Fall 2025 CS543 / ECE549
Computer Vision

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

And follow links

Outline

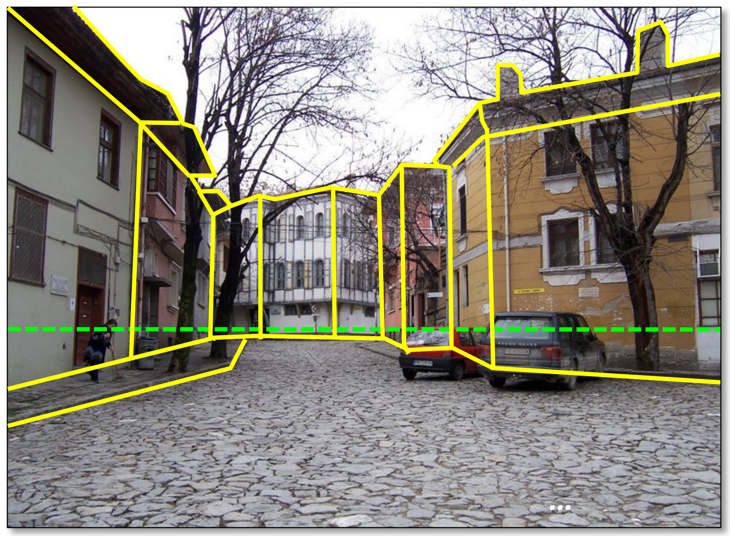
- Logistics, requirements
- Key tasks
- Why it is hard
- History of computer vision
- Current state of the art
- Topics covered in class

Logistics

Look at web page!

What kind of information can be extracted from an image?

What kind of information can be extracted from an image?



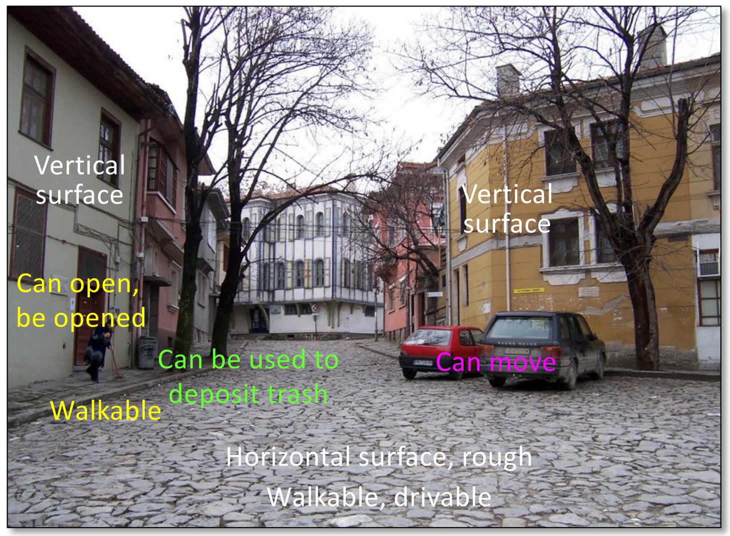
Geometric information

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

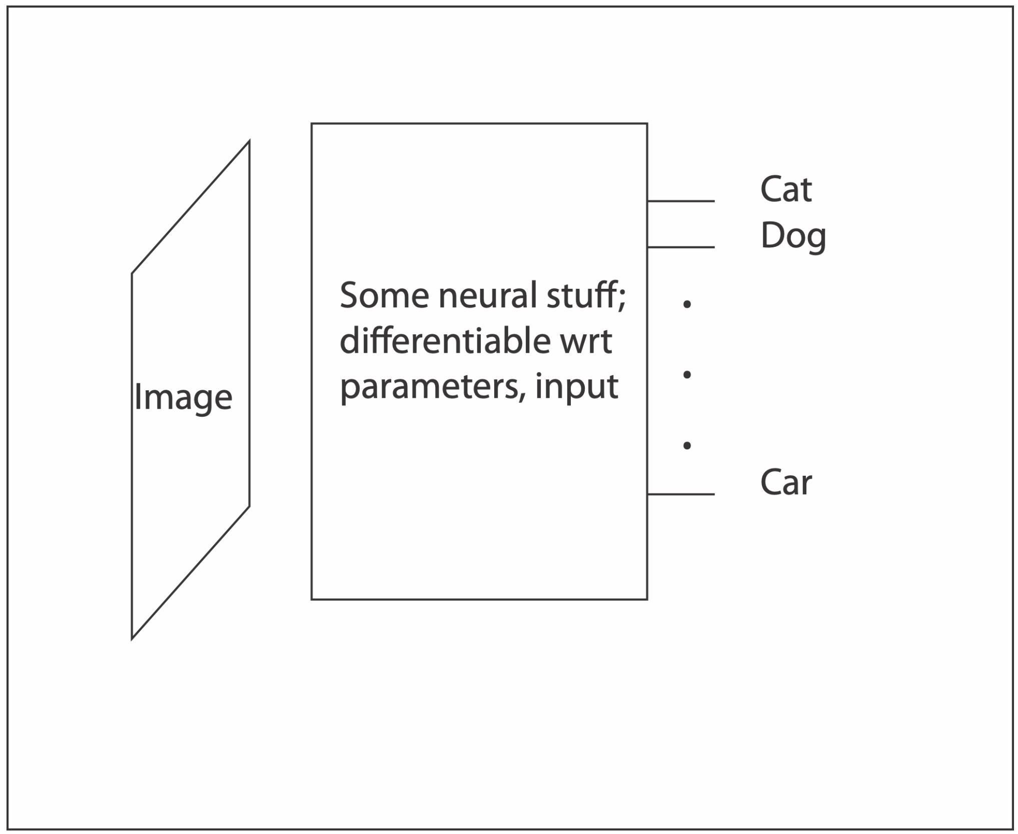
Semantic information

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

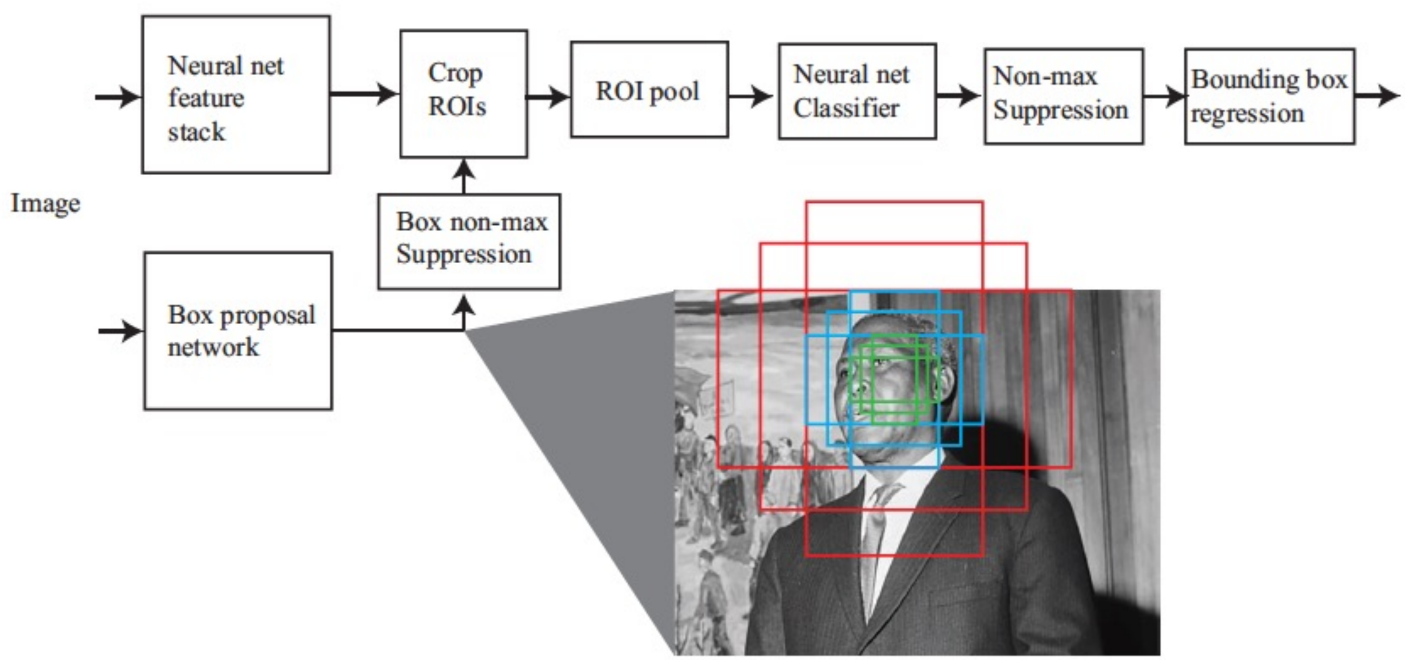
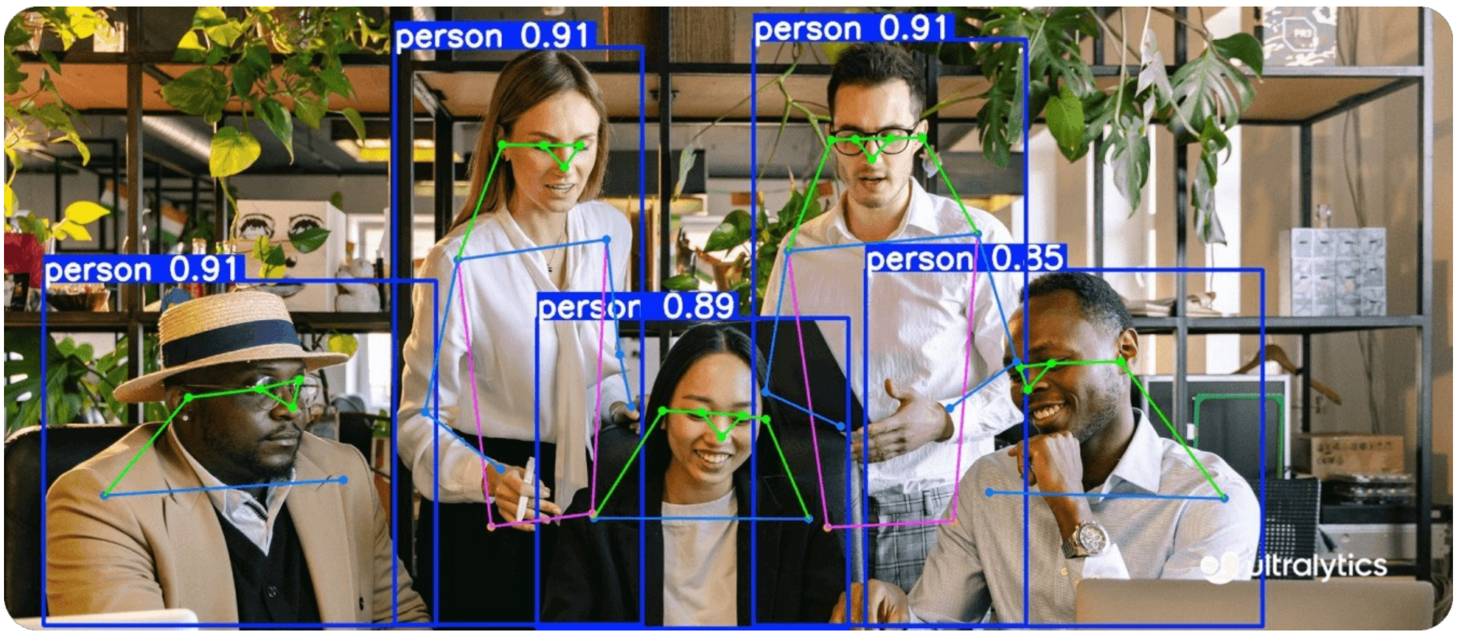


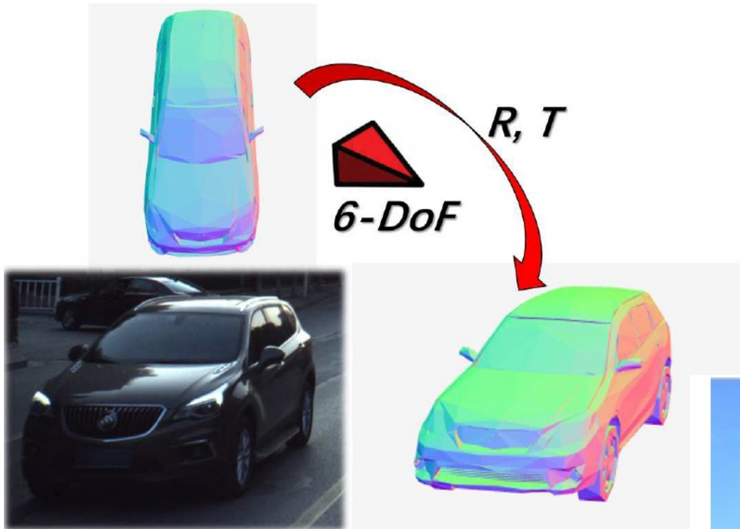
FIGURE 18.8: *Faster RCNN uses two networks. One uses the image to compute "objectness" scores for a sampling of possible image boxes. The samples (called "anchor boxes") are each centered at a grid point. At each grid point, there are nine boxes (three scales, three aspect ratios). The second is a feature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

