Fall 2022 CS543 / ECE549
Computer Vision

Course webpage URL: http://slazebni.cs.illinois.edu/fall22

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

Goal: To extract useful information from pixels

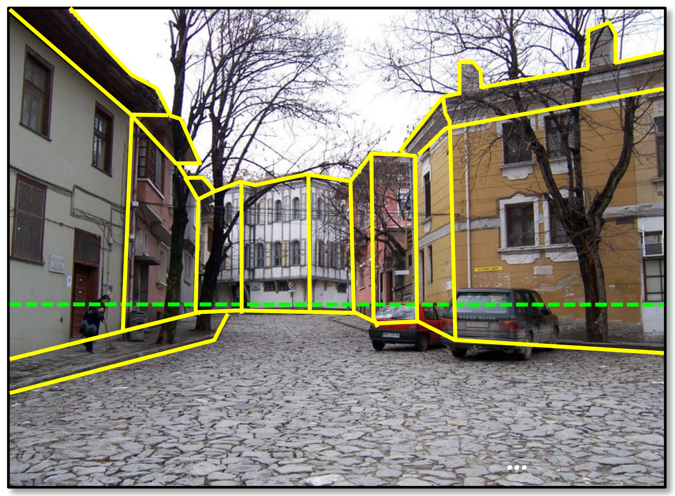
What we see

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

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*

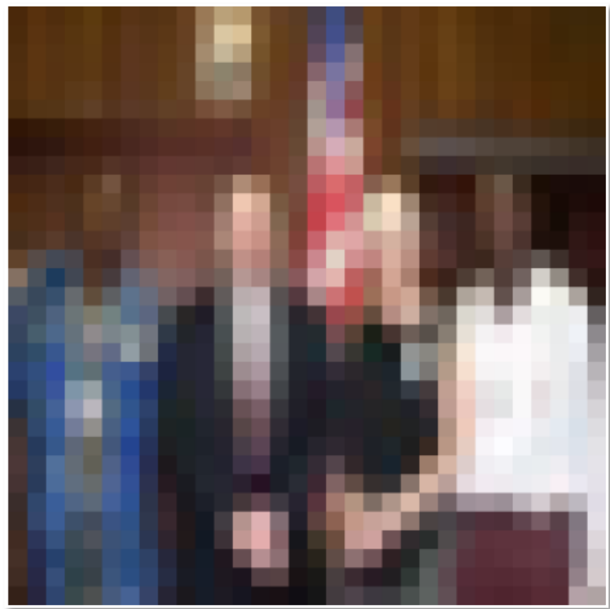
What kind of information can be extracted from an image?

Geometric information

Semantic information

Vision for action

Humans are remarkably good at vision…



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

…still, vision is hard even for humans

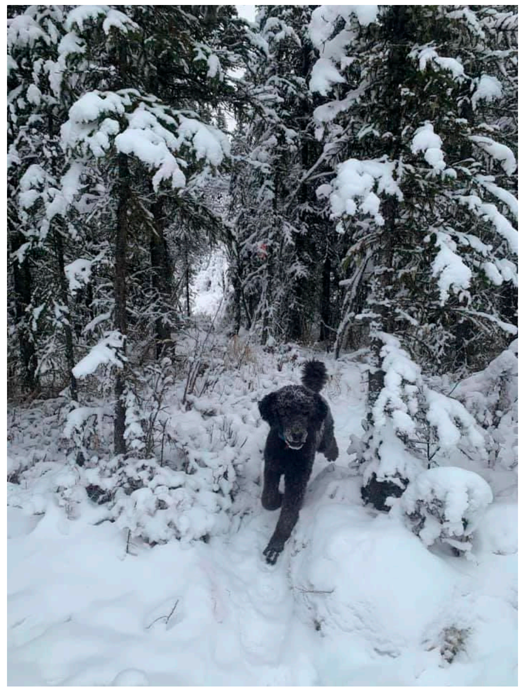


Image source

…still, vision is hard even for humans

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

…still, vision is hard even for humans

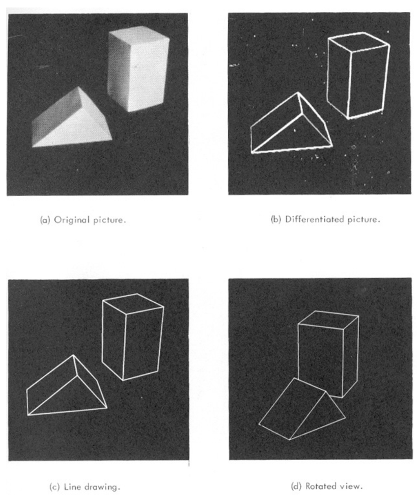
What color is this dress?

Images are fundamentally ambiguous!

Outline

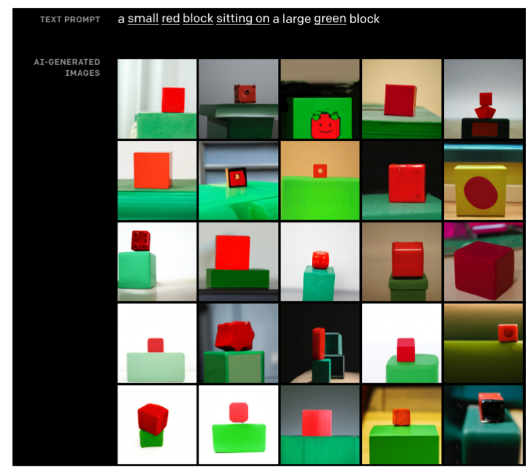
- Logistics, requirements
- Goal of computer vision and why it is hard
- History of computer vision

