

Computer Vision

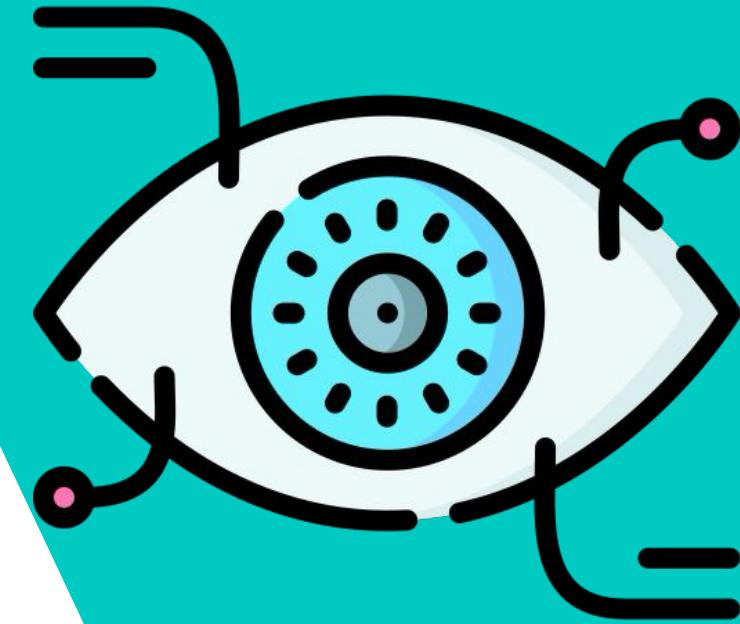
A.A. 2024-20245

Lecture 1: Introduction to Computer Vision

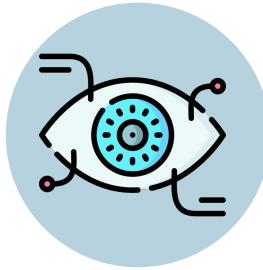


SAPIENZA
UNIVERSITÀ DI ROMA

ALCQR Lab



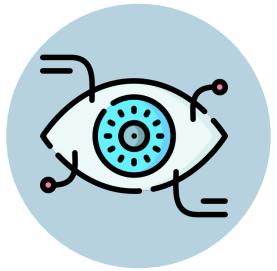
Overview



Introduction



History of
Computer Vision



Introduction

Artificial Intelligence

"An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."

[John McCarthy]

- Machine Learning
- Computer Vision
- Computer Graphics
- Natural Language Processing
- Robotics & Control
- Art, Industry 4.0, Education ...

Computer Vision

- Goal of Computer Vision is to **convert light into meaning** (geometric, semantic, . . .)

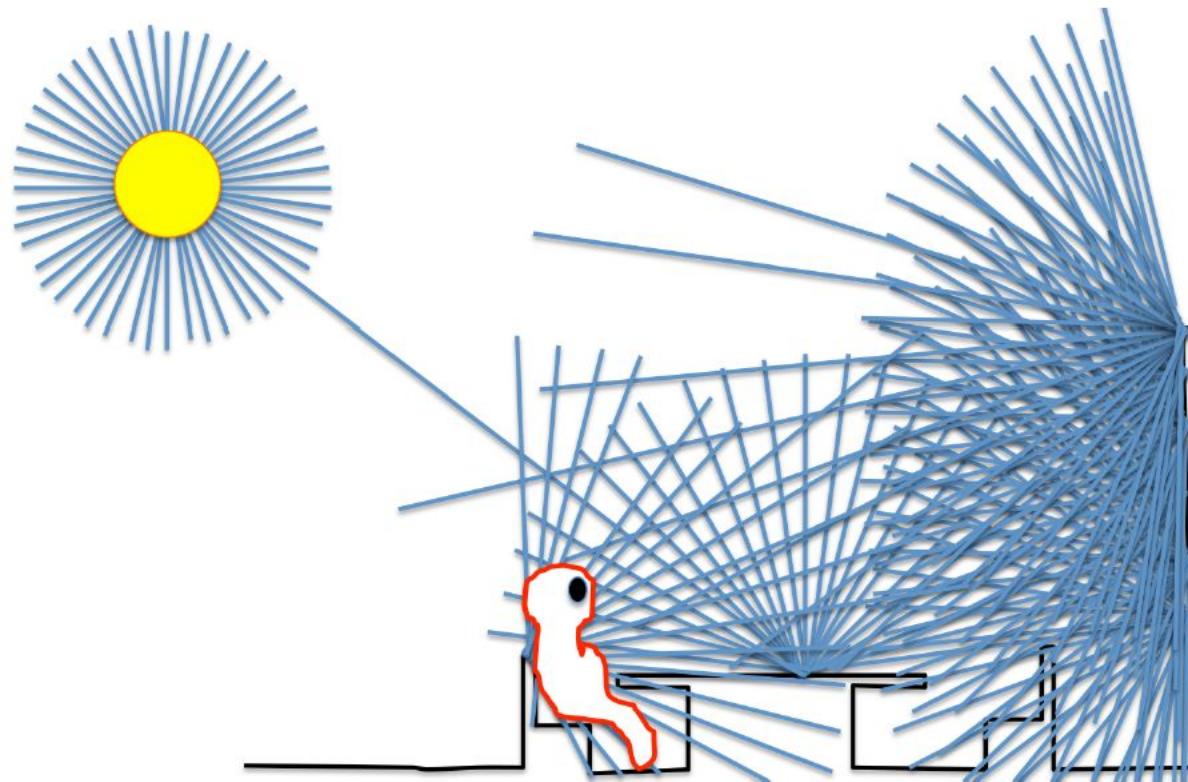
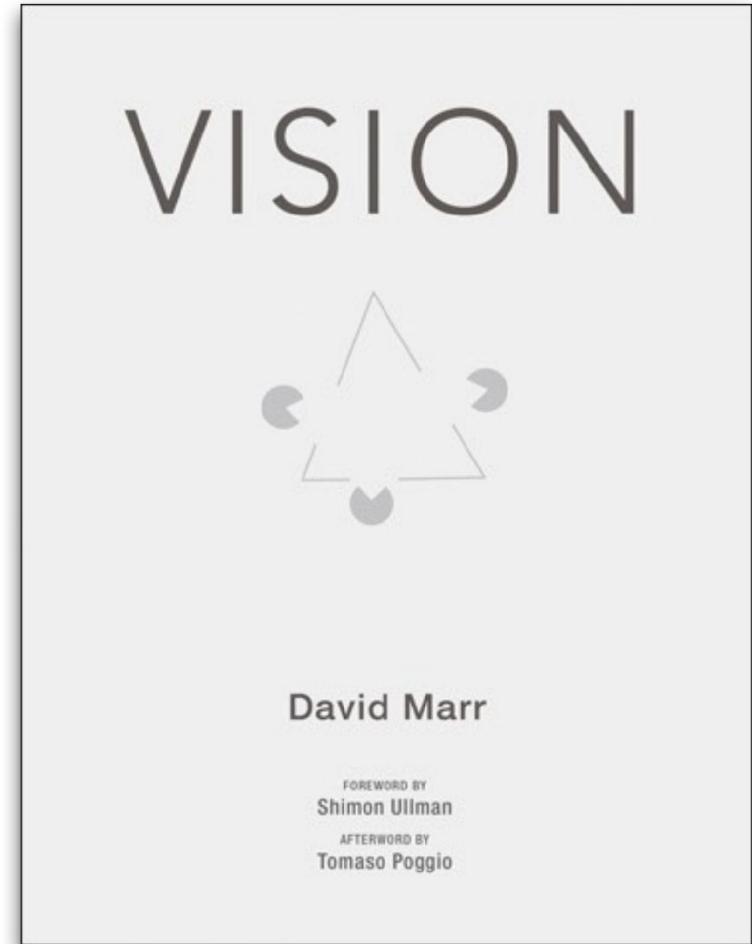


Image Credits: Antonio Torralba

Computer Vision

- What does it mean, to see? The answer for a common man (and Aristotle's, too) would be, to know what is where by looking.
- To discover from images what is present in the world, where things are, what actions are taking place, to predict and anticipate events in the world.



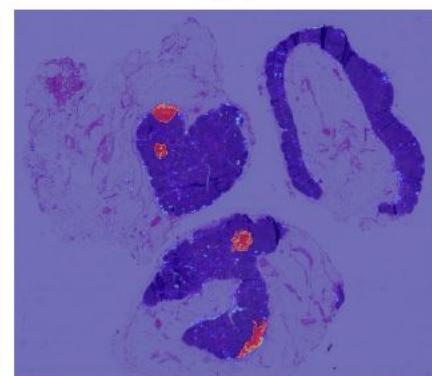
Slide Credit: Torralba, Freeman, Isola 21

Computer Vision Applications

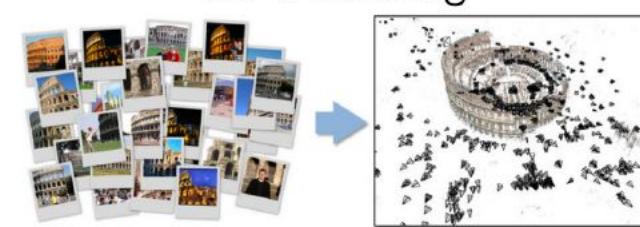
Robotics



Medical applications



3D modeling



Driving



Mobile devices



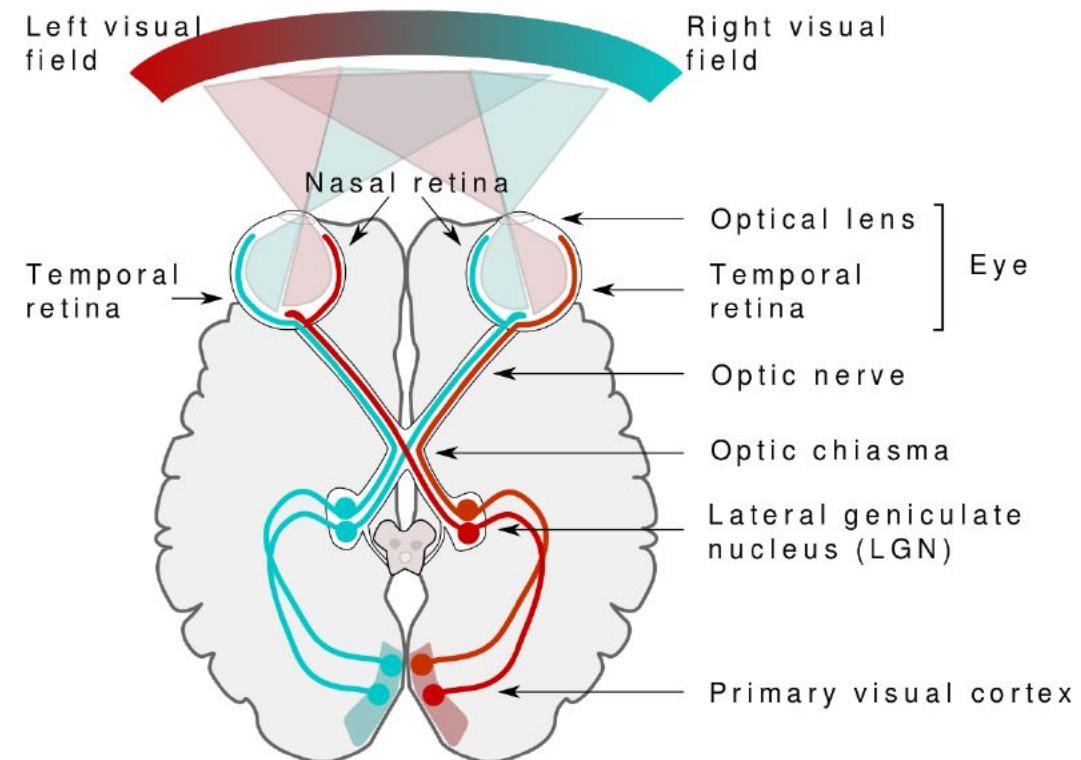
Accessibility



Slide Credit: Torralba, Freeman, Isola

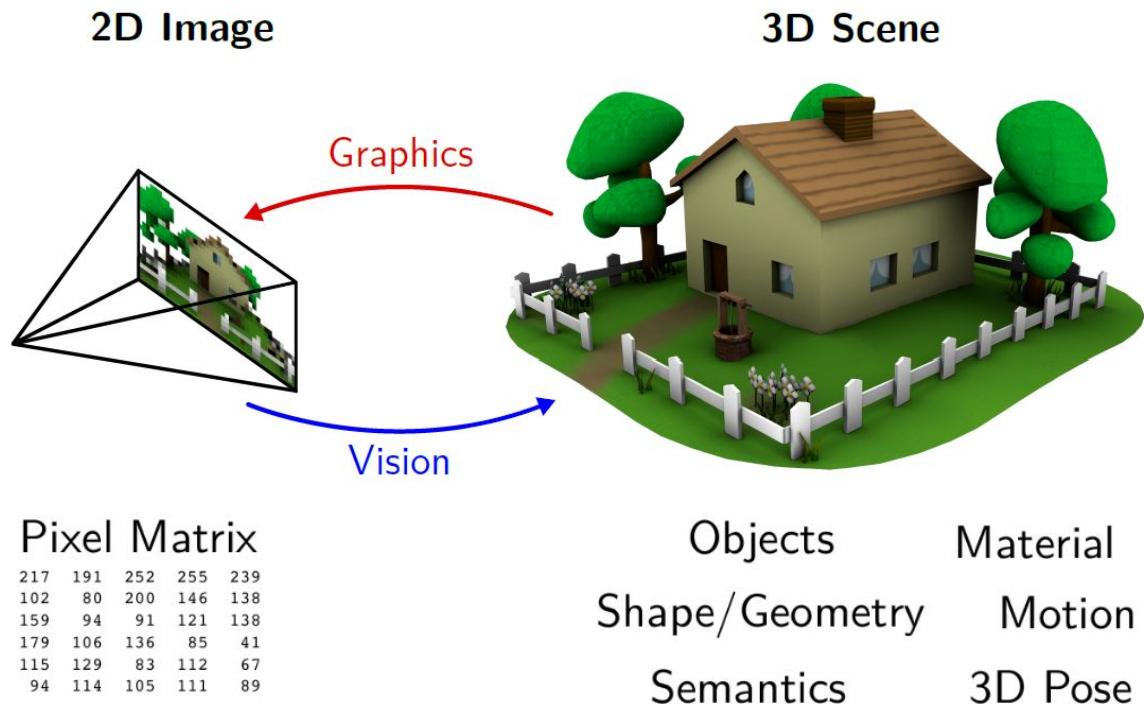
Computer Vision vs Biological Vision

- Over **50%** of the processing in the **human brain** is dedicated to **visual information**