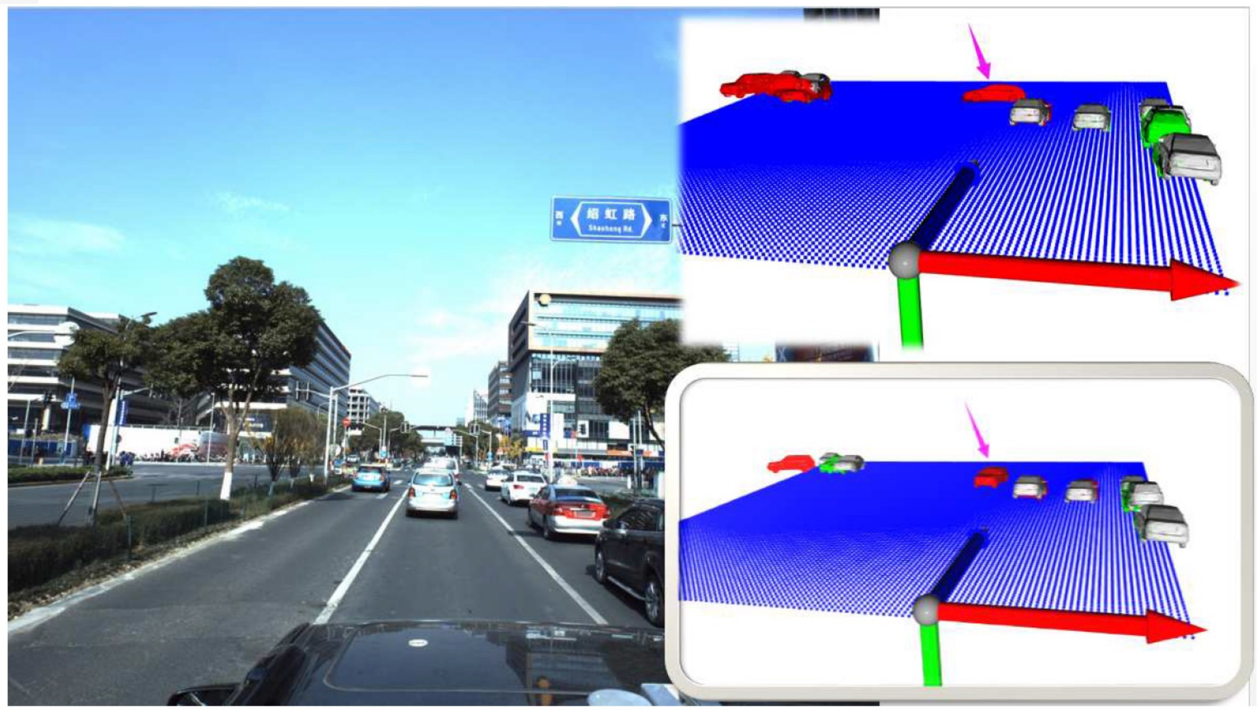
Detection and localization in 2D



YOLOv11 documentation by Ultralytics

Localization in 3D from detection





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

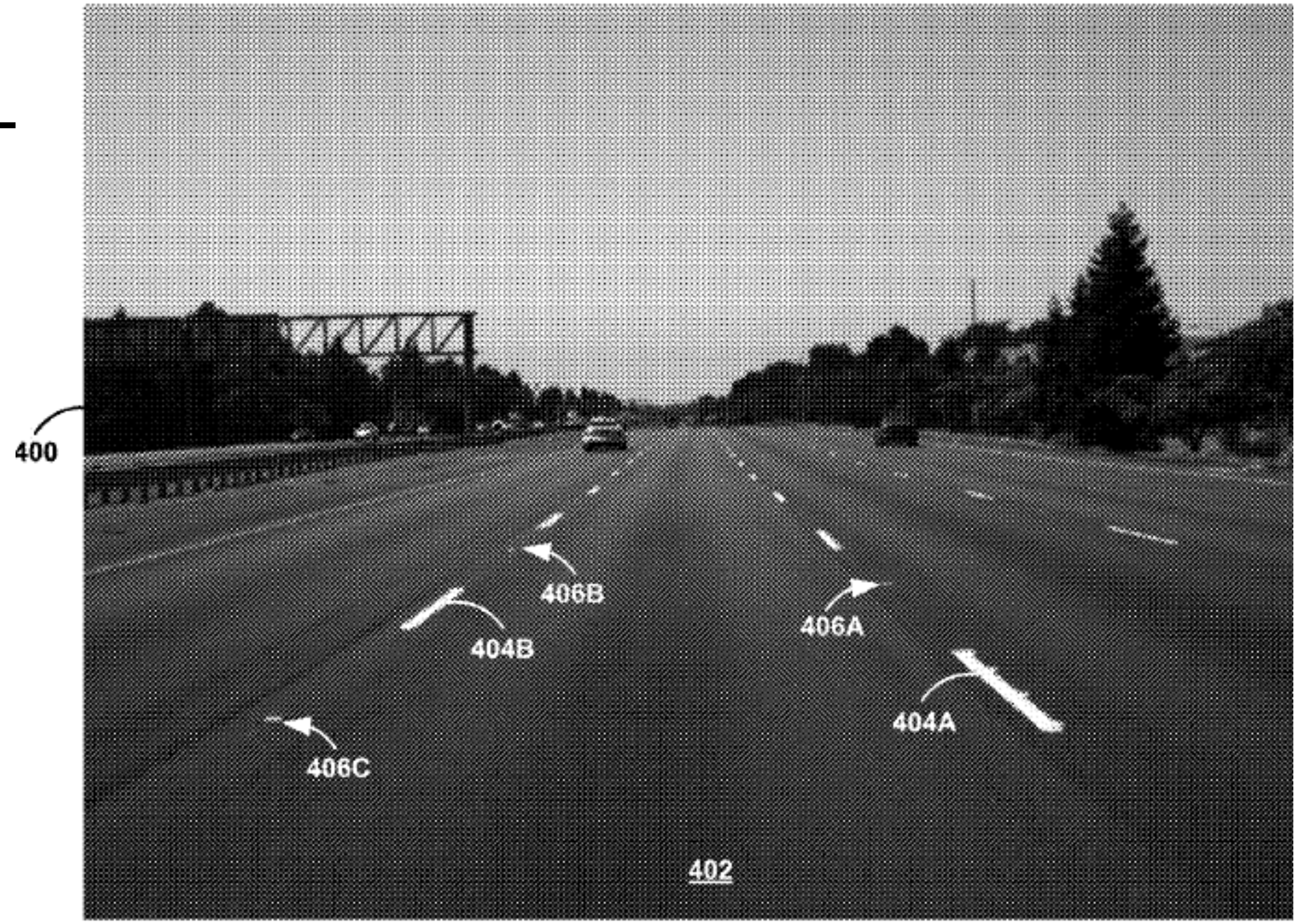
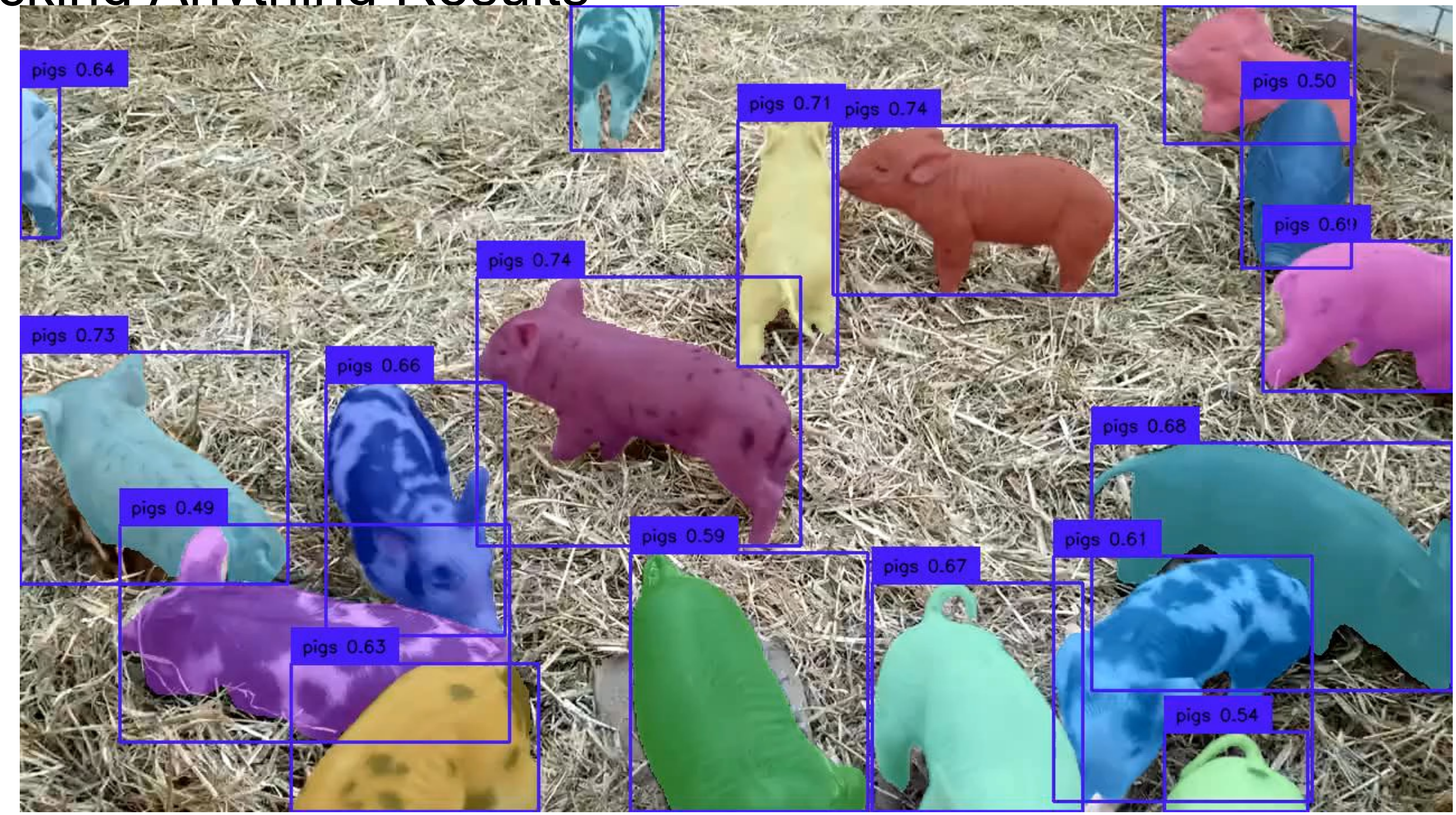


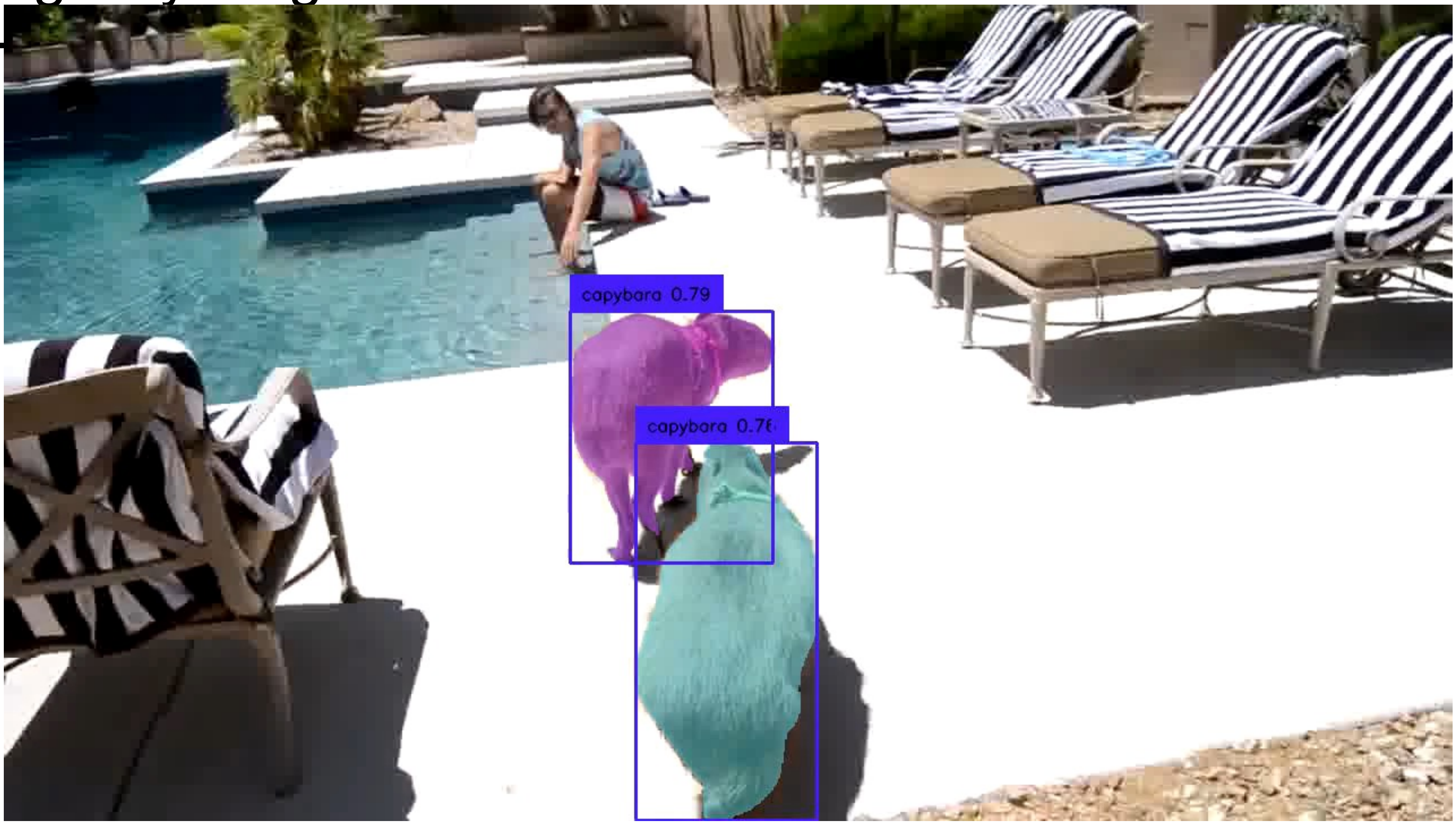
FIGURE 4A

Tracking Anything Results



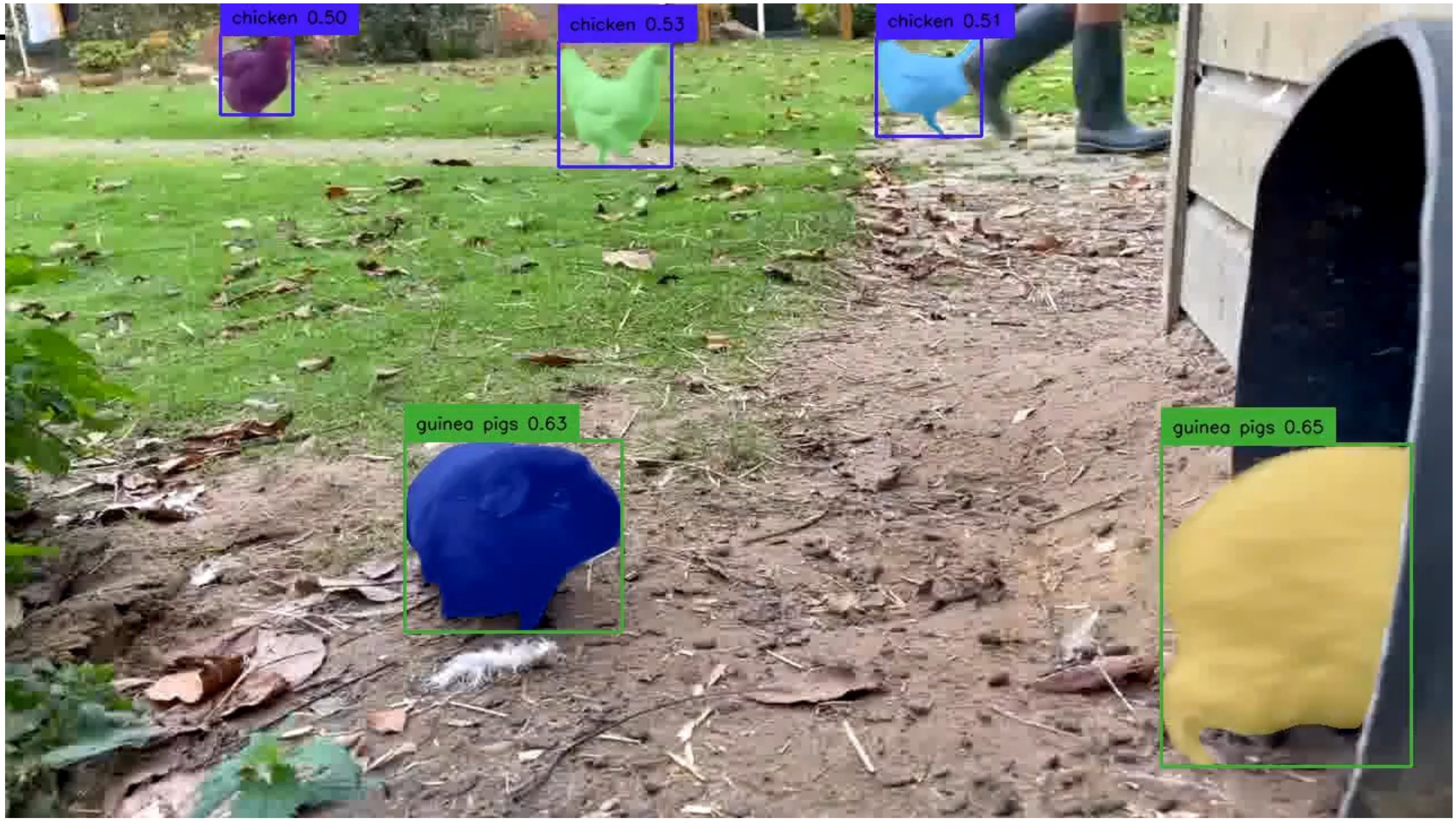
Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

