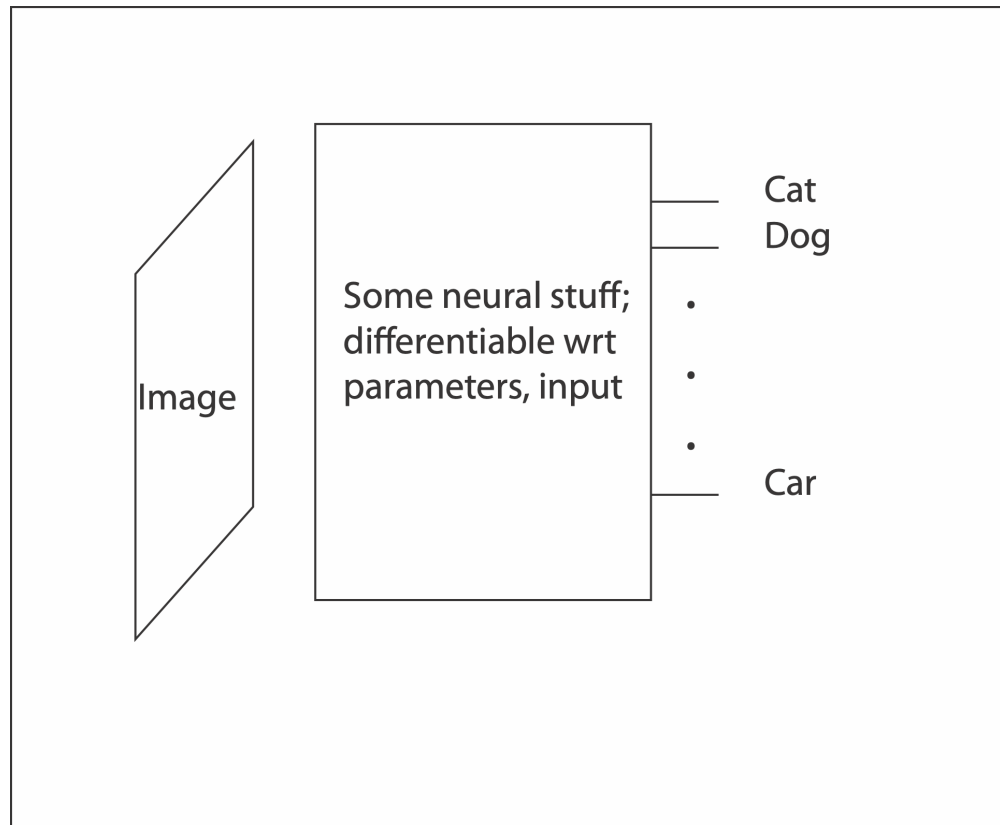


Geometric information

Semantic information

Vision for action

Classification



Detection

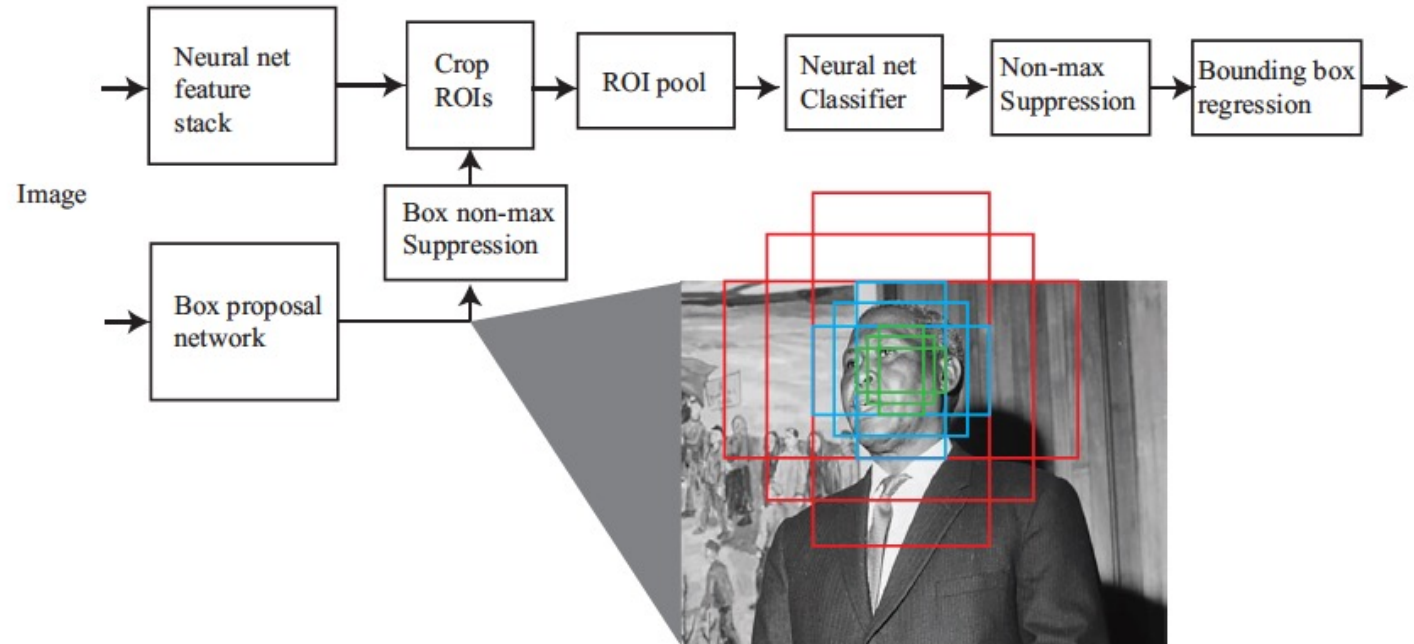
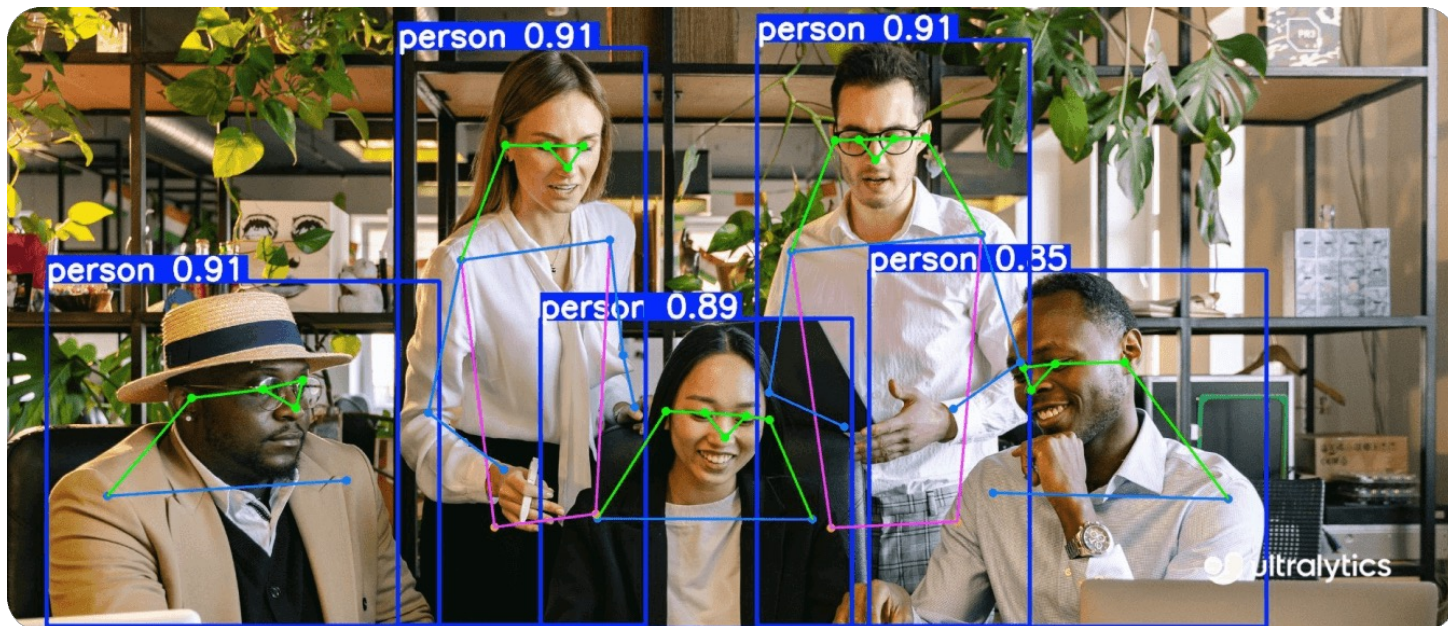


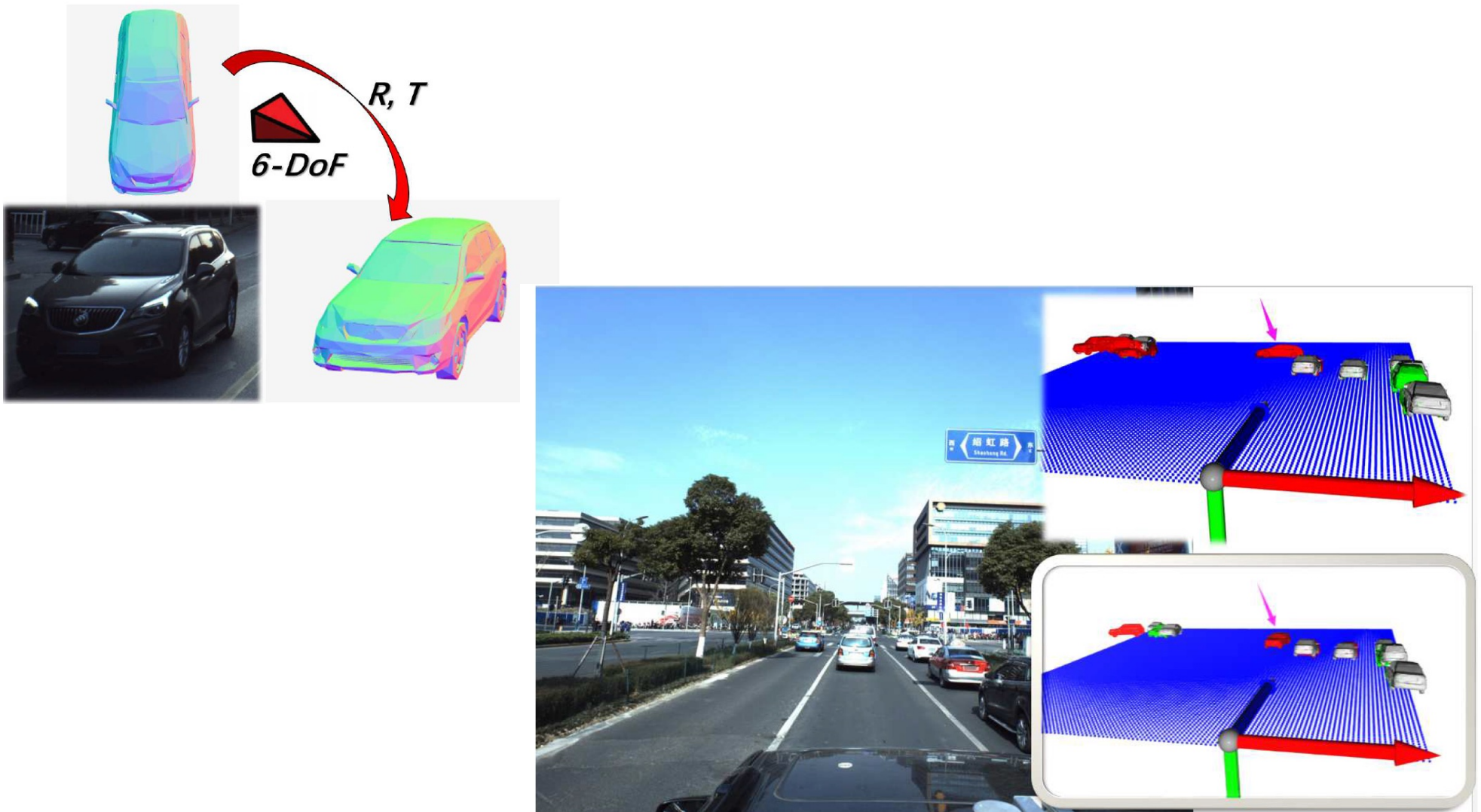
FIGURE 18.8: *Faster RCNN uses two networks. One uses the image to compute “objectness” scores for a sampling of possible image boxes. The samples (called “anchor boxes”) are each centered at a grid point. At each grid point, there are nine boxes (three scales, three aspect ratios). The second is a feature 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



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

Lane detection

US 9081385

Waymo and Google 2012

Strategy: detect markers (reflective paint), join up

exercise in robust fitting of curves

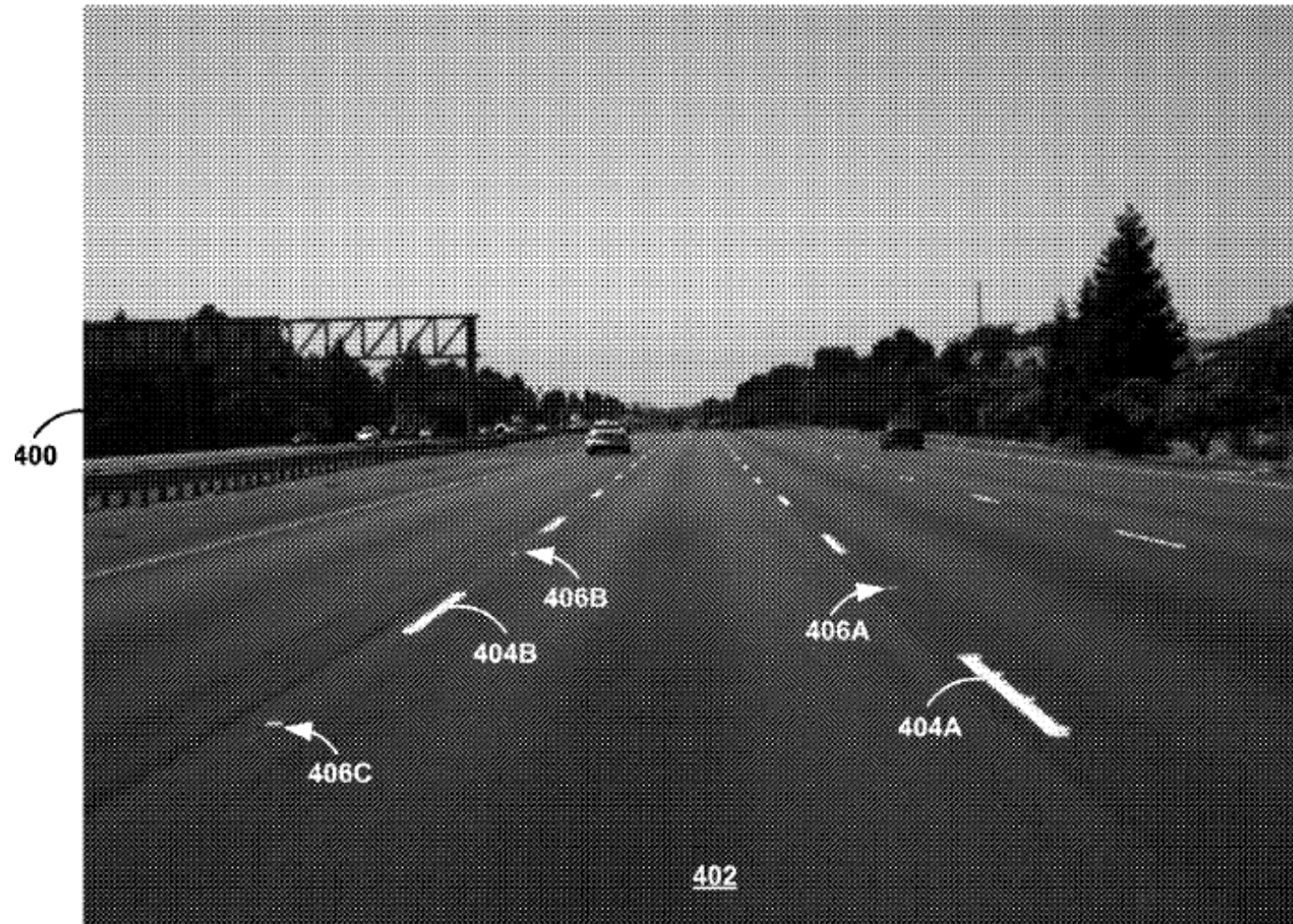
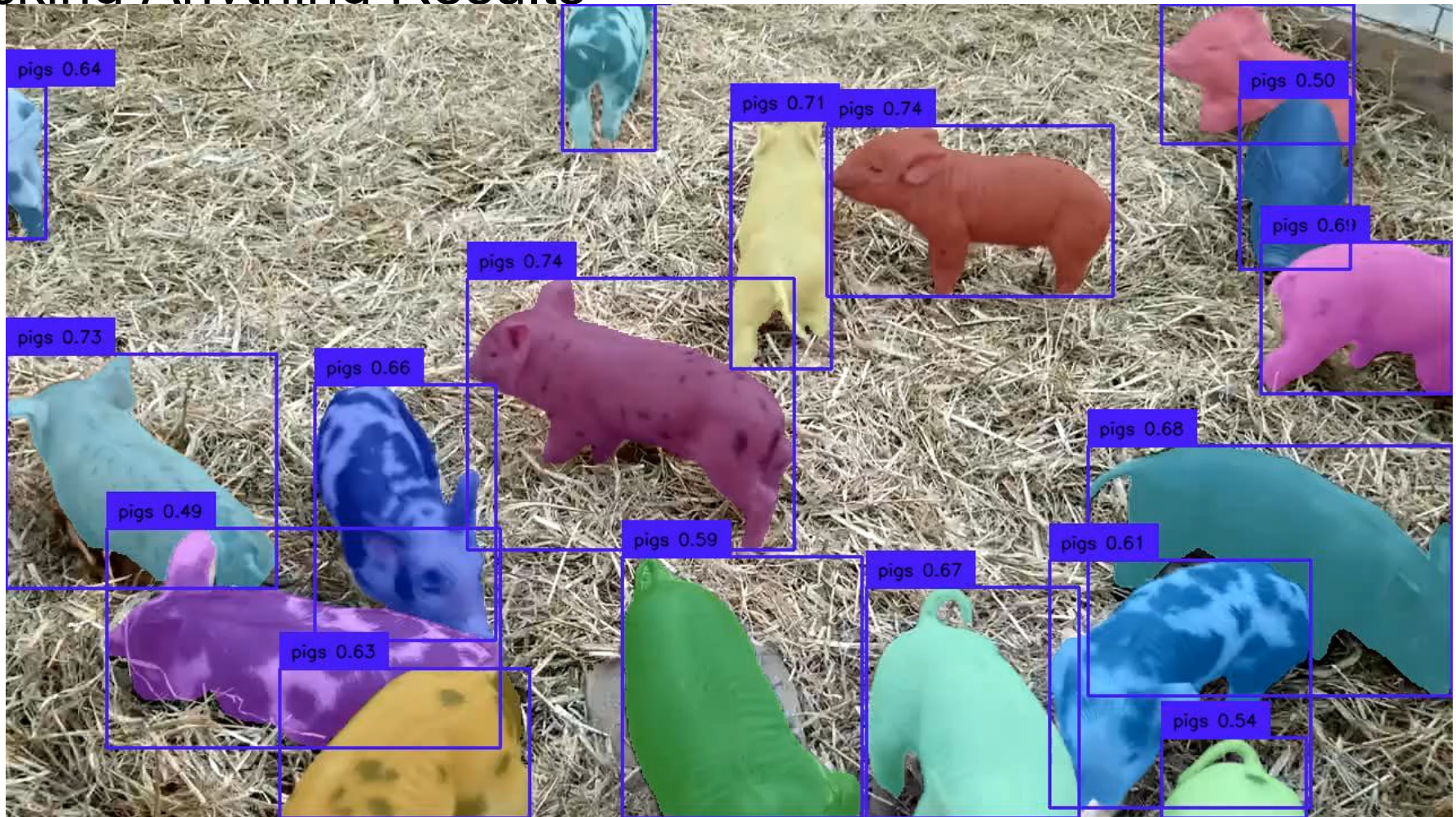