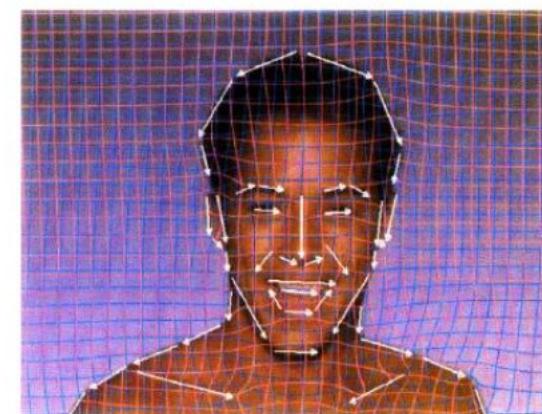
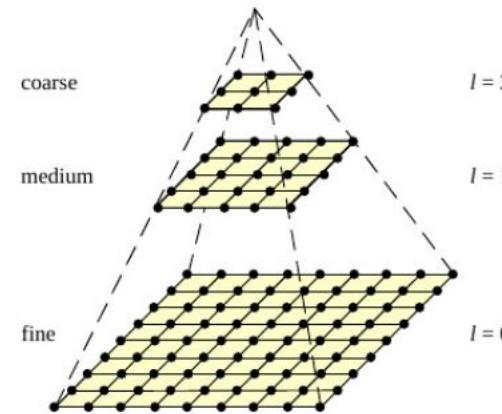
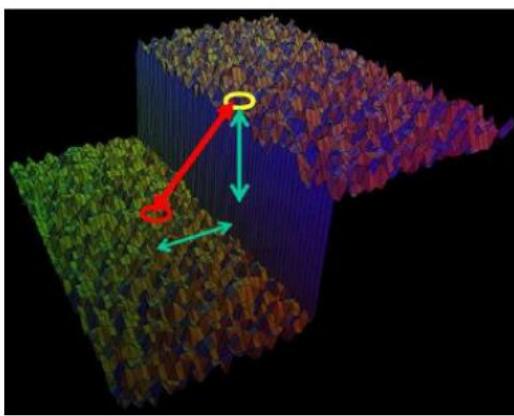
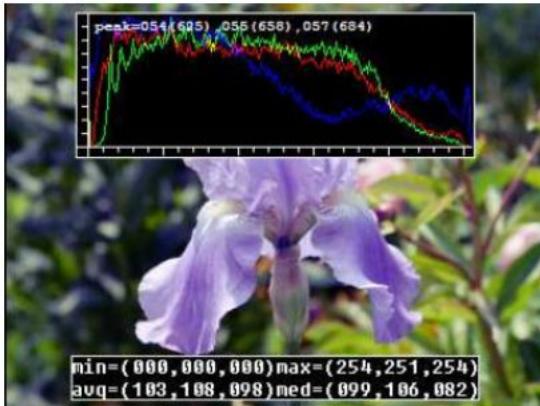


Computer Vision vs Computer Graphics

- **Computer Vision is an ill-posed inverse problem:**
 - Many 3D scenes yield the same 2D image
 - Additional constraints (knowledge about world) required

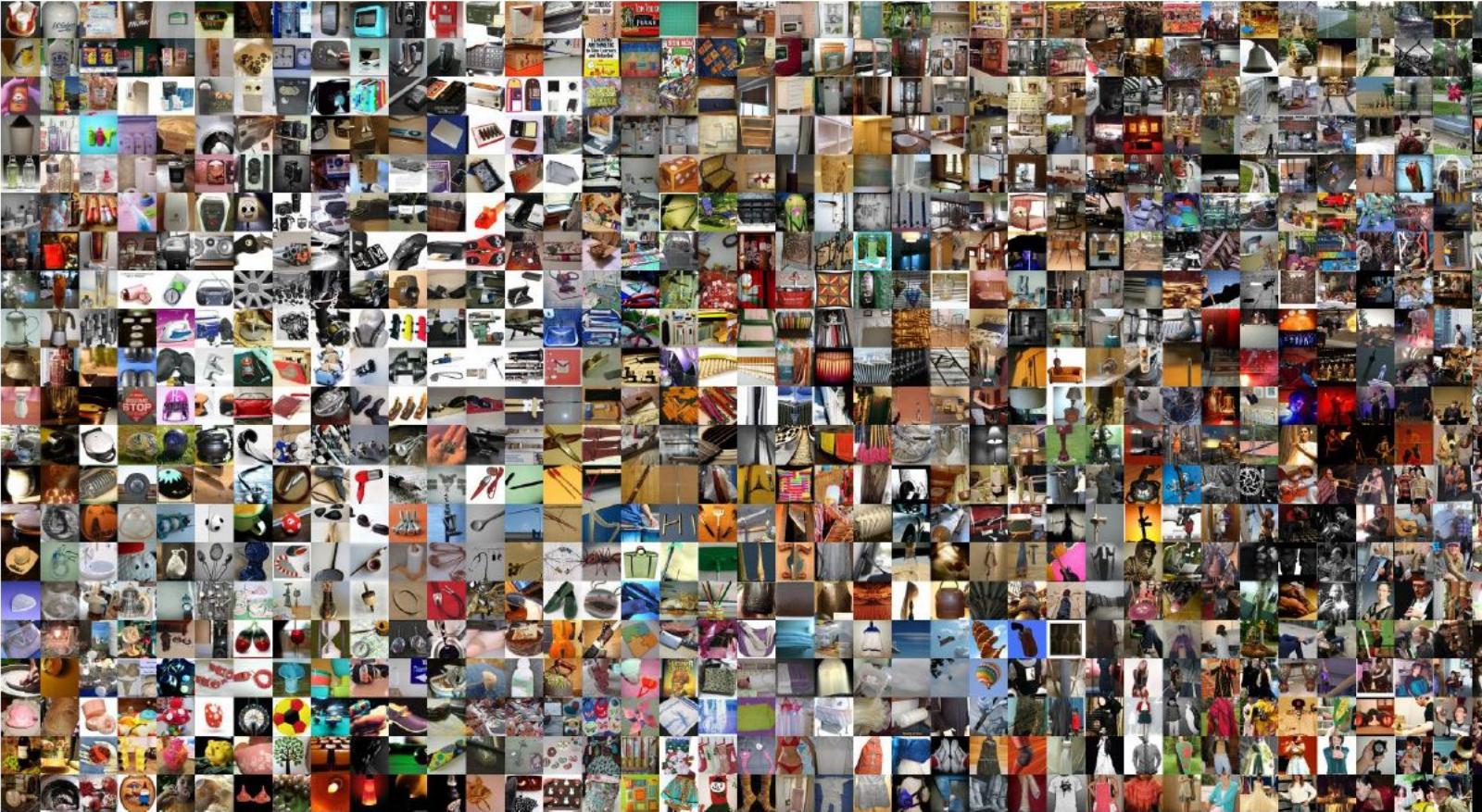


Computer Vision vs Image Processing



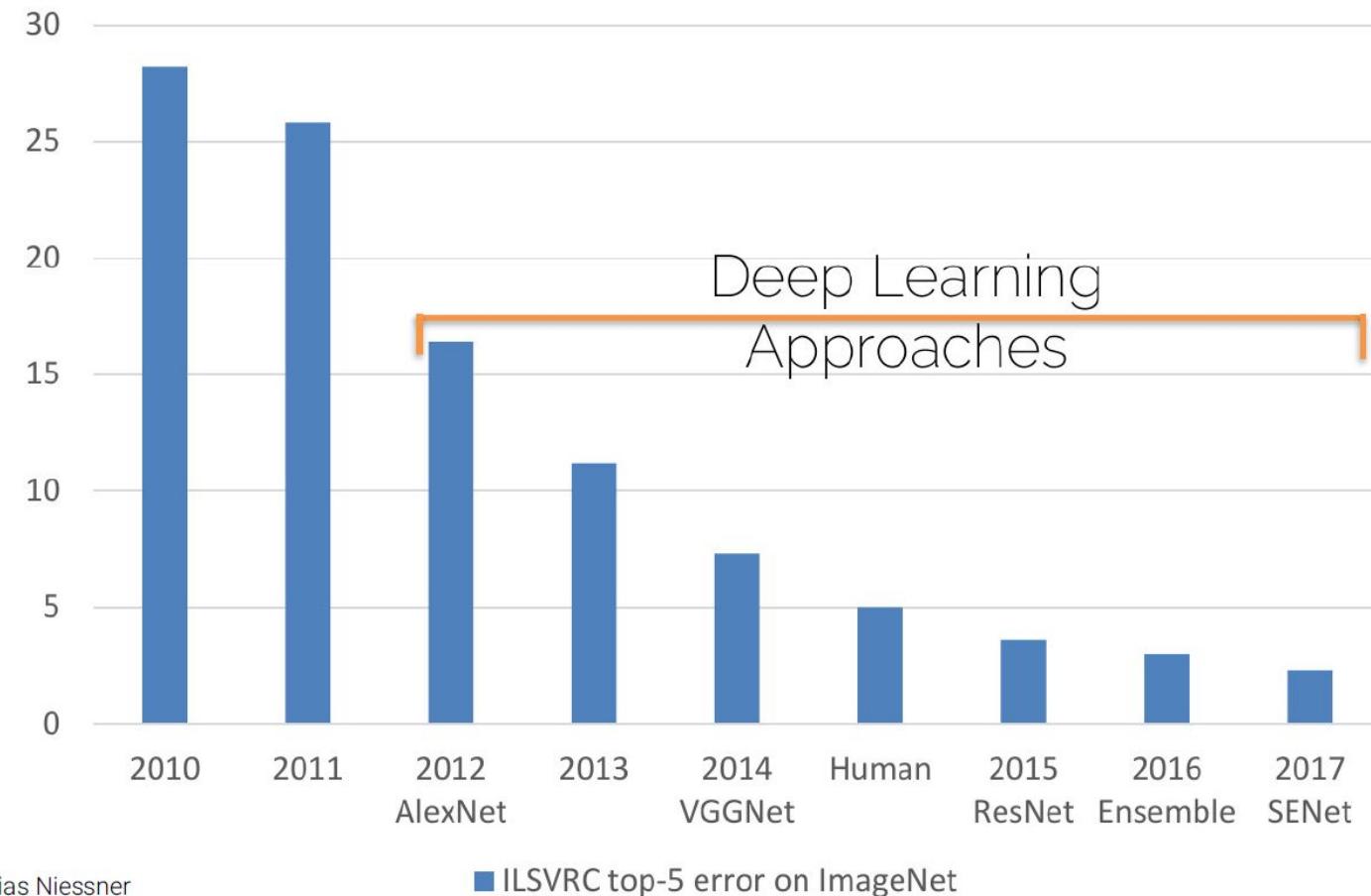
Slide Credits: Rick Szeliski

Computer Vision vs Deep Learning



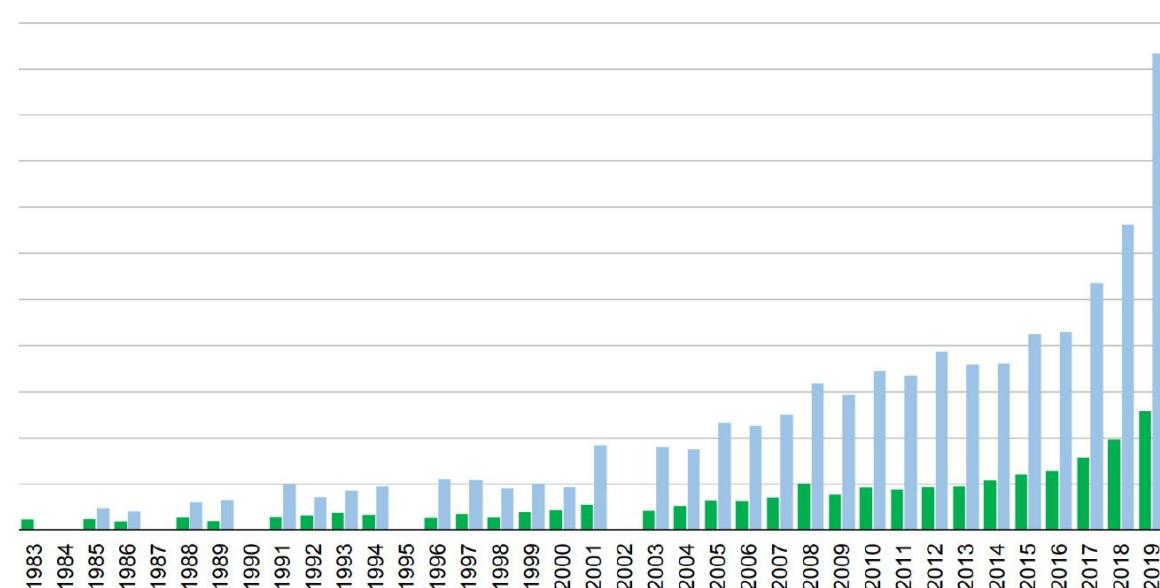
Deng, Dong, Socher, Li, Li and Li: ImageNet: A large-scale hierarchical image database. CVPR, 2009.