Tracking Anything Results



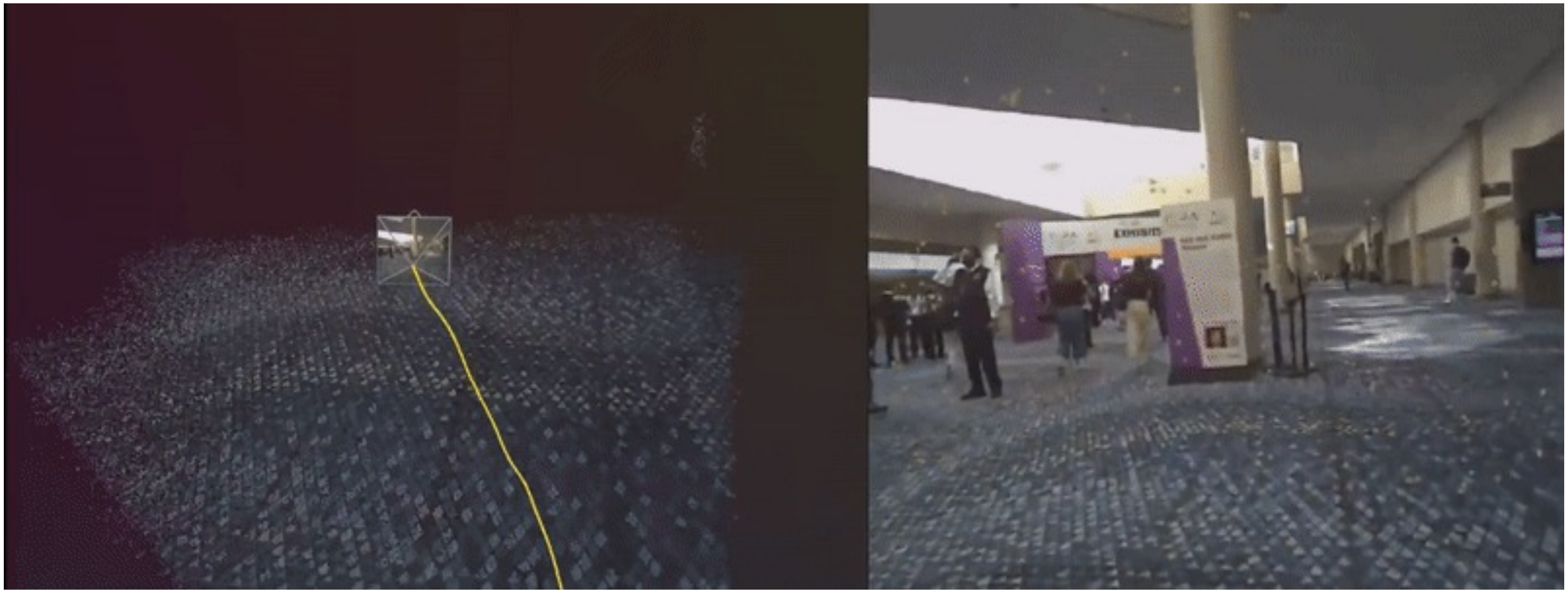
Cheng et al.; DEVA: Tracking anything with decoupled video segmentation; 2023

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=fBiataDpGIo

Goal: To extract useful information from pixels

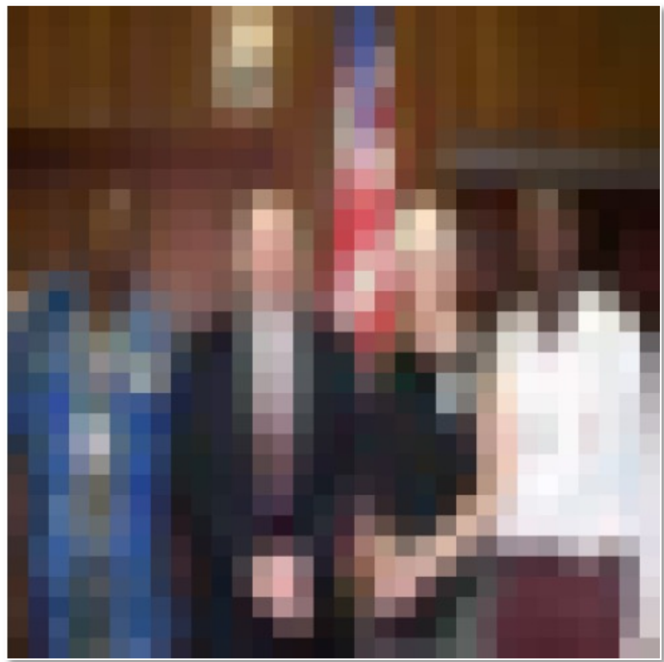
What we see

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

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

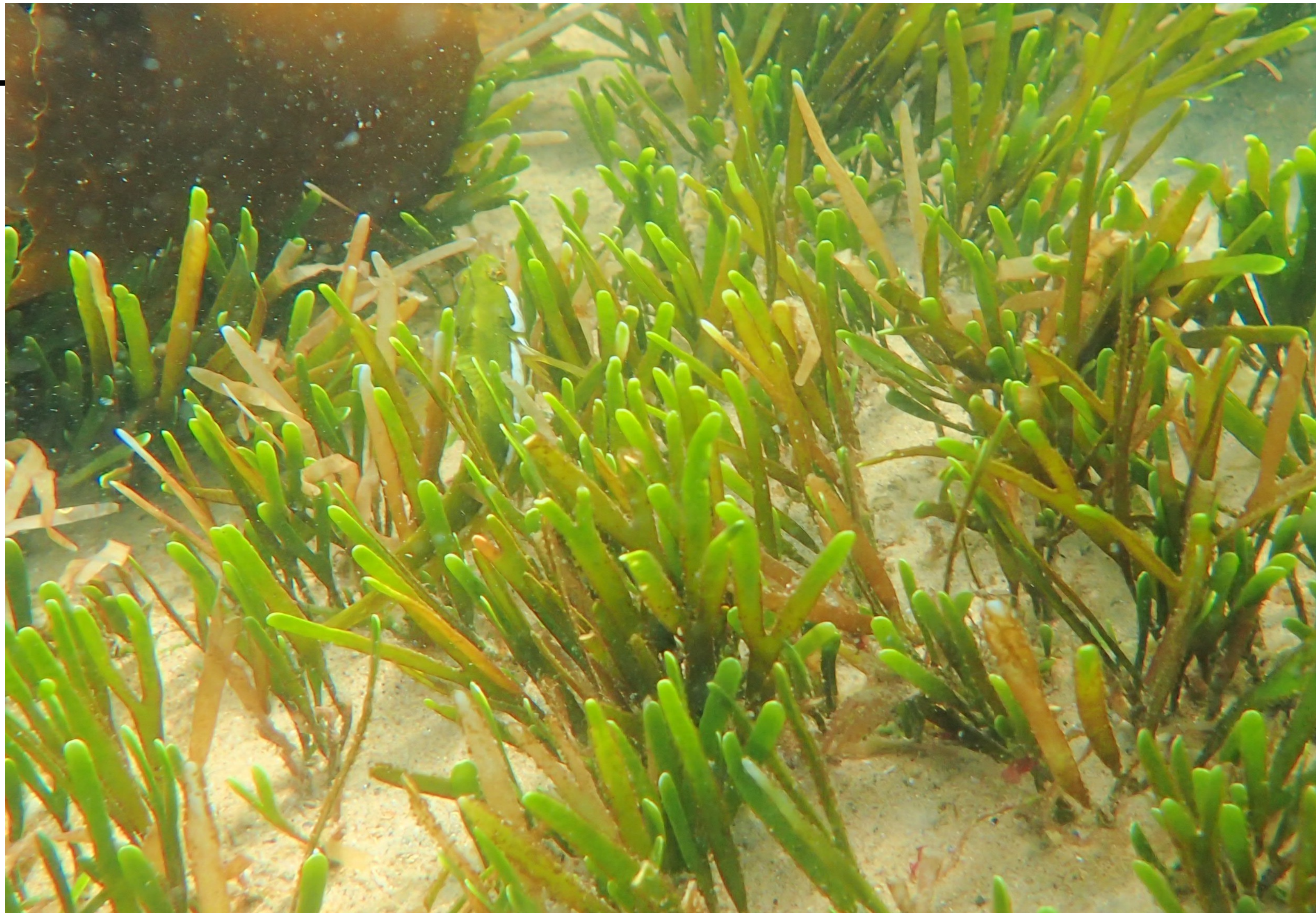
Image source

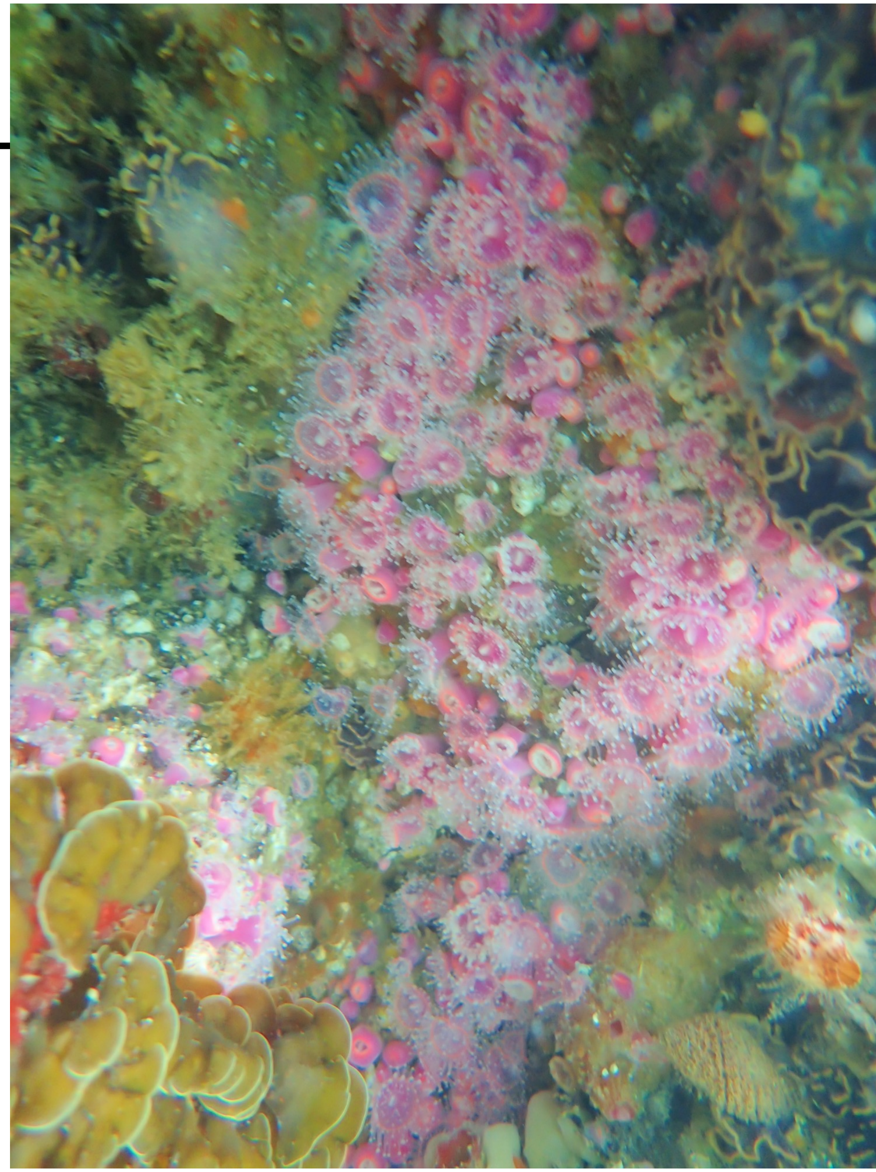
Image source

…still, vision is hard even for humans

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

Is





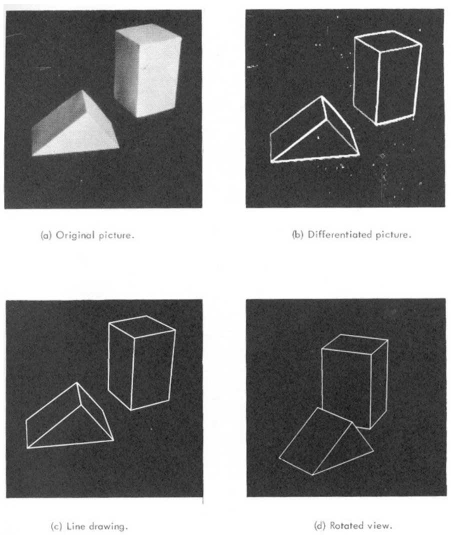
…still, vision is hard even for humans

What color is this dress?

Outline

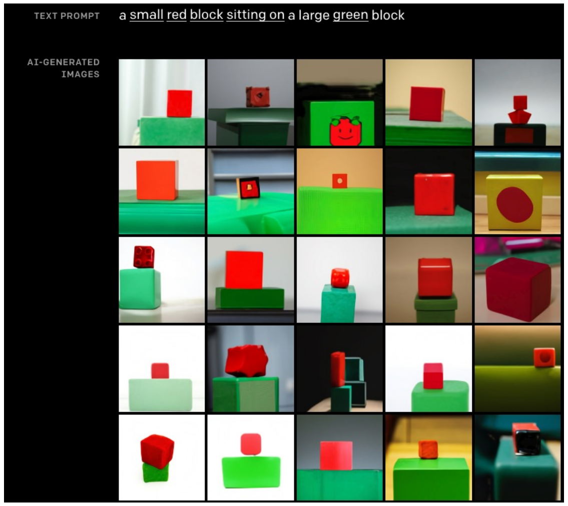
- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision

How it started



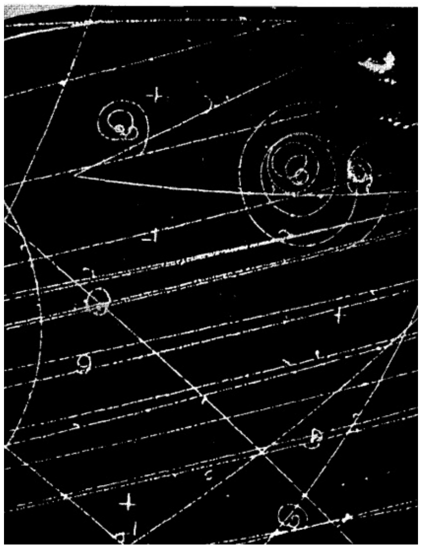
L. G. Roberts, 1963

How it's going

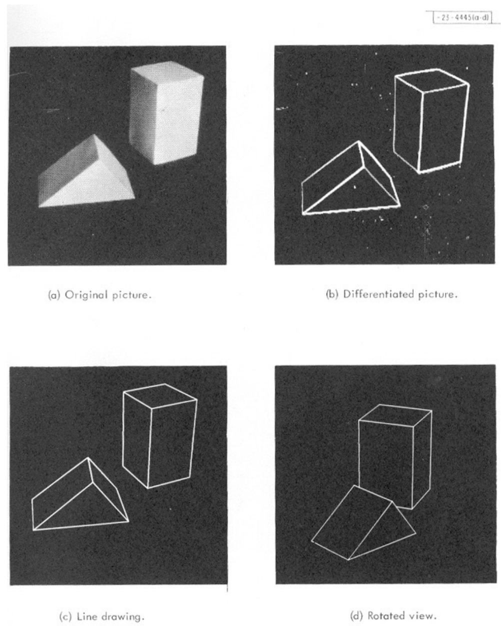


OpenAI DALL-E, 2020

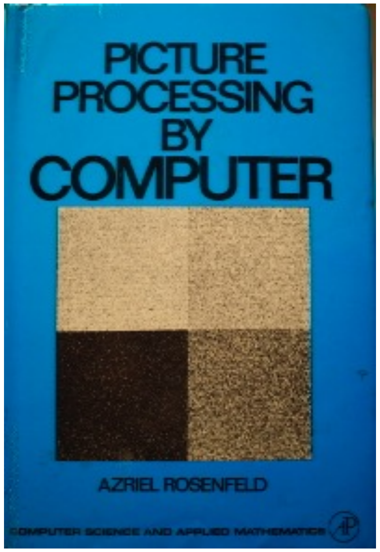
Origins



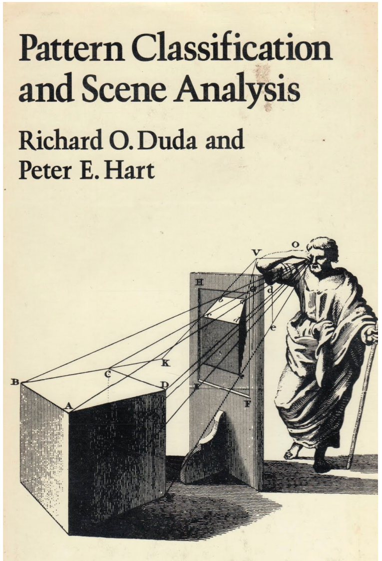
Hough, 1959



Roberts, 1963



Rosenfeld, 1969



Duda & Hart, 1972

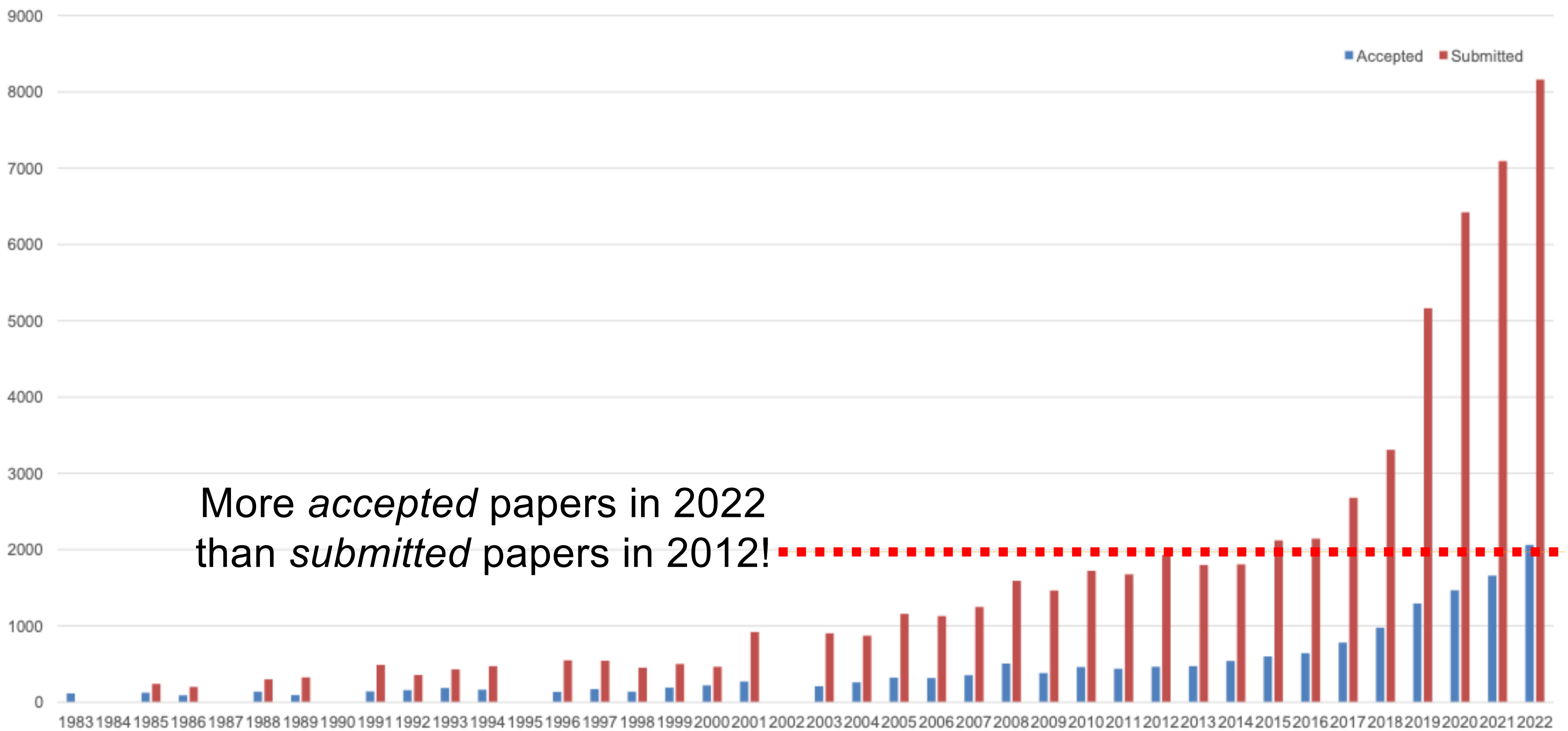
Decade by decade

- **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 Lazebnik's historical overview

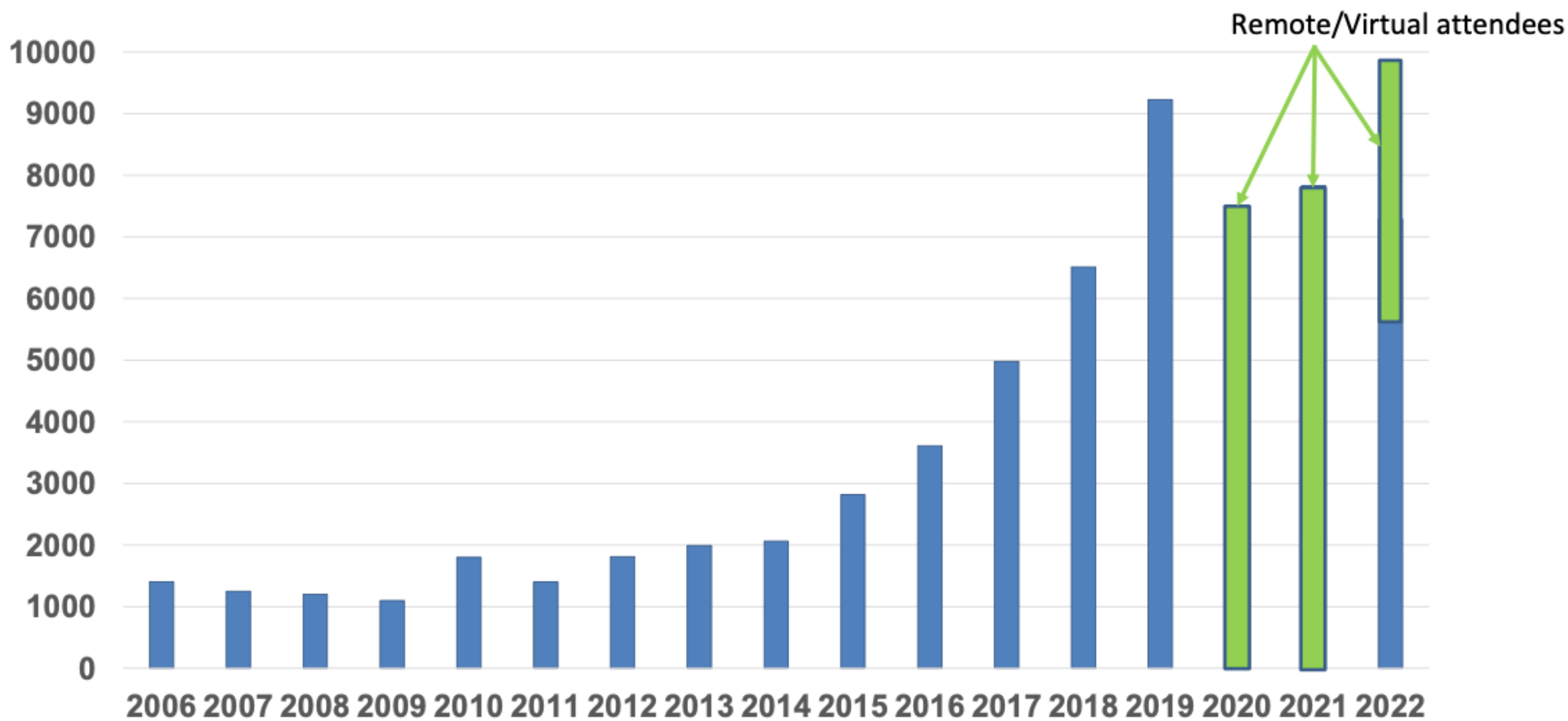
Adapted from J. Malik

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

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
Vision	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
Vision	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
Vision	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 : AAAI Conference on Artificial Intelligence** Feb 2, 2021 - Feb 9, 2021 - Vancouver , Canada https://aaai.org/Conferences/AAAI-21/	126	25.57

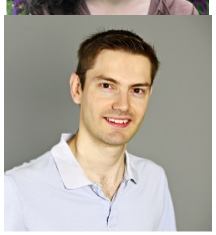
Vision group at Illinois

David Forsyth
- Marr prize, 1993; 2 ex students with Marr prizes; IEEE Tech. Achievement, Fellow; ACM Fellow; EIC IEEE TPAMI

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

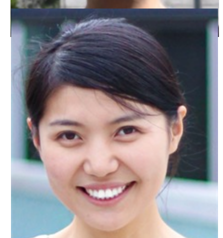
Derek Hoiem
- best paper, CVPR 2006; ACM Doctoral Dissertation honorable mention; Sloan Fellow;PAMI-TC Young Researcher

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



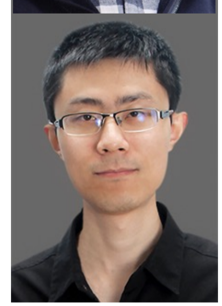
Alex Schwing
- Visual learning, segmentation and GAN models

Saurabh Gupta
- Linking visual sensing to motion



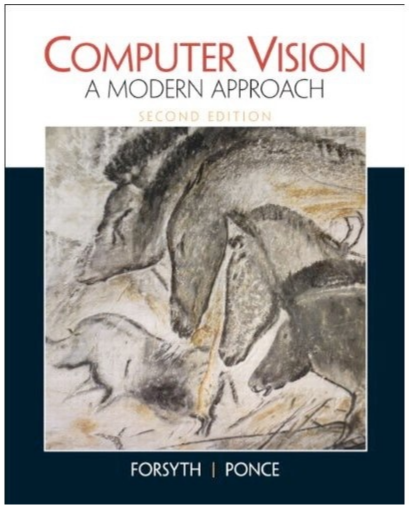
Liangyan Gui
- Understanding human movement

Shenlong Wang
- Simulation and sensing for autonomous vehicles



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

Vision group



The New Computer Vision

D.A. Forsyth

Likely about ~~2024~~ 2026

Cover design opportunity!

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.)

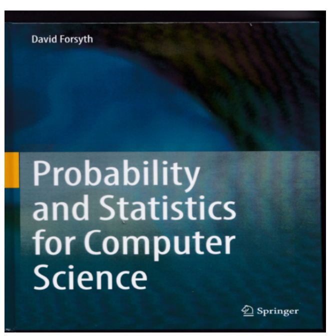
Startups:

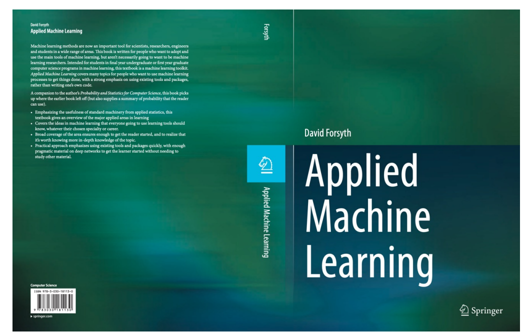
Lightform

Revery.ai

Reconstruct

Depix

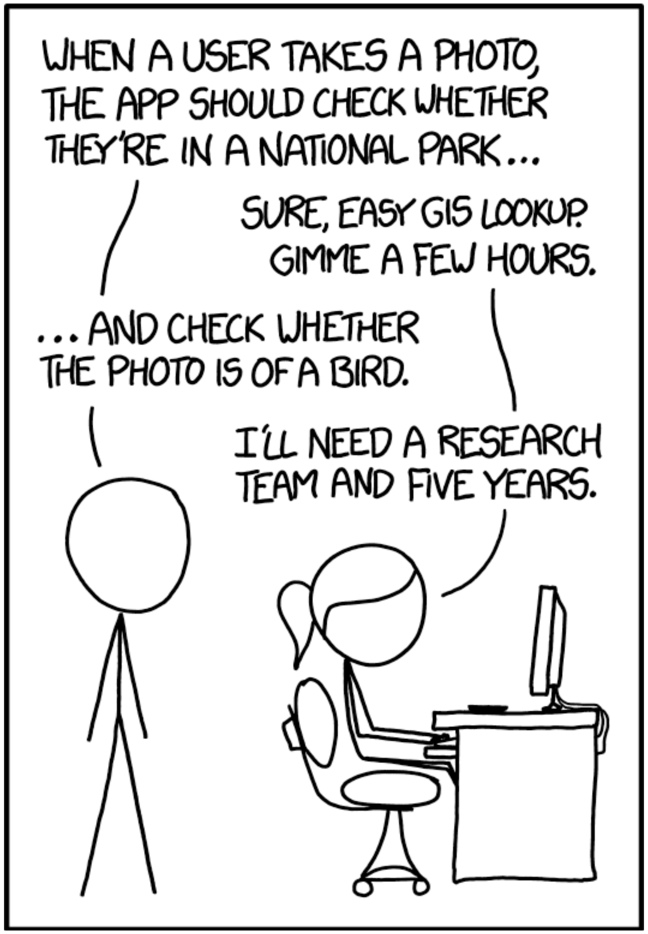




Introduction: Outline

- 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?



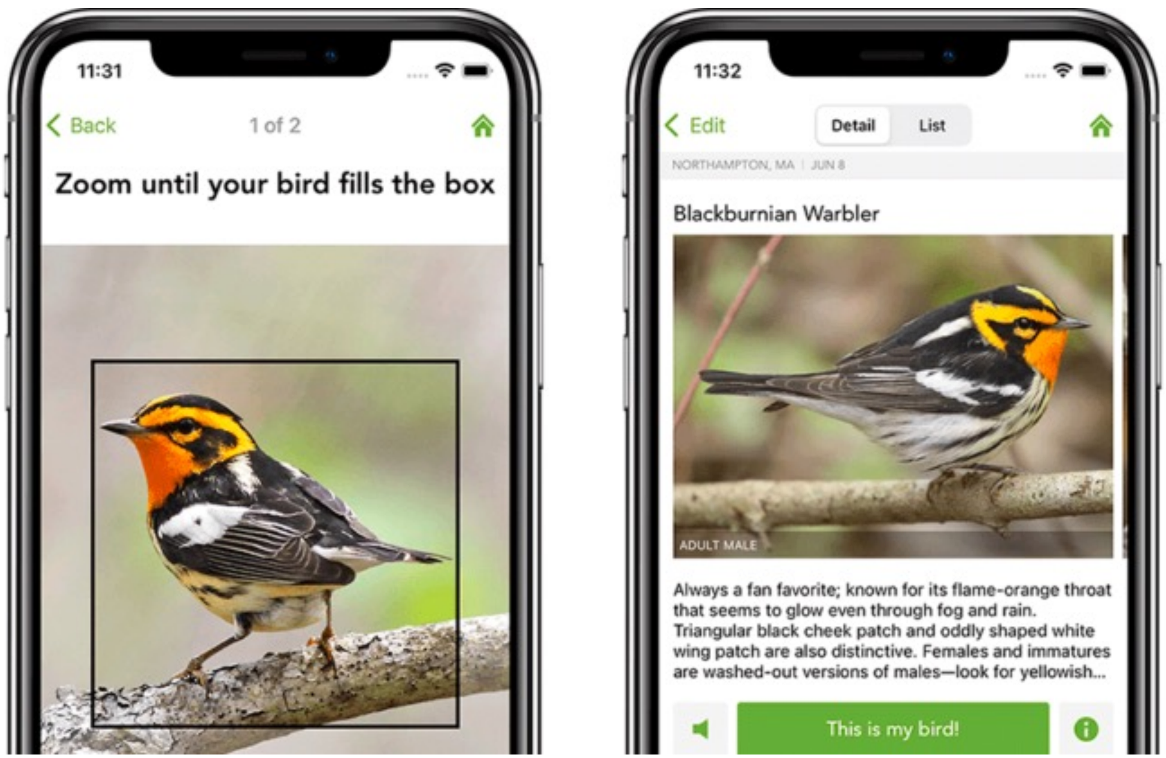
In the 60s, Marvin Minsky assigned a couple of undergrads to spend the summer programming a computer to use a camera to identify objects in a scene. He figured they'd have the problem solved by the end of the summer. Half a century later, we're still working on it.

https://xkcd.com/1425/

(September 24, 2014)

What can computer vision do today?