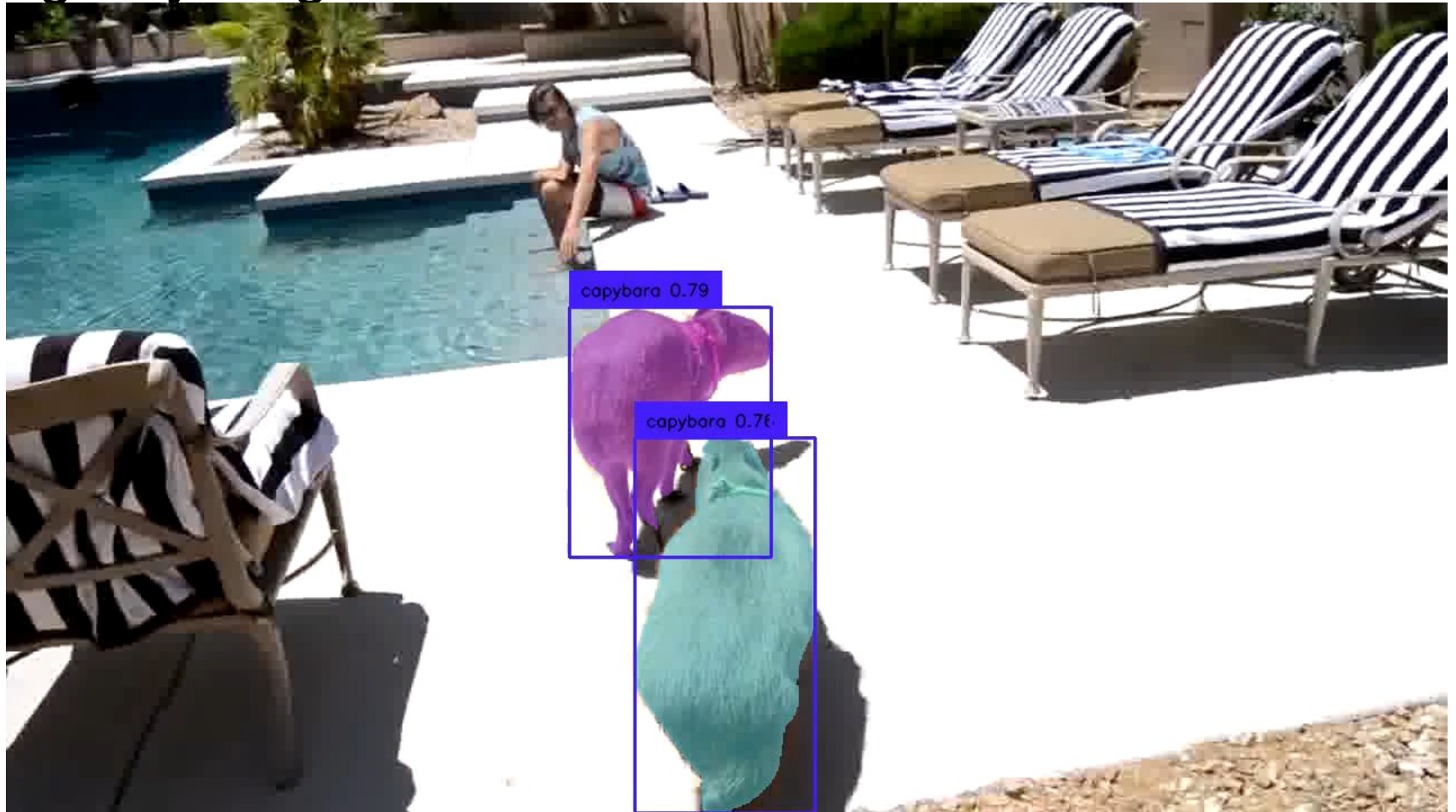
FIGURE 4A

Tracking Anything Results



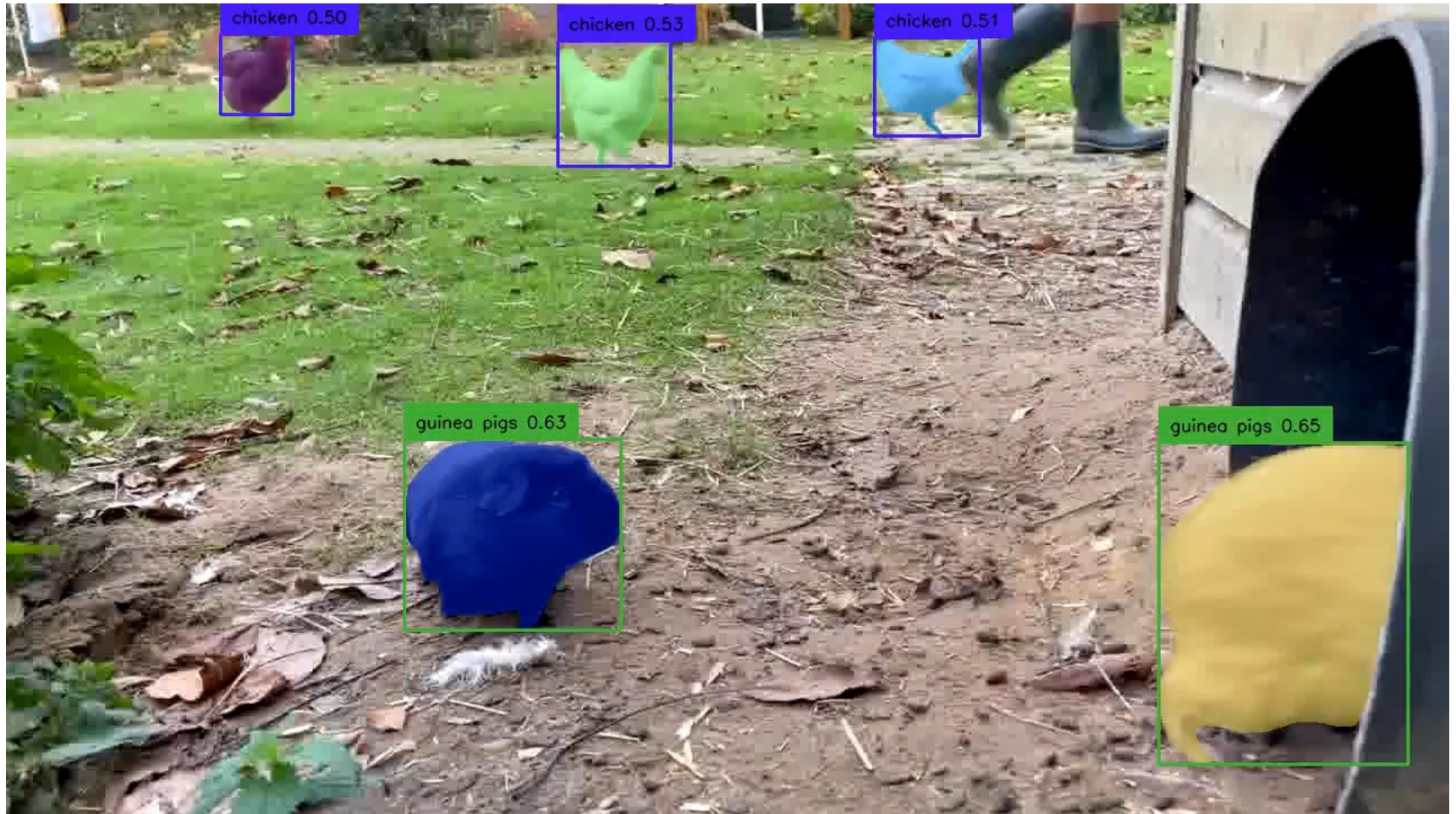
Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

Tracking Anything Results



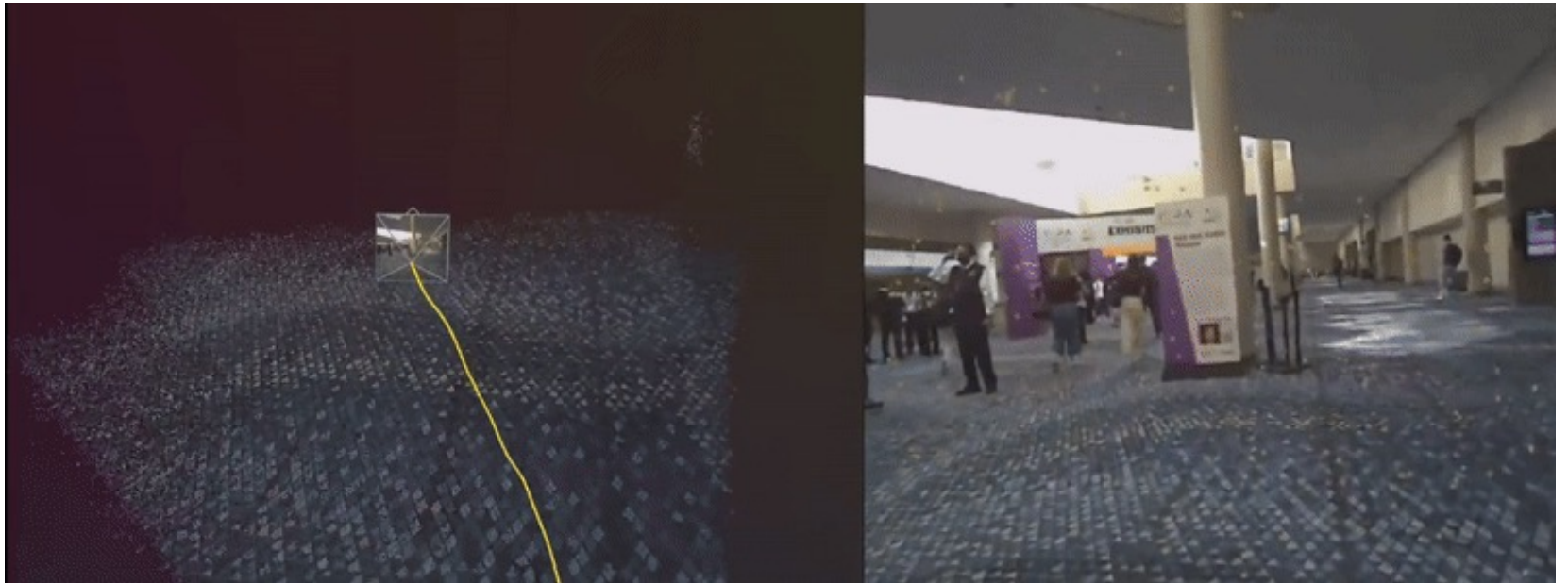
Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

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=fBiataDpGlo>

Goal: To extract useful information from pixels



What we see

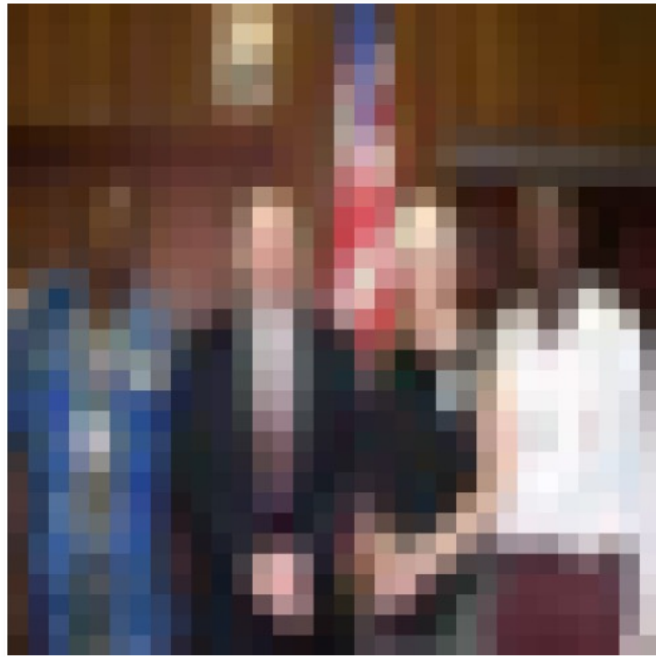
0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

What a computer sees

Trad. View: Images are fundamentally ambiguous!

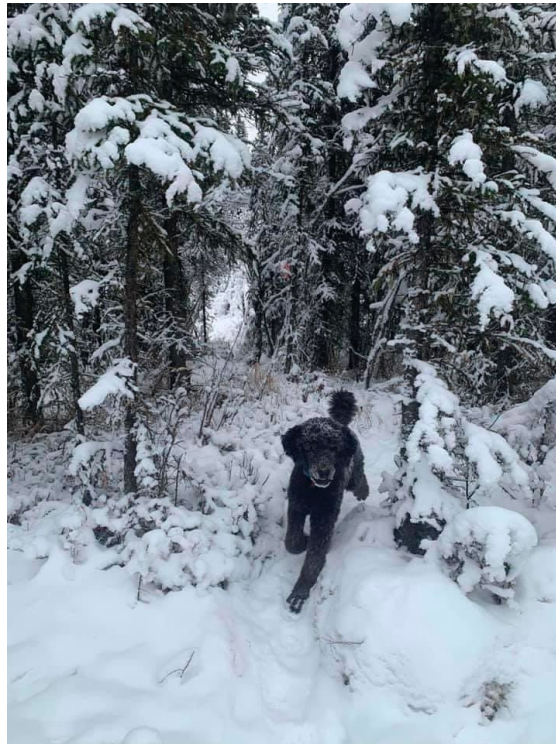


Humans are remarkably good at vision...



Source: "80 million tiny images" by Torralba et al.

...still, vision is hard even for humans



[Image source](#)



[Image source](#)

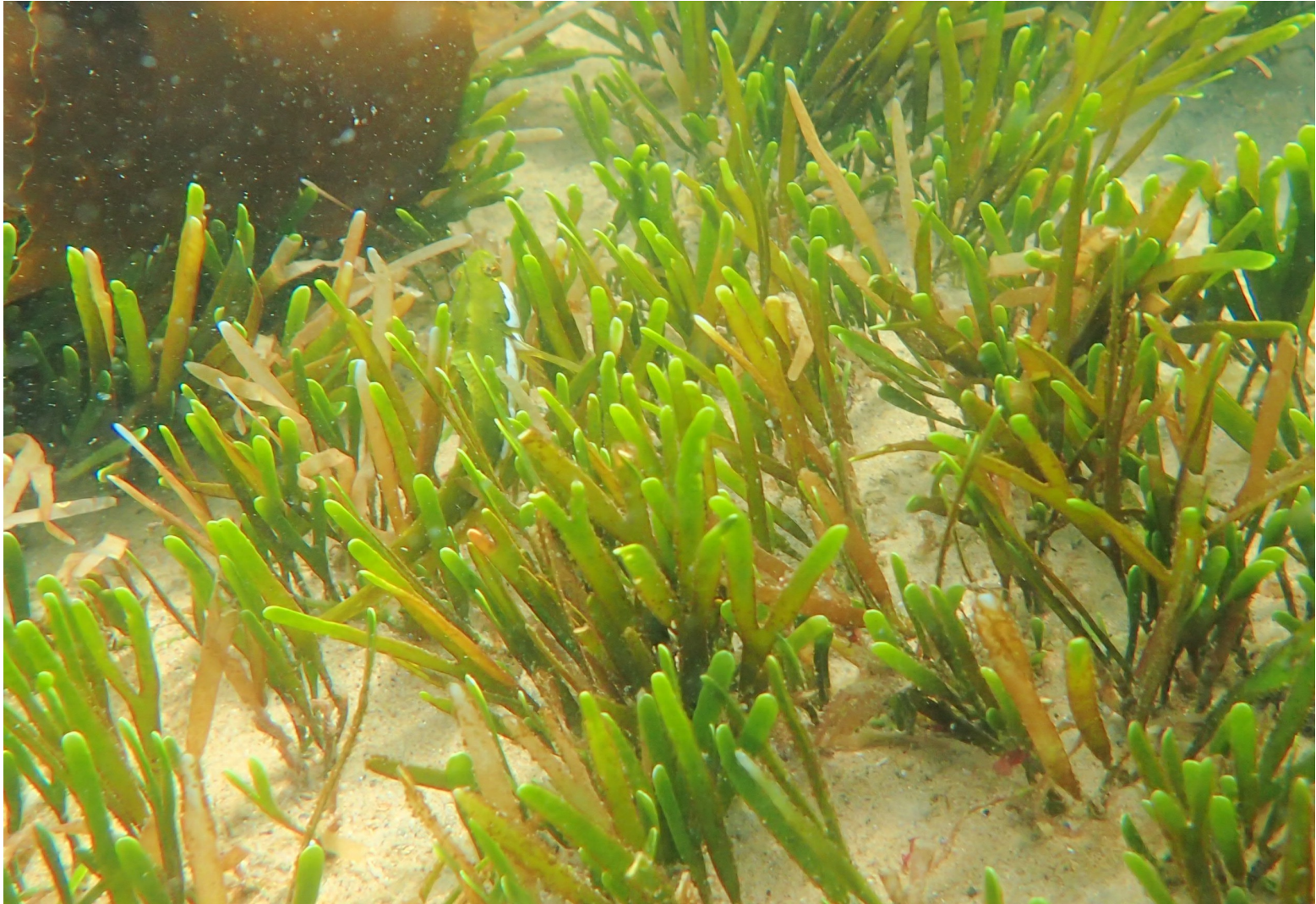
...still, vision is hard even for humans



Figure from Marr (1982), attributed to R. C. James

Is there a fi

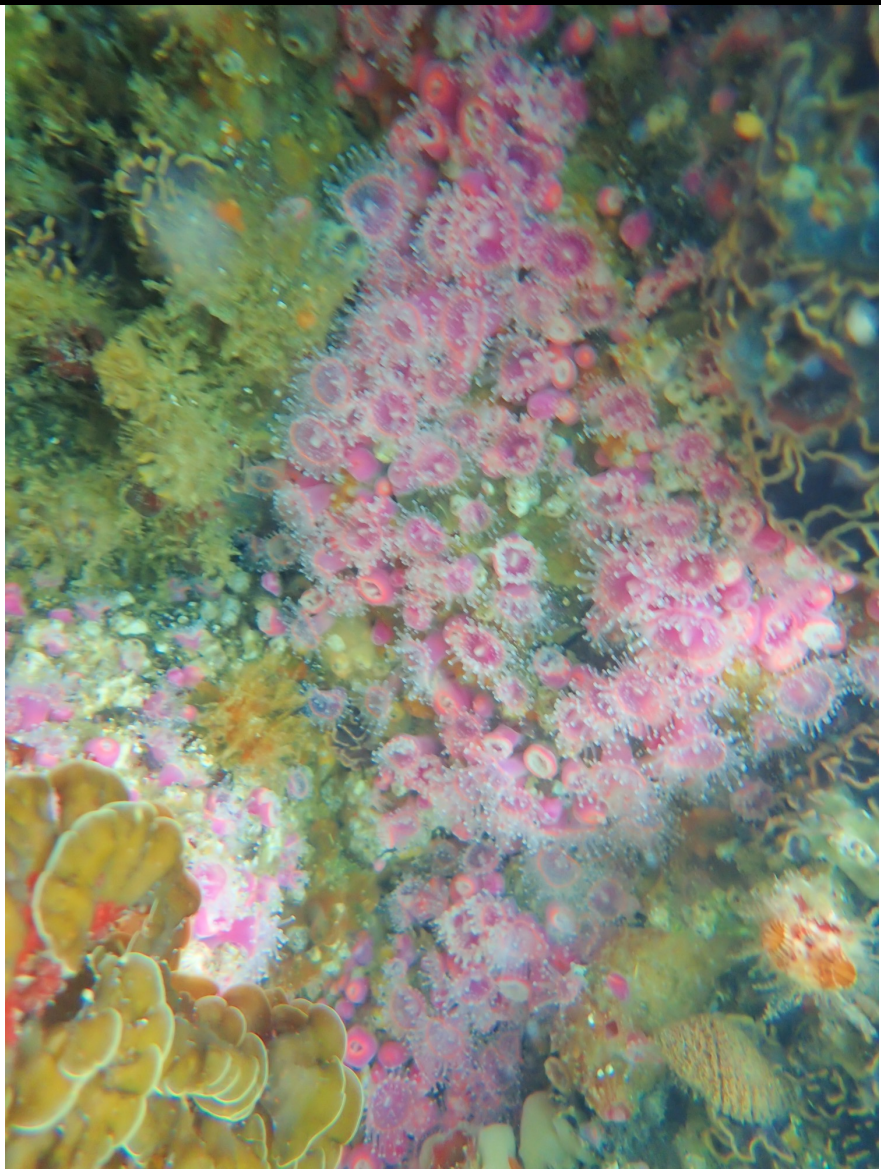








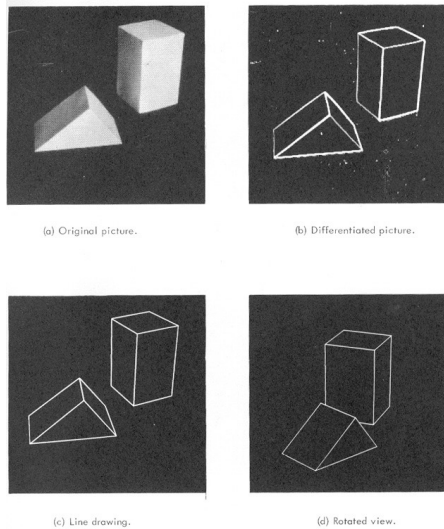




Outline

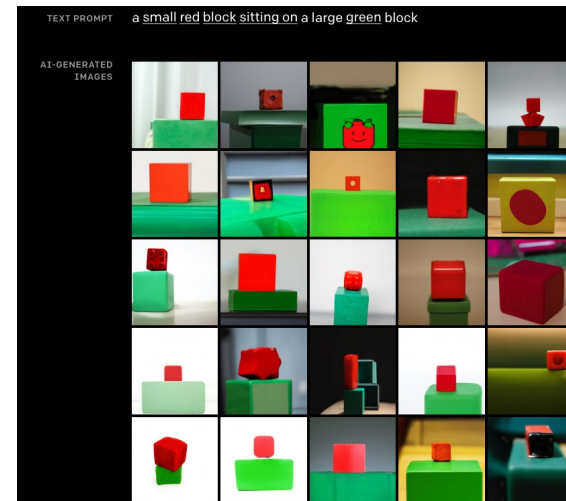
- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision

How it started



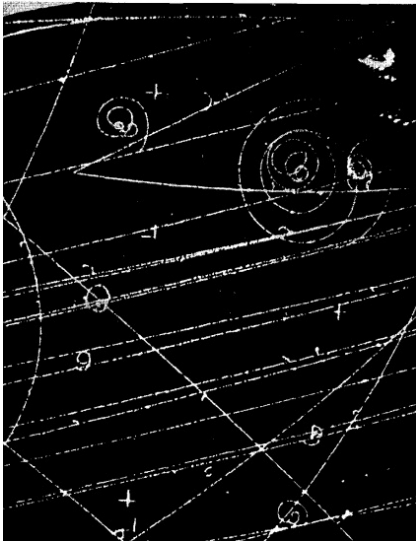
[L. G. Roberts](#), 1963

How it's going

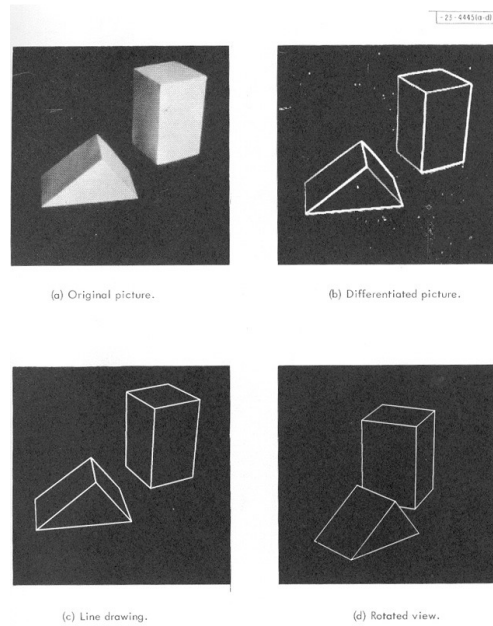


[OpenAI DALL-E](#), 2020

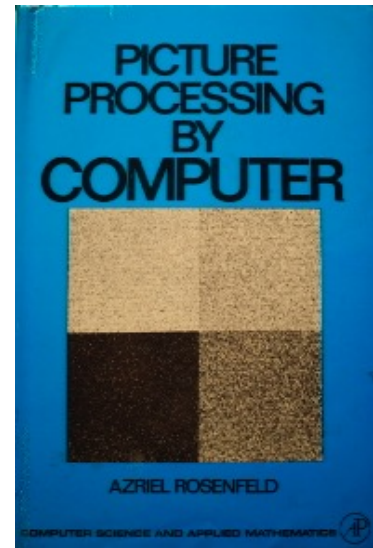
Origins



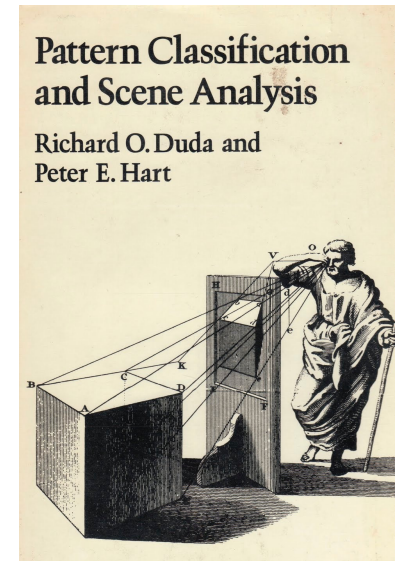
[Hough, 1959](#)



[Roberts, 1963](#)



Rosenfeld, 1969

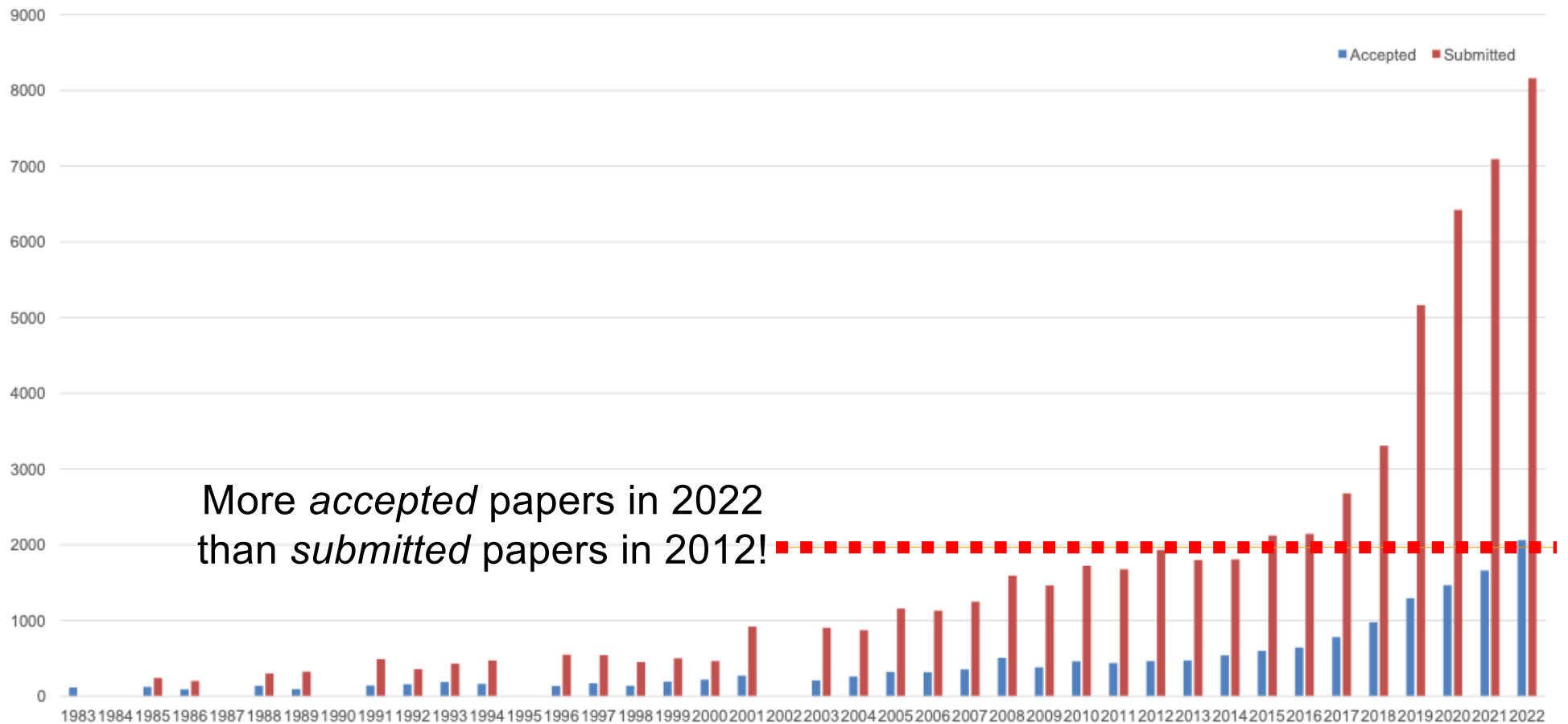


Duda & Hart, 1972

Decade by decade

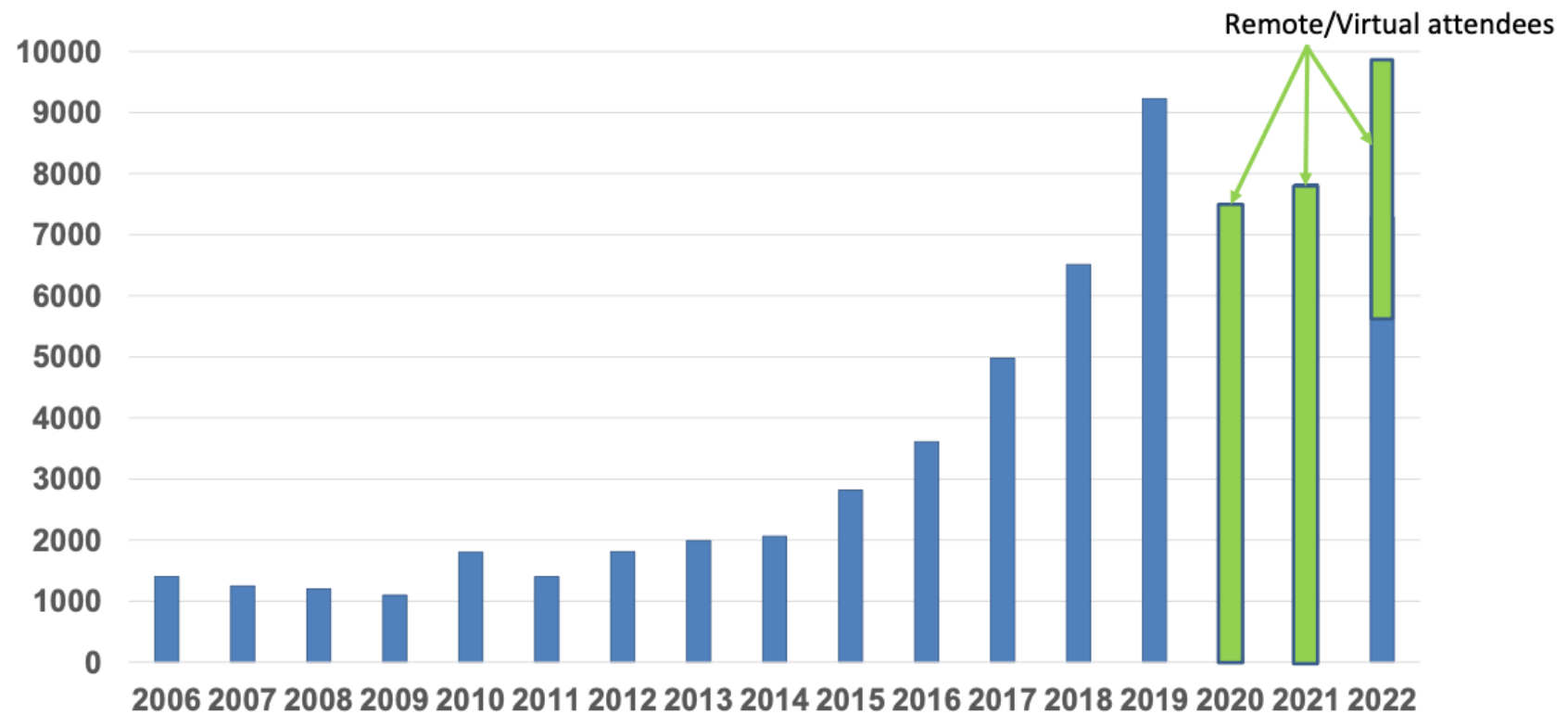
- **1960s:** Blocks world, image processing and pattern recognition
- **1970s:** Key recovery problems defined: structure from motion, stereo, shape from shading, color constancy. Attempts at knowledge-based recognition
- **1980s:** Fundamental and essential matrix, multi-scale analysis, corner and edge 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 Lazechnik's [historical overview](#)

Growth of the field: CVPR papers



Source: [CVPR 2022 opening sides](#)

Growth of the field: CVPR attendance

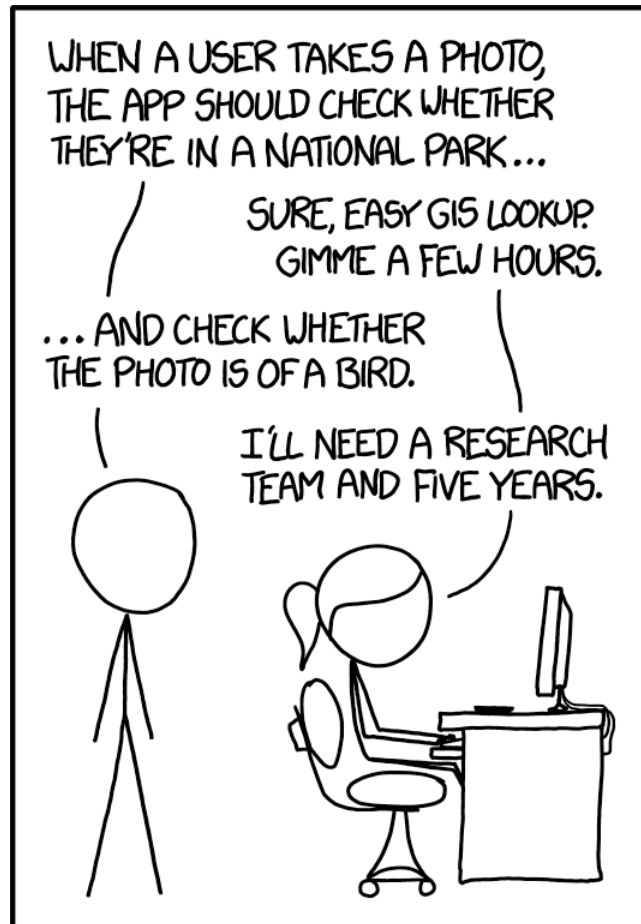


Source: [CVPR 2022 opening sides](#)

Introduction: Outline

- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision
- Current(ish) state of the art

What can computer vision do today?



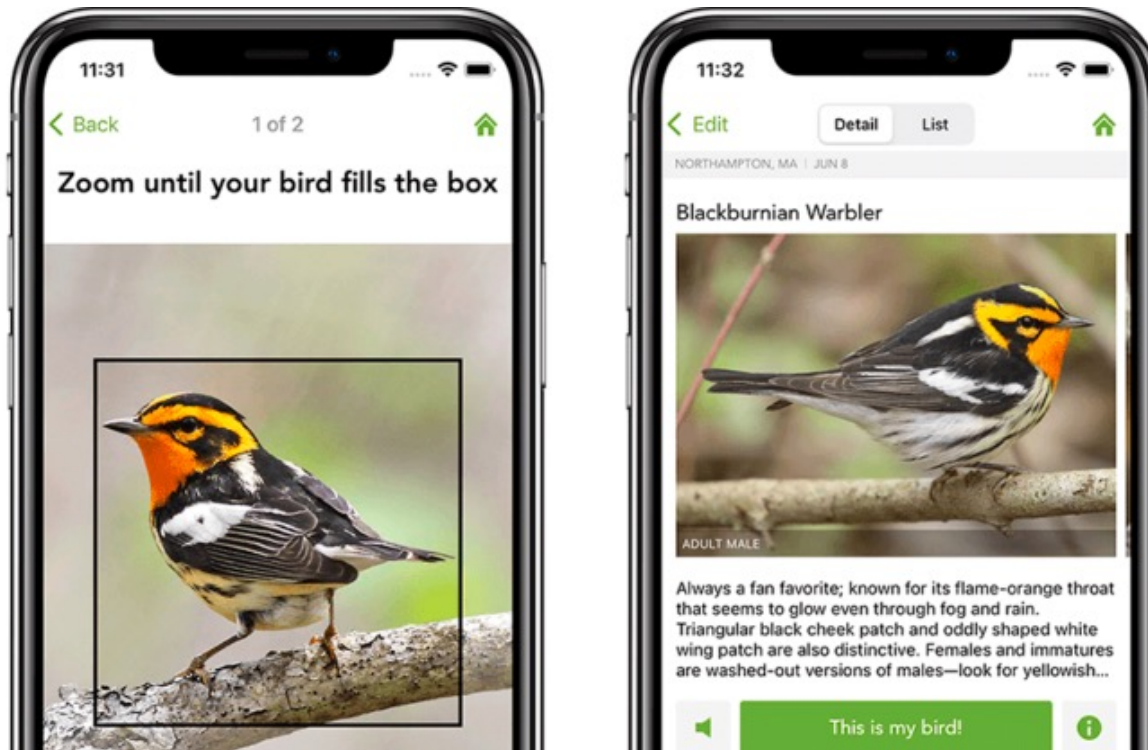
In the 60s, Marvin Minsky assigned a couple of undergrads to spend the summer programming a computer to use a camera to identify objects in a scene. He figured they'd have the problem solved by the end of the summer. Half a century later, we're still working on it.

<https://xkcd.com/1425/>

(September 24, 2014)

What can computer vision do today?

- It's 2025 now...



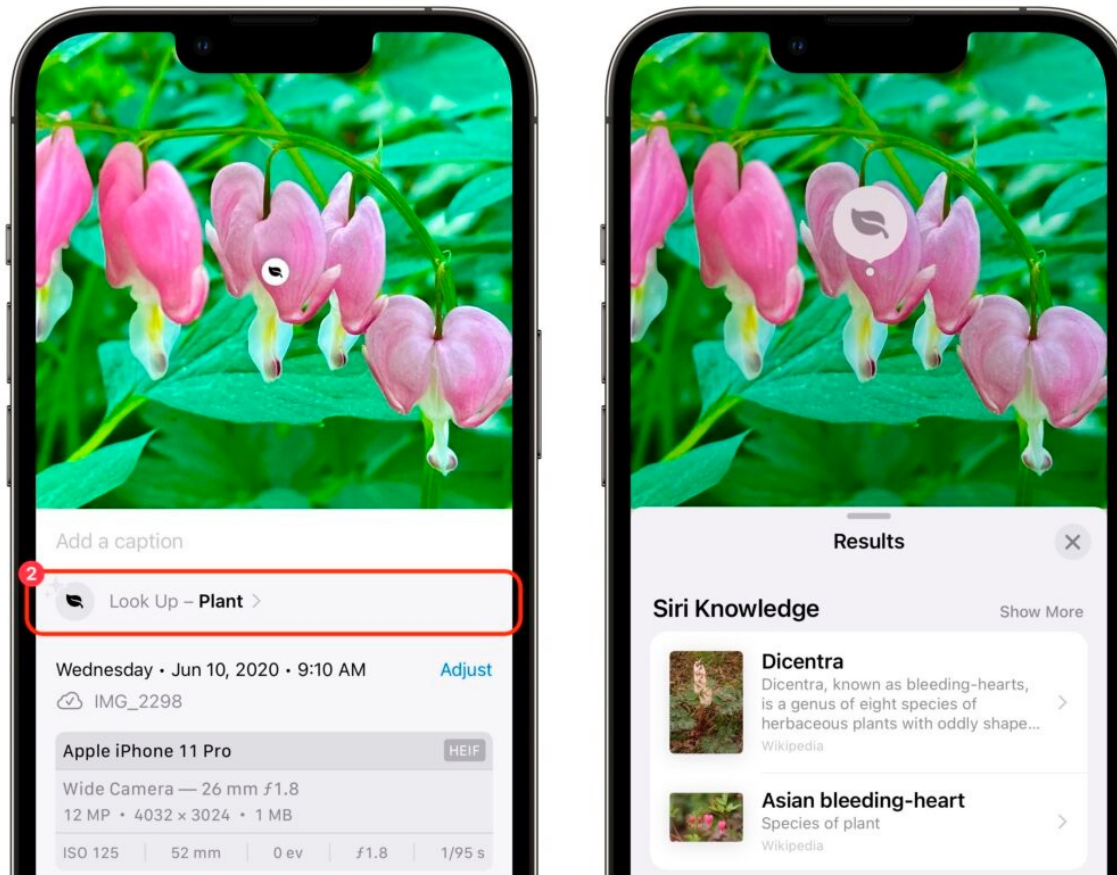
TheCornellLab 

Merlin[®]

<https://merlin.allaboutbirds.org/>

What can computer vision do today?

- It's 2025 now...



[Image source](#)

What can computer vision do today?