How it started



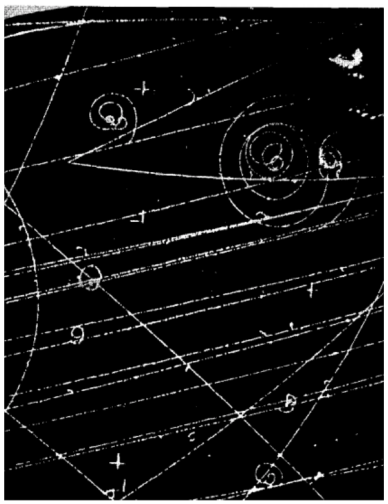
L. G. Roberts, 1963

How it's going

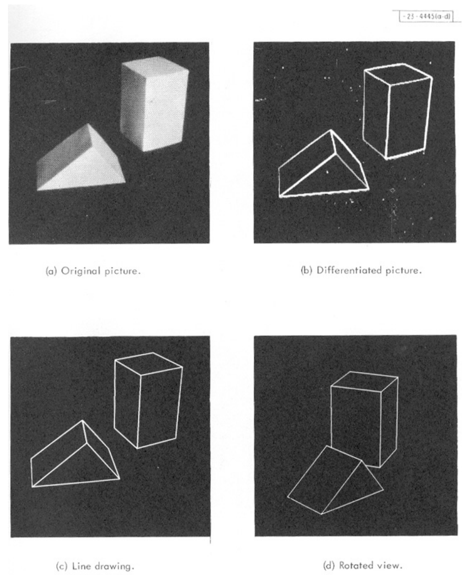


OpenAI DALL-E, 2020

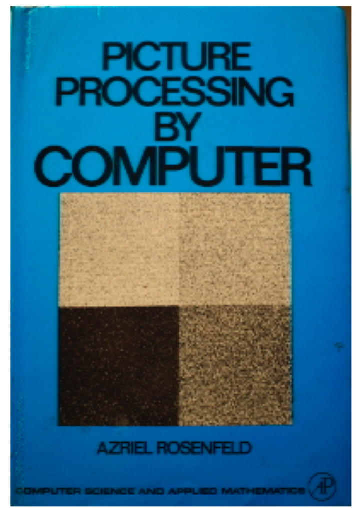
Origins



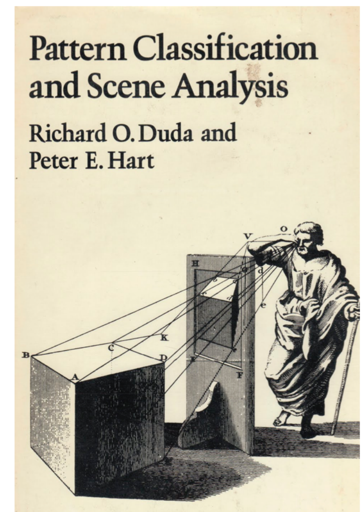
Hough, 1959



Roberts, 1963



Rosenfeld, 1969



Duda & Hart, 1972

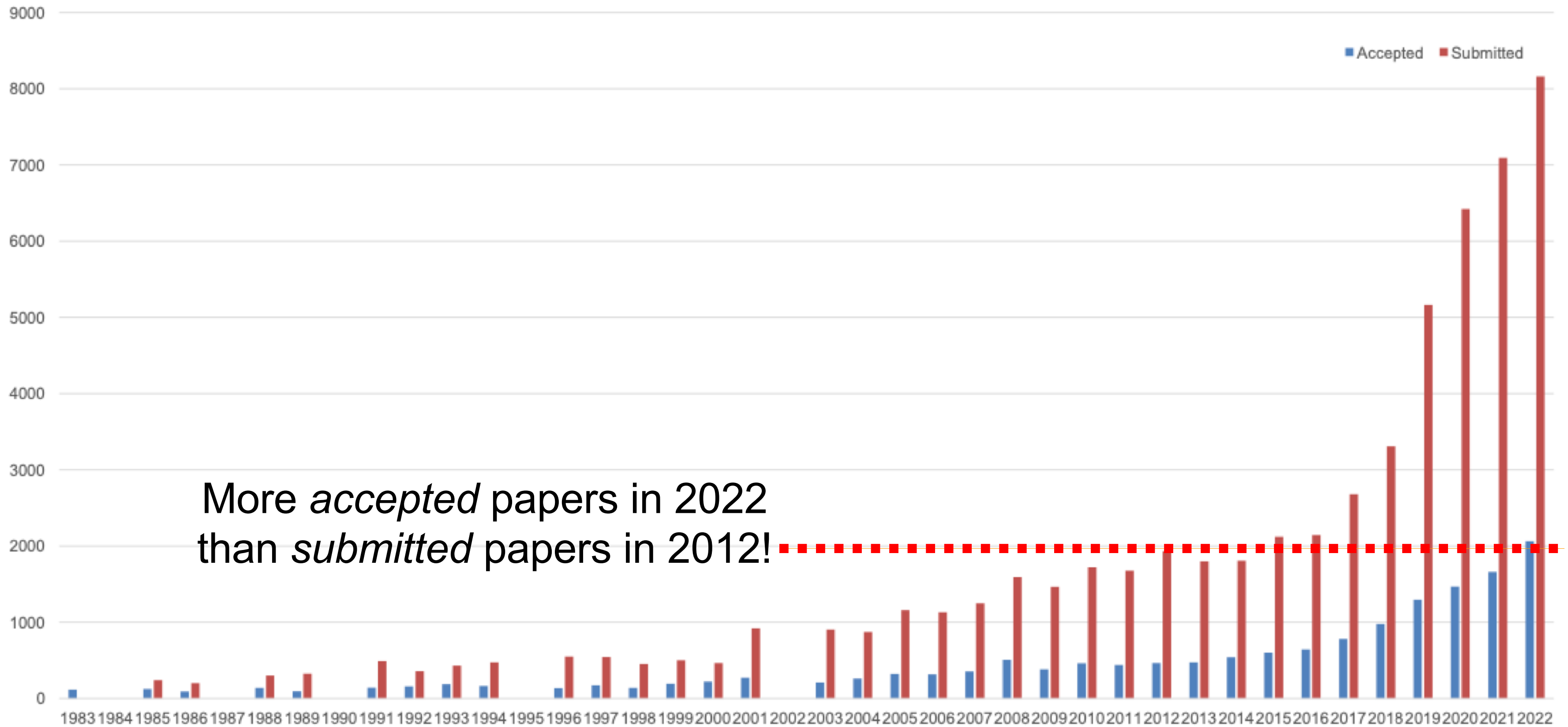
Decade by decade

- **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 my historical overview

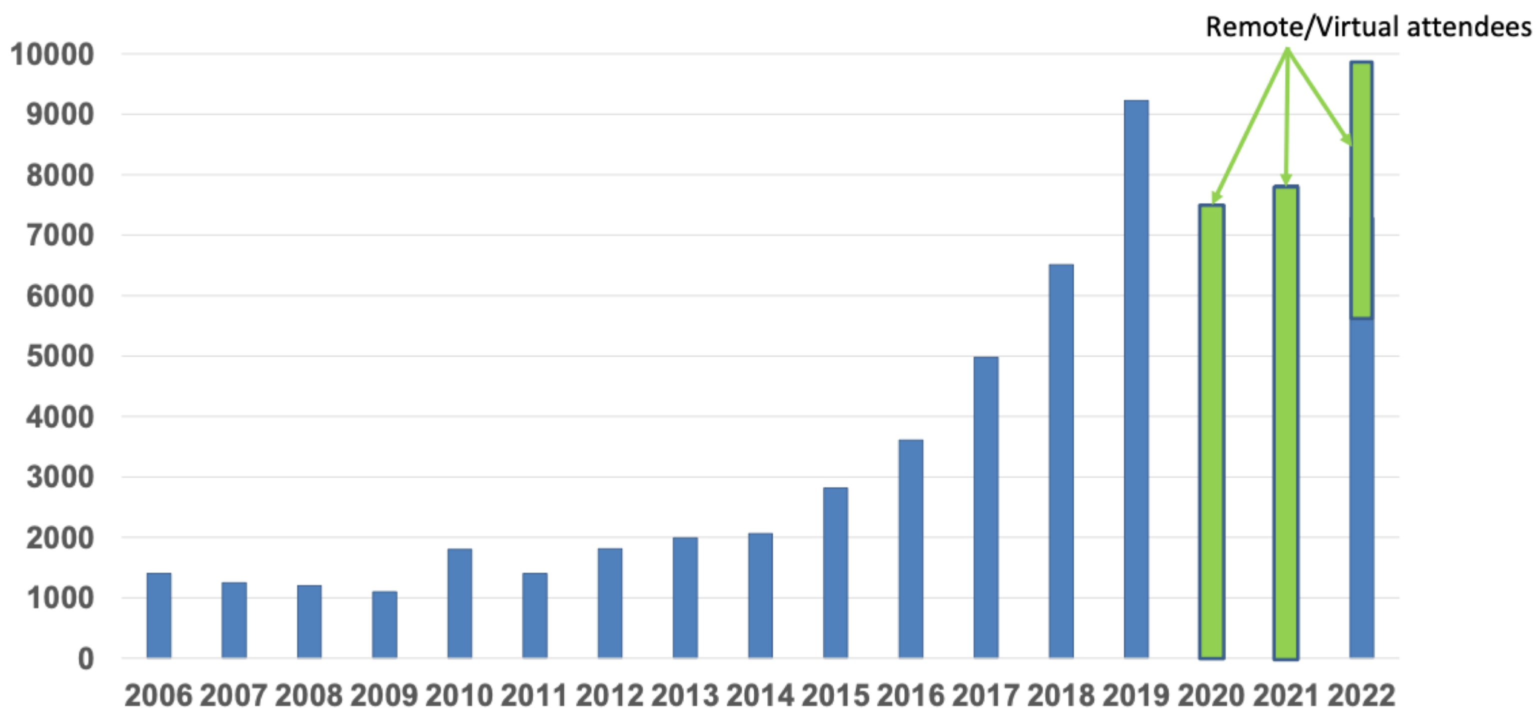
Adapted from J. Malik

Growth of the field: CVPR papers



Source: [CVPR 2022 opening sides]