- It’s 2025 now…



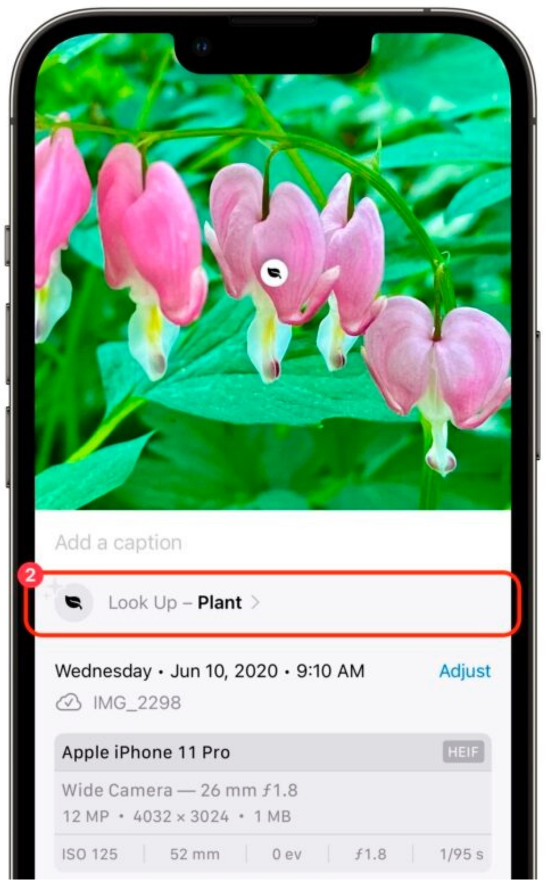
TheCornellLab

Merlin®

https://merlin.allaboutbirds.org/

What can computer vision do today?

- It’s 2025 now…



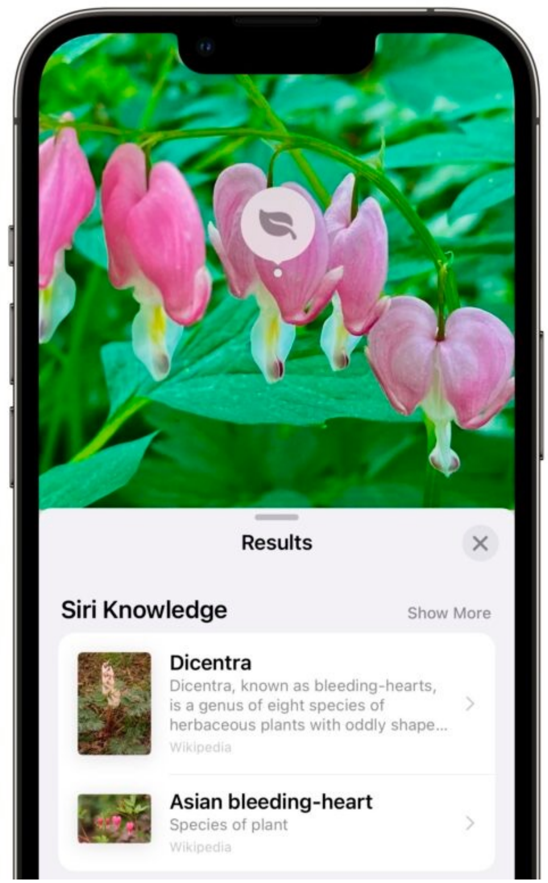


Image source

What can computer vision do today?

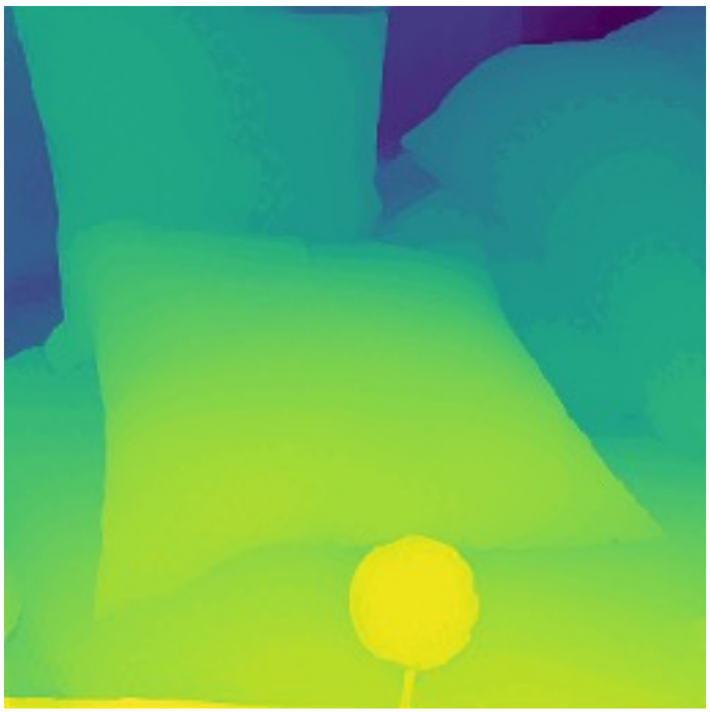
- Reconstruction
- Recognition
- *Reconstruction meets recognition, or 3D scene understanding*
- *Image generation*
- *Vision for action*

Regression

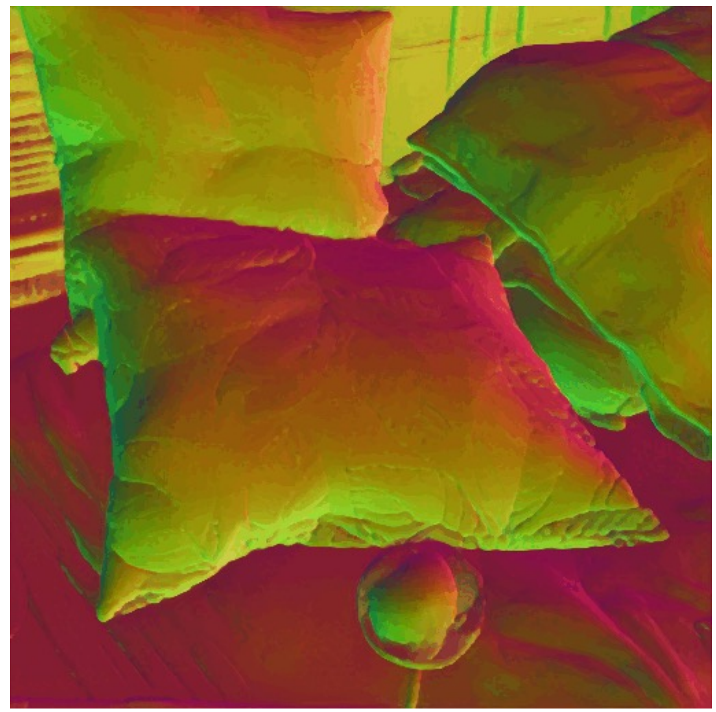
- We must make image-like things from images
- Examples:
  - depth map from image
  - normal map from image
  - derained image from rainy image
  - defogged image from foggy image
- Train with pairs (image, depth)
  - or (image, normal), etc
  - Loss
    - Squared error +abs value of error+other terms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

Ex

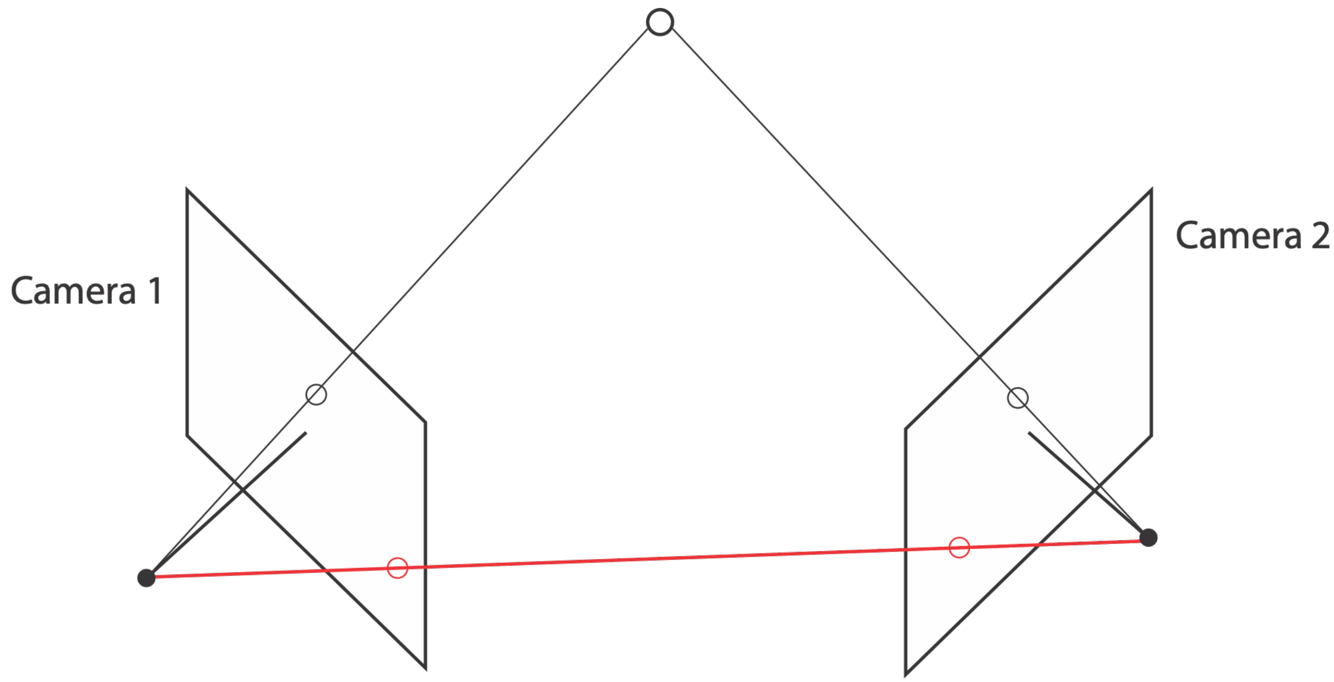
Depth (omnimap, current best depth est)



Normal (omnimap, current best normal est)

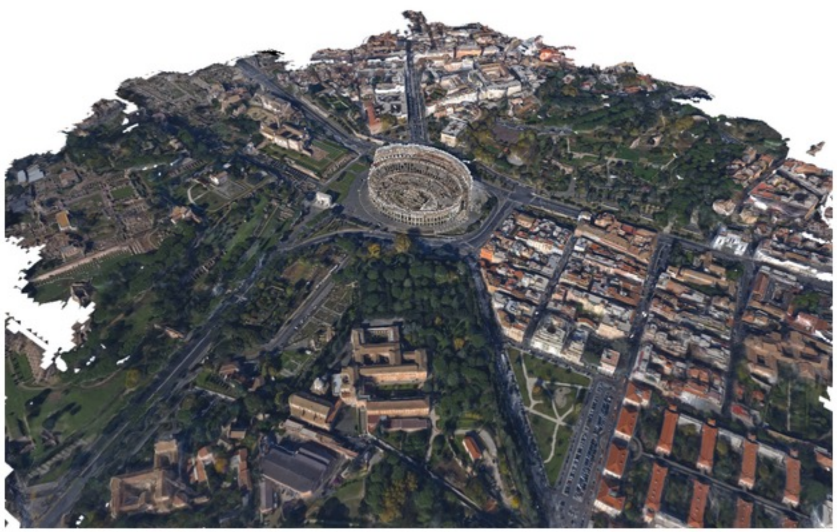


Correspondence yields 3D configuration



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

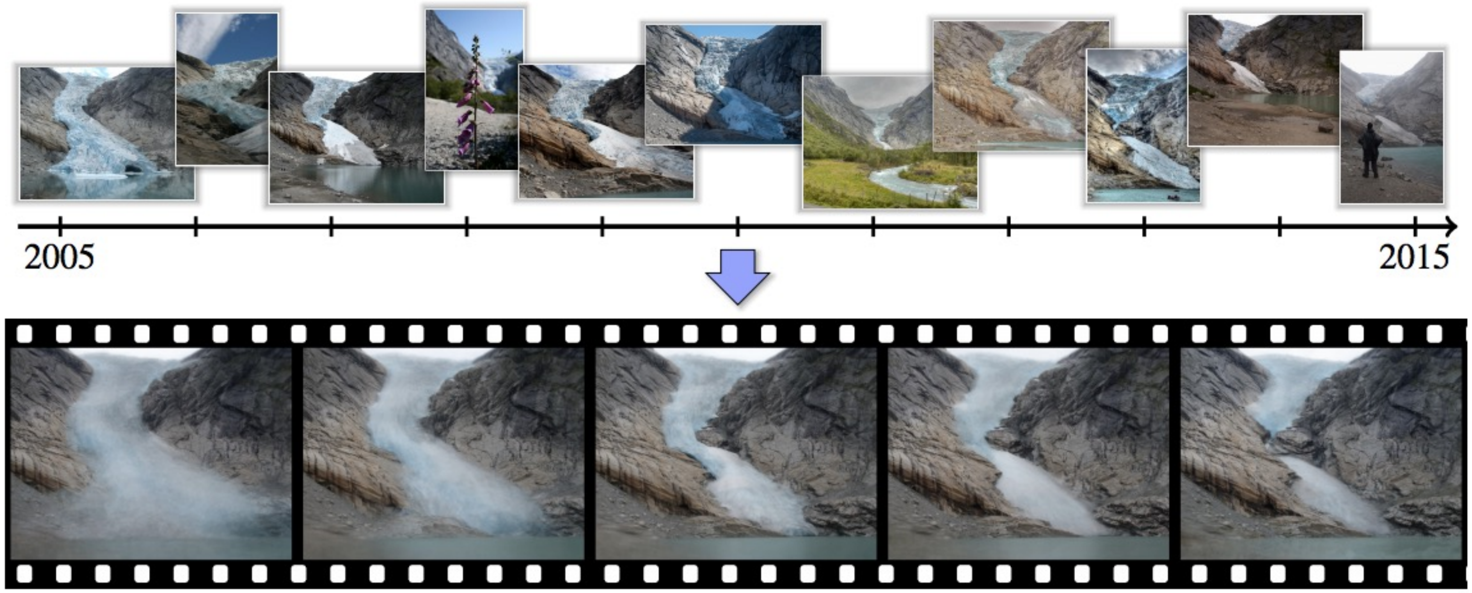


Figure 1: *We mine Internet photo collections to generate time-lapse videos of locations all over the world. Our time-lapses visualize a multitude of changes, like the retreat of the Briksdalsbreen Glacier in Norway shown above. The continuous time-lapse (bottom) is computed from hundreds of Internet photos (samples on top).* *Photo credits:* `Aliento Más Allá`, `jirihnidek`, `mcxurxo`, `elka_cz`, *Juan Jesús Orío, Klaus Wißkirchen,* `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

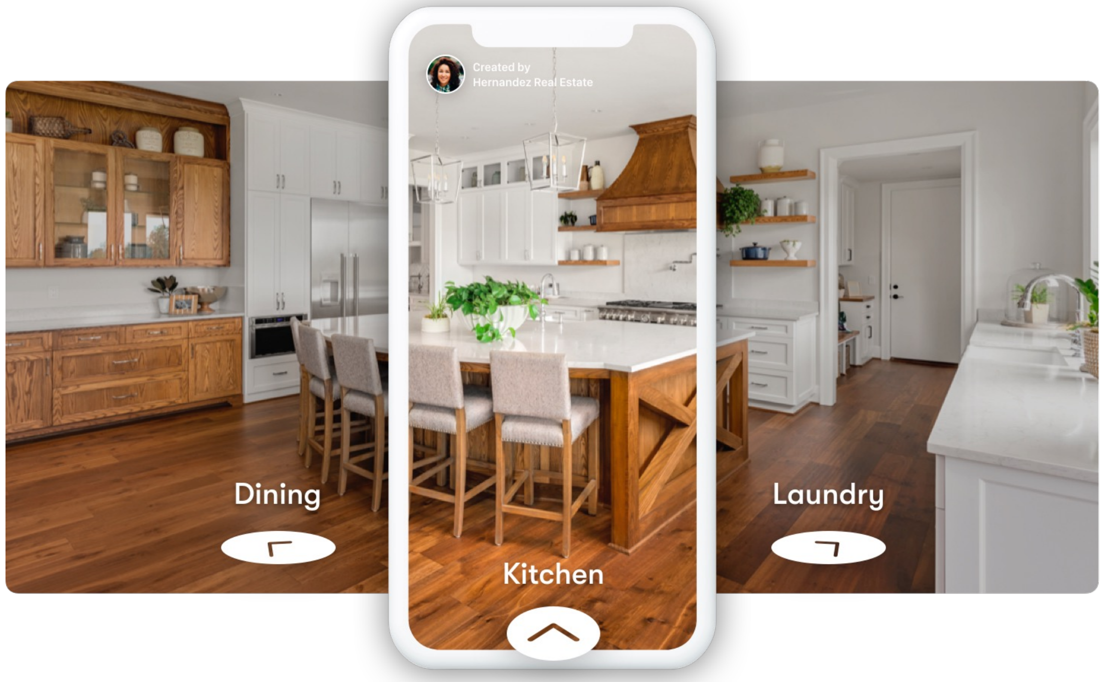
Figure 1: Real-time reconstructions of a moving scene with DynamicFusion; both the person and the camera are moving. The initially noisy and incomplete model is progressively denoised and completed over time (left to right).