The Deep Learning revolution



Slide Credit: Matthias Niessner

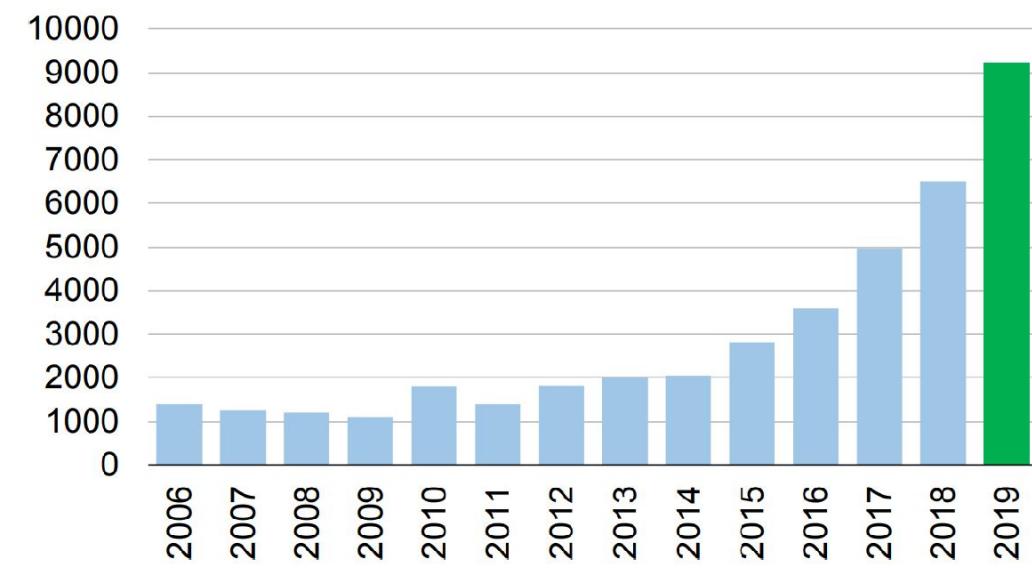
CVPR Submitted and Accepted Papers



Slide Credit: CVPR 2019 Welcome Slides

<https://cvpr.thecvf.com/>

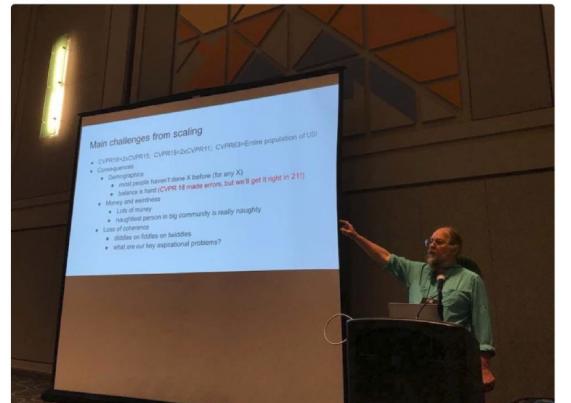
CVPR Attendance



Ira Kemelmacher
@kemelmi

Following

The entire population of the USA will attend #CVPR in 2063 :-)



Computer Vision vs LLMs

662 Views • April 19

...

Why are the people laughing in this picture?



LLAVA

The people in the image, including President Obama, are seen laughing and enjoying themselves as they look at a funny picture. The content of the picture is not provided, but it must be amusing for the group to be sharing laughter and smiles while viewing it. This scene demonstrates the power of humor in bringing people together and creating a light-hearted atmosphere, even during more formal events like a gathering with the President of the United States and other politicians.

Andrej Karpathy blog

About

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

Oct 22, 2012

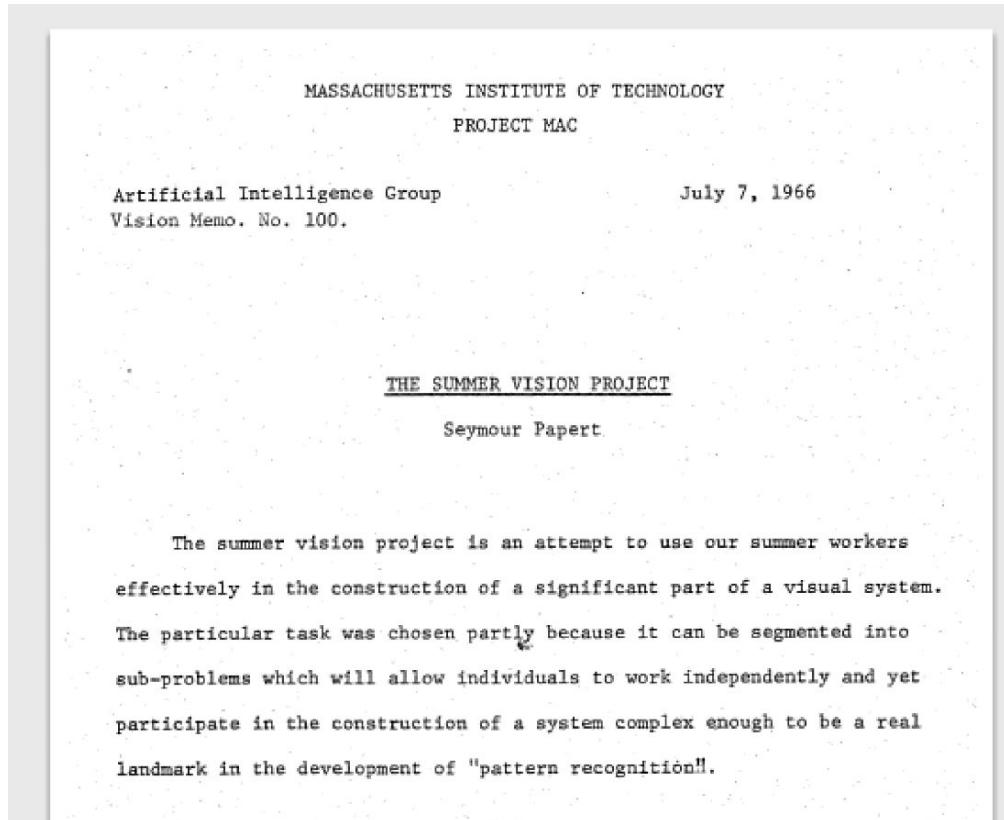


Here we are in September 2023 and the problem is solved. This is the output from ChatGPT's new Vision add-on:

In the image, Barack Obama, the former U.S. President, seems to be playfully posing as if he's trying to add weight while another official, who appears to be former UK Prime Minister David Cameron, is standing on a scale. Obama's gesture, where he's putting his foot forward as though trying to press down on the scale, suggests a playful attempt to make Cameron appear heavier. The lightheartedness of such a playful gesture, especially in the context of world leaders typically engaged in serious discussions, is a break from formality, which is likely why others in the vicinity are laughing. The scene captures a candid, informal moment amidst what might have been a formal setting or meeting.

Why Visual Perception is hard?

- Papert: The Summer Vision Project. MIT AI Memos, 1966.



Why Visual Perception is hard?



What we see

200	133	110	103	117	90	47	30	32	79	66	65
197	122	123	138	98	100	46	45	22	11	43	55
140	116	165	159	90	56	58	47	26	13	54	102
132	148	119	108	123	57	64	46	21	22	79	94
125	121	80	143	101	55	61	38	20	21	81	65
50	71	74	63	52	39	41	39	32	26	97	66
51	59	62	44	40	40	36	28	27	31	29	44
59	62	70	50	48	35	34	35	26	21	24	32
49	59	65	64	58	34	40	28	26	21	23	124
39	45	47	64	54	34	40	24	19	47	133	207
37	42	39	38	39	50	75	74	105	170	197	167
37	47	33	35	50	108	162	184	184	157	125	112
45	48	35	37	75	148	183	156	83	91	91	116
49	48	54	50	75	158	110	66	74	128	155	149
48	51	57	50	65	91	79	92	101	105	132	132
51	58	66	55	58	52	91	91	88	115	158	174
57	60	61	52	56	61	60	55	92	146	188	190
65	50	54	56	57	51	54	56	80	115	177	187
67	40	40	61	65	48	39	30	36	75	151	181
53	32	36	35	61	43	37	26	29	35	126	189
29	42	107	20	28	41	40	26	30	36	113	200
30	21	32	24	34	37	33	23	25	39	105	171
32	28	19	23	29	36	47	89	132	169	183	128
31	25	62	54	47	44	81	190	227	231	206	155
44	66	99	72	67	63	89	128	127	115	109	157
53	47	47	41	29	32	25	20	41	81	89	175
38	44	61	73	54	48	37	87	90	111	126	189
39	41	83	97	86	91	74	134	131	153	143	185
42	56	98	102	112	111	94	137	121	141	146	181

What the computer sees

Every image tells a story



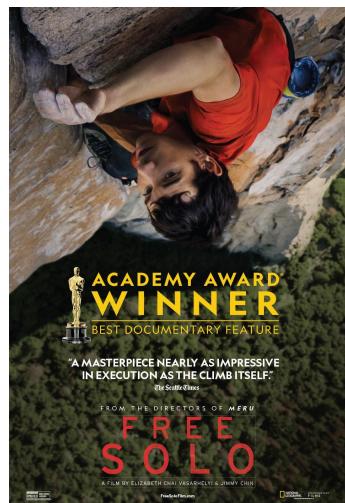
Alex Honnold

person/rock climber-□
free soloist

Rock wall/ Mountain
El Capitan

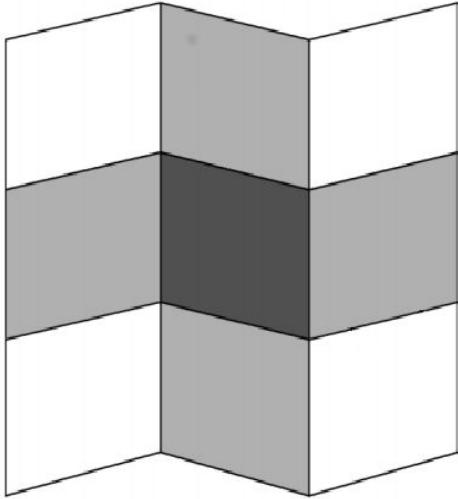
Trees / Yosemite
Park

A frame of the
documentary
«Free Solo»

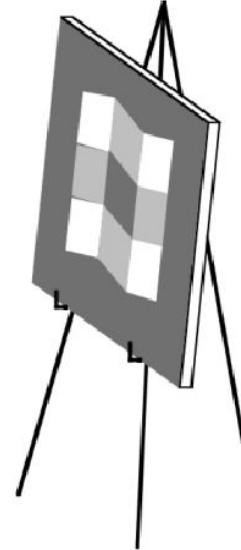


Challenges: Images are 2D Projections of the 3D World

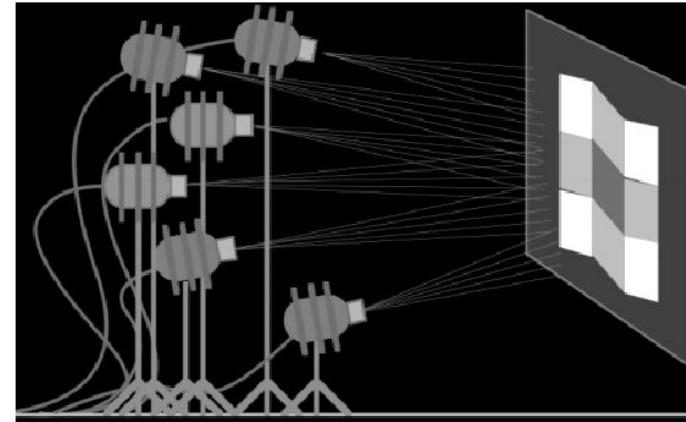
- **Adelson and Pentland's workshop metaphor:**
- To explain an image (a) in terms of reflectance, lighting and shape, (b) a painter, (c) a light designer and (d) a sculptor will design three different, but plausible, solutions.