Growth of the field: CVPR attendance

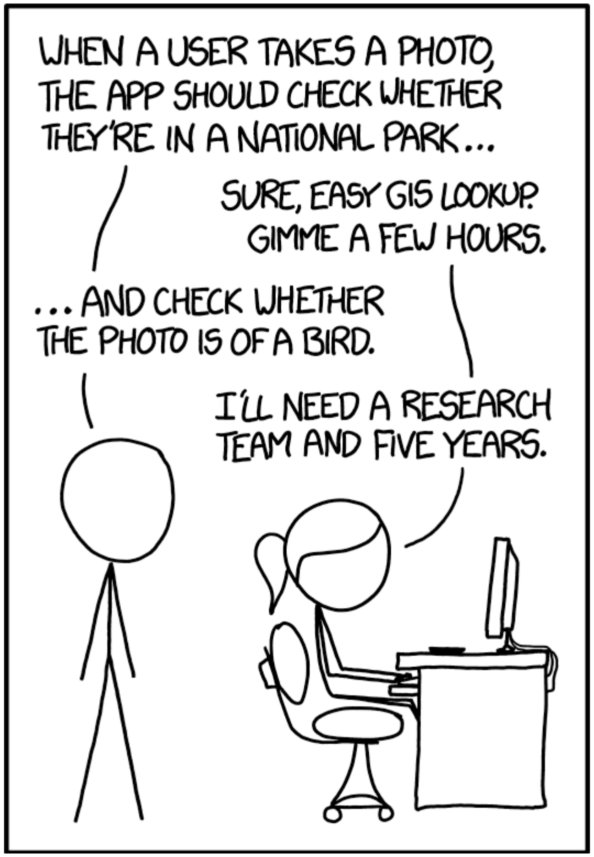


Source: [CVPR 2022 opening sides]

Introduction: Outline

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

What can computer vision do today?



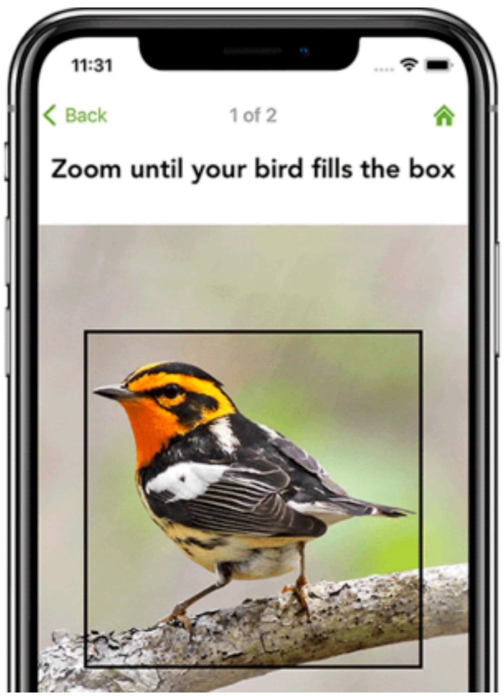
In the 60s, Marvin Minsky assigned a couple of undergrads to spend the summer programming a computer to use a camera to identify objects in a scene. He figured they'd have the problem solved by the end of the summer. Half a century later, we're still working on it.

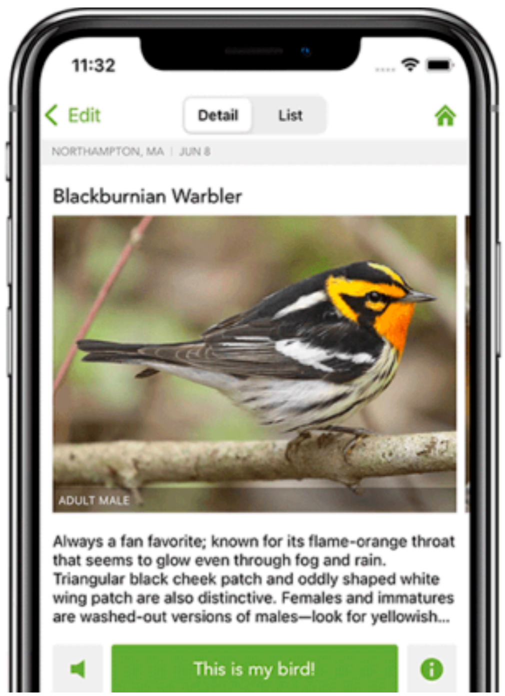
https://xkcd.com/1425/

(September 24, 2014)

What can computer vision do today?

- It’s 2022 now…





TheCornellLab

Merlin®

https://merlin.allaboutbirds.org/

What can computer vision do today?