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

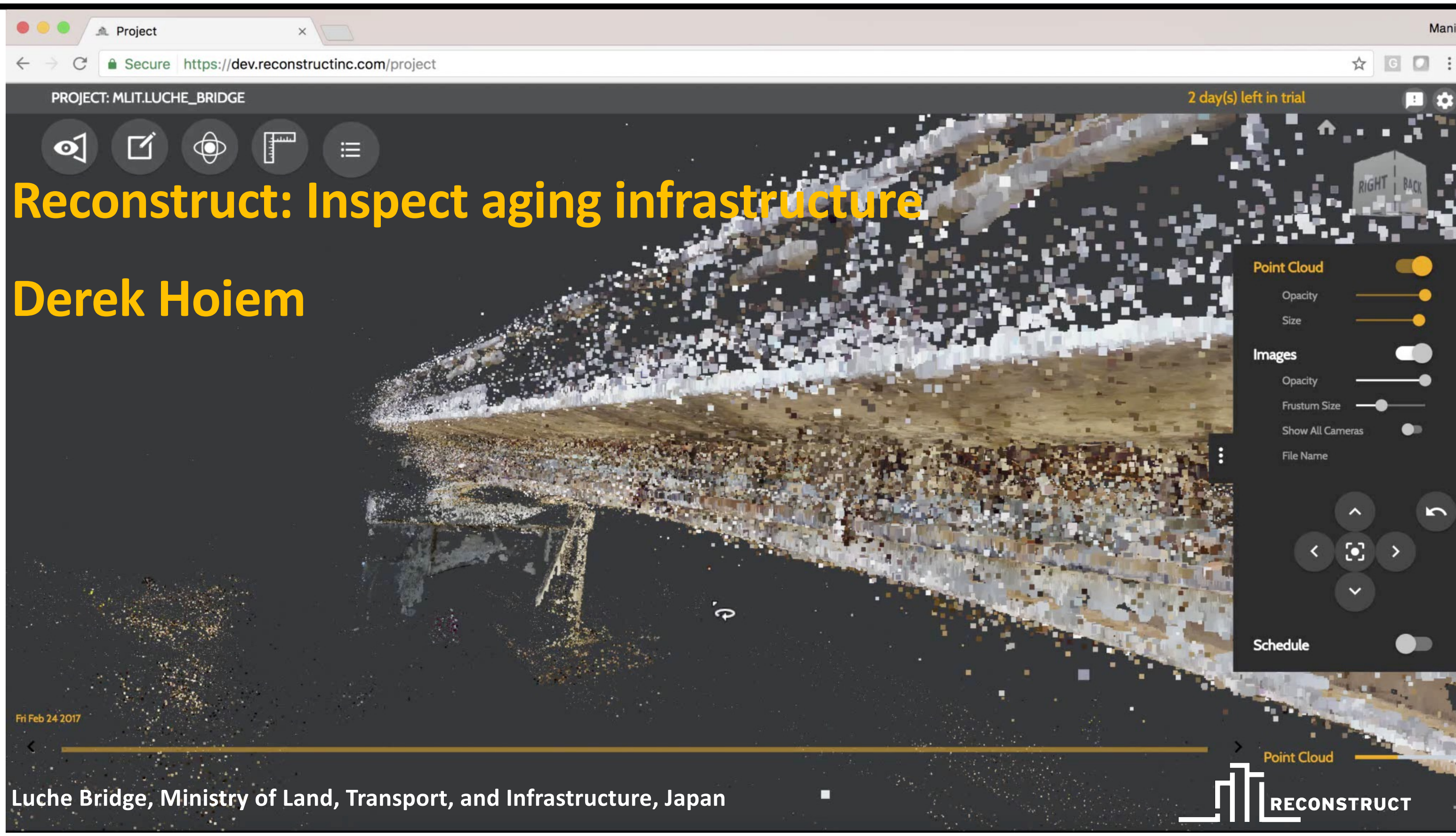
Make your listing pop with Zillow 3D Home® tours



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

Reconstruct: Inspect aging infrastructure

Derek Hoiem



Luche Bridge, Ministry of Land, Transport, and Infrastructure, Japan

Reconstruct: Align reality to plans for construction management

Derek Hoiem

Reconstruction: Commercial applications

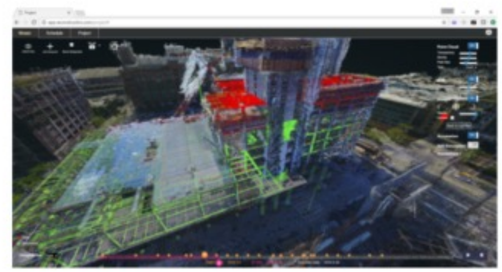
RECONSTRUCT INTEGRATES REALITY AND PLAN

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

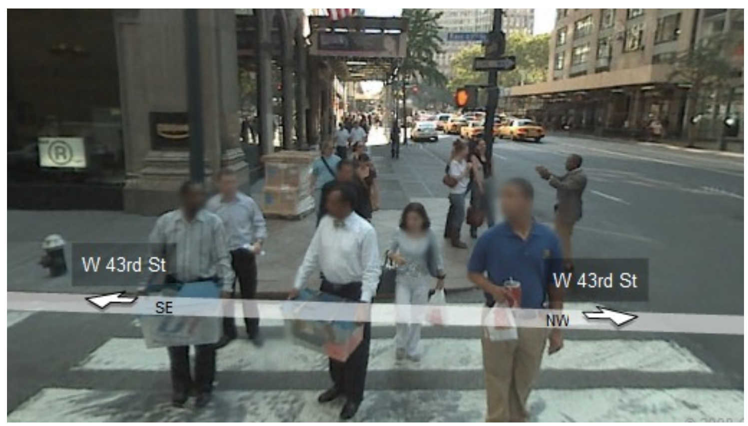
Analyze actual progress deviations by comparing Reality and Plan and predict risk with respect to the execution of the look-ahead schedule for each project location, to offer your project team with an opportunity to tap off potential delays before they surface on your jobsite.

reconstructinc.com

Source: D. Hoiem

Recognition: “Simple” patterns

Recognition: Faces



Recognition: Faces

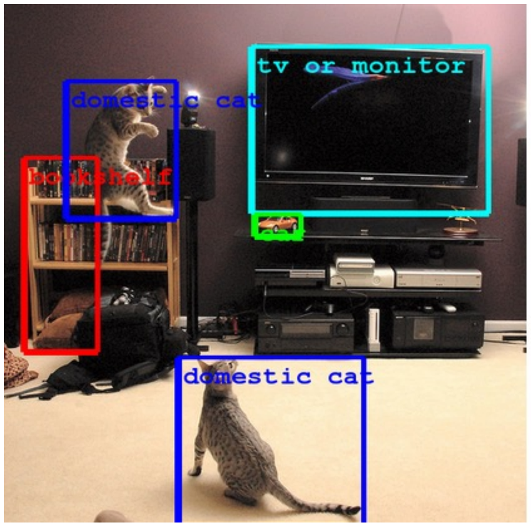
How China Uses High-Tech Surveillance to Subdue Minorities – New York Times, 5/22/2019

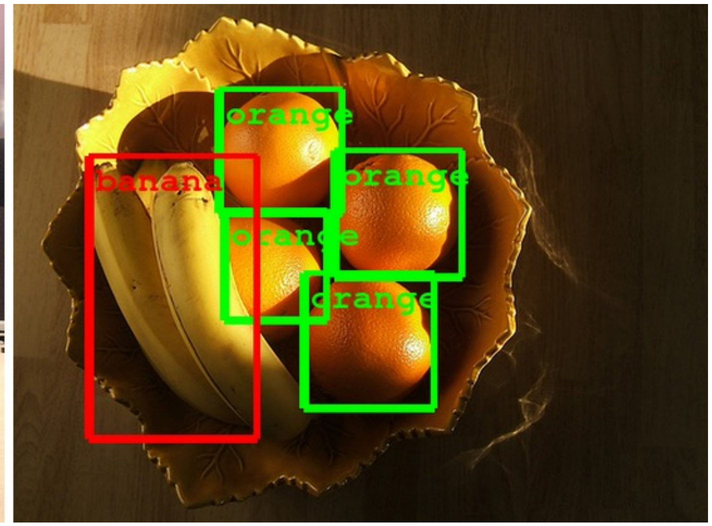
The Secretive Company That Might End Privacy As We Know It – New York Times, 1/18/2020

Wrongfully Accused by an Algorithm – New York Times, 6/24/2020

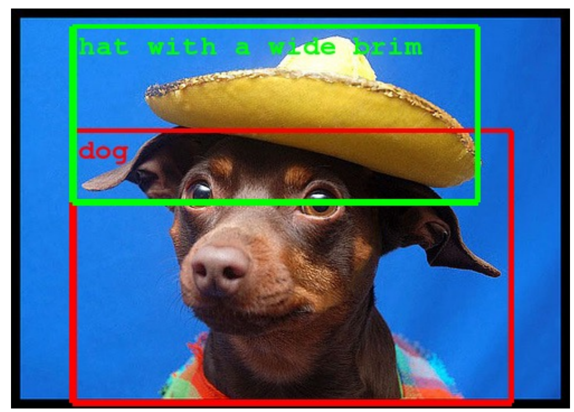
Facial Recognition Goes to War – New York Times, 4/7/2022

Recognition: General categories



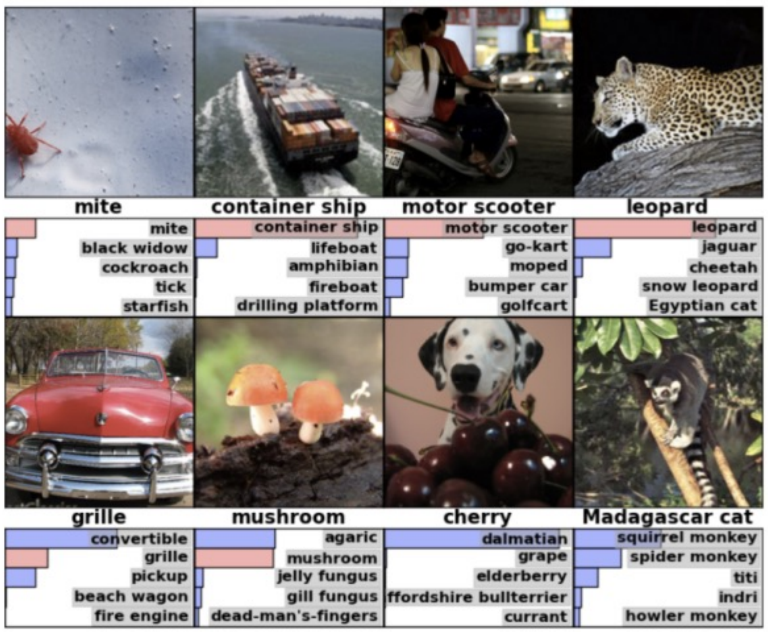


- [Computer Eyesight Gets a Lot More Accurate](), NY Times Bits blog, August 18, 2014
- [Building A Deeper Understanding of Images](), Google Research Blog, September 5, 2014



Recognition: General categories

ILSVRC



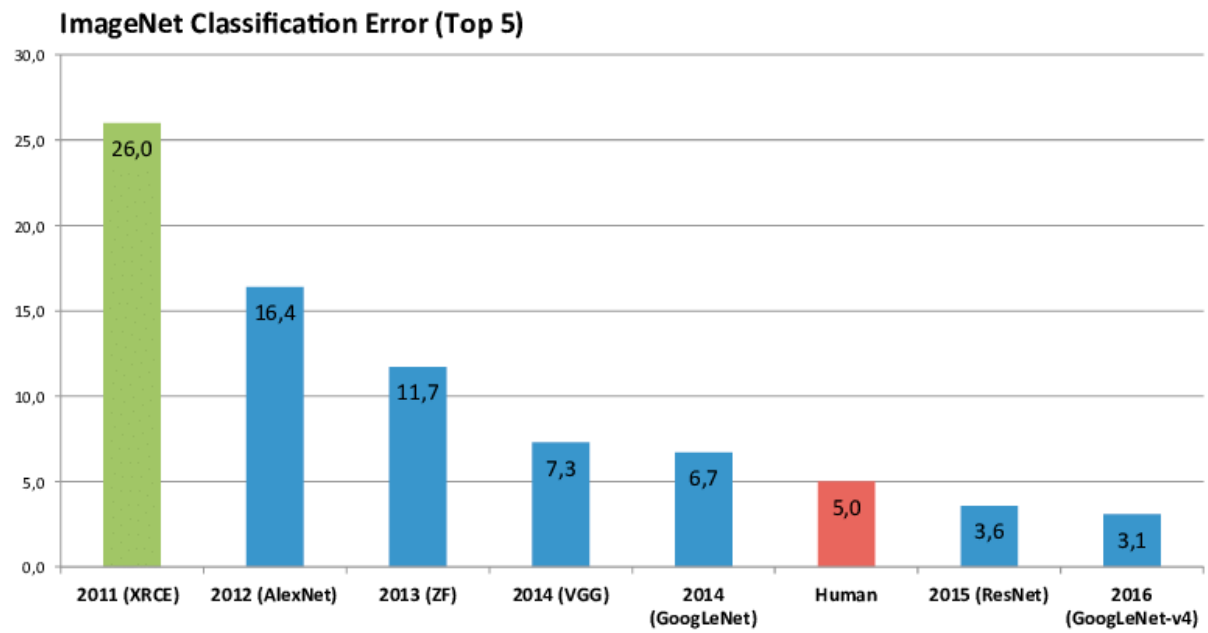
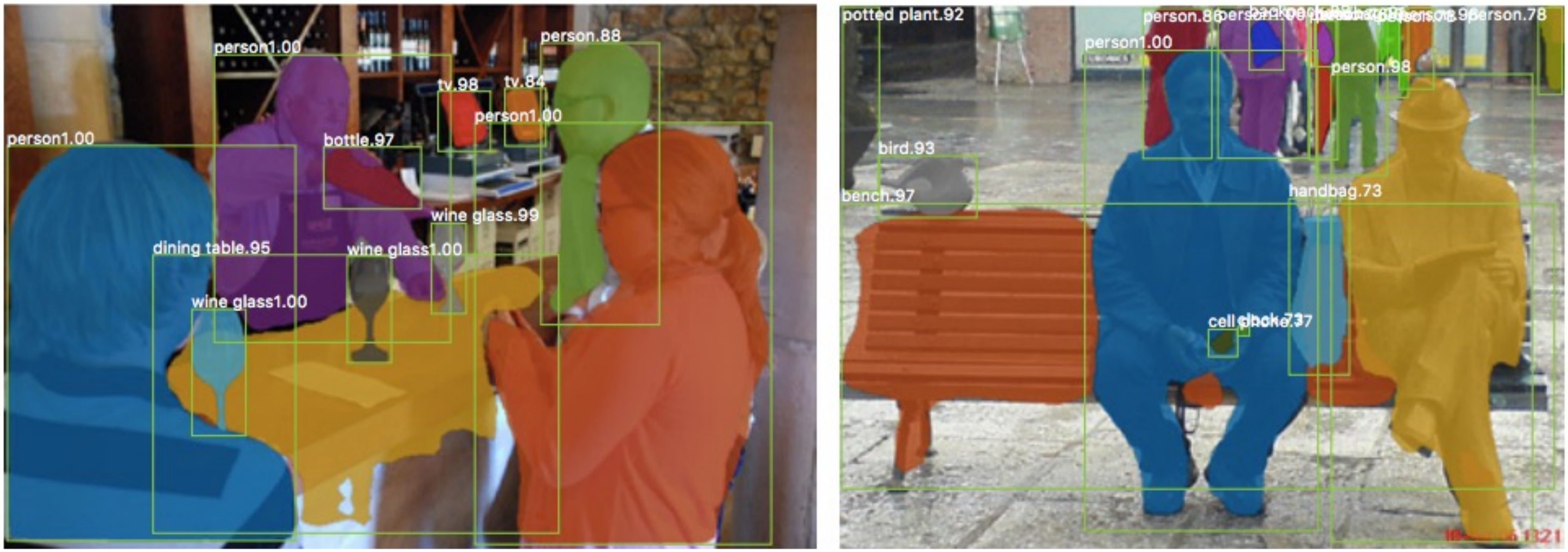