- Reconstruction
- Recognition
- *Reconstruction meets recognition, or 3D scene understanding*
- *Image generation*
- *Vision for action*

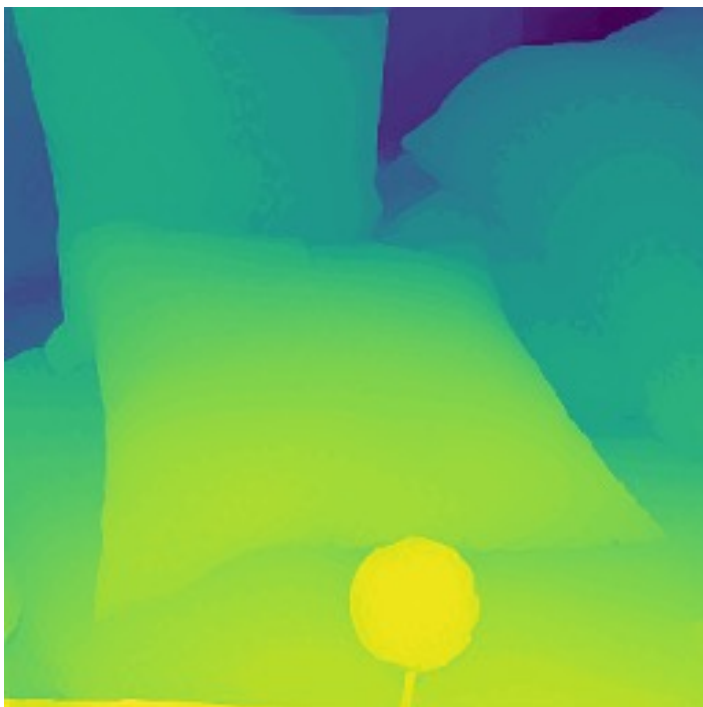
Regression

- We must make image-like things from images
- Examples:
 - depth map from image
 - normal map from image
 - derained image from rainy image
 - defogged image from foggy image
- Train with pairs (image, depth)
 - or (image, normal), etc
 - Loss
 - Squared error +abs value of error+other terms as required

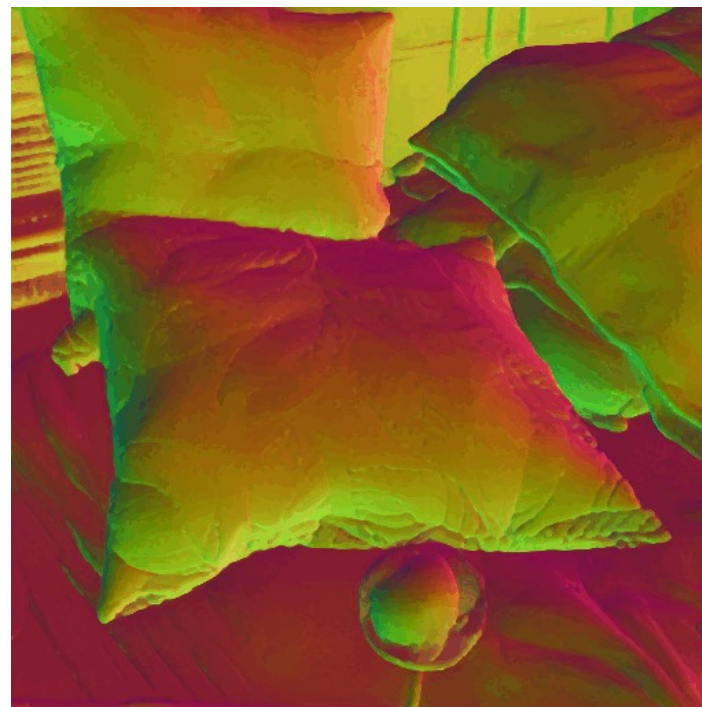
E



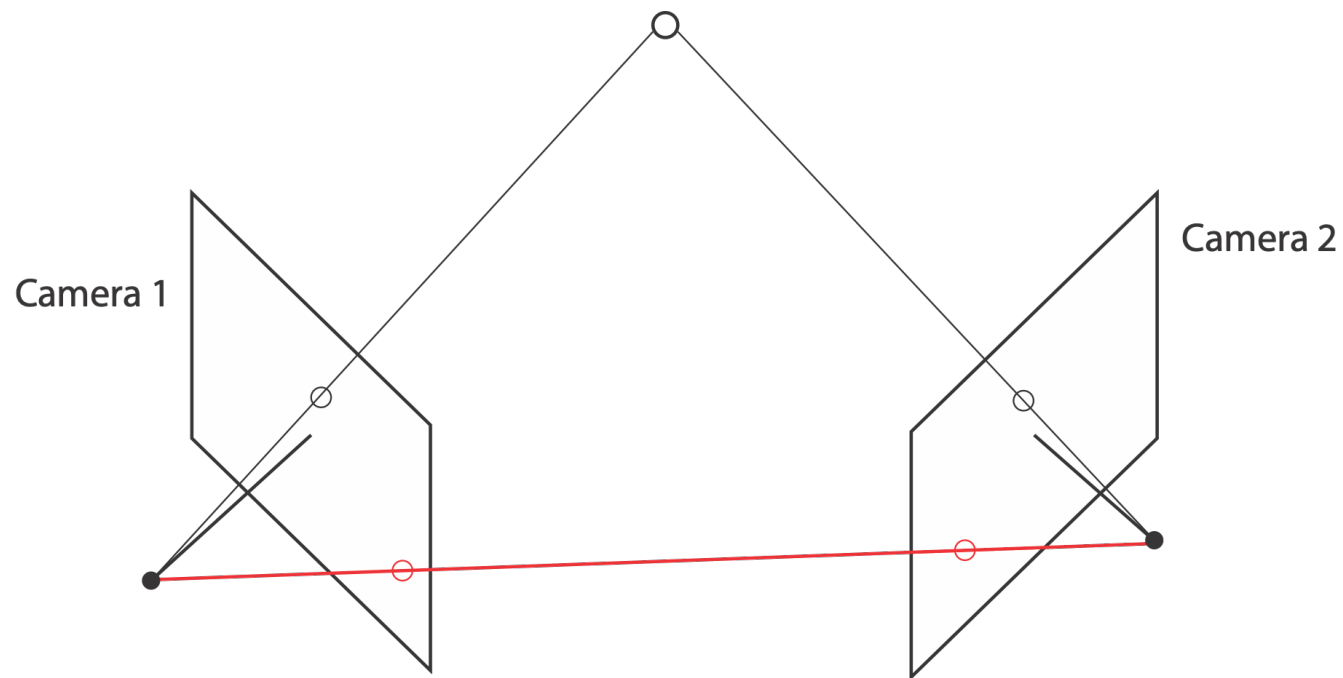
Depth (omnimap, current best depth est)



Normal (omnimap, current best normal est)



Correspondence yields 3D configuration



Reconstruction: 3D from photo collections

Colosseum, Rome, Italy



San Marco Square, Venice, Italy



Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. Seitz, [The Visual Turing Test for Scene Reconstruction](#), 3DV 2013

[YouTube Video](#)

Reconstruction: 4D from photo collections

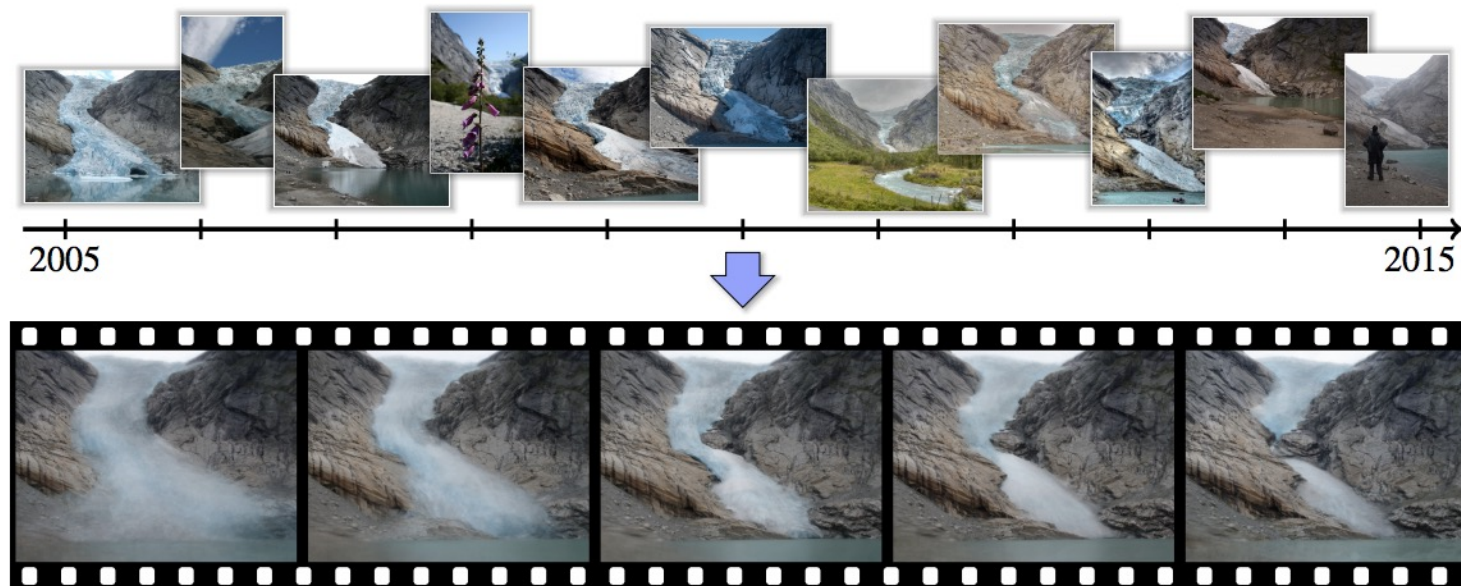


Figure 1: We mine Internet photo collections to generate time-lapse videos of locations all over the world. Our time-lapses visualize a multitude of changes, like the retreat of the Briksdalsbreen Glacier in Norway shown above. The continuous time-lapse (bottom) is computed from hundreds of Internet photos (samples on top). Photo credits: Aliento Más Allá, jirihndek, mcxurxo, elka.cz, Juan Jesús Orío, Klaus Wifßkirchen, Daikrieg, Free the image, dration and Nadav Tobias.

R. Martin-Brualla, D. Gallup, and S. Seitz, [Time-Lapse Mining from Internet Photos](#), SIGGRAPH 2015

[YouTube Video](#)

Reconstruction: 4D from depth cameras



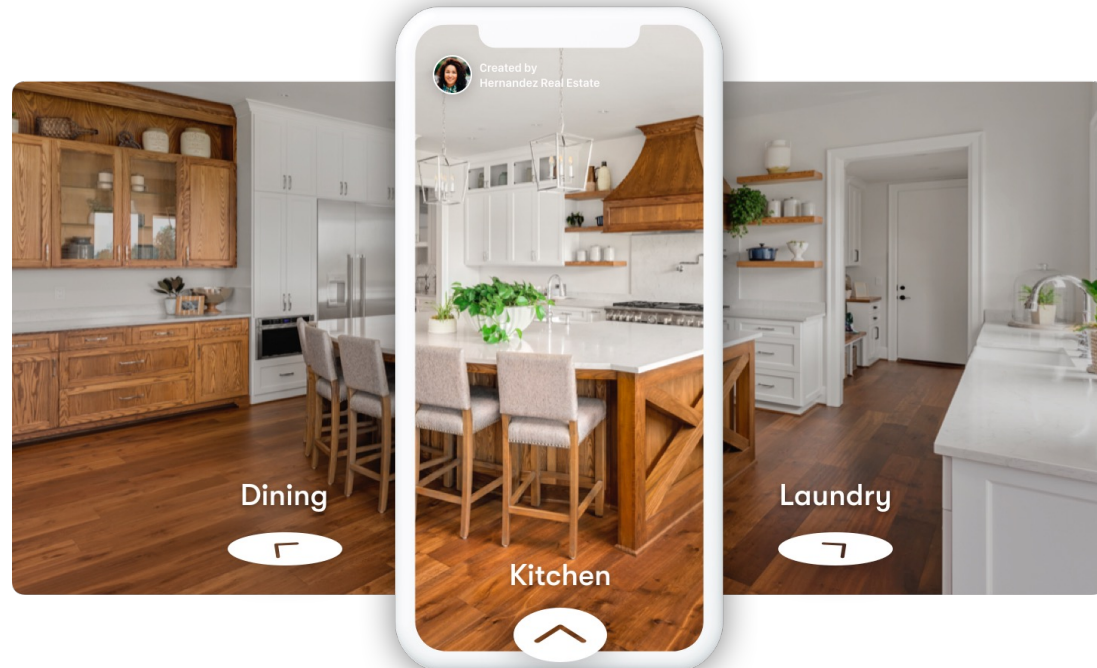
Figure 1: Real-time reconstructions of a moving scene with DynamicFusion; both the person and the camera are moving. The initially noisy and incomplete model is progressively denoised and completed over time (left to right).

R. Newcombe, D. Fox, and S. Seitz, [DynamicFusion:
Reconstruction and Tracking of Non-rigid Scenes in Real-Time](#),
CVPR 2015

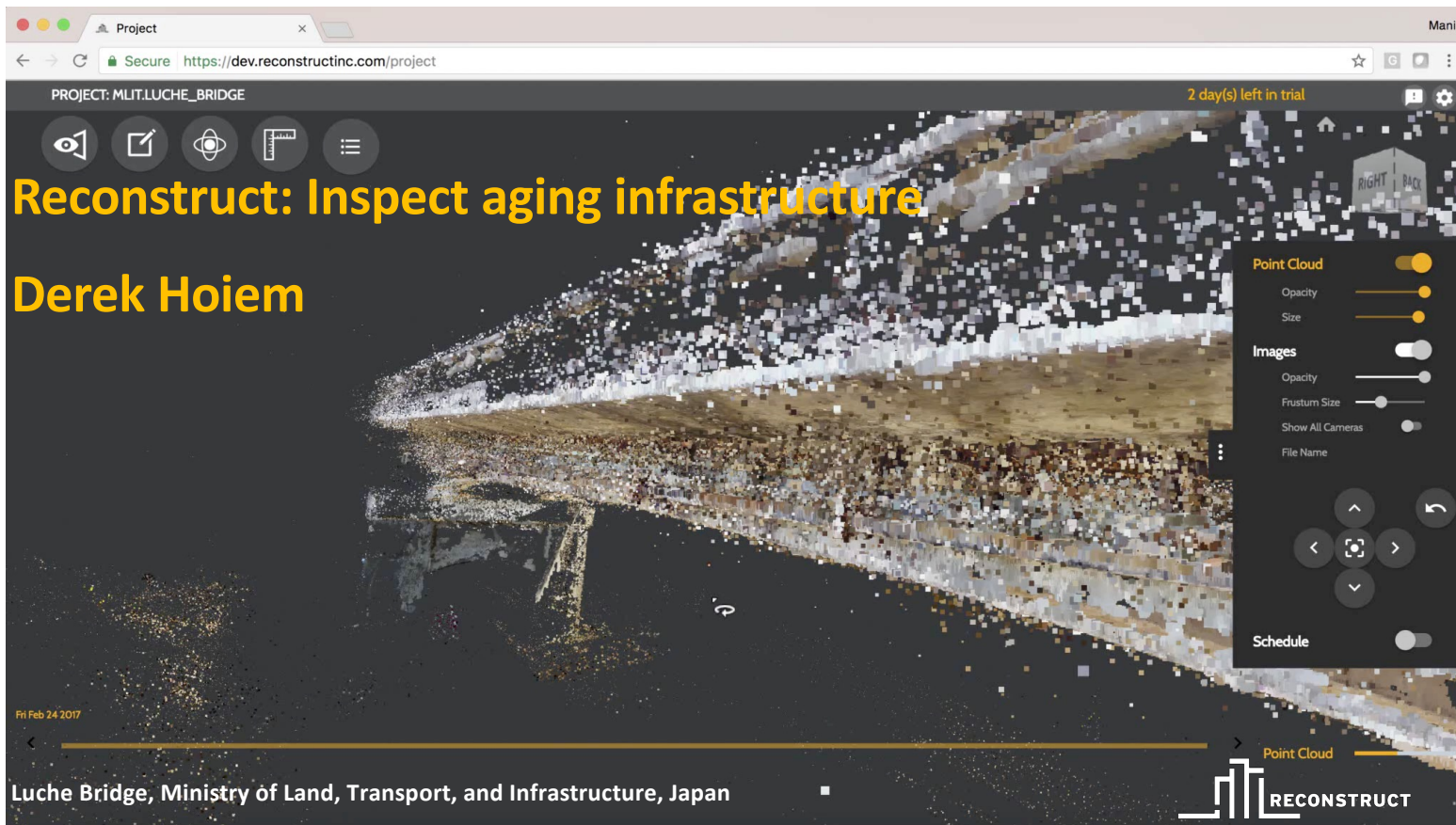
[YouTube Video](#)

Reconstruction: Commercial applications

**Make your listing pop with Zillow
3D Home® tours**



<https://www.zillow.com/z/3d-home/>





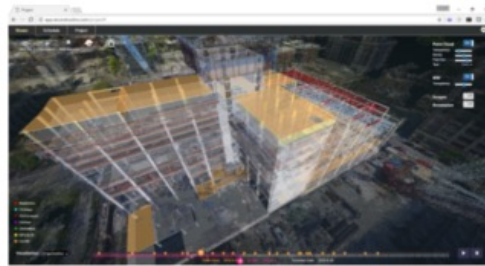
Reconstruction: Commercial applications

RECONSTRUCT INTEGRATES REALITY AND PLAN



Visual Asset Management

Reconstruct 4D point clouds and organize images and videos from smartphones, time-lapse cameras, and drones around the project schedule. View, annotate, and share anywhere with a web interface.



4D Visual Production Models

Integrate 4D point clouds with 4D BIM, review "who does what work at what location" on a daily basis and improve coordination and communication among project teams.



Predictive Visual Data Analytics

Analyze actual progress deviations by comparing Reality and Plan and predict risk with respect to the execution of the look-ahead schedule for each project location, to offer your project team with an opportunity to tap off potential delays before they surface on your jobsite.

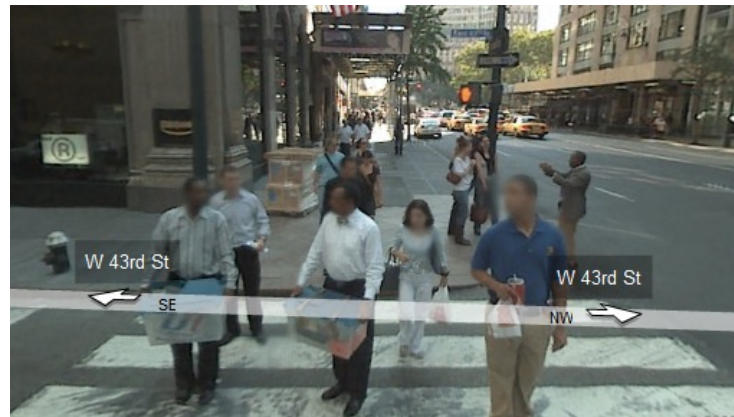
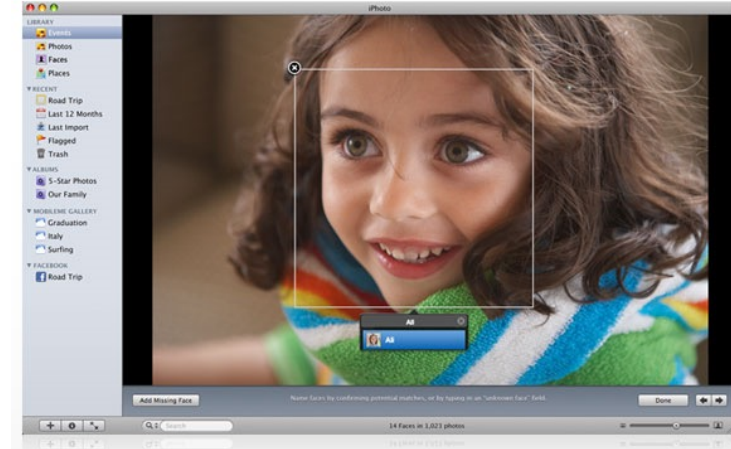
reconstructinc.com