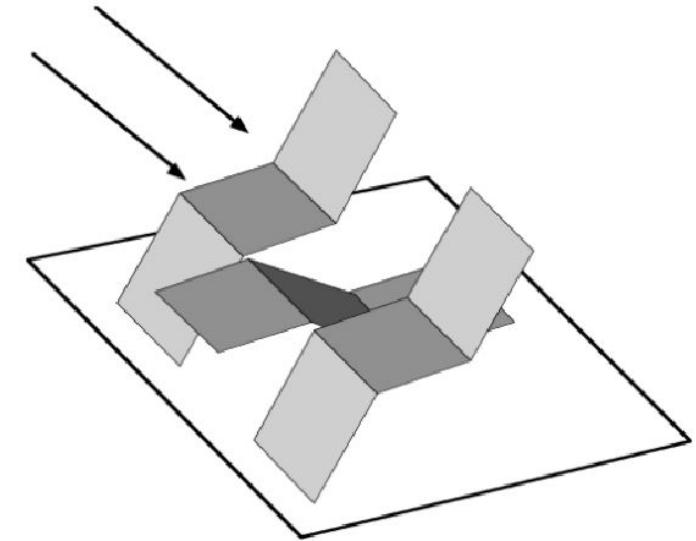
(a)



(b)



(c)



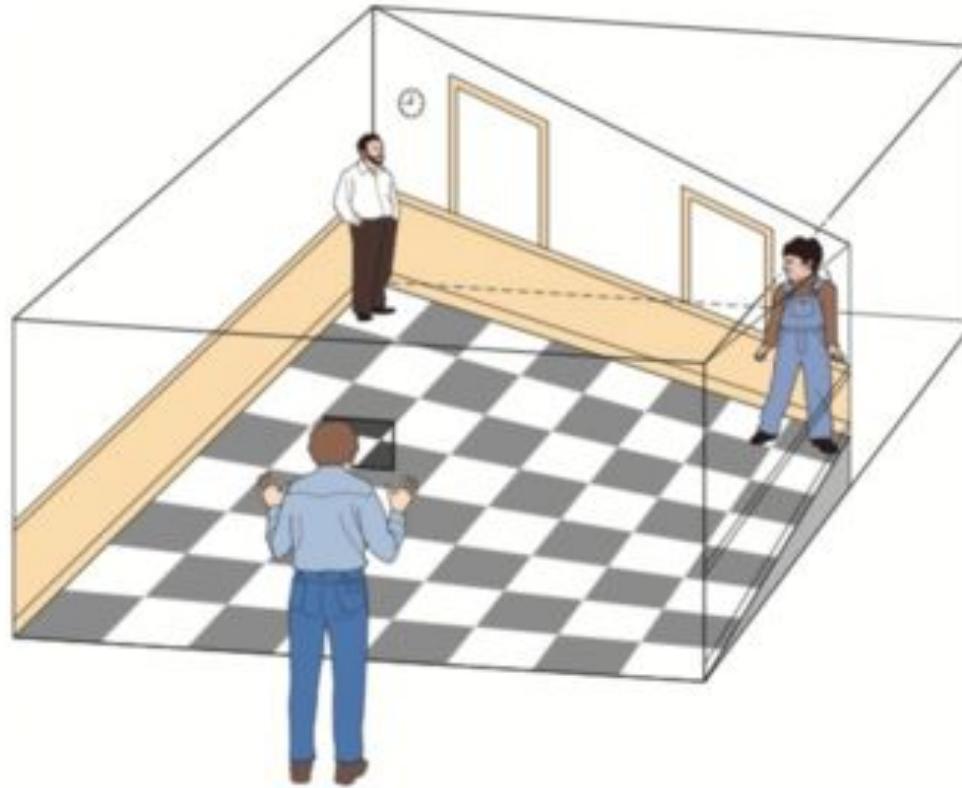
(d)

Adelson and Pentland: The perception of shading and reflectance. Perception as Bayesian inference, 1996.

Ames Room Illusion

Figure 5.39

Although not apparent to the viewer, the right side of the Ames room is one-half as deep as the left, the floor is higher and the ceiling is lower on the right, and the window on the right is smaller. All of these cues create the impression that the person on the right is much larger than the person on the left in the Ames room.



Challenges in CV

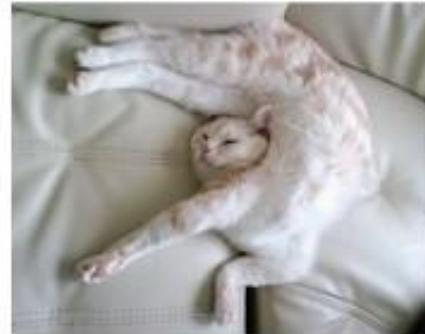
Viewpoint variation



Scale variation



Deformation



Occlusion



Illumination conditions



Background clutter



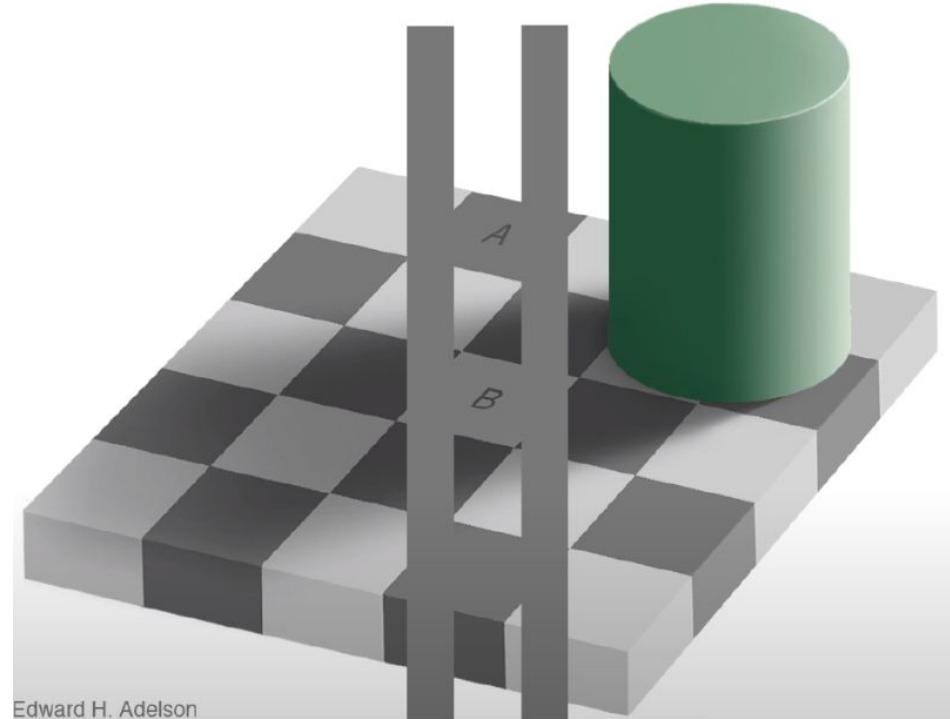
Intra-class variation



Challenges: Motion



Challenges: Perception vs Measurement



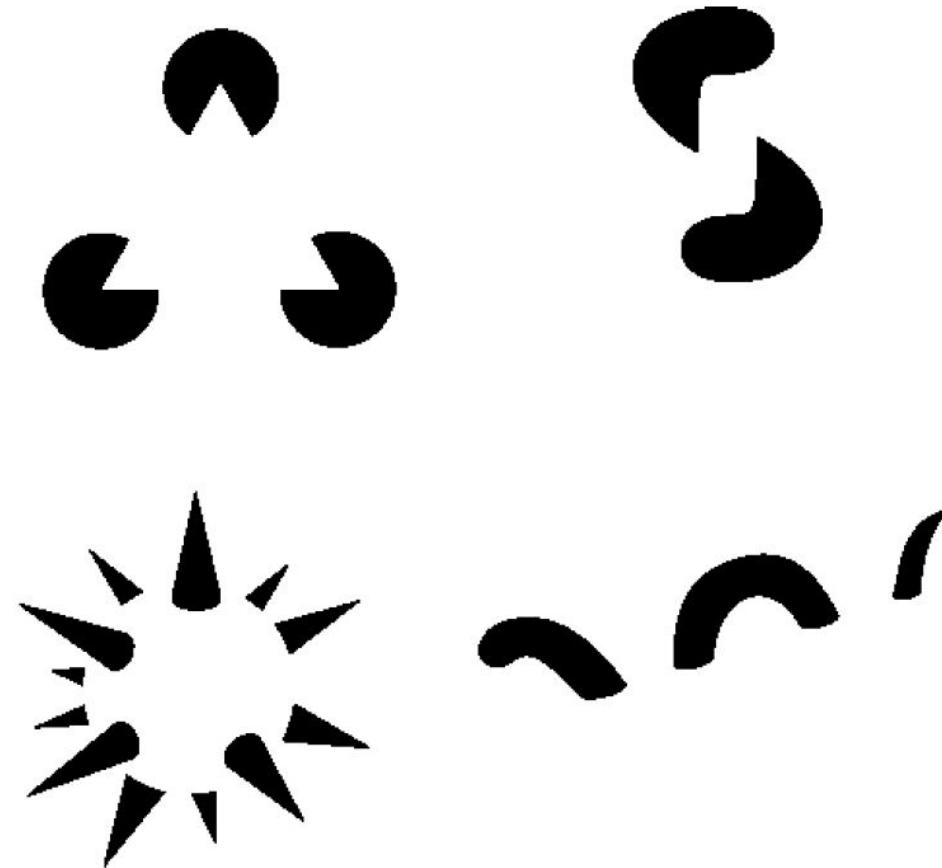
Edward H. Adelson

<http://persci.mit.edu/gallery/checkershadow>

Image Credits: Edward H. Adelson



Perception vs Measurement

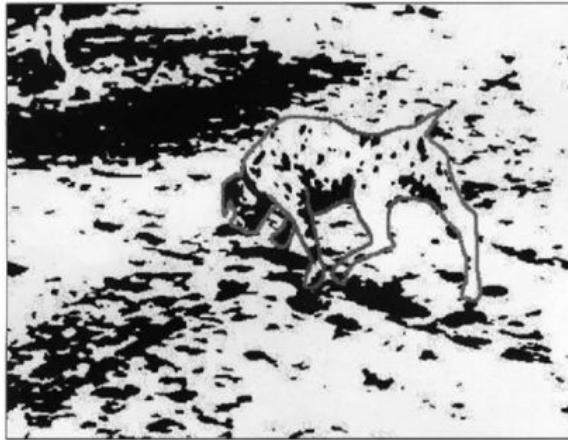


Gaetano Kanizsa (1976)

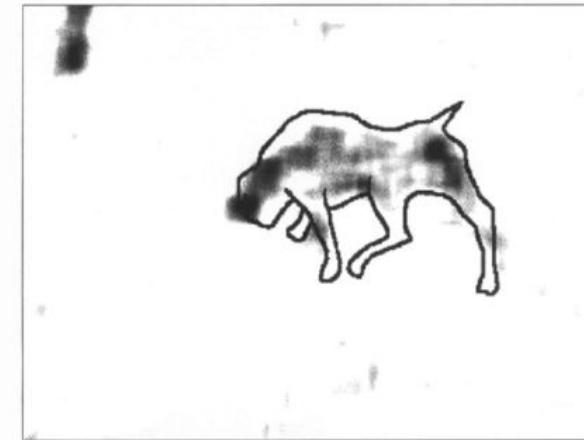
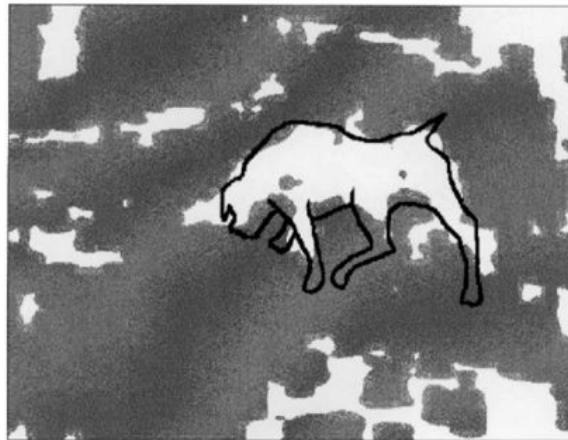
Perception vs Measurement



Perception vs Measurement



(a)



Perception vs Measurement



Challenges: Local Ambiguity



Image Credits: Antonio Torralba

Challenges: Local Ambiguity



Image Credits: Antonio Torralba

Challenges: number of object categories



Image Credits: Antonio Torralba

Challenges: explaining the semantic





History of Computer Vision

Old days



Perspective
Leonardo da Vinci
(1452–1519)

Image credits Andreas Geiger



Photometry
Johann Heinrich Lambert
(1728–1777)



Least Squares
Carl Friedrich Gauss
(1777–1855)