- It’s 2022 now…

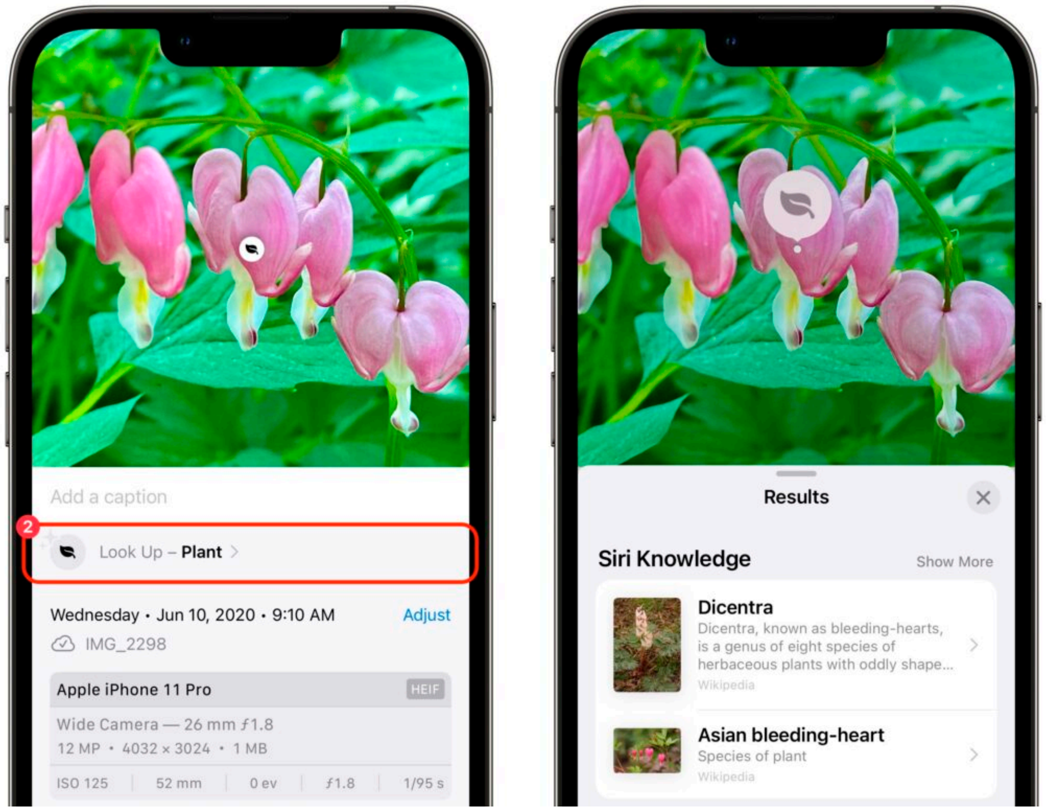


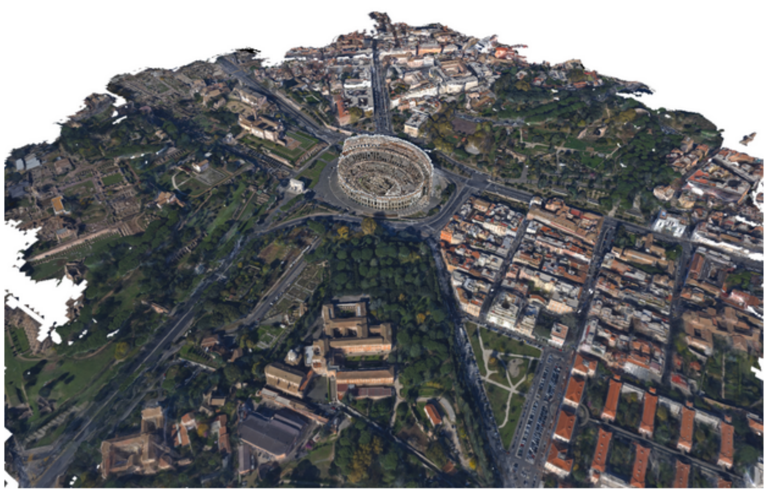
Image source

What can computer vision do today?

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

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

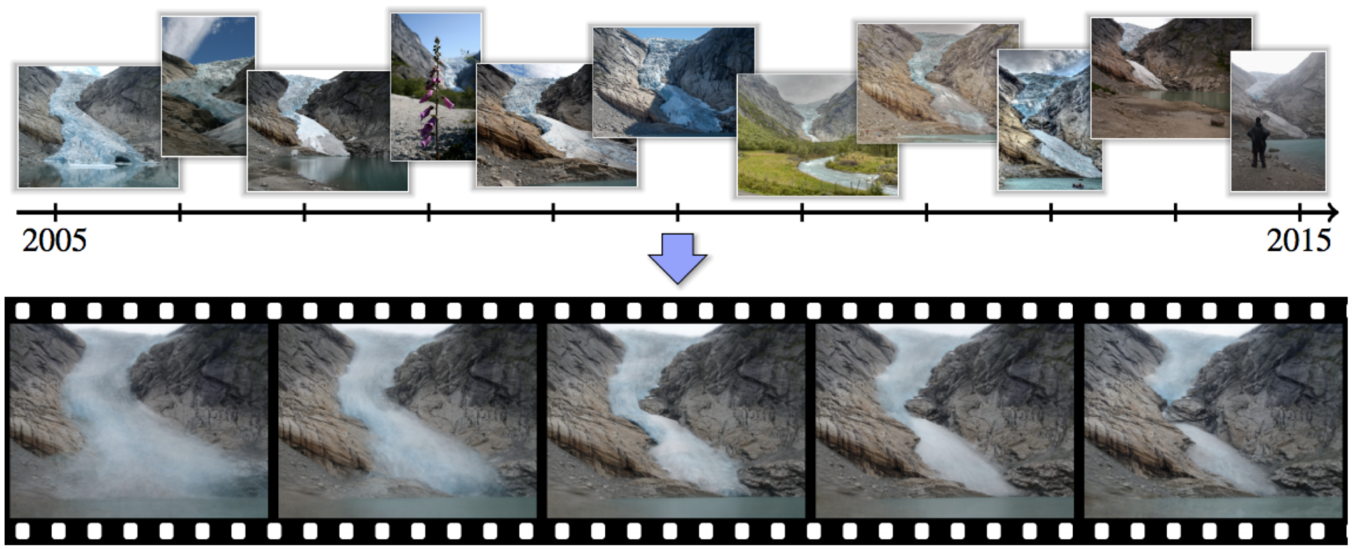


Figure 1: *We mine Internet photo collections to generate time-lapse videos of locations all over the world. Our time-lapses visualize a multitude of changes, like the retreat of the Briksdalsbreen Glacier in Norway shown above. The continuous time-lapse (bottom) is computed from hundreds of Internet 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

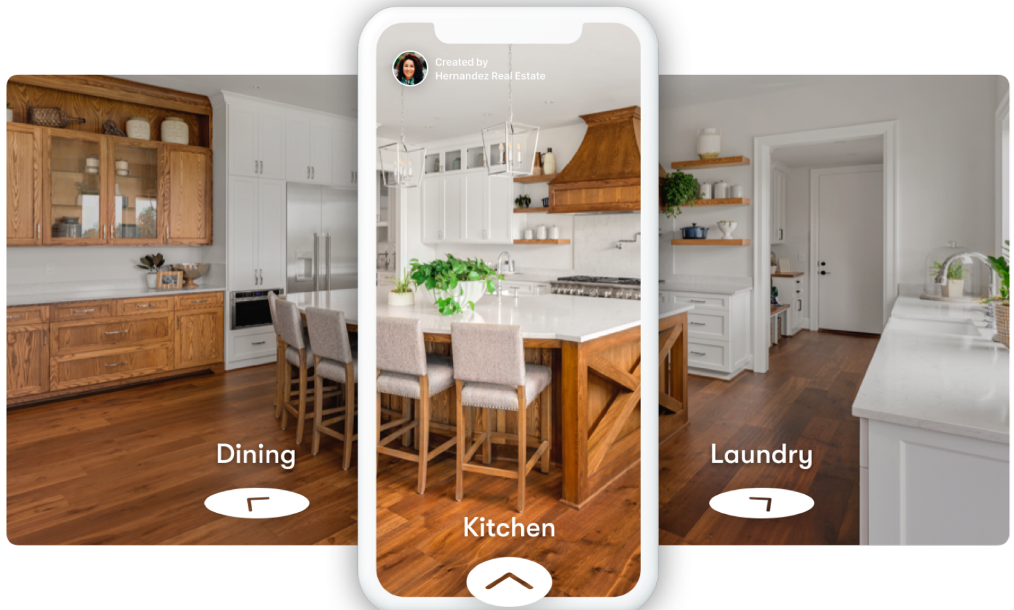
Figure 1: Real-time reconstructions of a moving scene with DynamicFusion; both the person and the camera are moving. The initially noisy and incomplete model is progressively denoised and completed over time (left to right).