Source: D. Hoiem

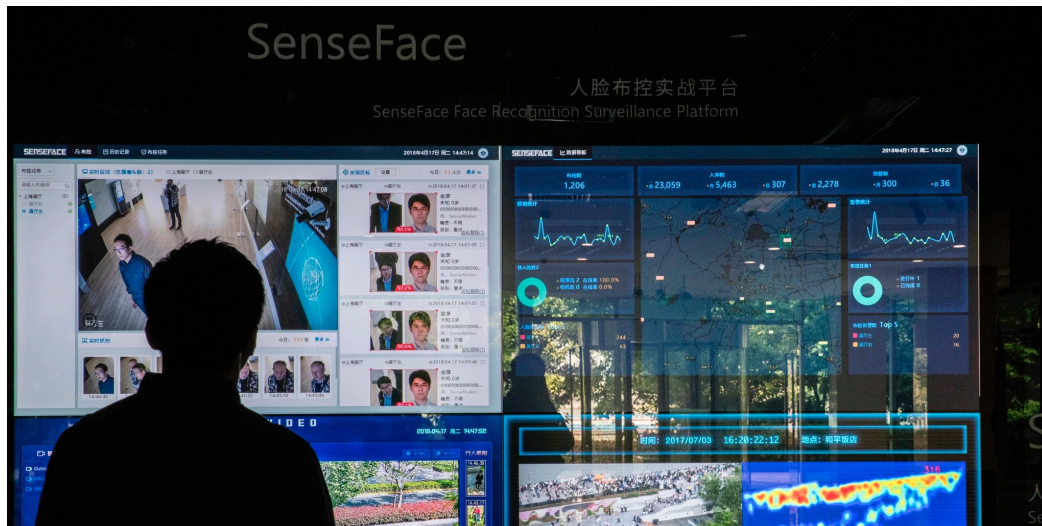
Recognition: “Simple” patterns



Recognition: Faces



Recognition: Faces



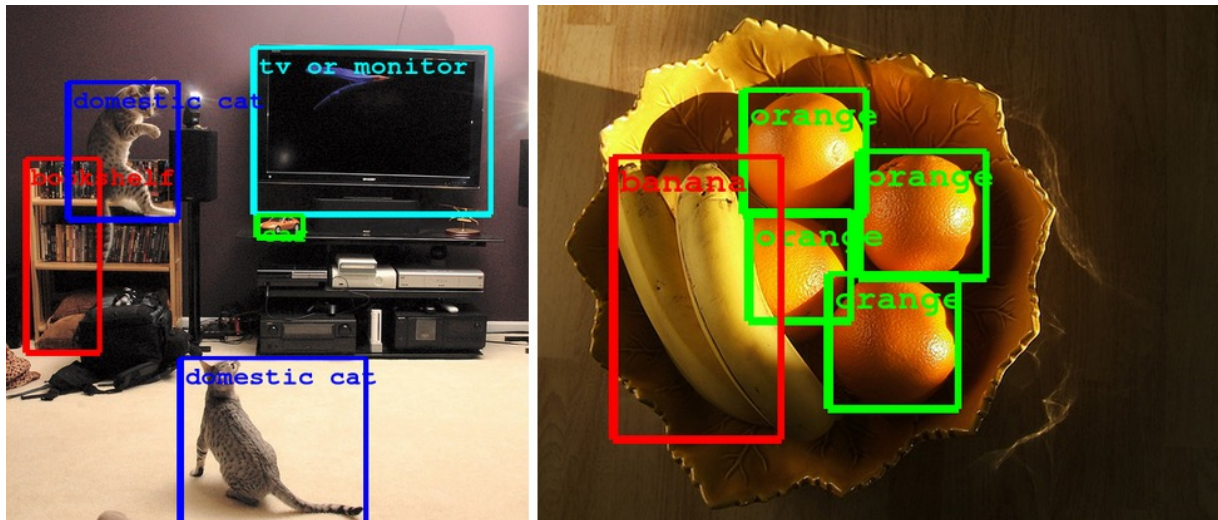
[How China Uses High-Tech Surveillance to Subdue Minorities](#) – New York Times, 5/22/2019

[The Secretive Company That Might End Privacy As We Know It](#) – New York Times, 1/18/2020

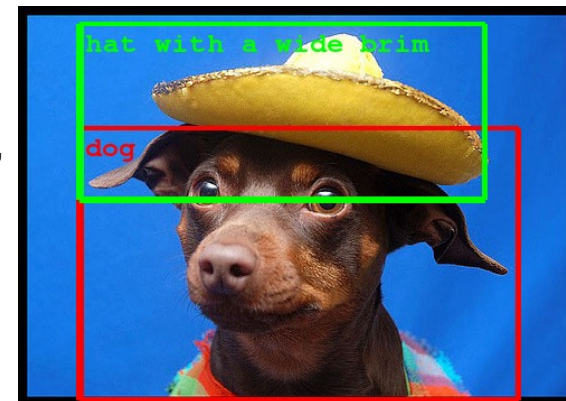
[Wrongfully Accused by an Algorithm](#) – New York Times, 6/24/2020

[Facial Recognition Goes to War](#) – New York Times, 4/7/2022

Recognition: General categories



- [Computer Eyesight Gets a Lot More Accurate](#), NY Times Bits blog, August 18, 2014
- [Building A Deeper Understanding of Images](#), Google Research Blog, September 5, 2014

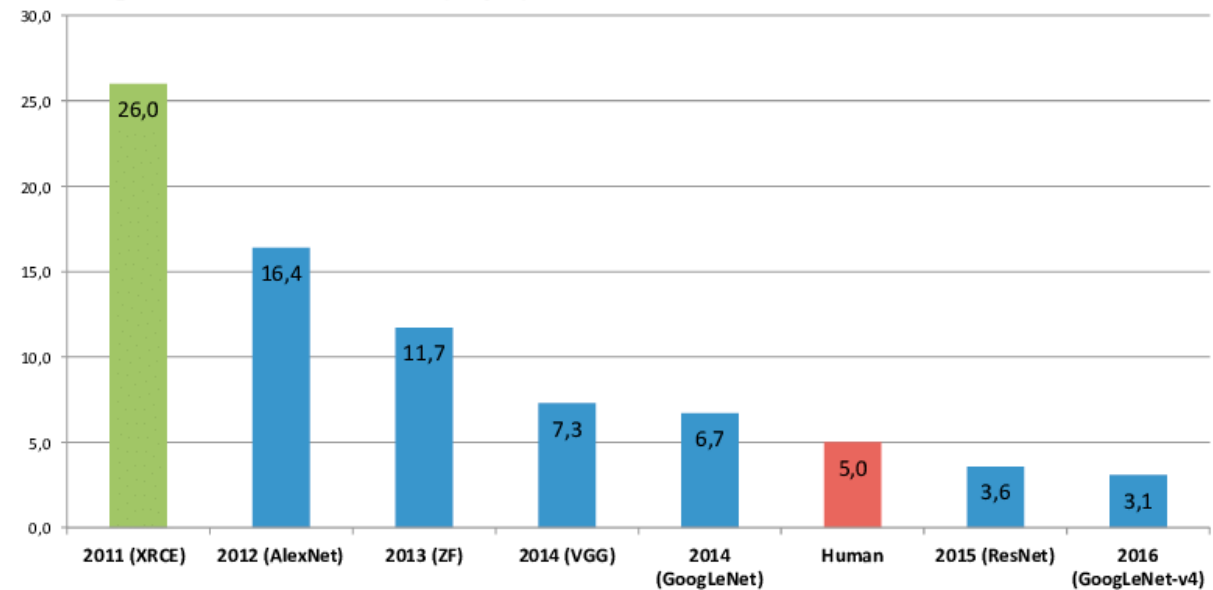


Recognition: General categories

ILSVRC

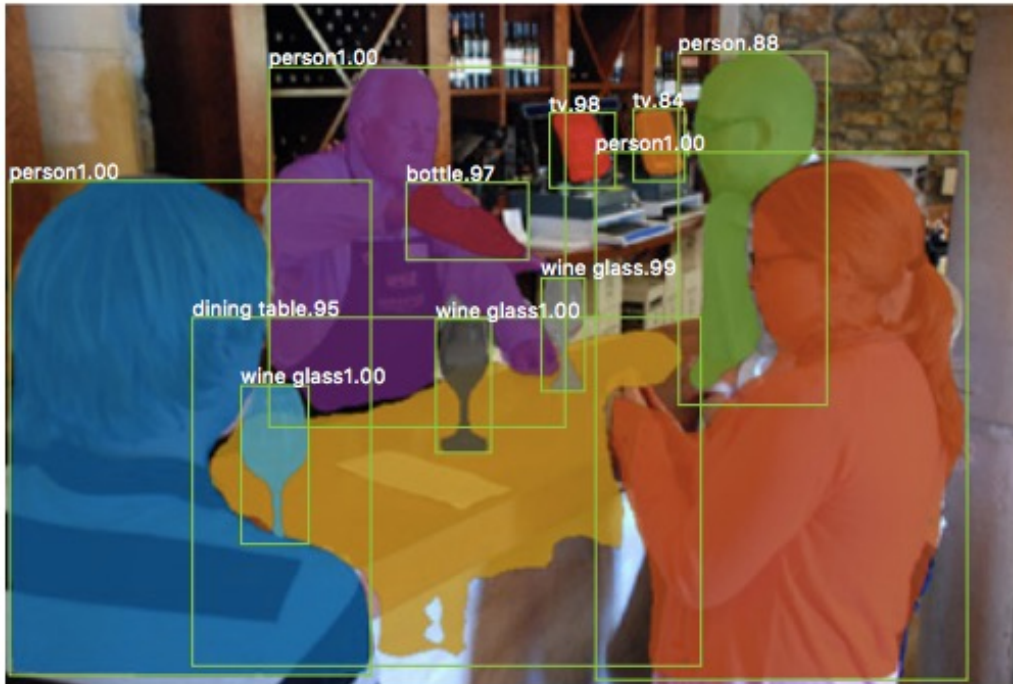


ImageNet Classification Error (Top 5)



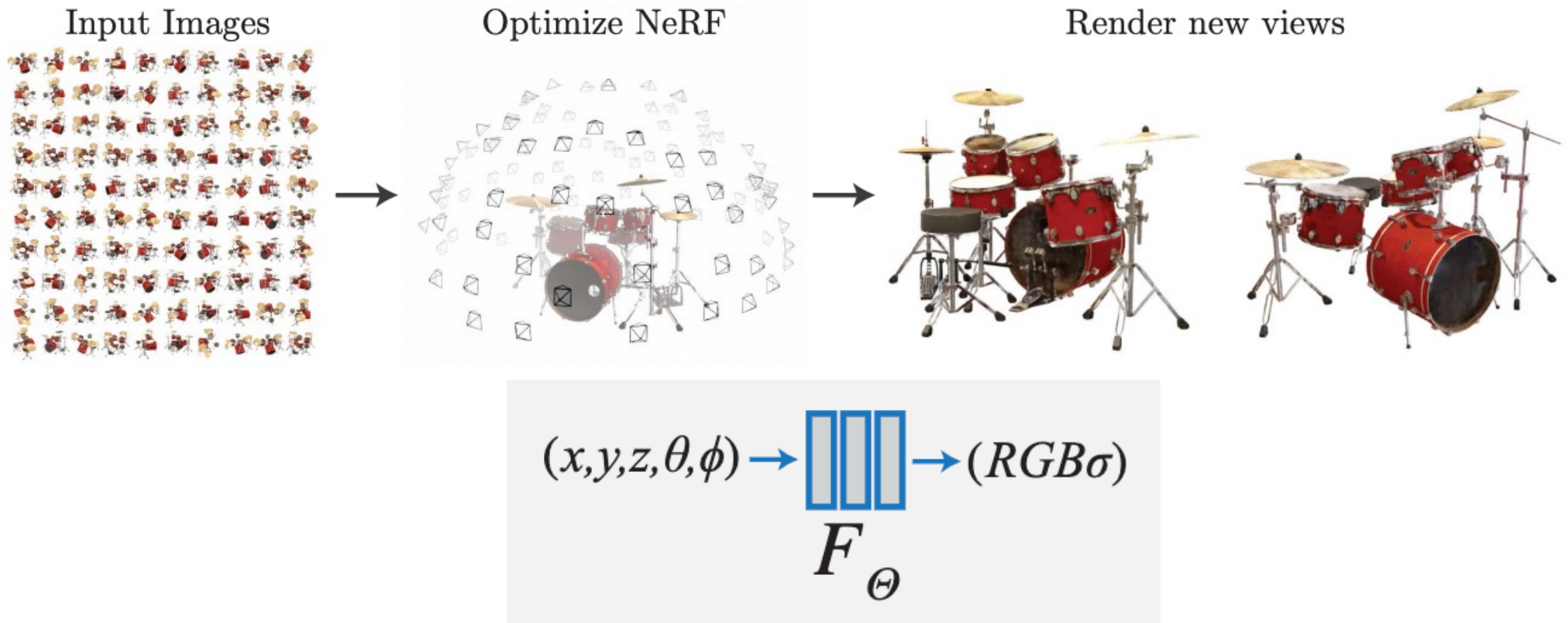
[Figure source](#)

Object detection, instance segmentation



K. He, G. Gkioxari, P. Dollar, and R. Girshick, [Mask R-CNN](#),
ICCV 2017 (Best Paper Award)

3D scene understanding: NERFs



B. Mildenhall et al., [Representing Scenes as Neural Radiance Fields for View Synthesis](#), ECCV 2020

3D scene understanding: NERFs



B. Mildenhall et al., [Representing Scenes as Neural Radiance Fields for View Synthesis](#), ECCV 2020

3D scene understanding: Single-view reconstruction

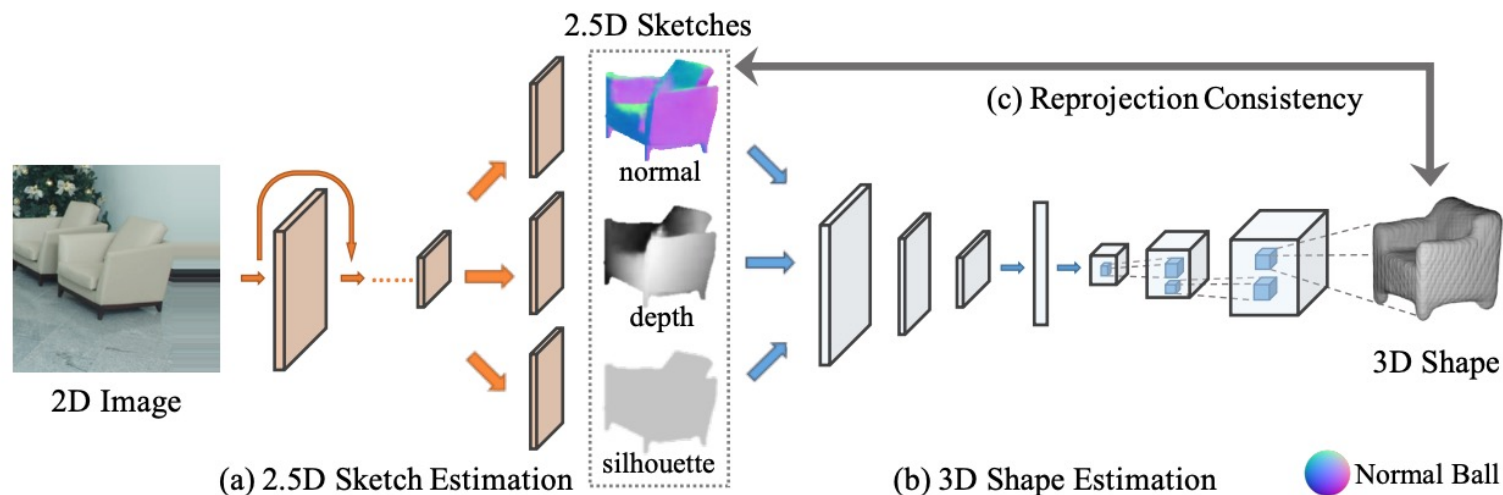


Figure 2: Our model (MarrNet) has three major components: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) a loss function for reprojection consistency. MarrNet first recovers object normal, depth, and silhouette images from an RGB image. It then regresses the 3D shape from the 2.5D sketches. In both steps, it uses an encoding-decoding network. It finally employs a reprojection consistency loss to ensure the estimated 3D shape aligns with the 2.5D sketches. The entire framework can be trained end-to-end.

Image generation: Faces

- 1024x1024 resolution, CelebA-HQ dataset



T. Karras, T. Aila, S. Laine, and J. Lehtinen, [Progressive Growing of GANs for Improved Quality, Stability, and Variation](#), ICLR 2018

[Follow-up work](#)

Image generation: DeepFakes

Harrison Ford Is Young Han In Solo Deepfake Video

Thanks to deepfake technology, the maligned Solo: A Star Wars Story now stars Harrison Ford instead of Alden Ehrenreich as the young Han.

BY DAN ZINSKI
2 DAYS AGO

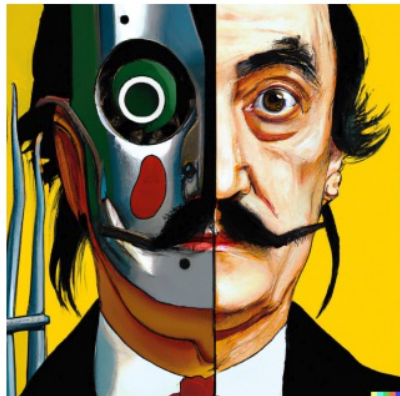


Just a random recent example...

<https://screenrant.com/star-wars-han-solo-movie-harrison-ford-video-deepfake/>
<https://www.youtube.com/watch?v=bC3uH4Xw4Xo>

<https://en.wikipedia.org/wiki/Deepfake>

Image generation: OpenAI DALL-E, DALL-E 2



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation

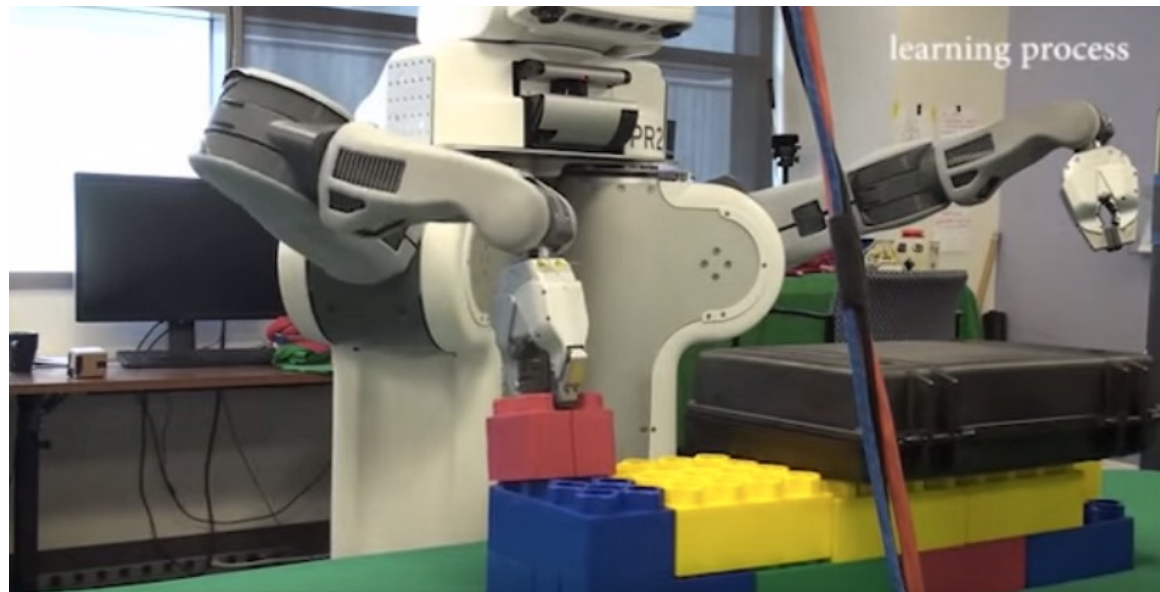


a corgi's head depicted as an explosion of a nebula

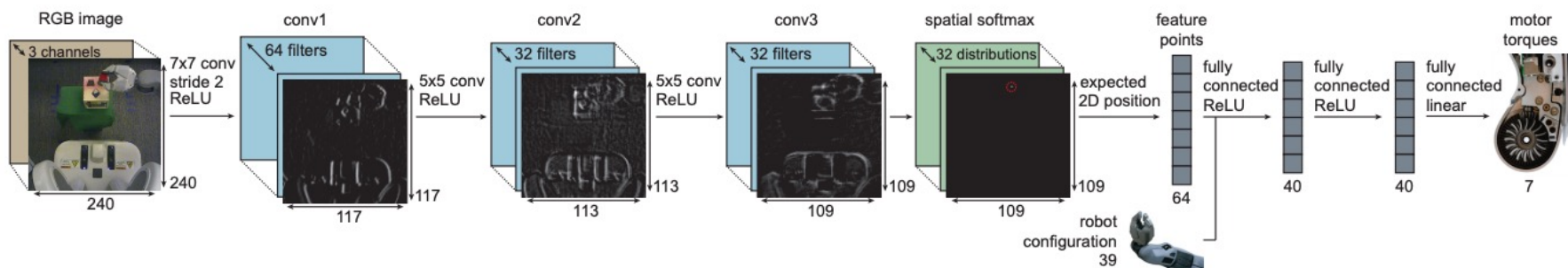
A. Ramesh et al., [Zero-Shot Text-to-Image Generation](https://openai.com/blog/dall-e/), ICML 2021. <https://openai.com/blog/dall-e/>

A. Ramesh et al., [Hierarchical Text-Conditional Image Generation with CLIP Latents](https://openai.com/dall-e-2/), arXiv 2022. <https://openai.com/dall-e-2/>