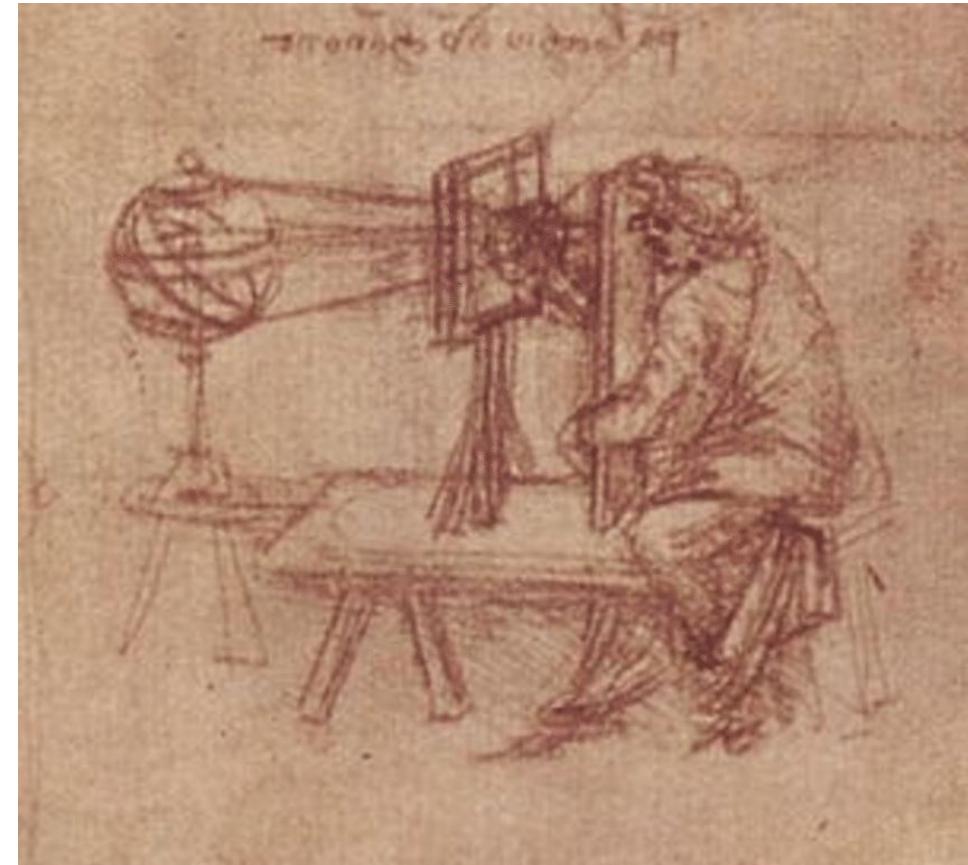


Stereopsis
Charles Wheatstone
(1802–1875)

1510: Perspectograph

- “Perspective is nothing else than the seeing of an object behind a sheet of glass, smooth and quite transparent, on the surface of which all the things may be marked that are behind this glass. All things transmit their images to the eye by pyramidal lines, and these pyramids are cut by the said glass. The nearer to the eye these are intersected, the smaller the image of their cause will appear.”

Leonardo da Vinci



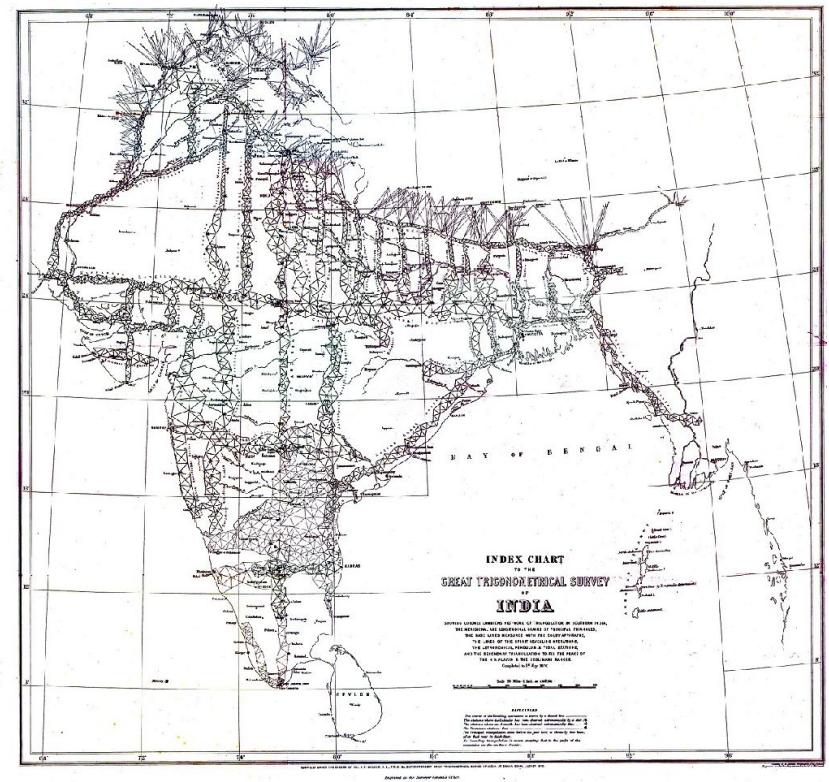
1839: Daguerreotype

- First publicly available photographic process invented by Louis Daguerre
- Widely used during the 1840s and 1850s
- Polish a sheet of silver-plated copper and treat with fumes to make light sensitive
- Make resulting latent image visible by fuming it with mercury vapor and remove its sensitivity to light by chemical treatment
- Rinse, dry and seal behind glass



1802-1871: Great Trigonometrical Survey

- Multi-decade project to measure the entire Indian subcontinent with scientific precision
- Under the leadership of George Everest, the project was made responsible of the Survey of India
- Manual bundle adjustment proves Mt. Everest highest mountain on earth mountain on earth



Overview

Waves of development:

- ▶ 1960-1970: Blocks Worlds, Edges and Model Fitting
- ▶ 1970-1981: Low-level vision: stereo, flow, shape-from-shading
- ▶ 1985-1988: Neural networks, backpropagation, self-driving
- ▶ 1990-2000: Dense stereo and multi-view stereo, MRFs
- ▶ 2000-2010: Features, descriptors, large-scale structure-from-motion
- ▶ 2010-now: Deep learning, large datasets, quick growth, commercialization

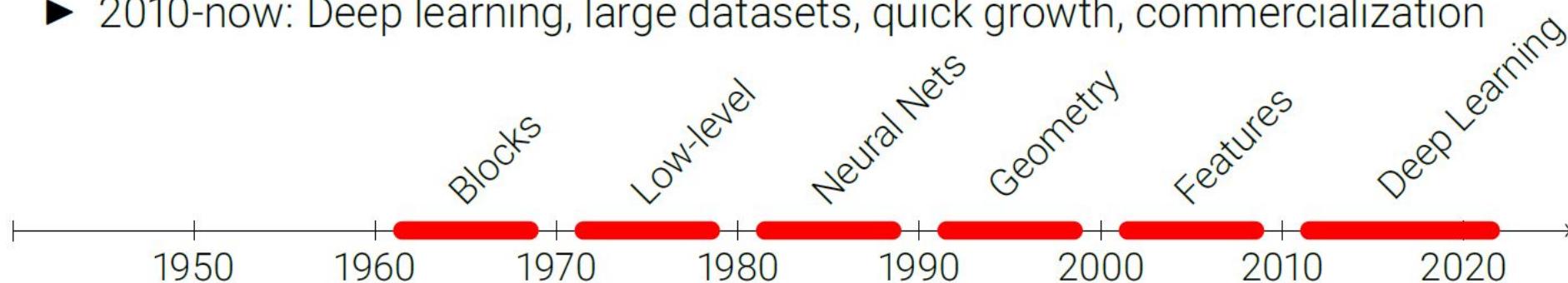
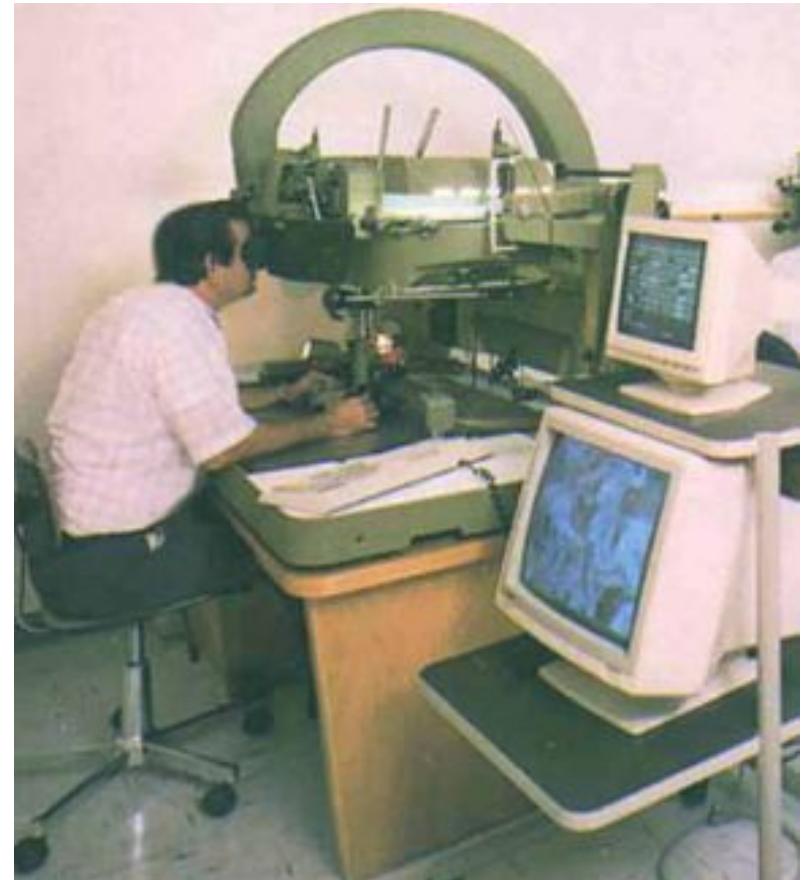


Image credits Andreas Geiger

A brief history of CV

1957: Stereo

- Gilbert Hobrough demonstrated an analog implementation of stereo image correlation
- This led to the creation of the Raytheon-Wild B8 Stereomat
- Used to create Elevation Maps (Photogrammetry, since 1840)

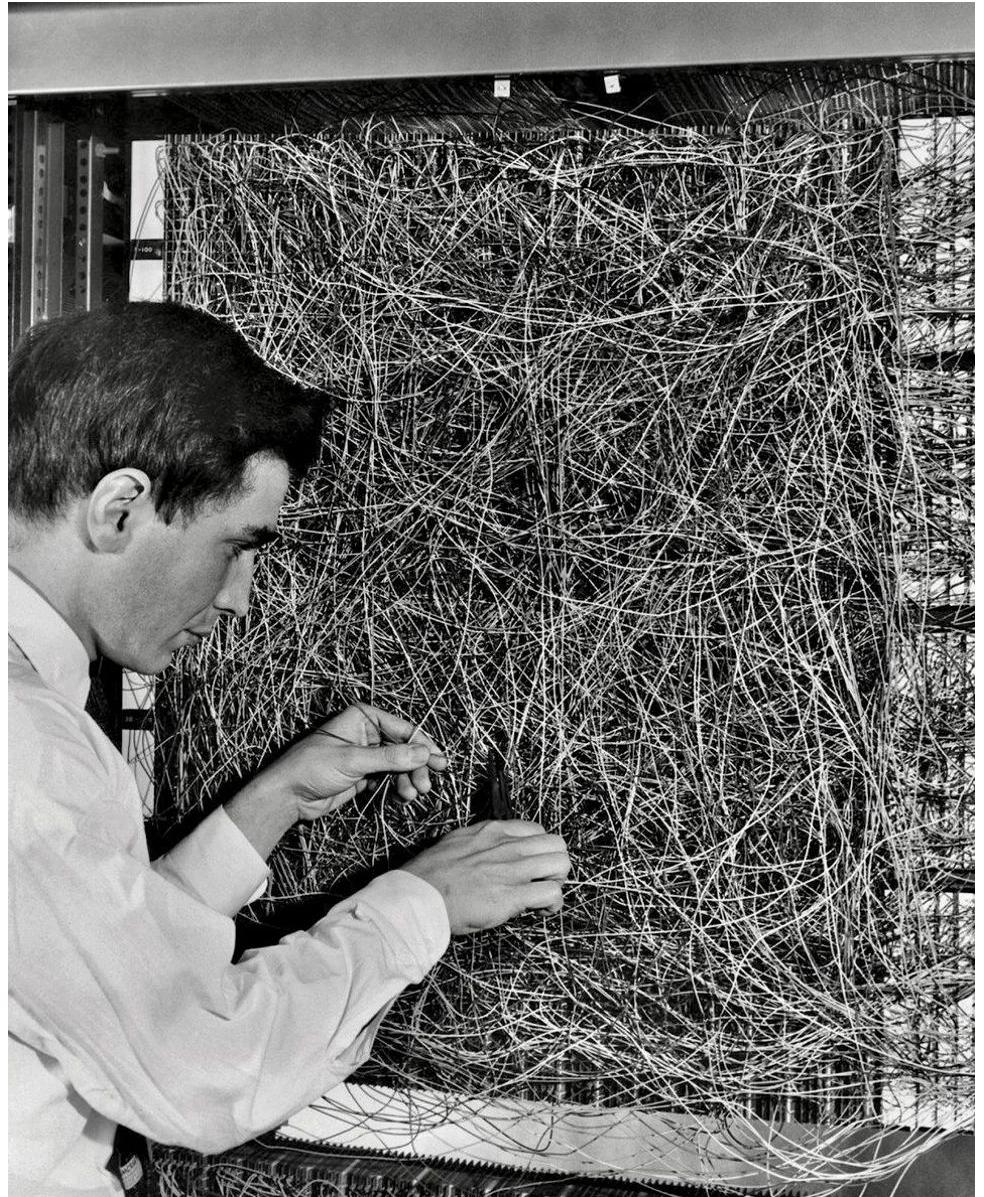


A brief history of CV

- **1958-1962: Rosenblatt's Perceptron**
- First algorithm and implementation to train single linear threshold neuron
- Optimization of perceptron criterion:

$$\mathcal{L}(\mathbf{w}) = - \sum_{n \in \mathcal{M}} \mathbf{w}^T \mathbf{x}_n y_n$$