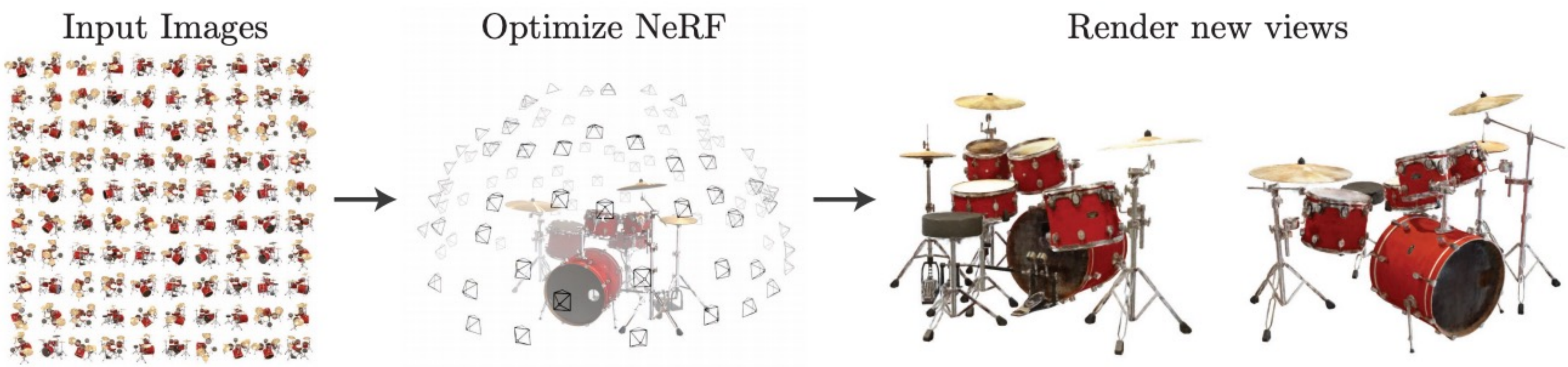
Figure source

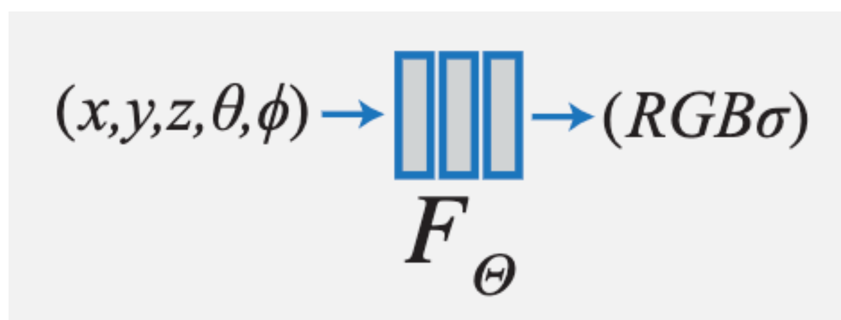
Object detection, instance segmentation



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

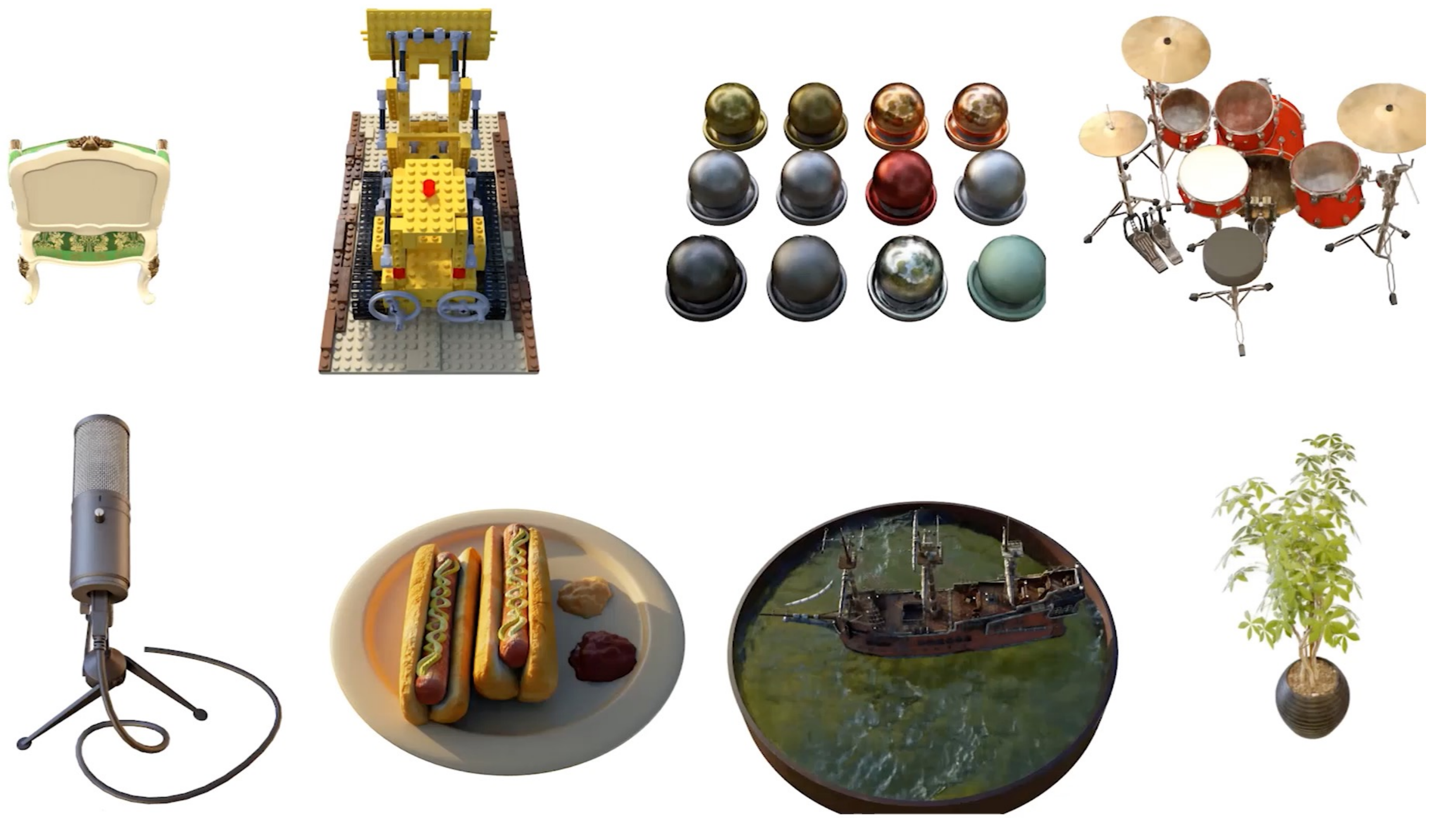
3D scene understanding: NERFs





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

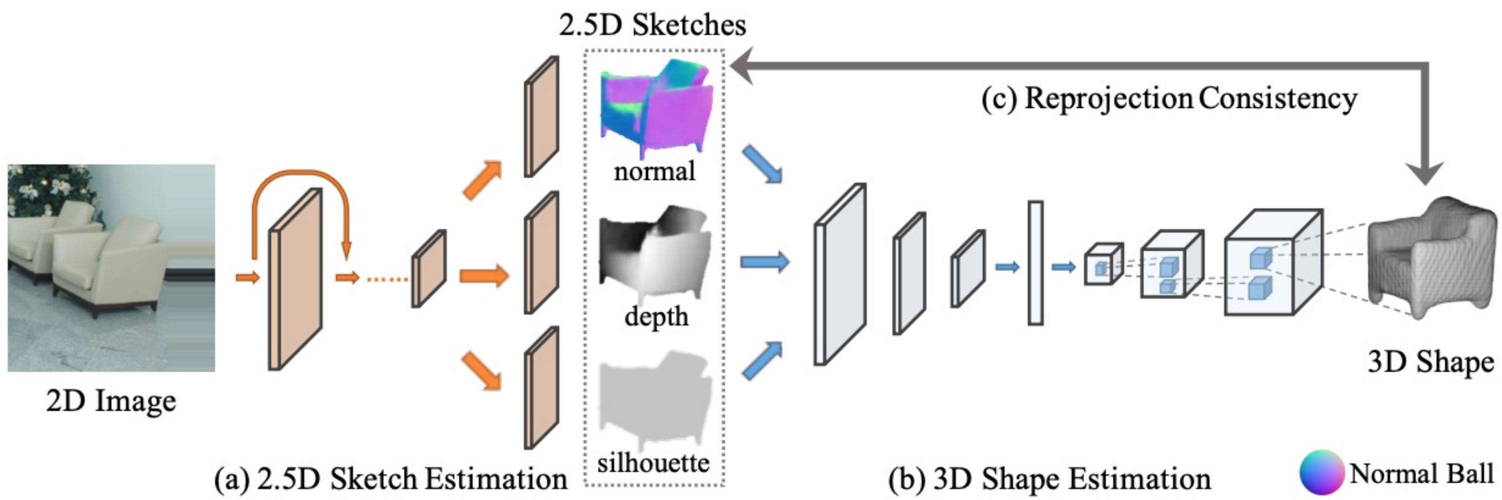


Figure 2: Our model (MarrNet) has three major components: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) a loss function for reprojection consistency. MarrNet first recovers object normal, depth, and silhouette images from an RGB image. It then regresses the 3D shape from the 2.5D sketches. 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.

J. Wu, Y. Wang, T. Xue, X. Sun, W. Freeman, J. Tenenbaum, MarrNet: 3D Shape Reconstruction via 2.5D Sketches, NeurIPS 2017

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

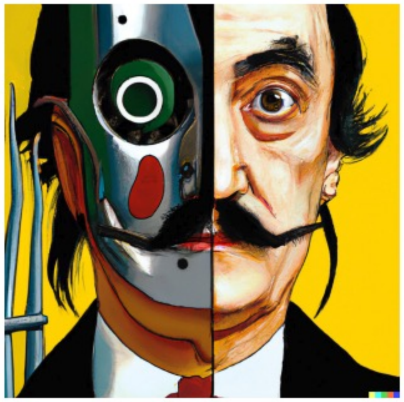
https://screenrant.com/star-wars-han-solo-movie-harrison-ford-video-deepfake/

https://www.youtube.com/watch?v=bC3uH4Xw4Xo

Just a random recent example…

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

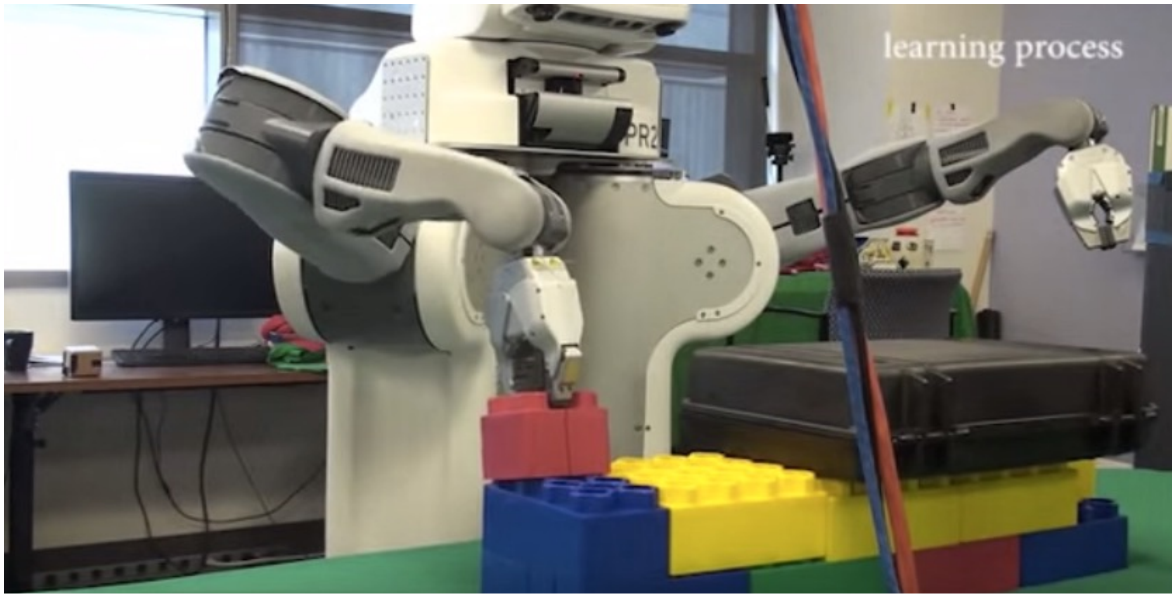
panda mad scientist mixing sparkling chemicals, artstation

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

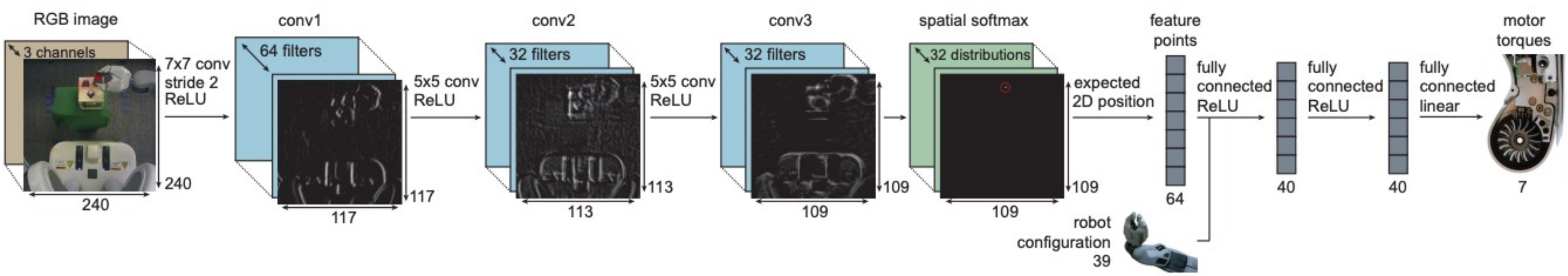
A. Ramesh et al., Zero-Shot Text-to-Image Generation, ICML 2021. https://openai.com/blog/dall-e/