Vision for action: Visuomotor learning

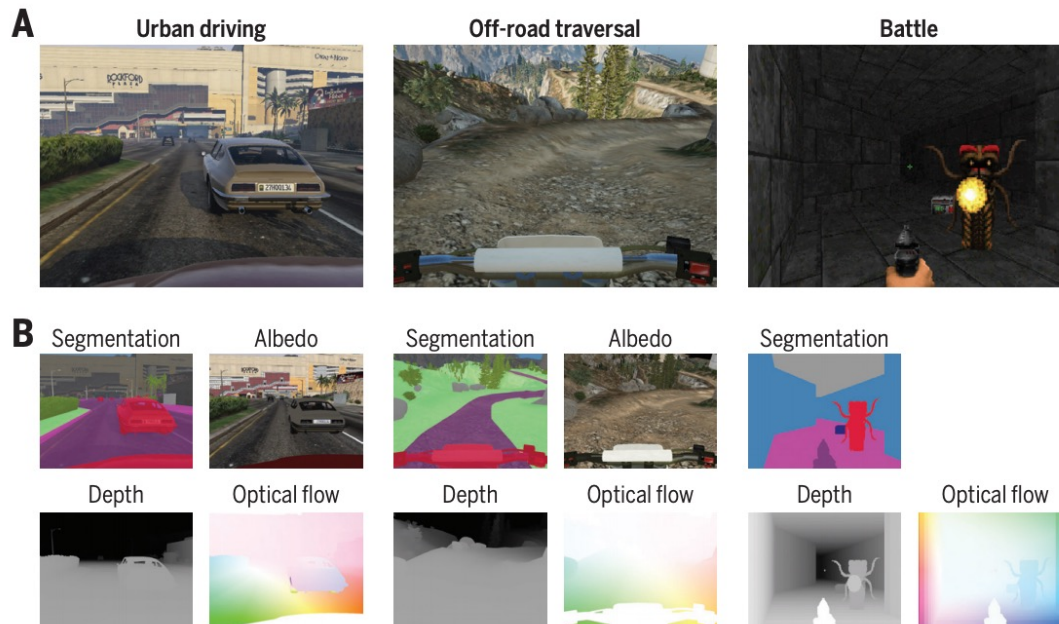


[Overview video](#),
[training video](#)

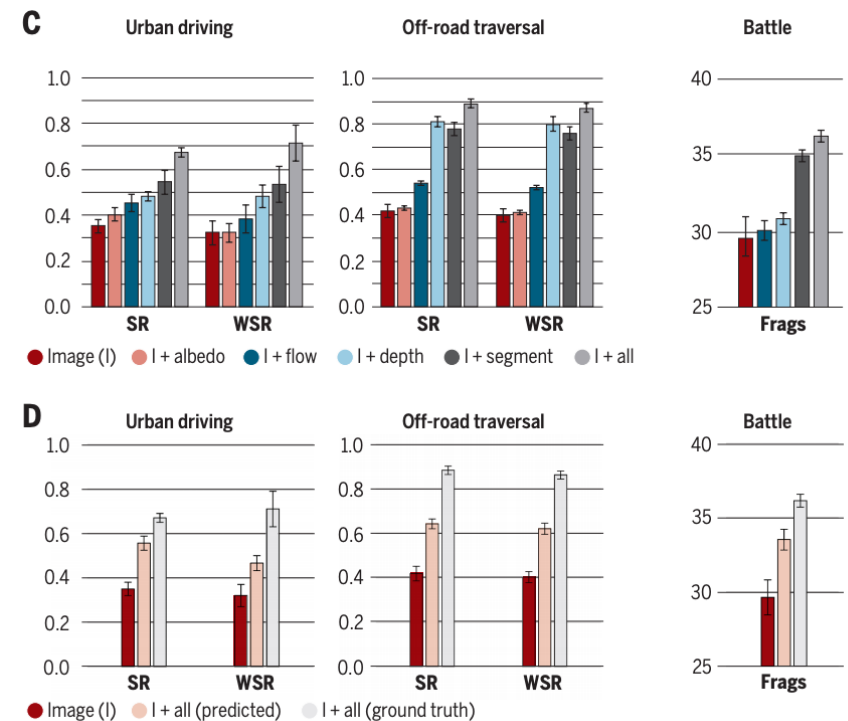


S. Levine, C. Finn, T. Darrell, P. Abbeel, [End-to-end training of deep visuomotor policies](#), JMLR 2016

Does computer vision matter for action?



“Our main finding is that computer vision does matter. Models equipped with intermediate representations train faster, achieve higher task performance, and generalize better to previously unseen environments.”



B. Zhou, P. Krähenbühl, and V. Koltun, [Does Computer Vision Matter for Action?](#) Science Robotics, 4(30), 2019 ([video](#))

Vision for action: Learning skills from video



Fig. 1. Simulated characters performing highly dynamic skills learned by imitating video clips of human demonstrations. **Left:** Humanoid performing cartwheel B on irregular terrain. **Right:** Backflip A retargeted to a simulated Atlas robot.

[Video](#)

X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, S. Levine, [SFV: Reinforcement Learning of Physical Skills from Videos](#), SIGGRAPH Asia 2018

Outline

- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision
- Current state of the art
- **Topics covered in class**

Topics covered in class

I. Elementary Image Representations:

Point transformations, geometric transformations, filters, denoising, edges, interest points

II. Mid-level vision:

Voting, Fitting, Registration

III. Learned Image Representations:

Learned denoising, Mapping images to images, Classification, Detection

IV. Image formation and geometric vision

Cameras, Light, Color, Calibration

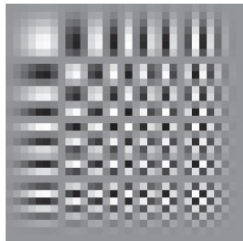
V. Pairs of Cameras and more:

Geometry, Odometry, Optic Flow, Stereopsis, Structure from Motion, Tracking

Elementary image representations



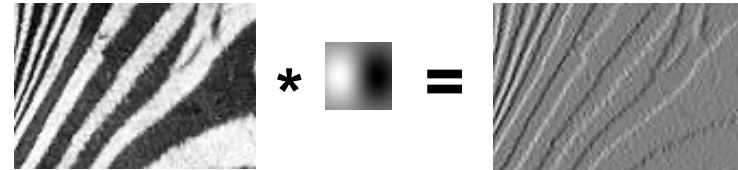
Basic image processing



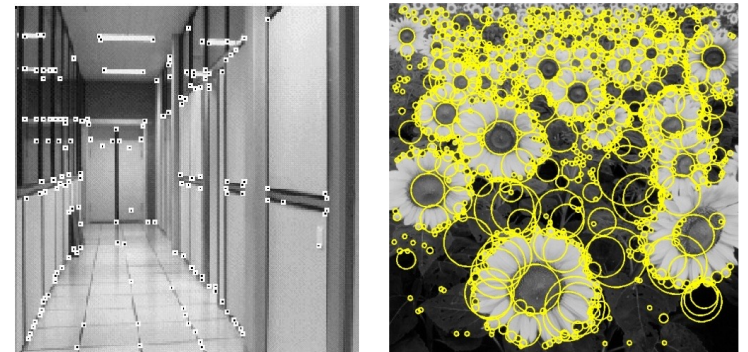
Upsampling and downsampling



downsampled by 8

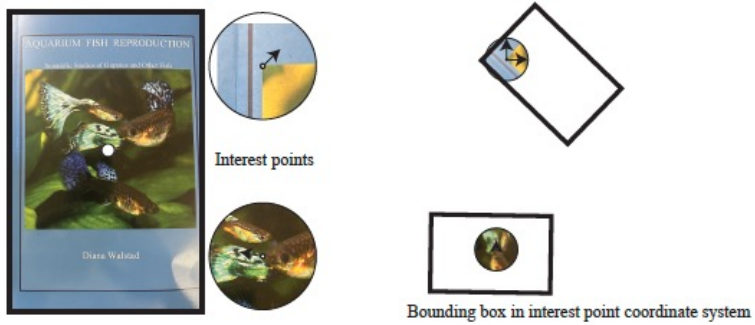


Linear filtering
Edge detection

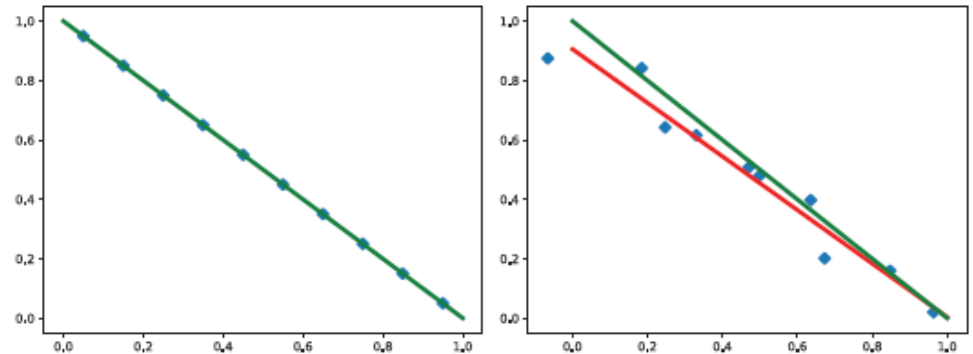


Feature extraction

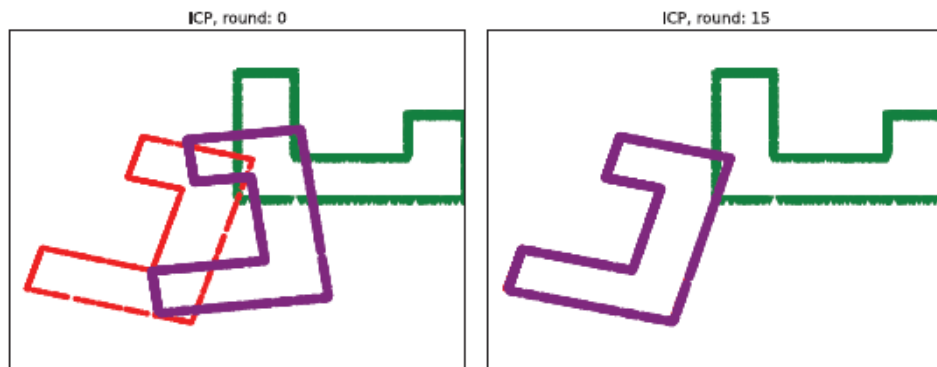
Mid-level Vision



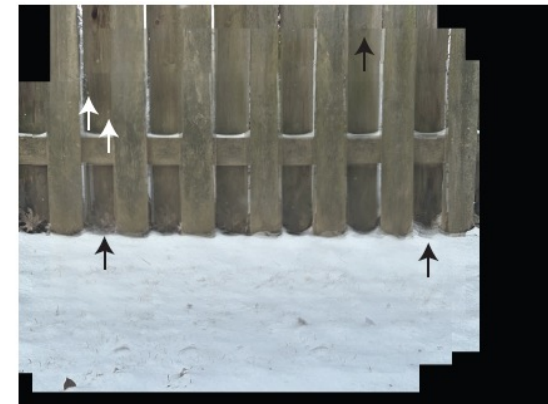
Voting



Fitting



Registration

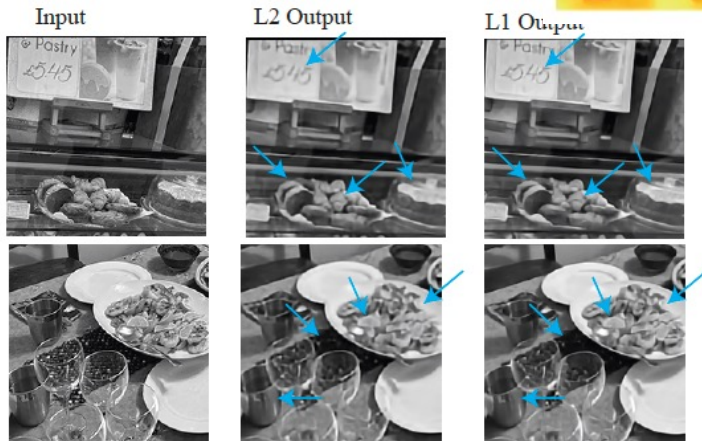


Mosaics

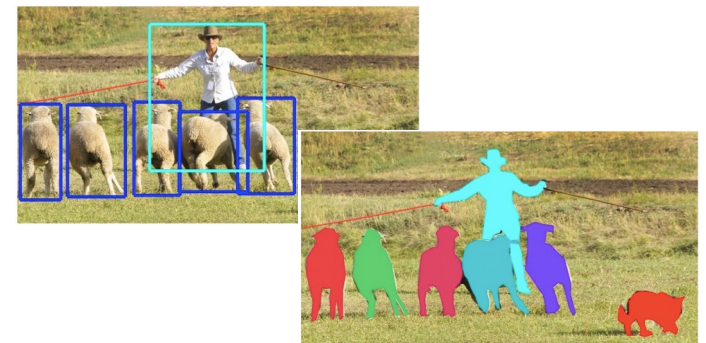
Learned Image Representations



Depth, normal, etc
From images

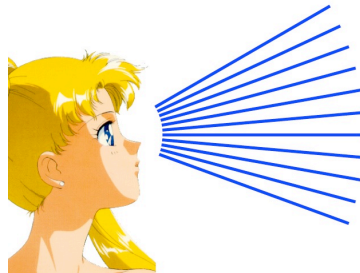


Learned denoising

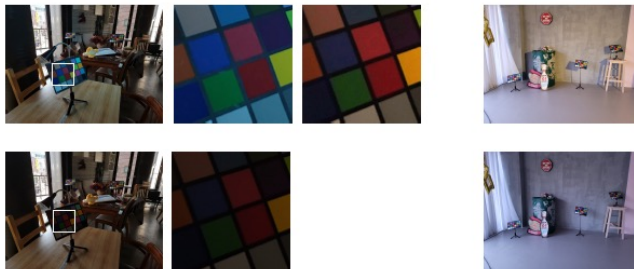
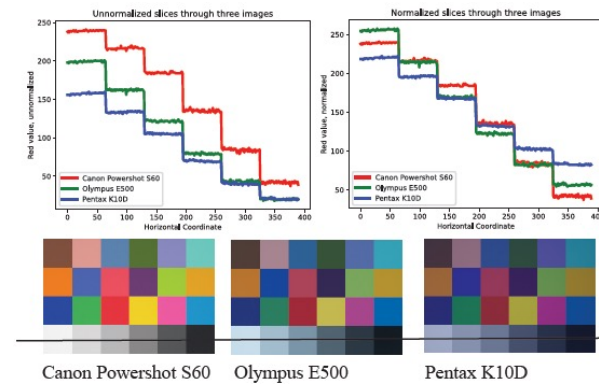


Object detection and segmentation

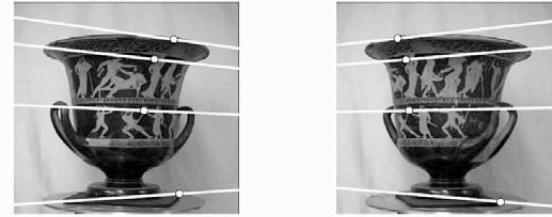
IV. Image formation and geometric vision



Cameras and sensors



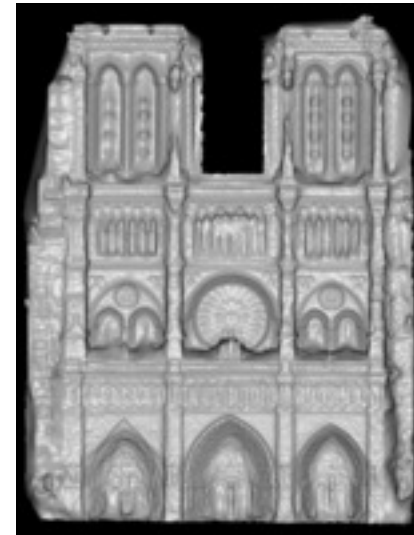
V. Pairs of cameras and more



Two-view geometry, stereo



Structure from motion



Multi-view stereo

Obtaining Images

D.A. Forsyth,
University of Illinois at Urbana Champaign

Some slides adapted from
Svetlana Lazebnik, who adapted from [Alyosha Efros](#), [Derek
Hoiem](#)

Model of imaging

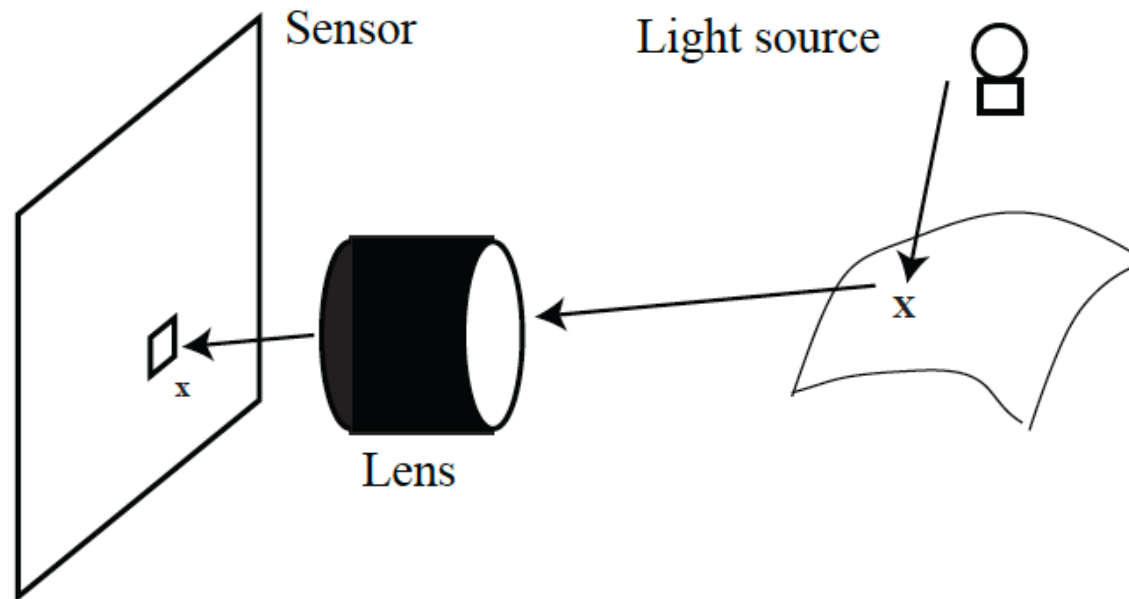


FIGURE 2.1: *A high-level model of imaging. Light leaves light sources and reflects from surfaces. Eventually, some light arrives at a camera and enters a lens system. Some of that light arrives at a photosensor inside the camera.*

A Pinhole Camera

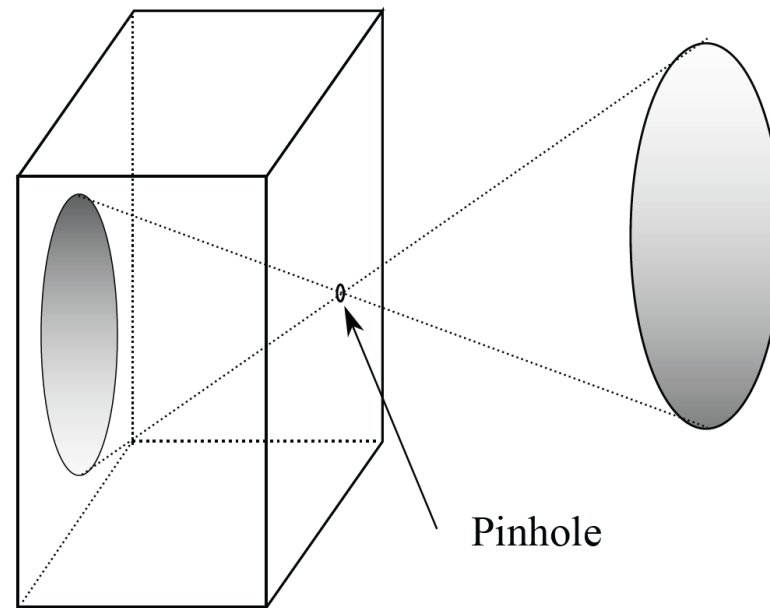


FIGURE 2.3: *In the pinhole imaging model, a light-tight box with a pinhole in it views an object. The only light that a point on the back of the box sees comes through the very small pinhole, so that an inverted image is formed on the back face of the box.*

The Pinhole Camera

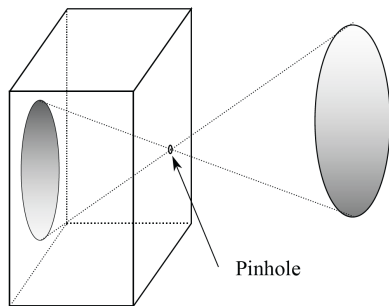


FIGURE 2.3: In the pinhole imaging model, a light-tight box with a pinhole in it views an object. The only light that a point on the back of the box sees comes through the very small pinhole, so that an inverted image is formed on the back face of the box.