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



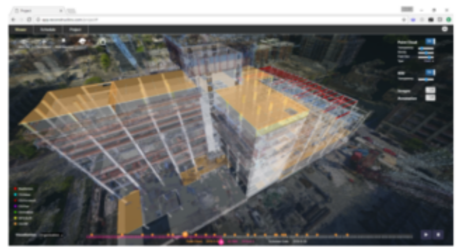
https://www.zillow.com/z/3d-home/

Reconstruction: Commercial applications

RECONSTRUCT INTEGRATES REALITY AND PLAN

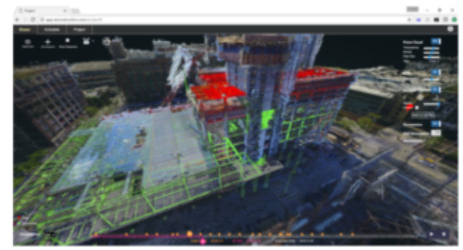
Visual Asset Management

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

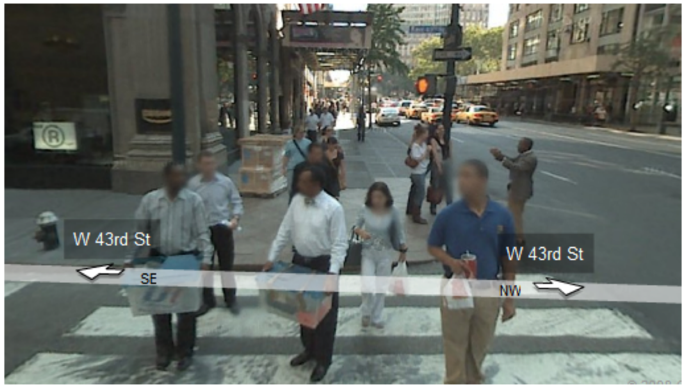
Analyze actual progress deviations by comparing Reality and Plan and predict risk with respect to the execution of the look-ahead schedule for each project location, to offer your project team with an opportunity to tap off potential delays before they surface on your jobsite.

reconstructinc.com

Source: D. Hoiem

Recognition: “Simple” patterns

Recognition: Faces



Recognition: Faces

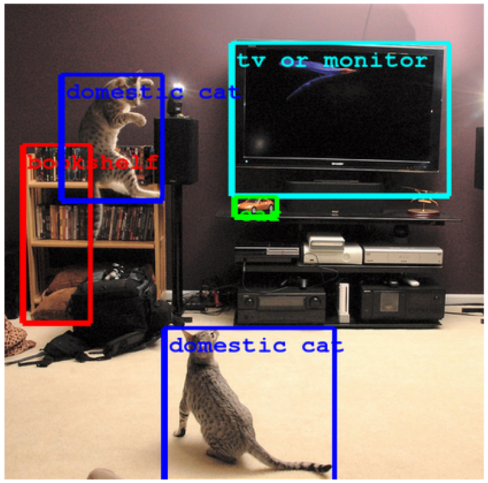
How China Uses High-Tech Surveillance to Subdue Minorities – New York Times, 5/22/2019

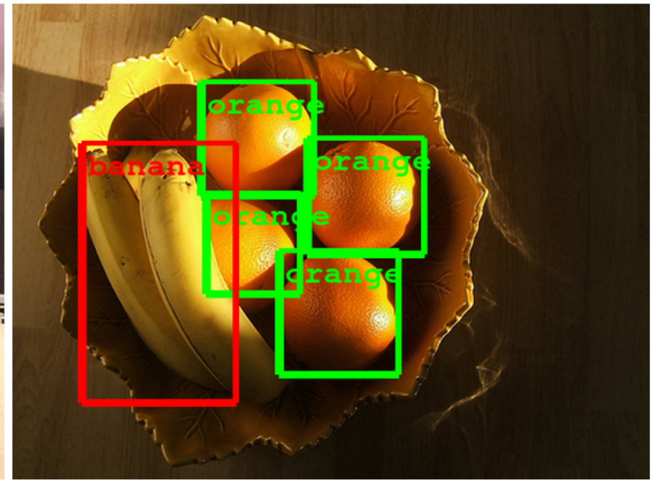
The Secretive Company That Might End Privacy As We Know It – New York Times, 1/18/2020

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

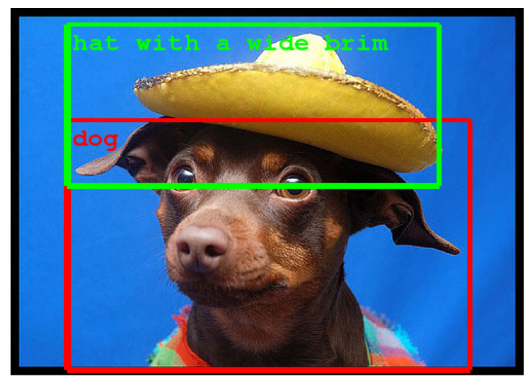
Facial Recognition Goes to War – New York Times, 4/7/2022

Recognition: General categories



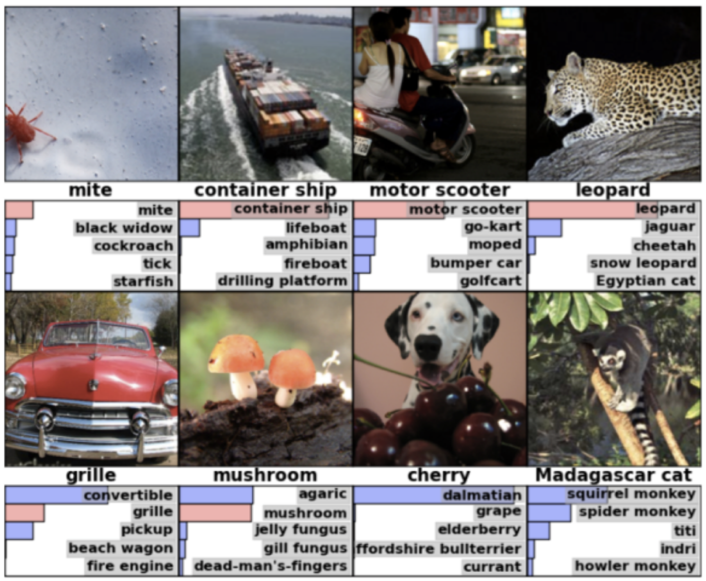


- [Computer Eyesight Gets a Lot More Accurate](), NY Times Bits blog, August 18, 2014
- [Building A Deeper Understanding of Images](), Google Research Blog, September 5, 2014