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

Vision for action: Visuomotor learning

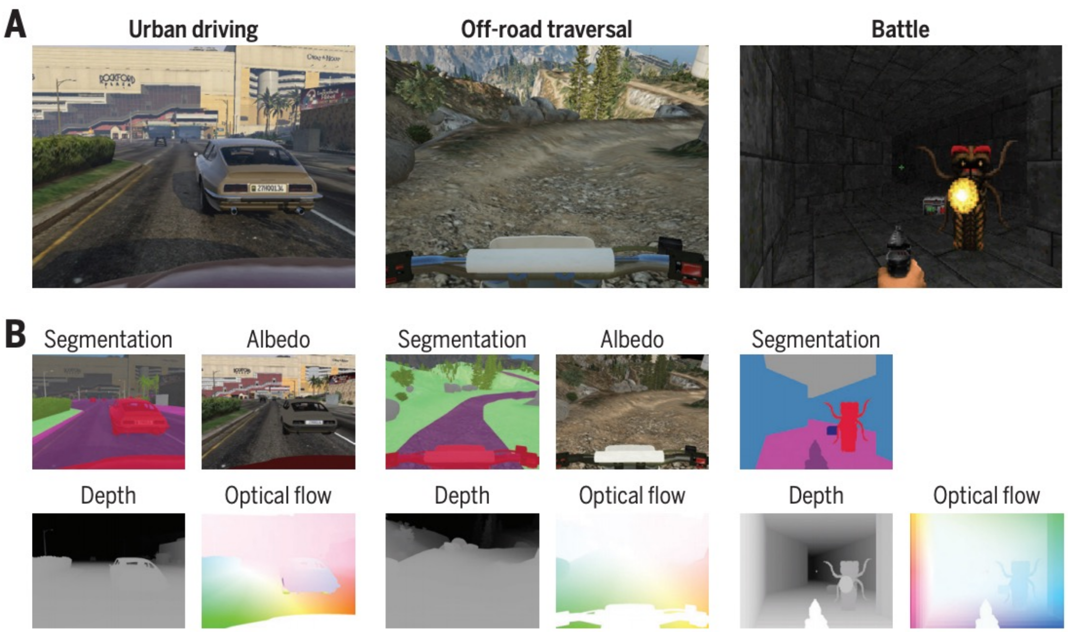


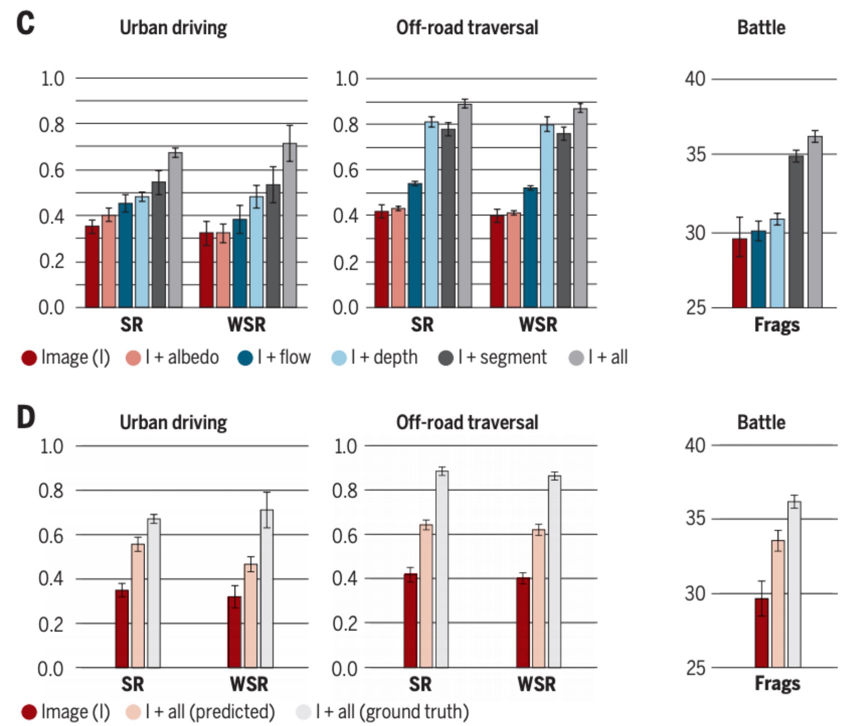
[Overview video](), [training video]()



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

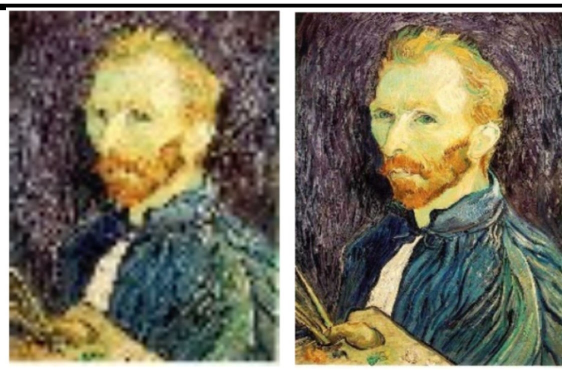
IV. Image formation and geometric vision

Cameras, Light, Color, Calibration

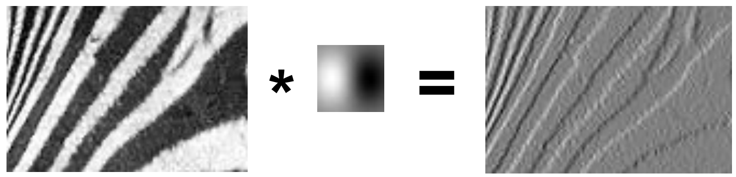
V. Pairs of Cameras and more:

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

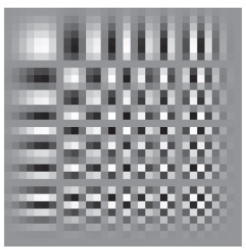
Elementary image representations



Basic image processing

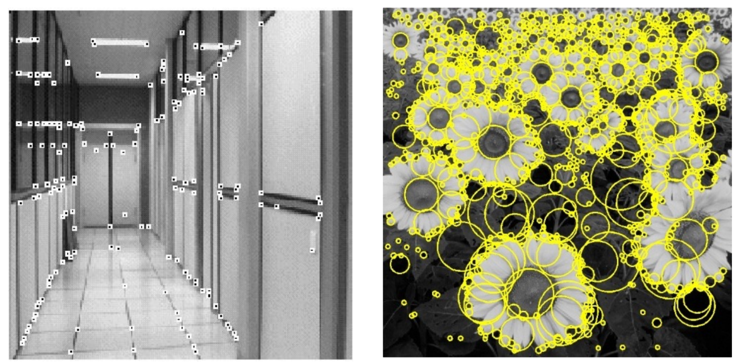


Linear filtering
Edge detection



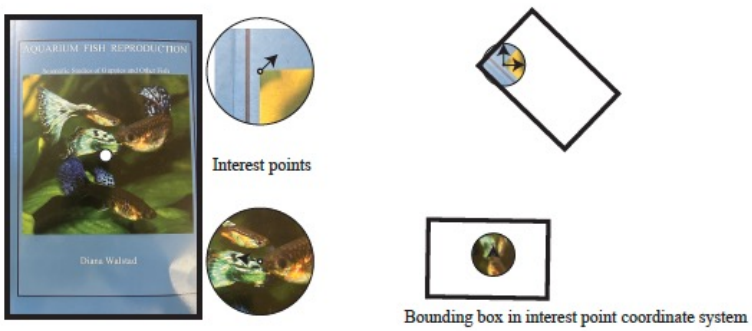
Upsampling and downsampling

downsampled by 8

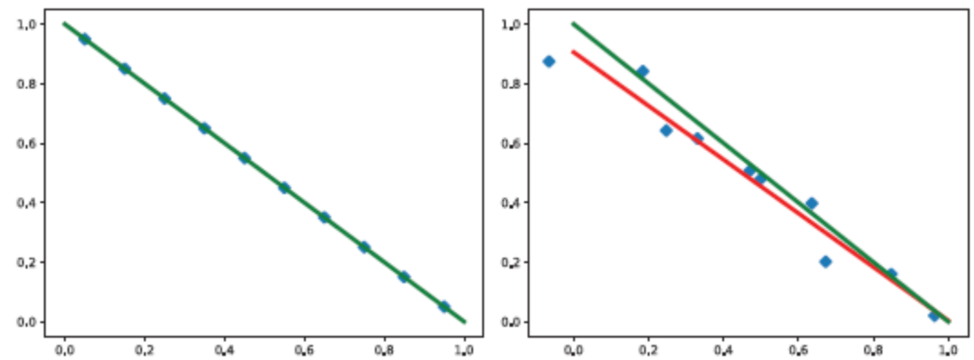


Feature extraction

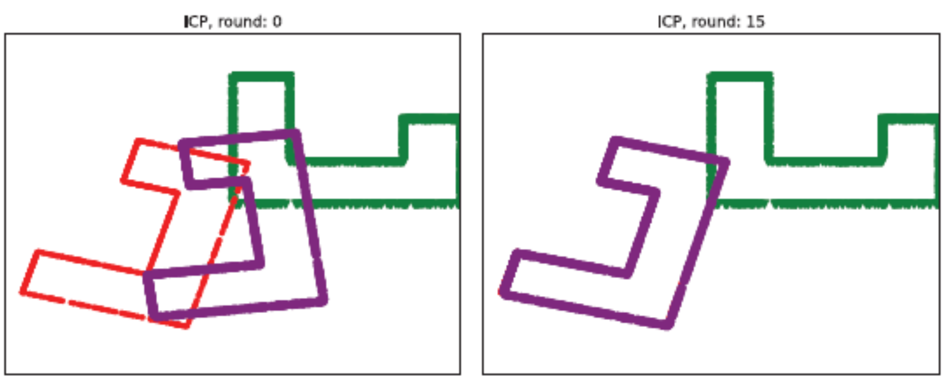
Mid-level Vision



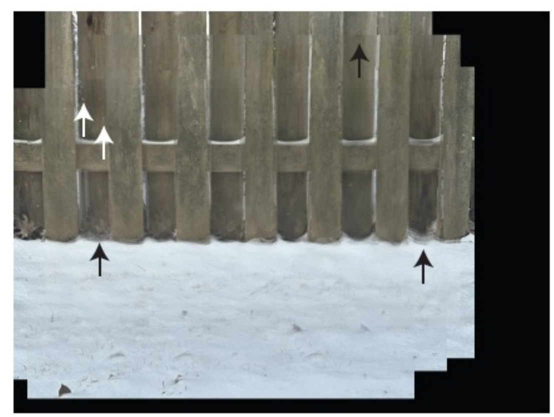
Voting



Fitting



Registration

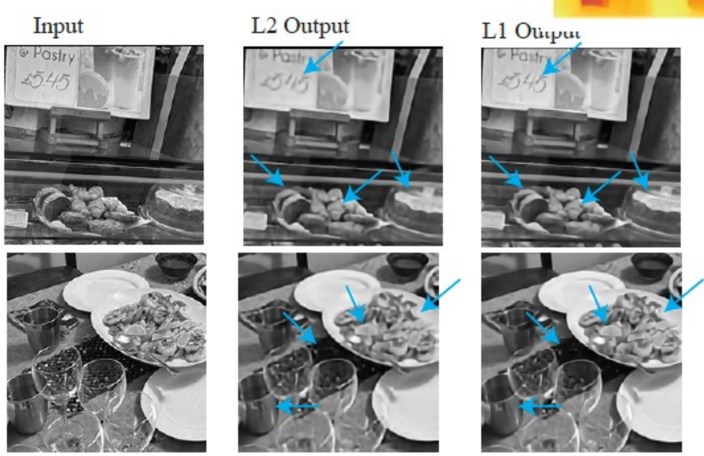


Mosaics

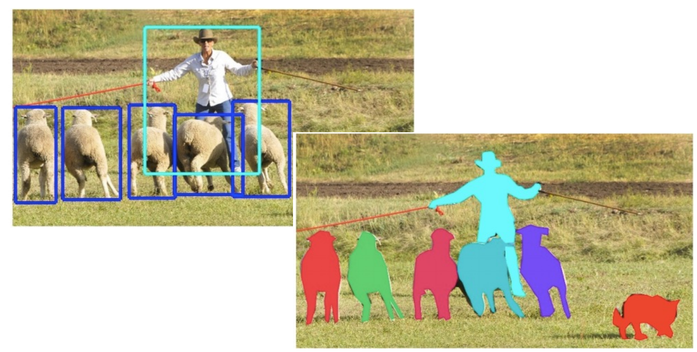
Learned Image Representations

Depth, normal, etc

From images

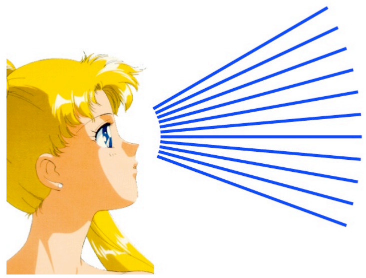


Learned denoising

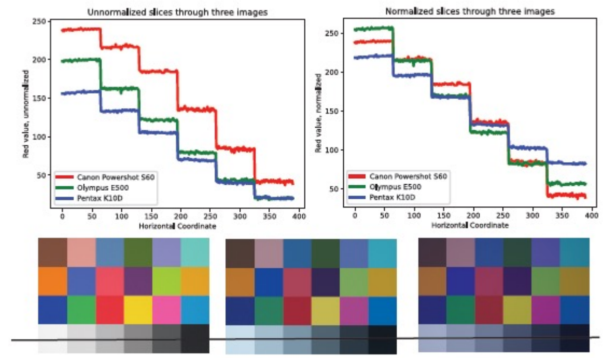


Object detection and segmentation

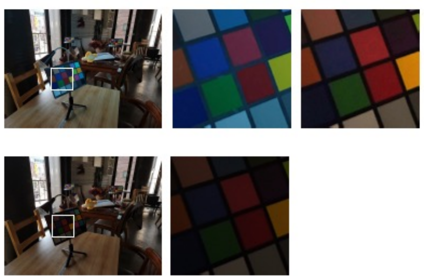
IV. Image formation and geometric vision



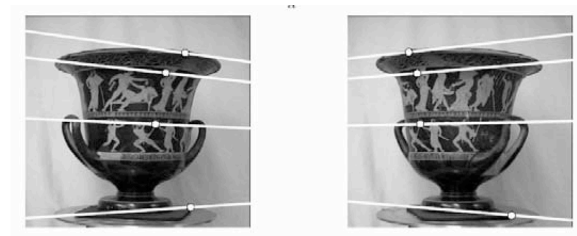
Cameras and sensors



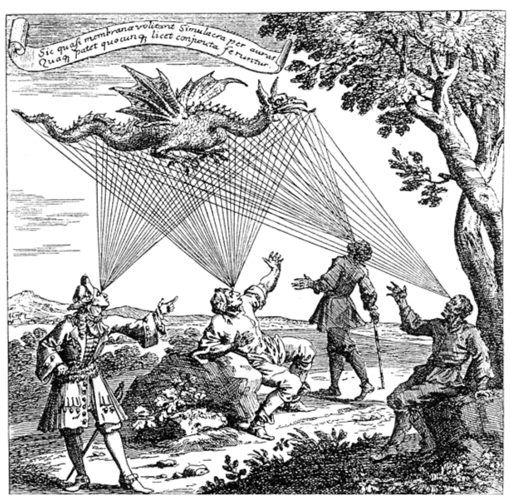
Canon Powershot S60 Olympus E500 Pentax K10D



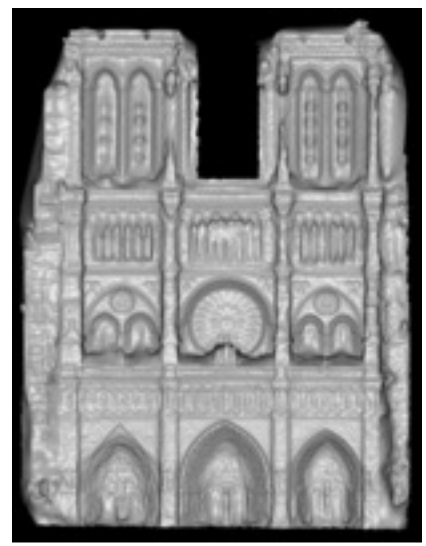
V. Pairs of cameras and more



Two-view geometry, stereo



Structure from motion



Multi-view stereo