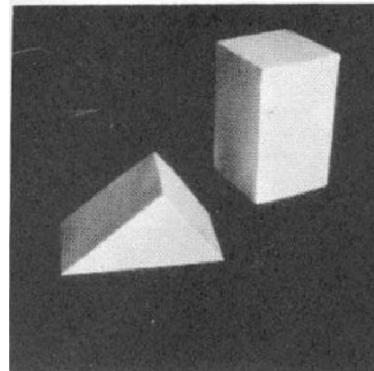
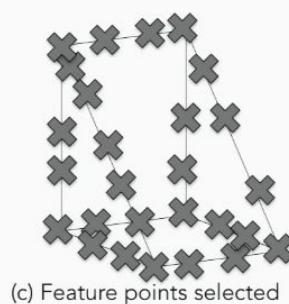
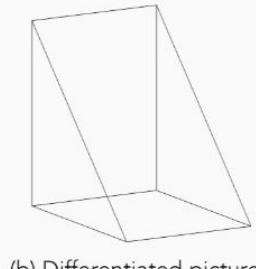
- Novikoff proved convergence
- Rosenblatt claimed that the perceptron will lead to computers that walk, talk, see, write, reproduce and are conscious of their existence



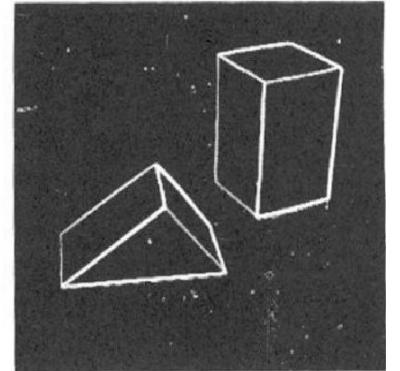
A brief history of CV

1963: Larry Robert's Blocks World

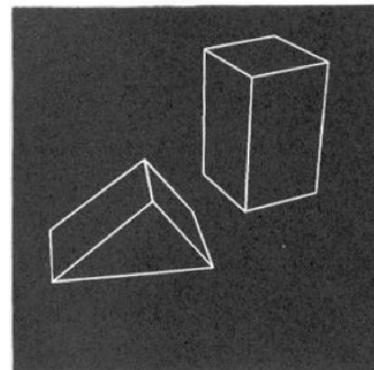
- Scene understanding for robotics
- Extracts edges as primitives
- Infers 3D structure of an object from topological structure of the 2D lines
- Interpret images as projections of 3D scenes, not 2D pattern recognition



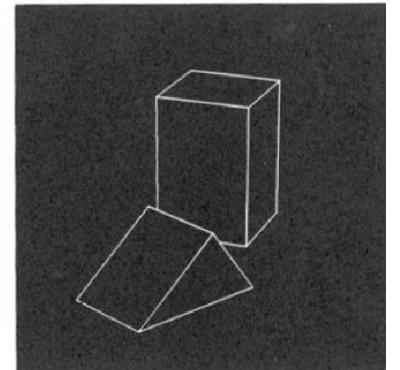
(a) Original picture.



(b) Differentiated picture.



(c) Line drawing.



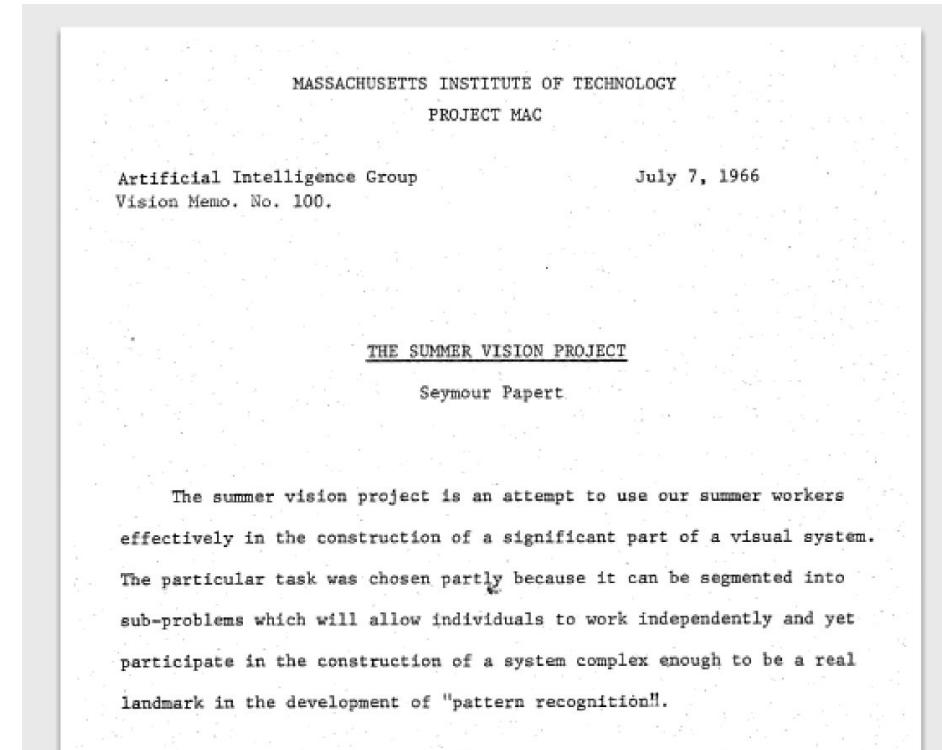
(d) Rotated view.

Roberts: Machine Perception of Three-Dimensional Solids. PhD Thesis, 1965

A brief history of CV

1966: MIT Summer Vision Project

- Underestimated the challenge of computer vision



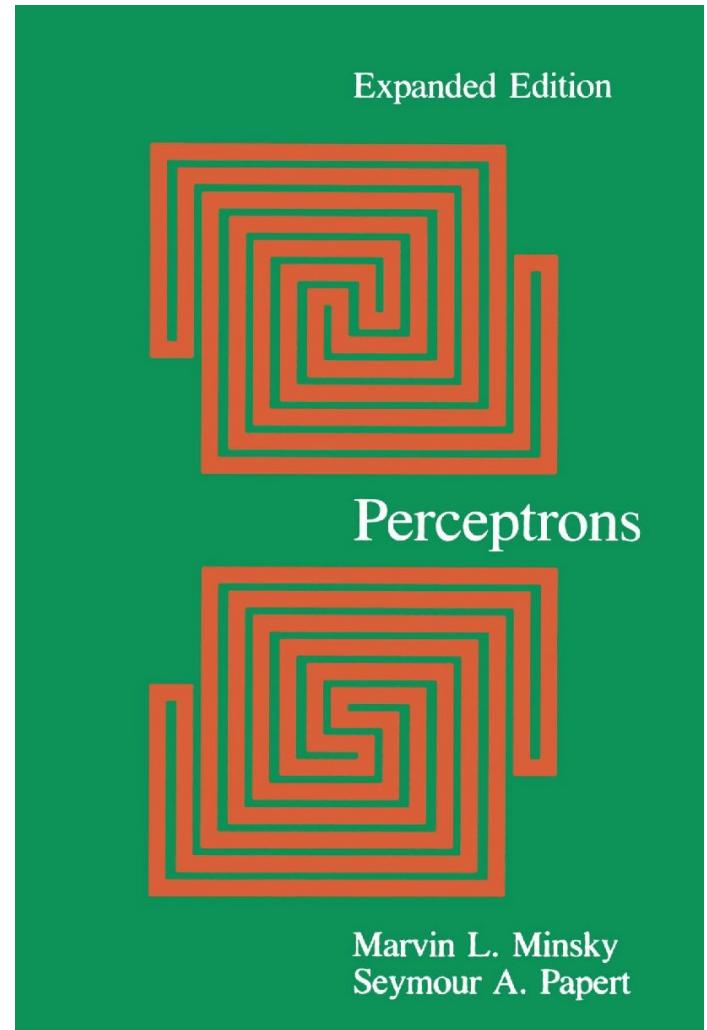
Papert: The Summer Vision Project. MIT AI Memos, 1966.

A brief history of CV

1969: Minsky and Papert publish book

- Several discouraging results
- Showed that single-layer perceptrons cannot solve some very simple problems (XOR problem, counting)
- Symbolic AI research dominates 70s

Minsky and Papert: Perceptrons: An introduction to computational geometry. MIT Press, 1969.



A brief history of CV

1970: MIT Copy Demo

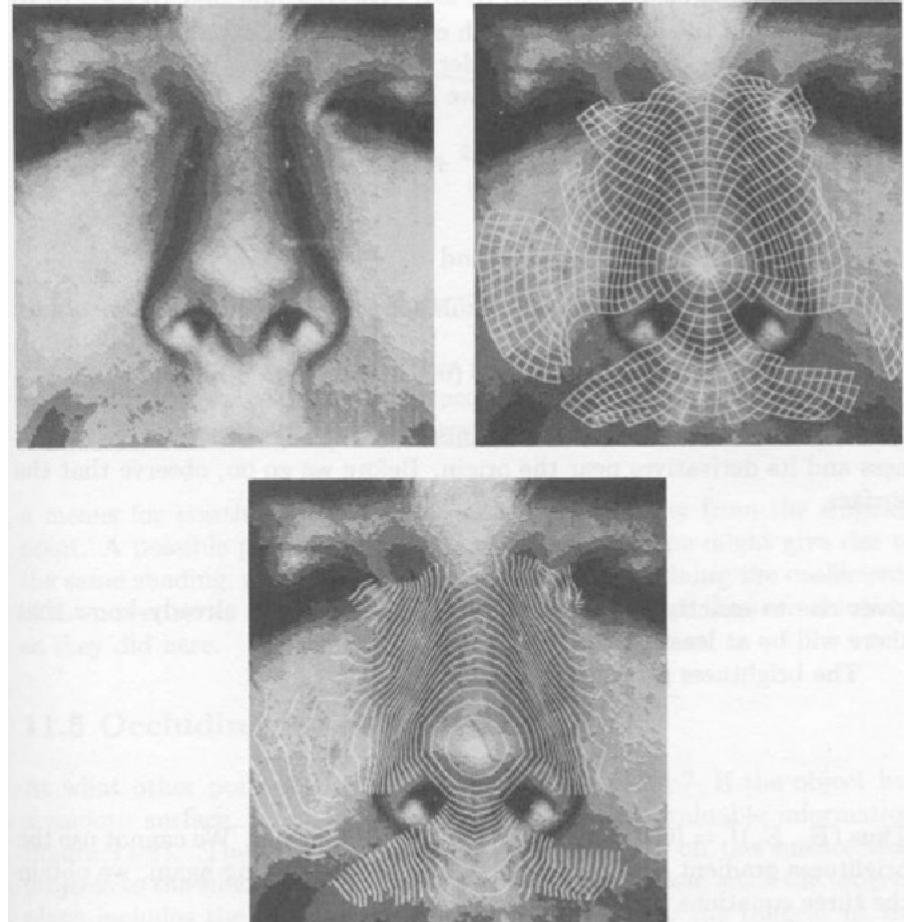
- Vision system recovers structure of a blocks scene, robot plans and builds copy from another set of blocks
- Vision, planning and manipulation
- But low-level edge finding not robust enough for task, led to attention on low-level vision

Papert: The Summer Vision Project. MIT AI Memos, 1966.

A brief history of CV

1970: Shape from Shading

- Recover 3D from single 2D image
- Assumes Lambertian surface and constant albedo (fraction of a sunlight that is diffusely reflected by a body)
- Applies smoothness regularization to constrain the ill-posed problem



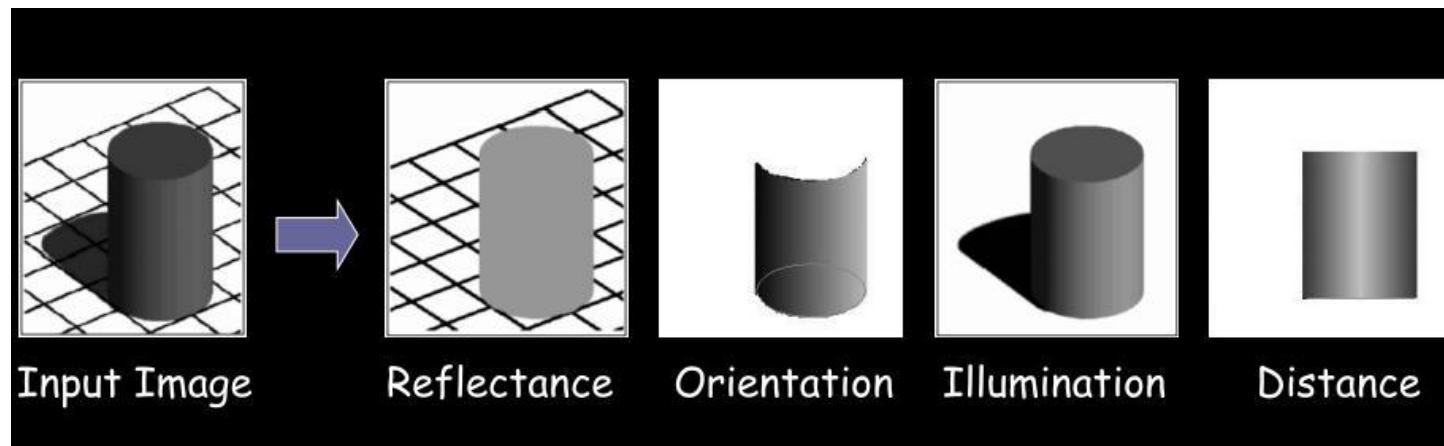
Horn: Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View. MIT TR, 1970.