Recognition: General categories

ILSVRC



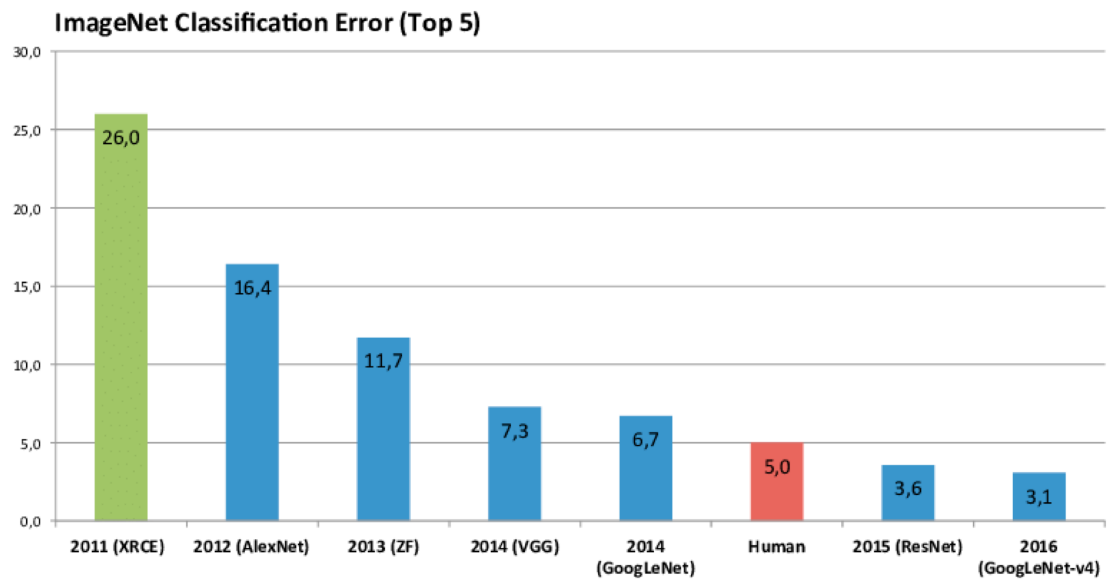
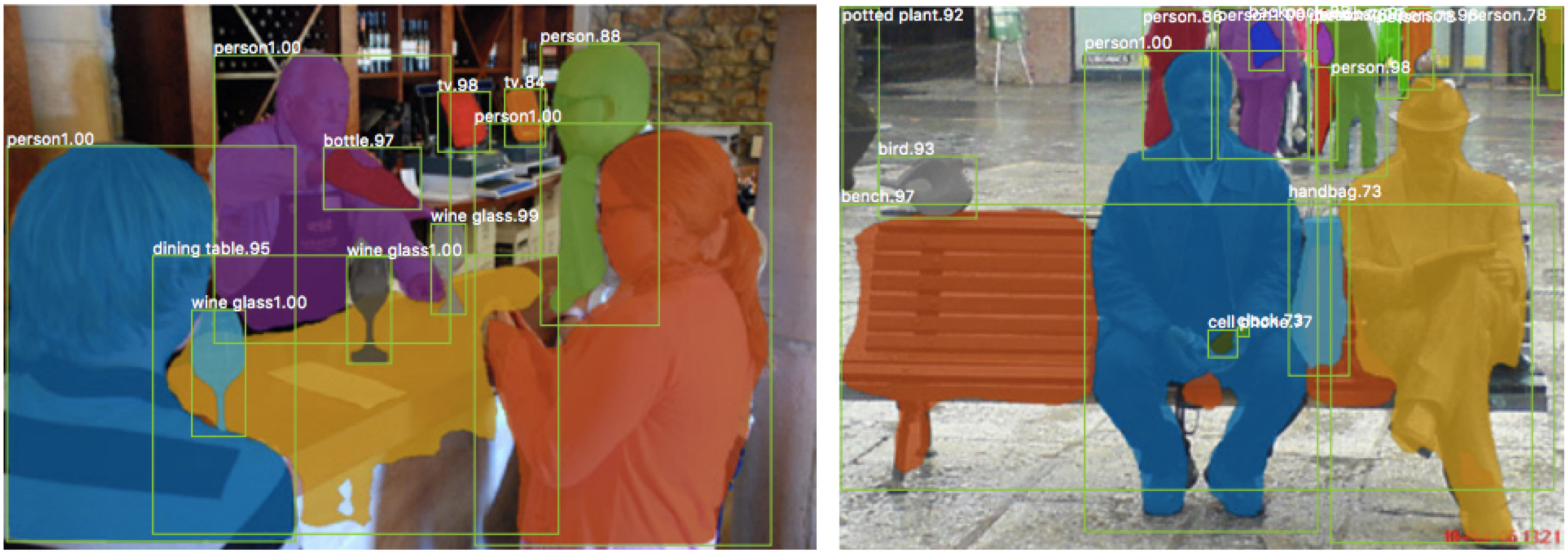


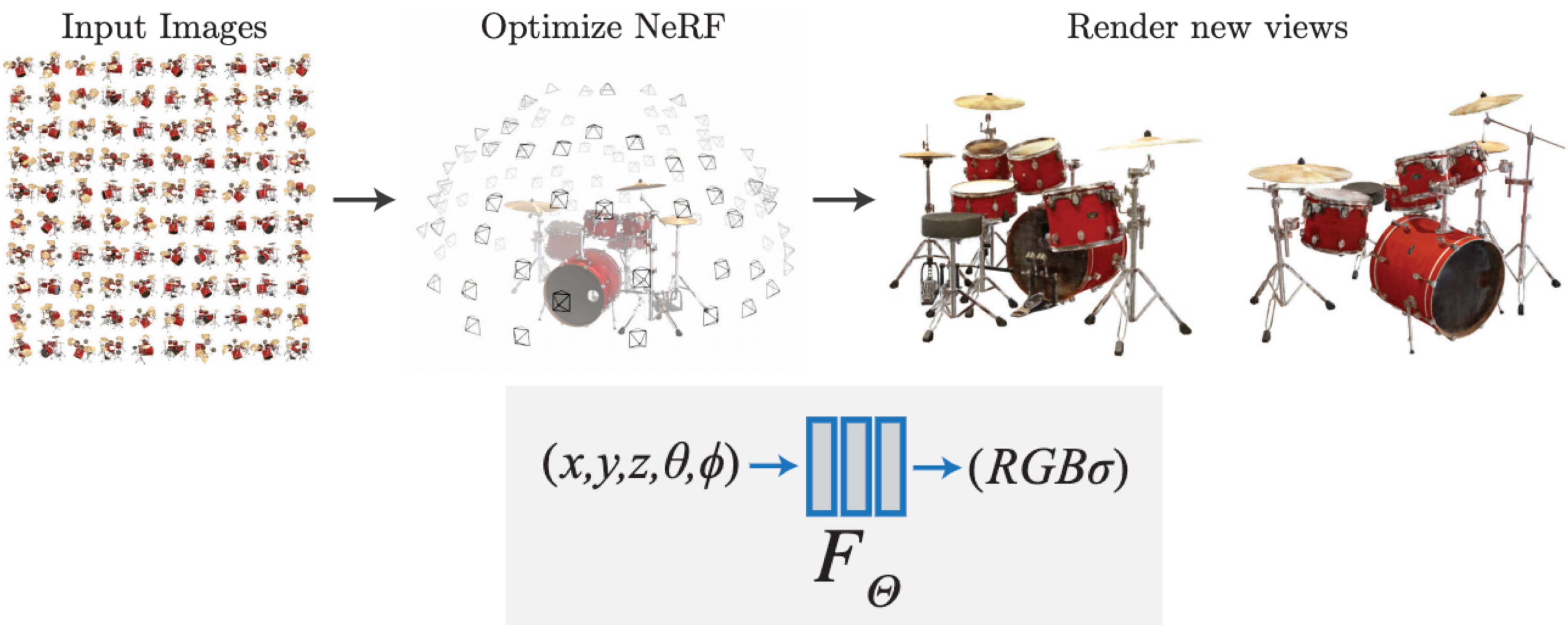
Figure source

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

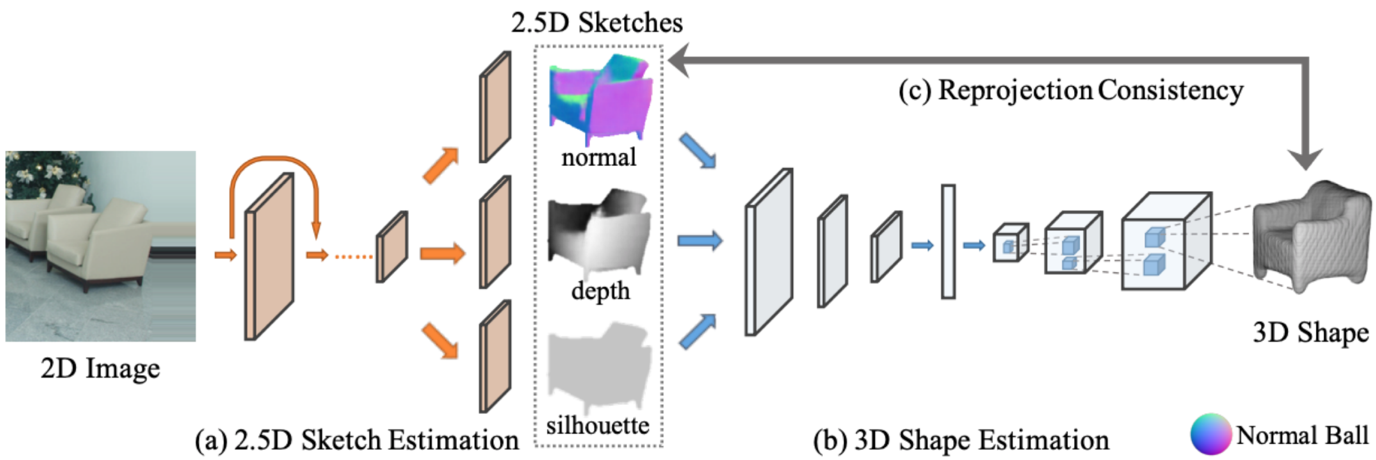


Figure 2: Our model (MarrNet) has three major components: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) a loss function for reprojection consistency. MarrNet first recovers object normal, depth, and silhouette images from an RGB image. It then regresses the 3D shape from the 2.5D sketches.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