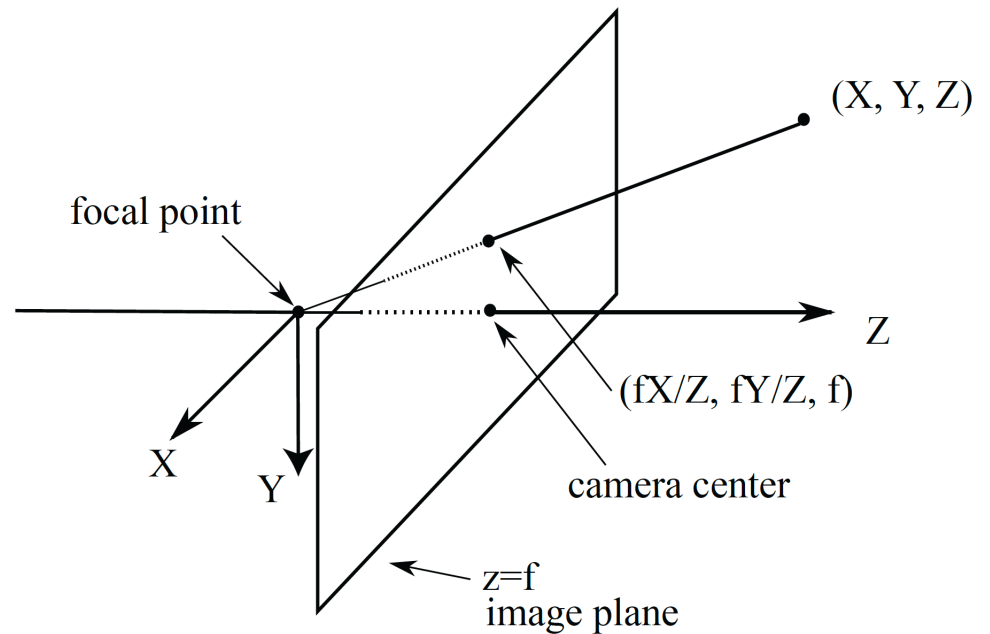


FIGURE 2.4: The usual geometric abstraction of the pinhole model. The box doesn't affect the geometry, and is omitted. The pinhole has been moved to the back of the box, so that the image is no longer inverted. The image is formed on the plane $z = f$, by convention. Notice the y-axis goes down in the image. This allows me to use a right handed coordinate system and also have z increase as one moves into the image.

The lens arranges that light arriving at $\mathbf{x}$ on the sensor all arrived from one point ($\mathbf{X}$ in Figure 2.1) on a surface in 3D, assuming it is in focus. At $\mathbf{x}$, the sensor collects power P for some period Δt , then passes the result on to the camera electronics. The sensor responds to energy $P\Delta t$, so collecting more power for a shorter period or less power for a longer period will result in indistinguishable results. The value of the pixel at i, j on the grid is a *sample* of a function of position (Figure 2.2).

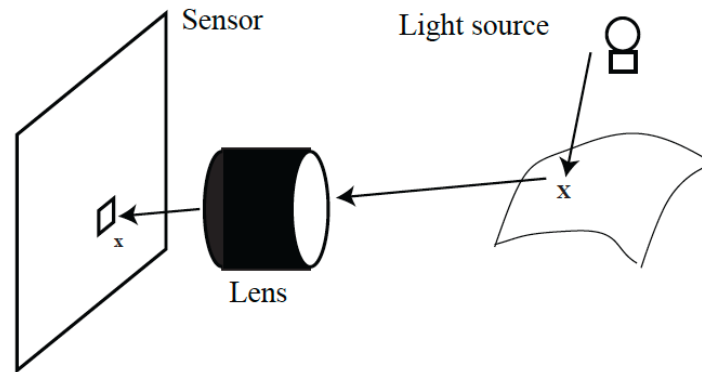
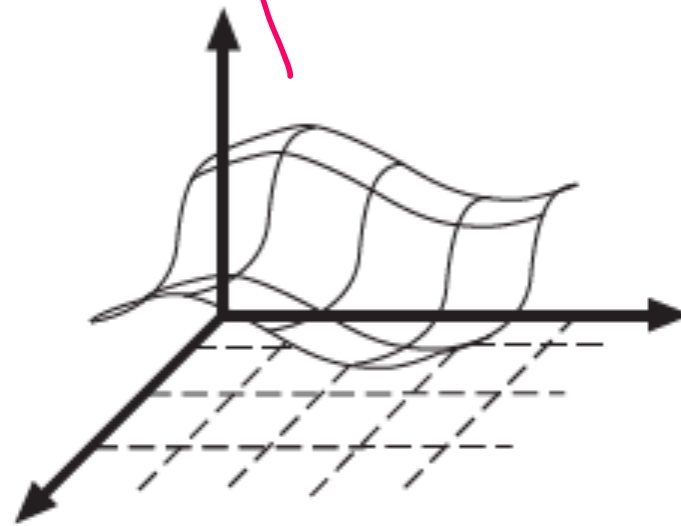
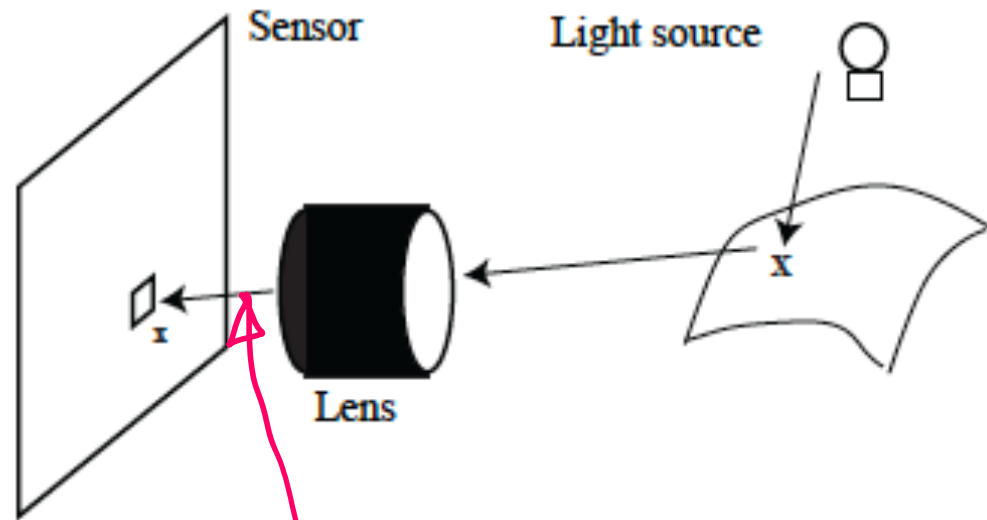


FIGURE 2.1: *A high-level model of imaging. Light leaves light sources and reflects from surfaces. Eventually, some light arrives at a camera and enters a lens system. Some of that light arrives at a photosensor inside the camera.*



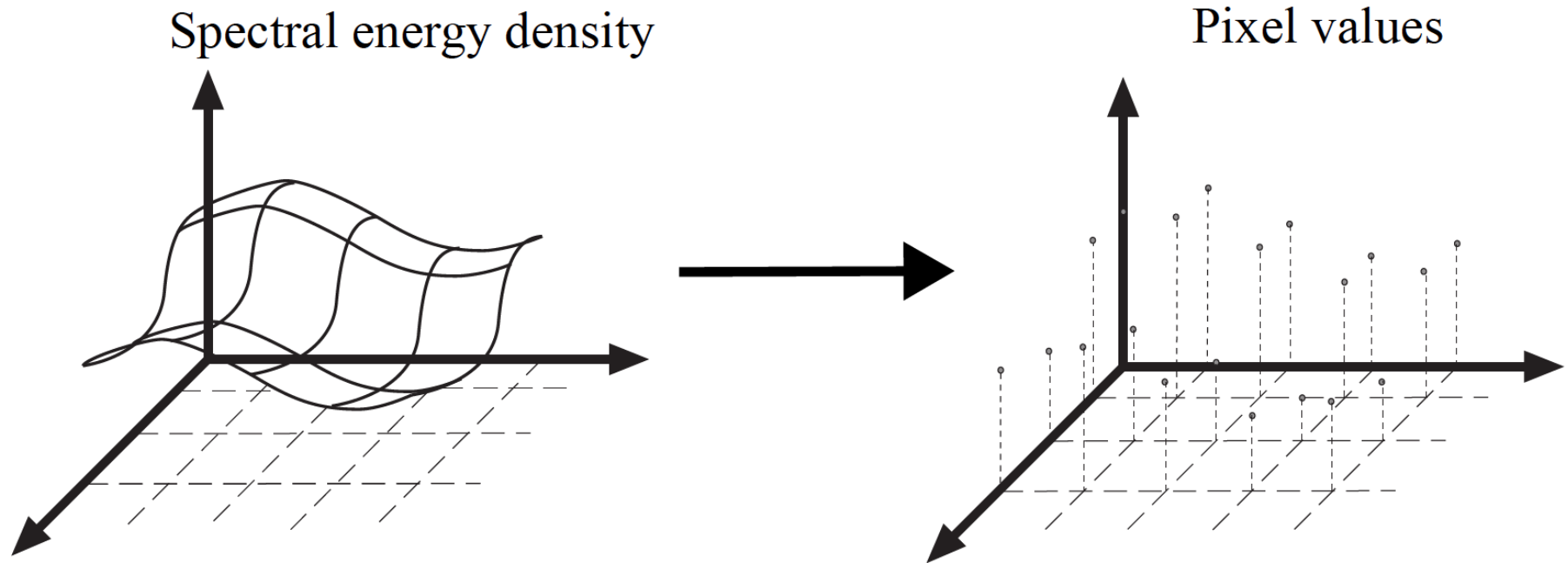


FIGURE 2.2: *Because each pixel in the sensor averages over a small range of directions and positions, the process mapping the input spectral energy distribution to pixel values can be thought of as sampling. On the **left**, is a representation of the energy distribution as a continuous function of position. The value reported at each pixel is the value of this function at the location of the pixel (**right**).*

Cameras aren't linear in input energy

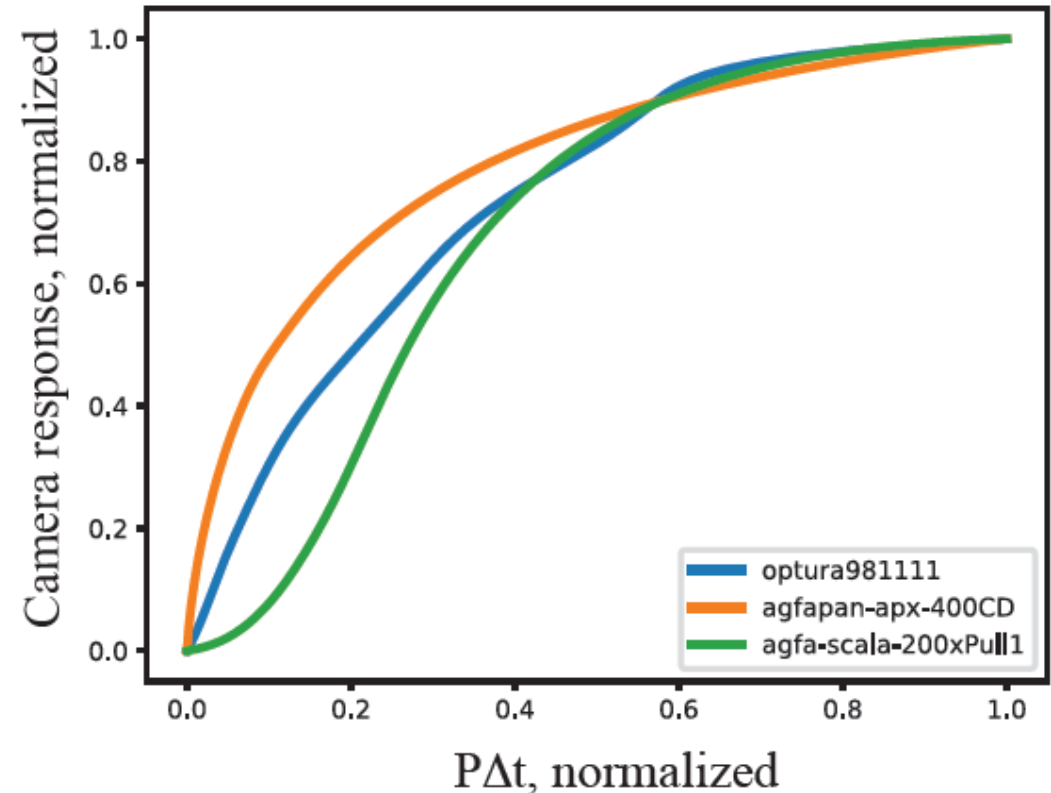
This is deliberate, and usually a property of camera electronics.

- Helps imitate film
- Increases dynamic range

It's not the sensor (which is usually linear)

- you can usually get linear pix from cameras with enough work

Different cameras have different nonlinearities



From wikipedia, https://commons.wikimedia.org/wiki/File:Raw_vs_jpg.jp



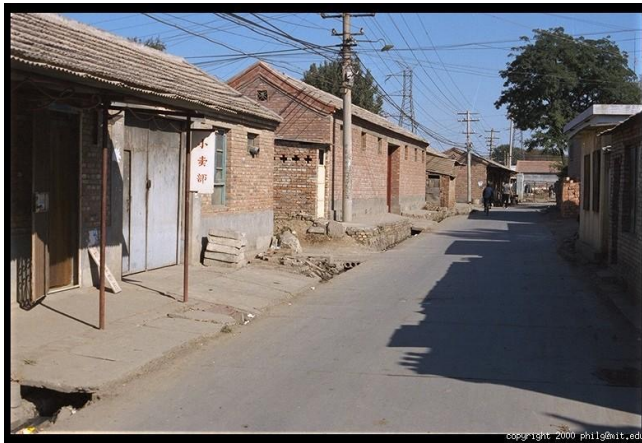
https://commons.wikimedia.org/wiki/File:Raw_vs_jpg.jpg

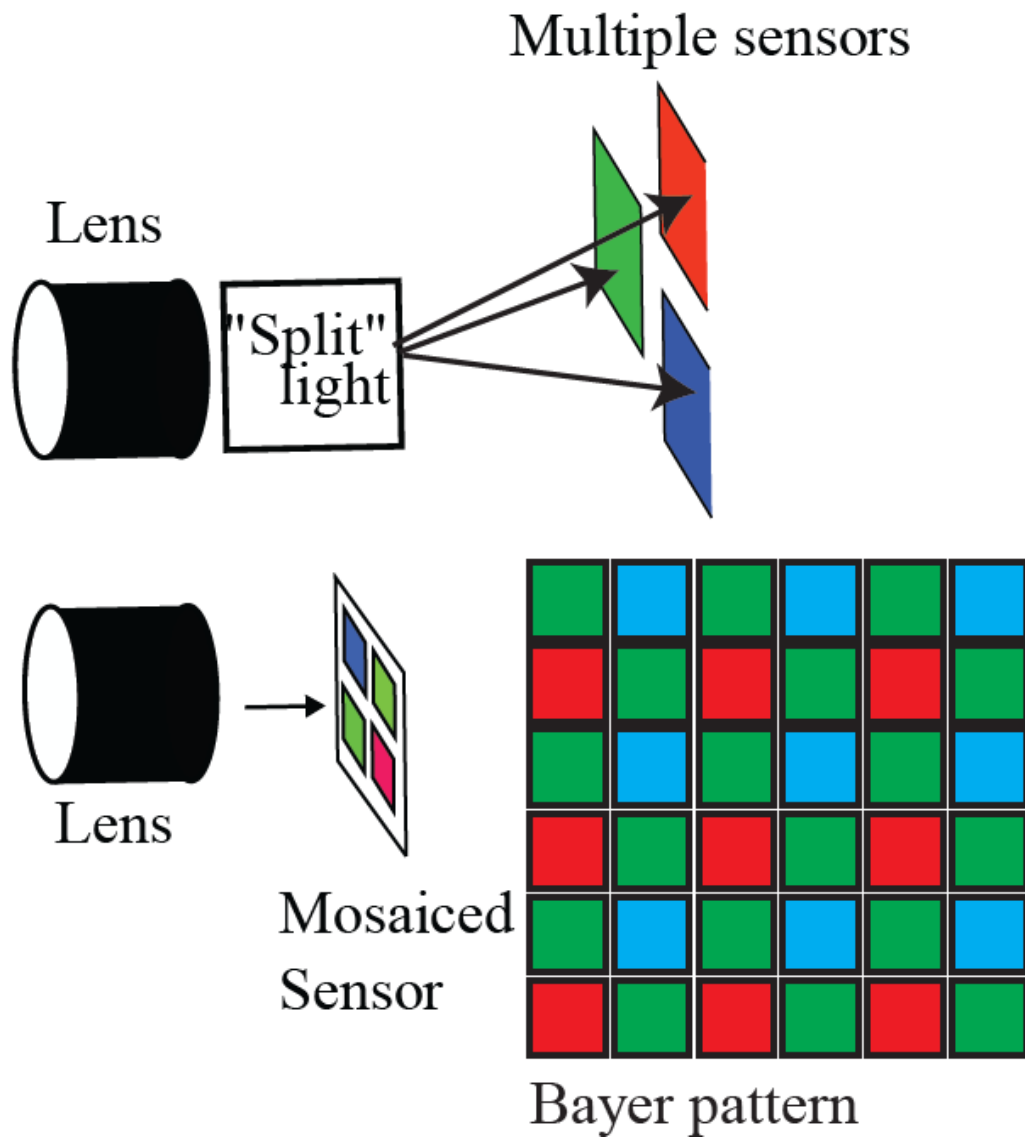




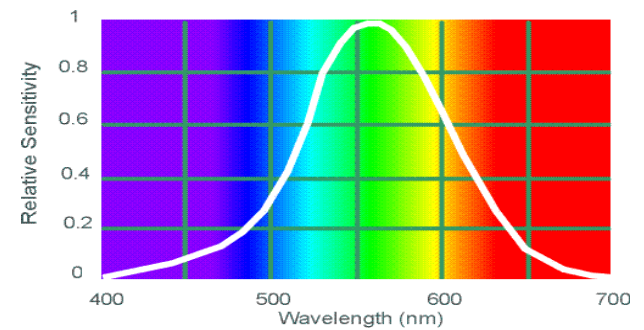
[https://commons.wikimedia.org/wiki/File:Raw_vs_jpg.
jpg](https://commons.wikimedia.org/wiki/File:Raw_vs_jpg.jpg)

Digital color image





Why more green?



Human Luminance Sensitivity Function

Images in Python

```
im = cv2.imread(filename)           # read image
im = cv2.cvtColor(im, cv2.COLOR_BGR2RGB) # order channels as RGB
im = im / 255                       # values range from 0 to 1
```

RGB image `im` is a `H x W x 3` matrix (numpy.ndarray)

`im[0, 0, 0]` is the top-left pixel value in R-channel

`im[y, x, c]` is the value $y+1$ pixels down, $x+1$ pixels to right in the c^{th} channel

`im[H-1, W-1, 2]` is the bottom-right pixel in B-channel

Pointwise image transformations

Apply the same function to each pixel value

- perhaps different for R, G, B

Simple applications

- remove camera non-linearity
- simulate different camera
- increase/reduce contrast

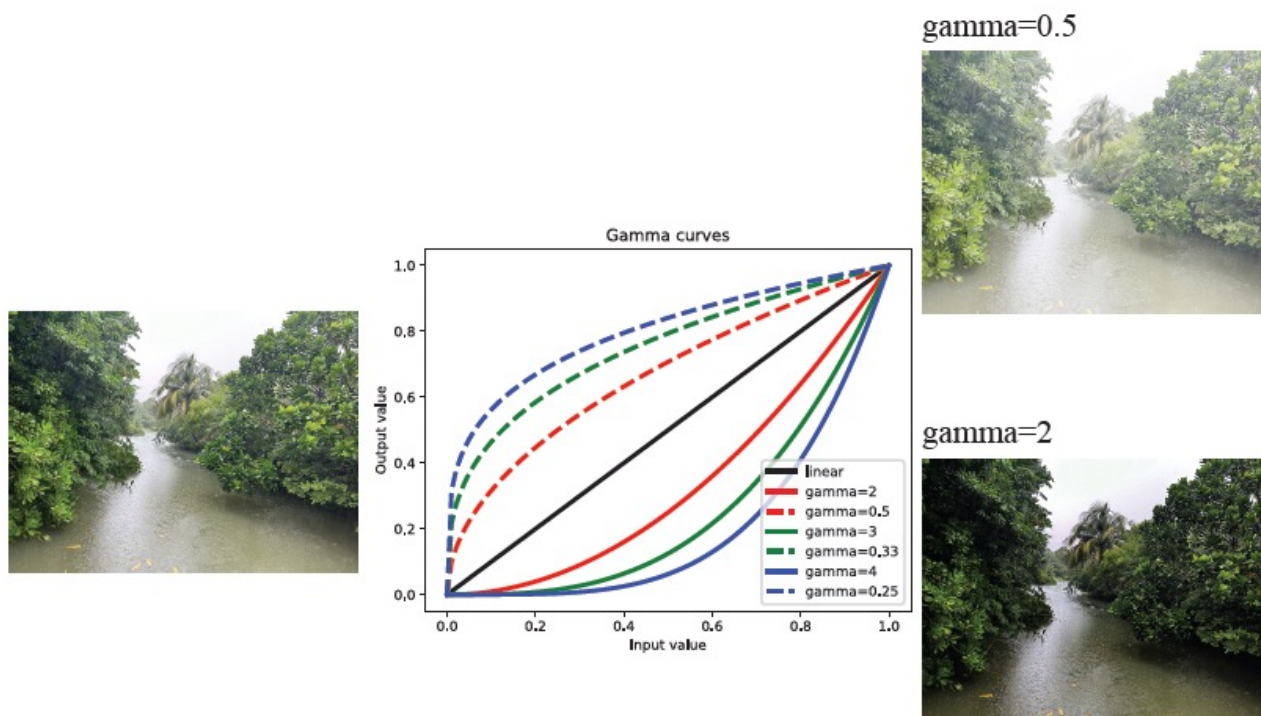


FIGURE 2.7: Many imaging and rendering devices have a response that is a power of the input, so that $\text{output} = C \text{input}^\gamma$, where γ is a parameter of the device. One can simulate this effect by applying a transform like those shown in the center (curves for several values of γ). Note that you can remove the effect of such a transform – gamma correct the image – by applying another such transform with an appropriately chosen γ . The image on the left is transformed to the two examples on the right with different γ values. Image credit: Figure shows my photograph of a river in Singapore.