Figure 11-7. The shape-from-shading method is applied here to the recovery of the shape of a nose. The first picture shows the (crudely quantized) gray-level image available to the program. The second picture shows the base characteristics superimposed, while the third shows a contour map computed from the elevations found along the characteristic curves.

A brief history of CV

1978: Intrinsic Images

- Decomposing an image into its different intrinsic 2D layers, such as reflectance, shading, shape and motion components
- Useful for downstream tasks, e.g., object detection independent from shadows and lighting

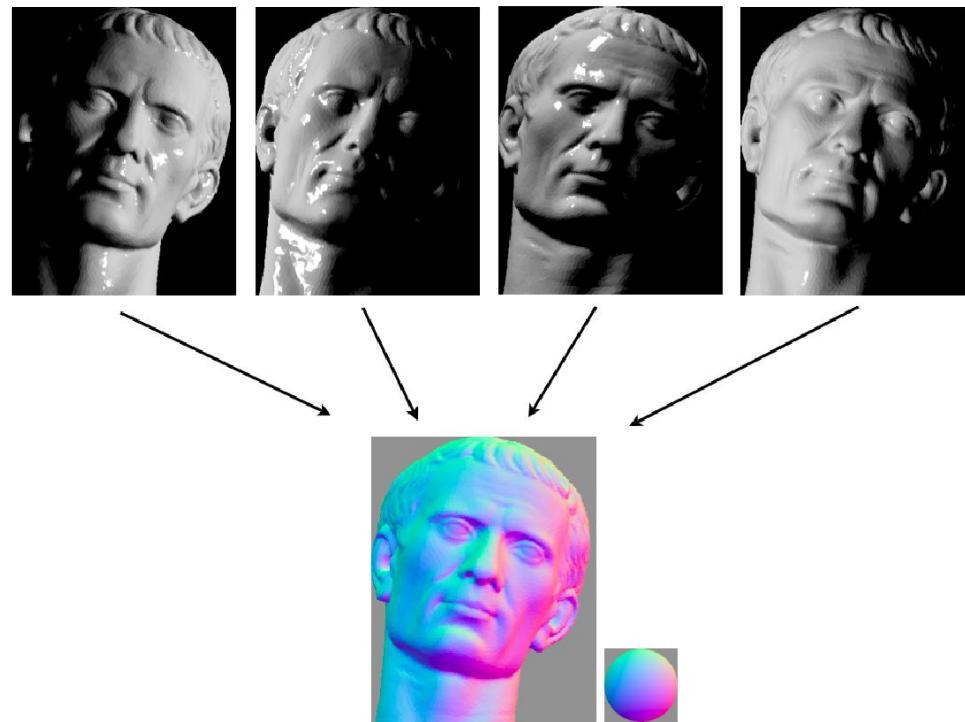


Barrow and Tenenbaum: Recovering intrinsic scene characteristics from images. Computer Vision Systems, 1978. 6

A brief history of CV

1980: Photometric Stereo

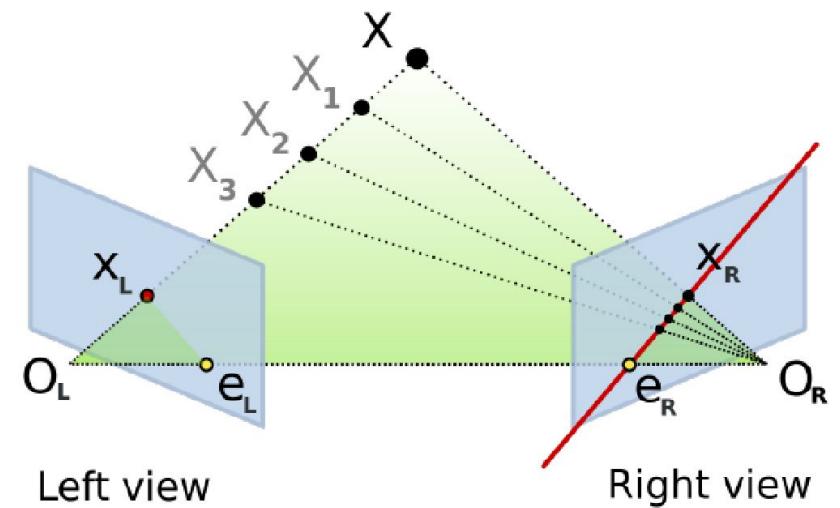
- Recover 3D from multiple 2D images, taken from the same viewpoint with different lighting conditions
- Requires at least 3 images
- Unprecedented detail and accuracy
- Lambertian assumption has been relaxed subsequently



A brief history of CV

1981: Essential Matrix

- Defines two-view geometry as matrix mapping points to epipolar lines
- Reduces correspondence search to a 1D problem
- Can be estimated from a set of 2D correspondences
- Key ideas known for 100 years

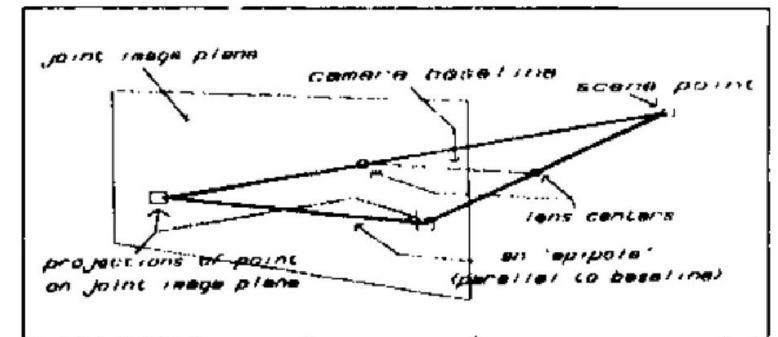
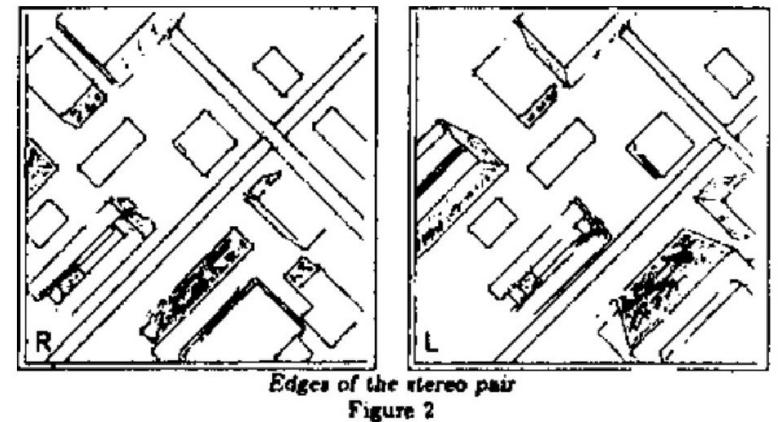


Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. Nature, 1981.

A brief history of CV

1981: Binocular Scanline Stereo

- Correlate points along epipolar lines
- Use dynamic programming to introduce constraints along scanlines (image rows)
- Allows for overcoming ambiguities, but streaking artifacts between rows

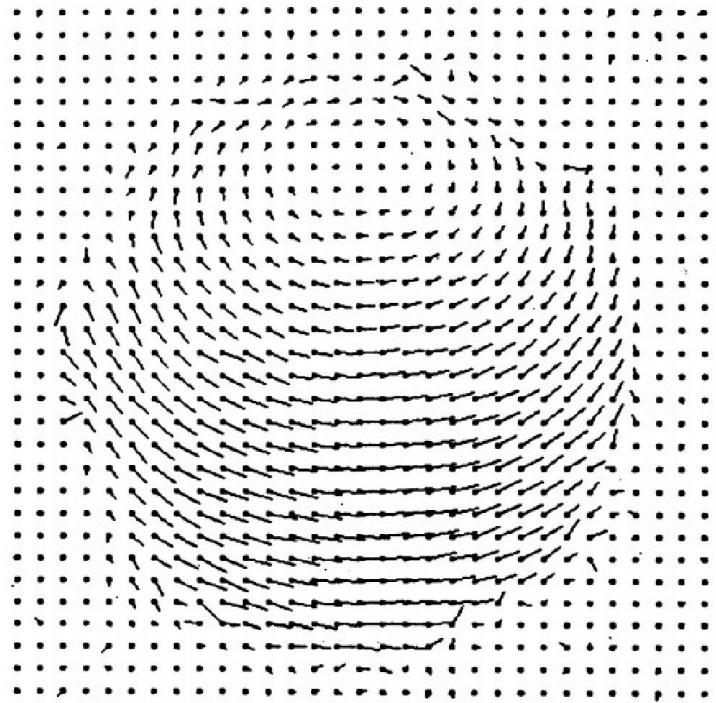


Baker and Binford: Depth from Edge and Intensity Based Stereo. IJCAI, 1981. 68

A brief history of CV

1981: Dense Optical Flow

- Pattern of apparent motion of objects, surfaces, and edges in a visual scene
- Measured by (densely) tracking pixels between two frames
- Investigated by Gibson to describe the visual stimulus of animals
- Horn-Schunck algorithm

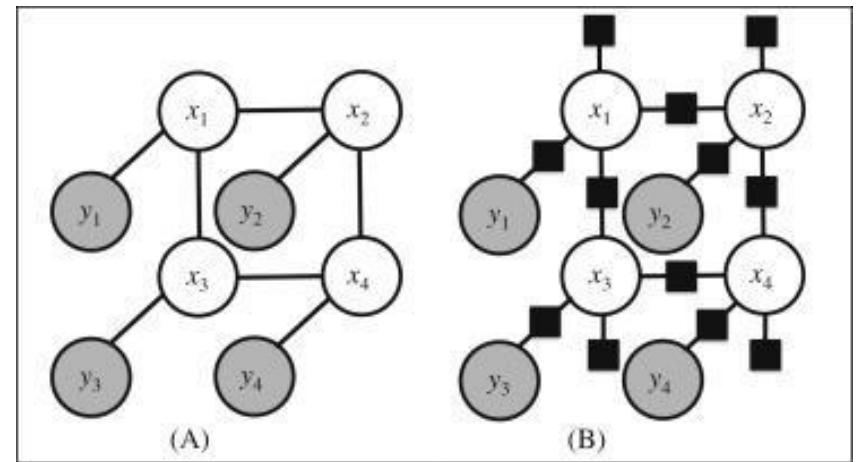


Horn and Schunck: Determining Optical Flow. Artificial Intelligence, 1981.

A brief history of CV

1984: Markov Random Fields

- MRFs for encoding prior knowledge (e.g., about smoothness)
- Resolves ambiguities in many ill-posed vision problems (e.g., stereo, flow, denoising)
- Global optimization (e.g., variational inference, sampling, belief propagation, graph cuts)

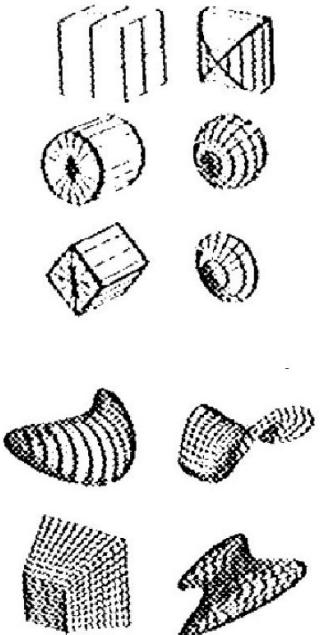
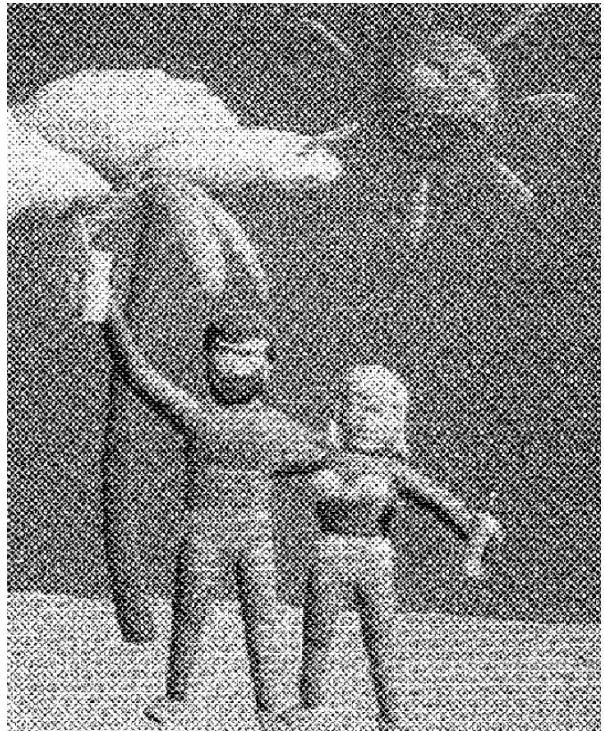


Geman and Geman: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. TPAMI, 1984.