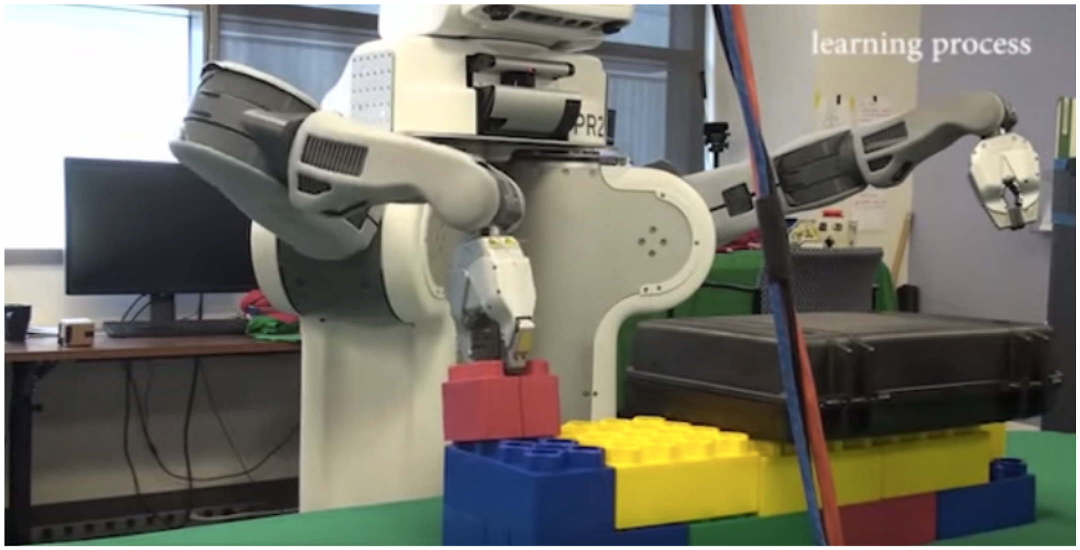
panda mad scientist mixing sparkling chemicals, artstation

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/

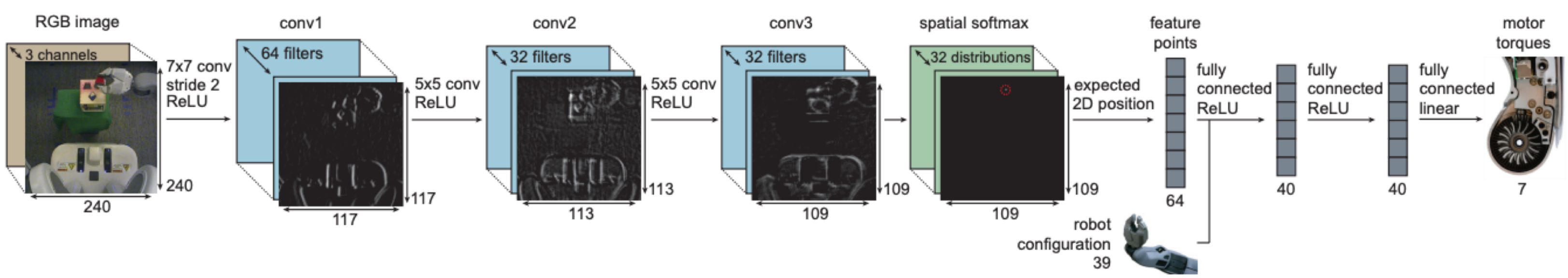
A. Ramesh et al., [Hierarchical Text-Conditional Image Generation with CLIP Latents](), arXiv 2022. https://openai.com/dall-e-2/

Vision for action: Visuomotor learning



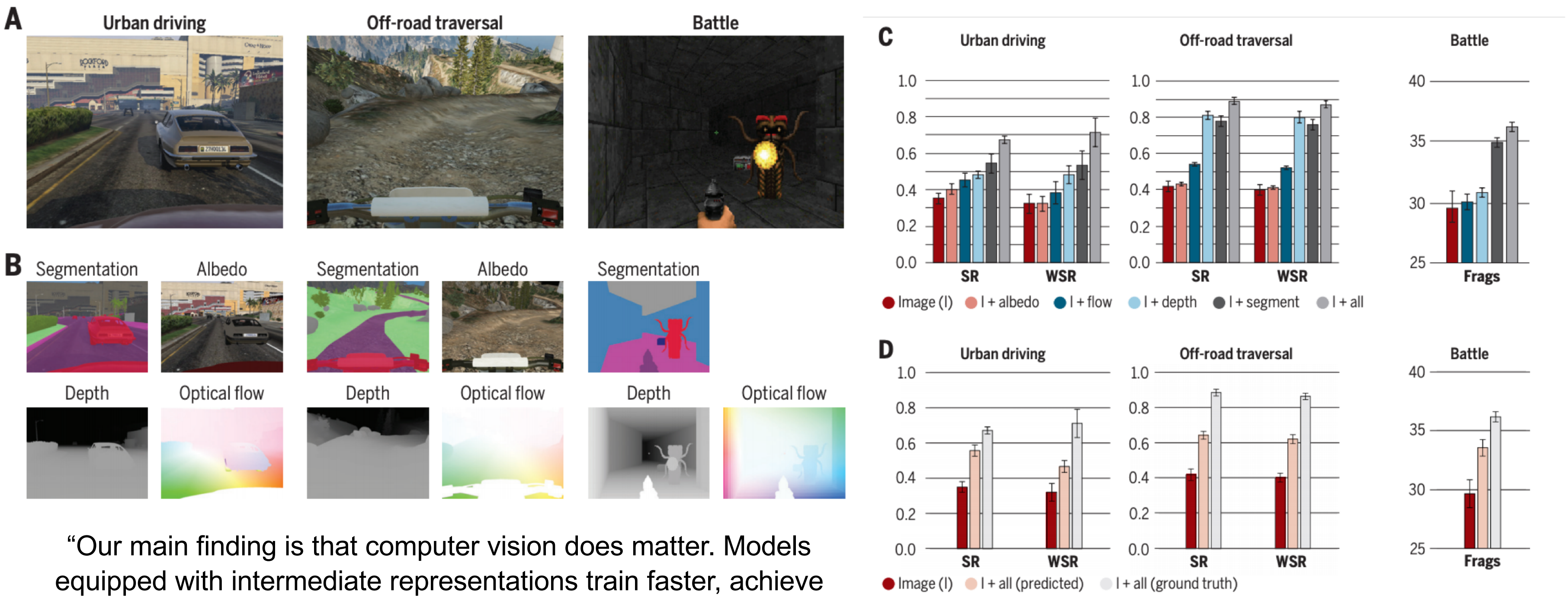
Overview video,

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

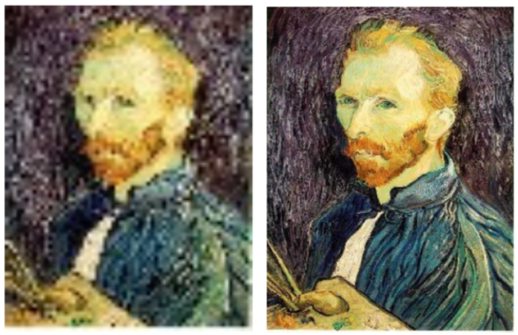
Outline

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

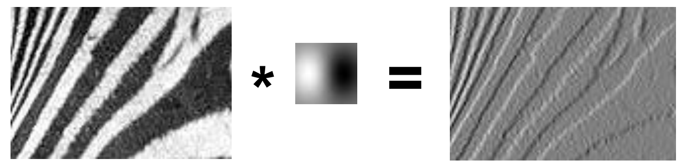
Topics covered in class

I. Early vision: Image processing and feature extraction
II. Mid-level vision: Grouping and fitting
III. Image formation and geometric vision
IV. Recognition