Other sensors...

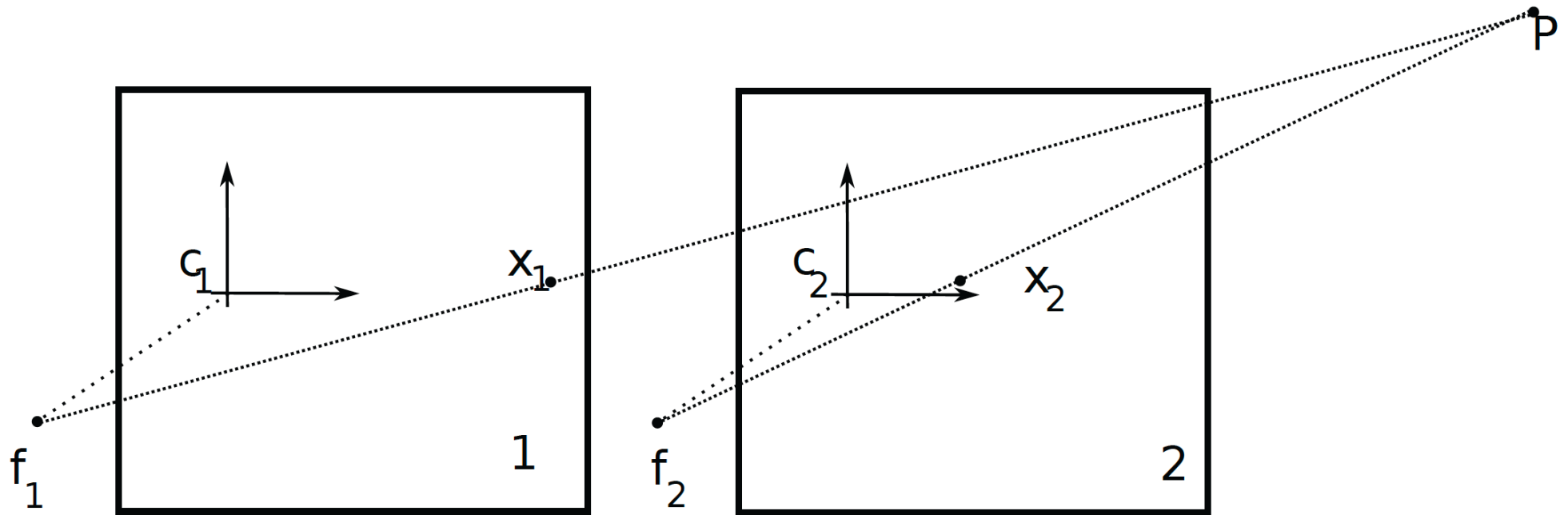
Basically, light cameras with extra tricks

- IR cameras
 - light cameras at different wavelengths
 - » but you might have to cool sensors
- Polarization cameras
 - specialized sensors to measure intensity at particular polarizations
- Multispectral cameras
 - various arrangements to image with more sensors than R, G, and B

Really different

- Depth cameras

Two cameras yield depth



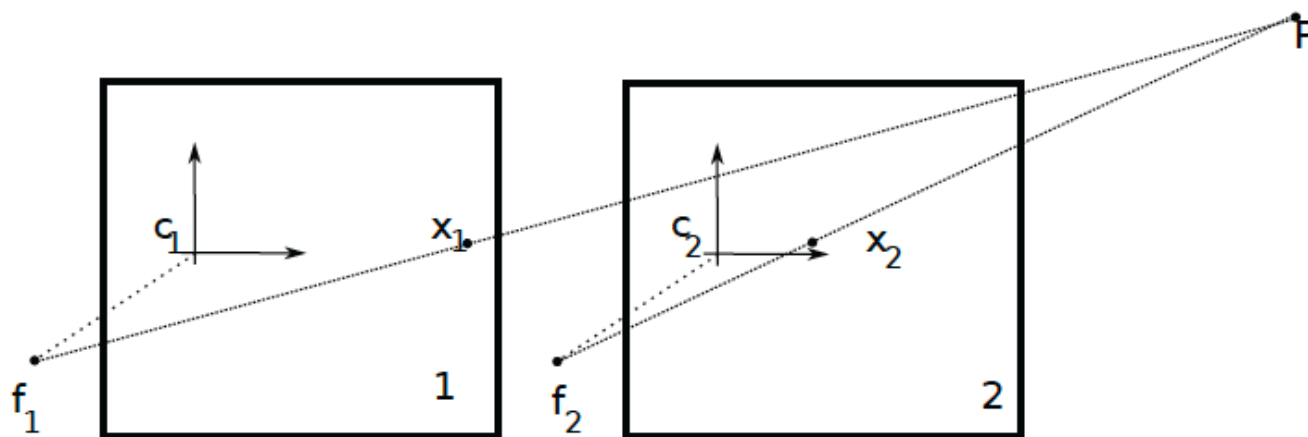


FIGURE 2.8: When two pinhole cameras view a point, the 3D coordinates of the point can be reconstructed from the two images of that point. This applies for almost every configurations of the cameras. It is an elementary exercise in trigonometry (exercises) to determine P from the positions of the two focal points, the locations of the point in the two images, and the distance between the focal points. Considerable work can be required to find appropriate matching points, but the procedures required are now extremely well understood (Chapters 35). One can now buy camera systems that use this approach to report 3D point locations (often known as RGBD cameras). Here we show a specialized camera geometry, chosen to simplify notation. The second camera is translated with respect to the first, along a direction parallel to the image plane. The second camera is a copy of the first camera, so the image planes are parallel. In this geometry, the point being viewed shifts somewhat to the left in the right camera.

If's, and's and but's...

But there are limits to stereopsis. Measuring large depths with two cameras that are close together requires highly accurate estimates of point positions in images. Figure 2.8 shows a simple geometry that illustrates the problem. The point **P** projects to $\mathbf{x}_1$ in camera 1, and to $\mathbf{x}_2$ in camera 2. Notice because of the carefully chosen camera geometry, the y -coordinates of $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same; only the x -coordinates differ. Write x_1 for the x -coordinate of $\mathbf{x}_1$; X for the x -coordinate of P , and so on. From the triangles in that figure, we have

$$d = x_2 - x_1 = f \frac{(X - B) - X}{Z} = -f \frac{B}{Z}$$

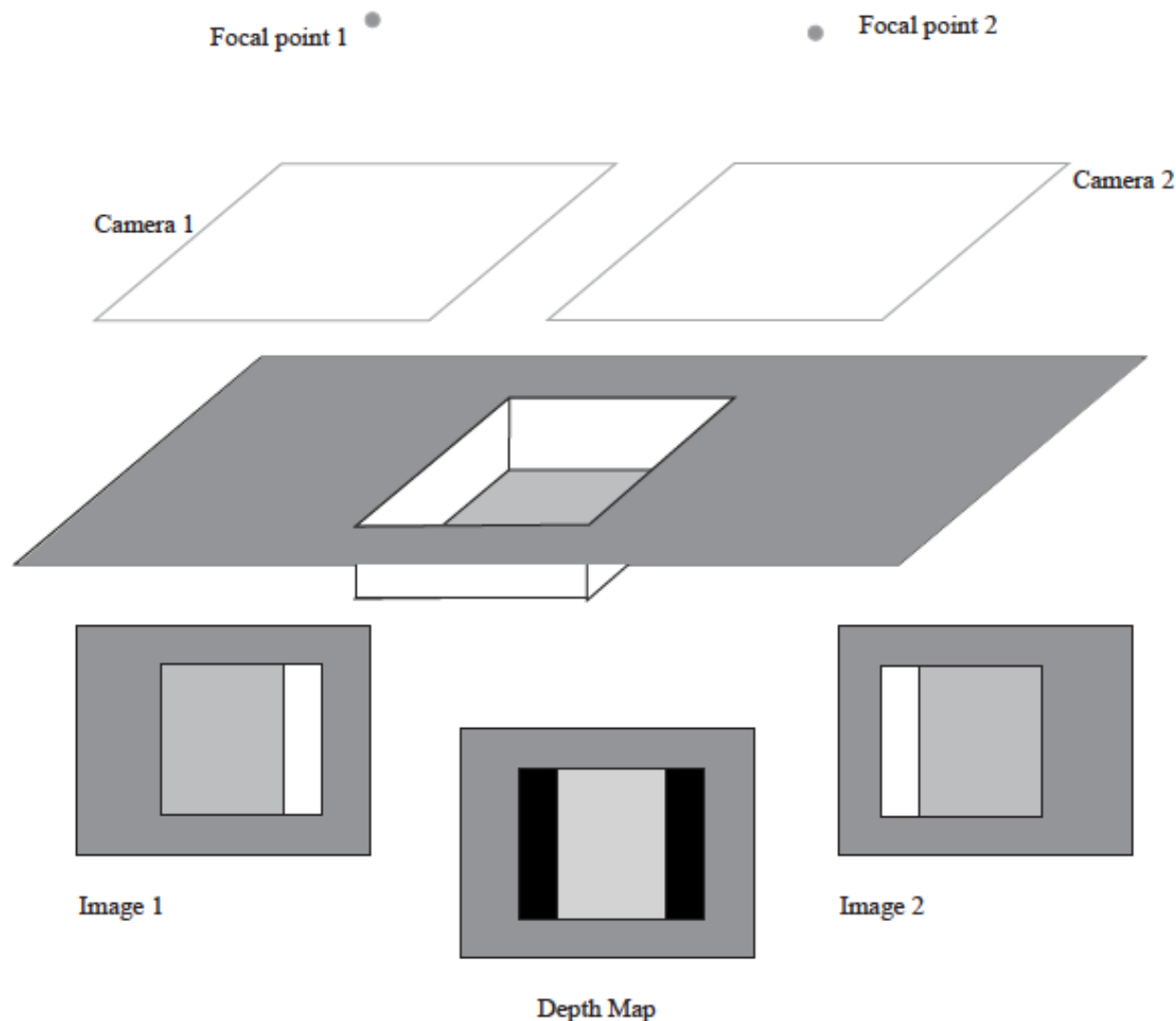
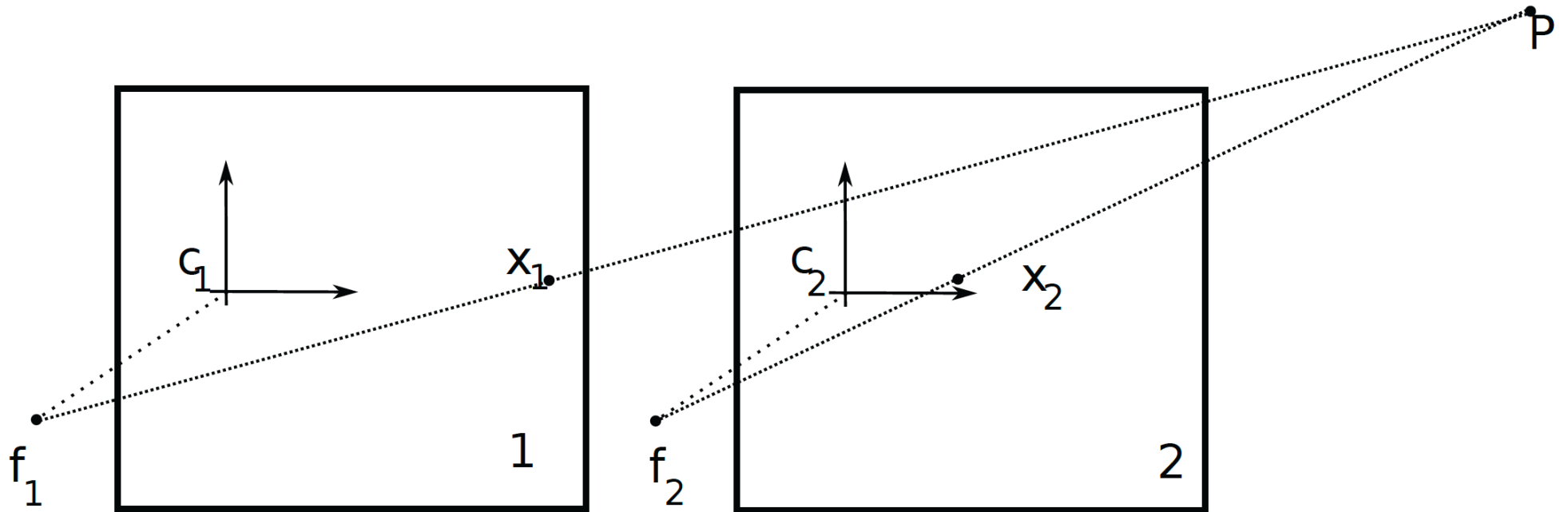


FIGURE 2.9: Top shows two pinhole cameras viewing a rectangular depression in a flat surface. As the images show, camera on the left can see the right wall, and that on the right can see the left wall. This means that these walls cannot be reconstructed directly using trigonometry, and so the depth map will have holes in it. The depth map here is shown with a fairly common convention, where nearer surfaces are lighter, farther surfaces are darker, and holes are “infinitely far away”.

Camera – projector stereo



Structured light

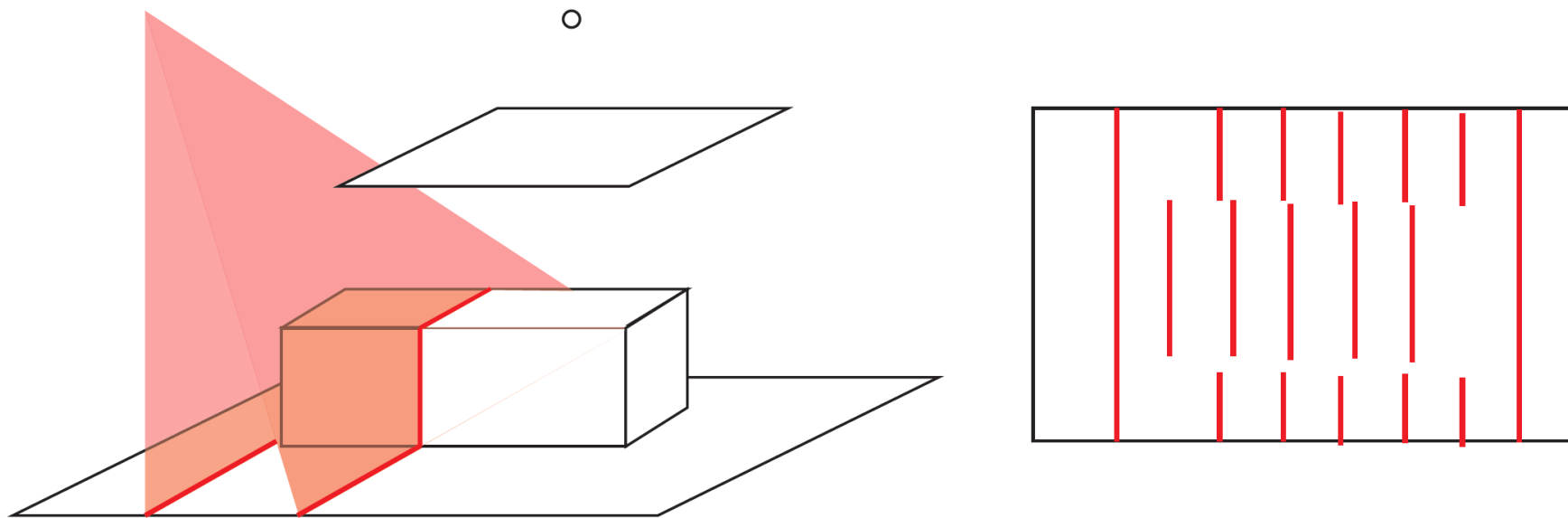


FIGURE 2.10: *The projector on the **left** casts planes of light into the scene. These are viewed by a camera. Their shape in the image (**right**) is a cue to the depth in the scene.*

LIDAR

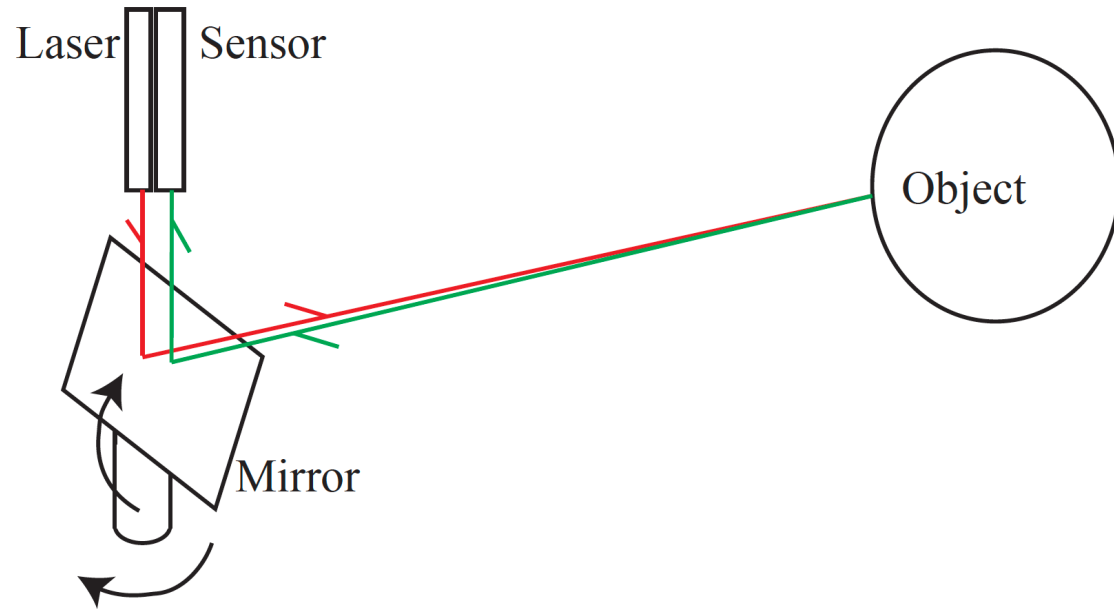


FIGURE 2.11: *A laser flashes a brief pulse of light onto a mirror. The light travels into a scene, is reflected from an object and returns to the mirror, and is reflected into a sensor. The time from flash to sensing reveals the distance to the object. The mirror can be tilted or rotated to flash the light in different directions, and so to measure depth along different rays.*

-
- 2.1. As the pinhole in a pinhole camera gets larger, the image on the image plane (a) gets brighter and (b) is less focused. Why?
 - 2.2. For a sufficiently small pinhole, diffraction effects cause the image to be defocused. Explain.
 - 2.3. Why do cameras have lenses?
 - 2.4. For smaller Δt , images will be darker; for larger Δt , moving objects may have blurred outlines. Why?
 - 2.5. A common form of camera response function is $\text{Intensity}(P\Delta t) = Ce^{\gamma P\Delta t}$. Do you expect γ to be positive or negative? Why?
 - 2.6. Why are mosaiced color CCD cameras more common than multiscan cameras?
 - 2.7. You want to increase the contrast in an image by a pointwise image transformation that (a) makes dark pixels darker *and* (b) makes bright pixels brighter. Why will $f(I) = Ce^{\gamma I}$ not achieve this? Sketch a transformation that will.
 - 2.8. Why is it hard to build a stereo camera that is (a) small and (b) good at measuring large depths?
 - 2.9. What happens in a scanning time of flight sensor if the mirror moves slowly and objects in the scene move quickly?
 - 2.10. An object is 30m away from a scanning time of flight sensor. How long does it take for the pulse to travel from camera to object to sensor? You can take the speed of light to be 3×10^8 meters per second.