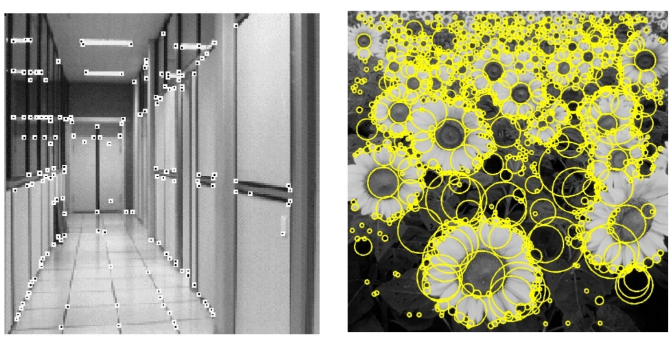
I. Image processing and feature extraction



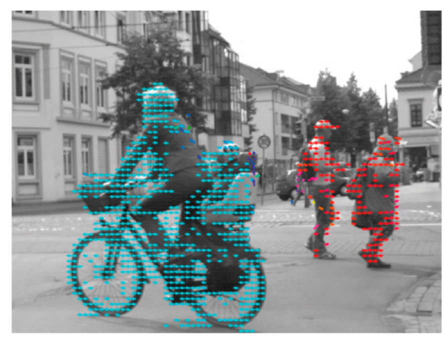
Basic image processing



Linear filtering
Edge detection

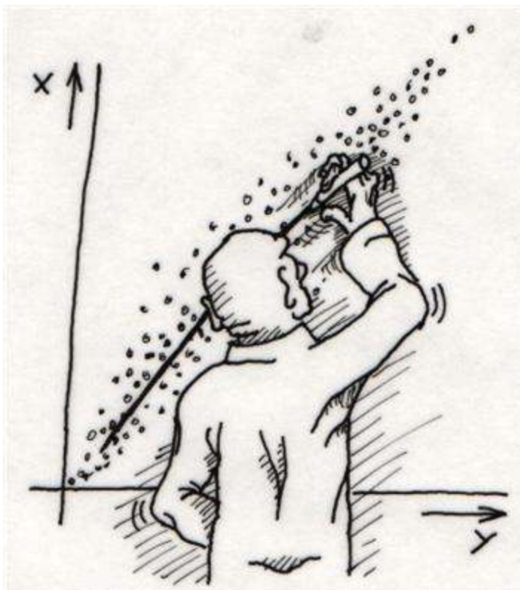


Feature extraction



Optical flow

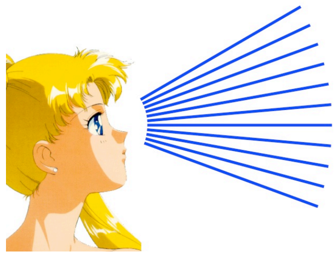
II. Grouping and fitting



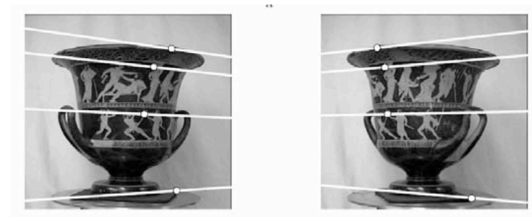
Fitting: Least squares
Voting methods

Alignment

III. Image formation and geometric vision

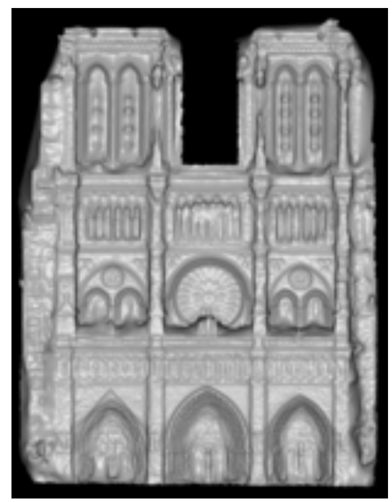


Cameras and sensors
Light and color



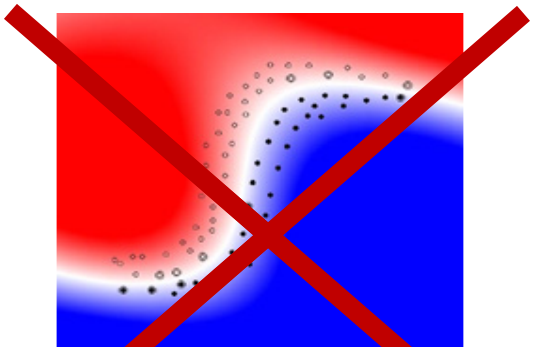
Two-view geometry, stereo

Structure from motion

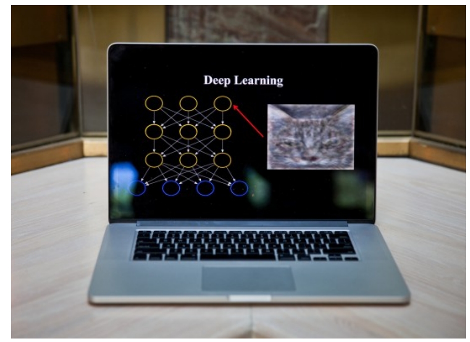


Multi-view stereo

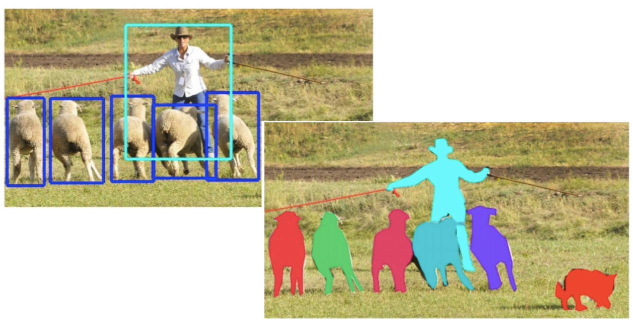
IV. Recognition



~~Intro to classification and neural networks~~



Deep learning architectures for images



Object detection and segmentation

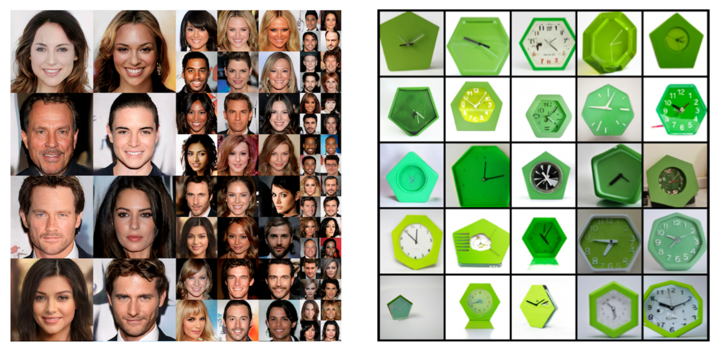


Image generation