A brief history of CV

1980s: Part-based Models

- 1973: Pictorial Structures
- 1976: Generalized Cylinders (solids of revolution, swept curves)
- 1986: Superquadrics (generalization of quadric surfaces)
- Express complex relationships
- Compact representation

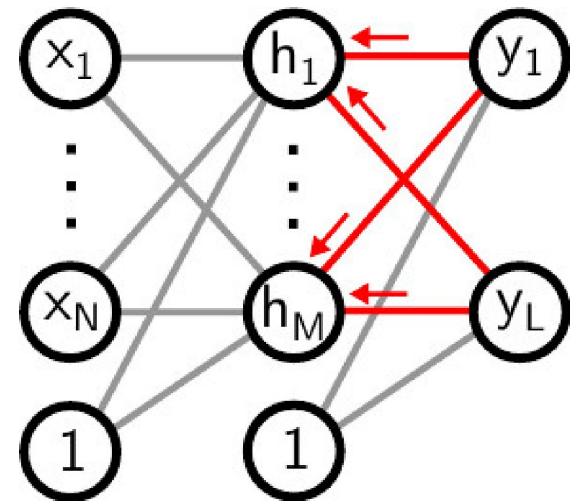


Pentland: Parts: Structured descriptions of shape. AAAI, 1986.

A brief history of CV

1986: Backpropagation Algorithm

- Efficient calculation of gradients in a deep network wrt. network weights
- Enables application of gradient based learning to deep networks
- Known since 1961, but first empirical success in 1986
- Remains main workhorse today

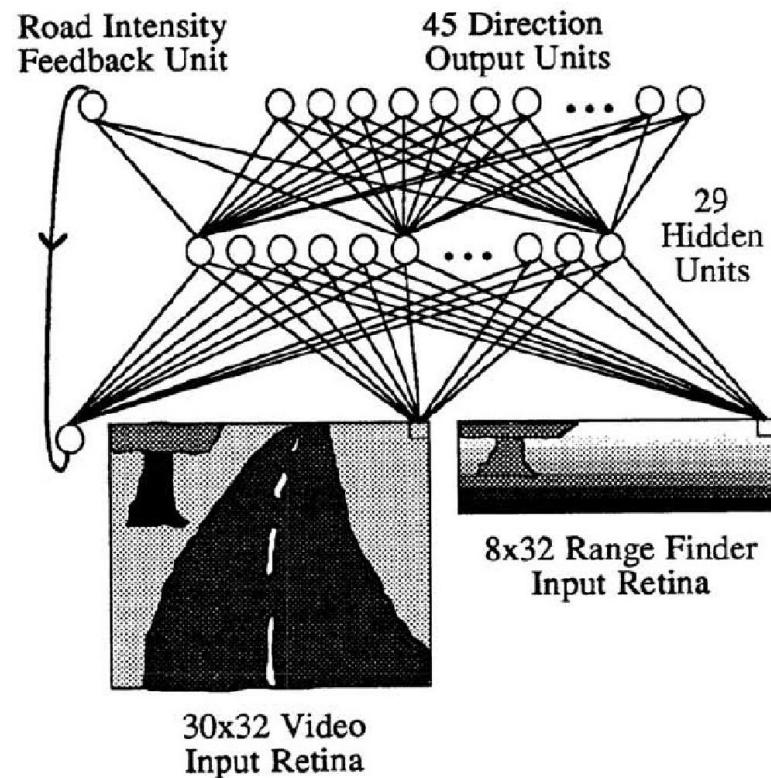


Rumelhart, Hinton and Williams: Learning representations by back-propagating errors. Nature, 1986.

A brief history of CV

1988: Self-Driving Car ALVINN

- Forward-looking, vision based driving
- Fully connected neural network maps road images to vehicle turn radius
- Trained on simulated road images
- Tested on unlined paths, lined city streets and interstate highways
- 90 consecutive miles at up to 70 mph

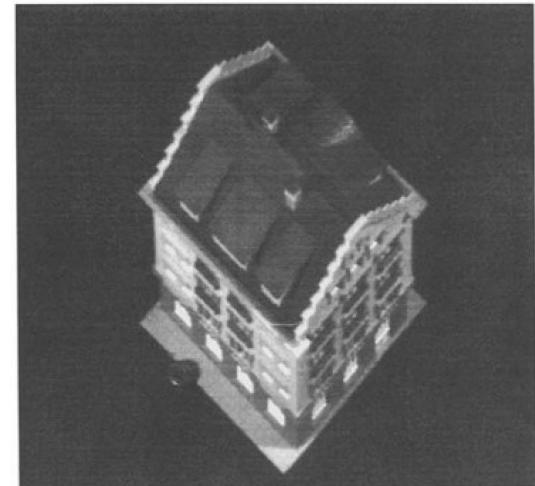


Pomerleau: ALVINN: An Autonomous Land Vehicle in a Neural Network. NIPS, 1988.

A brief history of CV

1992: Structure-from-Motion

- Estimating 3D structures from 2D image sequences of static scenes
- Requires only a single camera
- Tomasi-Kanade factorization provides closed-form (SVD-based) solution for orthographic case
- Today: non-linear least squares

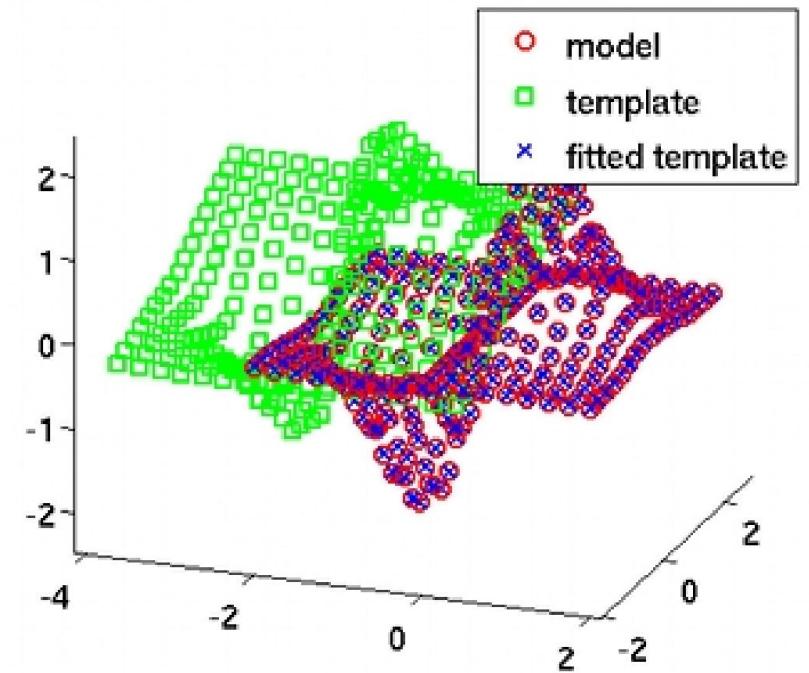


Tomasi and Kanade: Shape and motion from image streams under orthography: a factorization method. IJCV, 1992.

A brief history of CV

1992: Iterative Closest Points

- Registering two point clouds by iteratively optimizing a (rigid or non-rigid) transformation
- Used to aggregate partial 2D or 3D surfaces from different scans, to estimate relative camera poses from point clouds or to localize wrt. a map



Besl and McKay: A Method for Registration of 3-D Shapes. PAMI, 1992.

A brief history of CV

1996: Volumetric Fusion

- Aggregation of multiple implicitly represented surfaces by averaging signed distance values
- Mesh-extraction as post-processing



Photograph of
polychrome original

Range surface
from one scan

Reconstruction
before
hole-filling

Reconstruction
after
hole-filling

Hardcopy

Curless and Levoy: A Volumetric Method for Building Complex Models from Range Images. SIGGRAPH, 1996.

A brief history of CV

1998: Multi-View Stereo

- 3D reconstruction from multiple input images using level-set methods
- Reconstruction vs. image matching
- Proper model of visibility
- Flexible topology
- Provable convergence
- Other approaches (dead-ends): Voxel-coloring, space carving

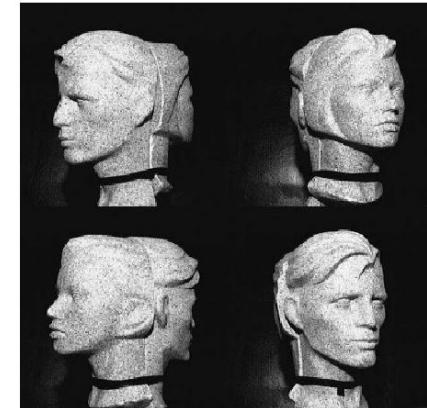
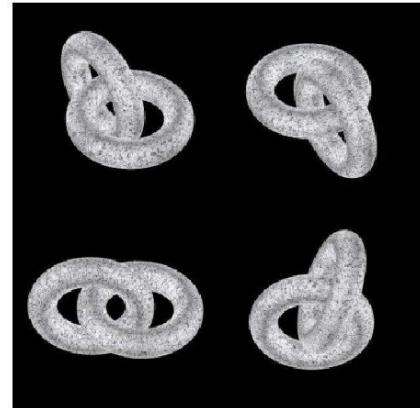


Fig. 3. Multicamera images of 3D objets. On the left hand side, two crossing synthetic tori (24 images). On the right hand side, real images: two human heads (18 images).

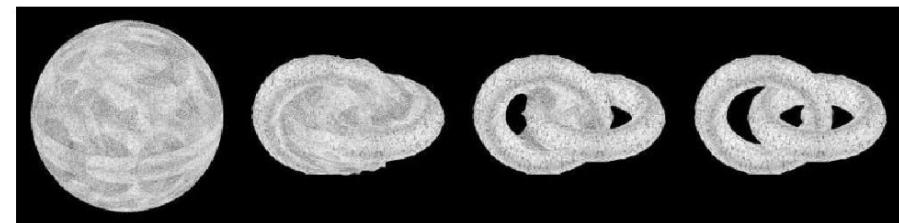


Fig. 4. Evolution of the surface for the two tori.

Faugeras and Keriven: Complete Dense Stereovision Using Level Set Methods. ECCV, 1998.

A brief history of CV

1998: Stereo with Graph Cuts

- Popular discrete MAP inference algorithm for Markov Random Fields
- First versions included unary and pairwise terms
- Later versions also included specific forms of higher-order potentials
- Global reasoning compared to scanline stereo

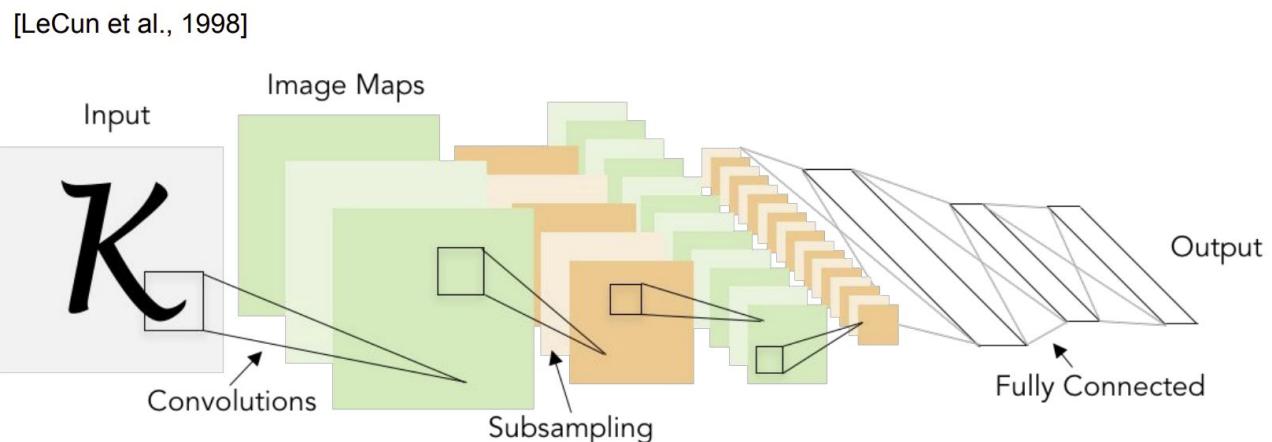


Boykov, Veksler and Zabih: Markov Random Fields with Efficient Approximations. CVPR, 1998.

A brief history of CV

1998: Convolutional Neural Networks

- Similar to Neocognitron, but trained end-to-end using backpropagation
- Implements spatial invariance via convolutions and max-pooling
- Weight sharing reduces parameters
- Tanh/Softmax activations
- Only good results on MNIST so far

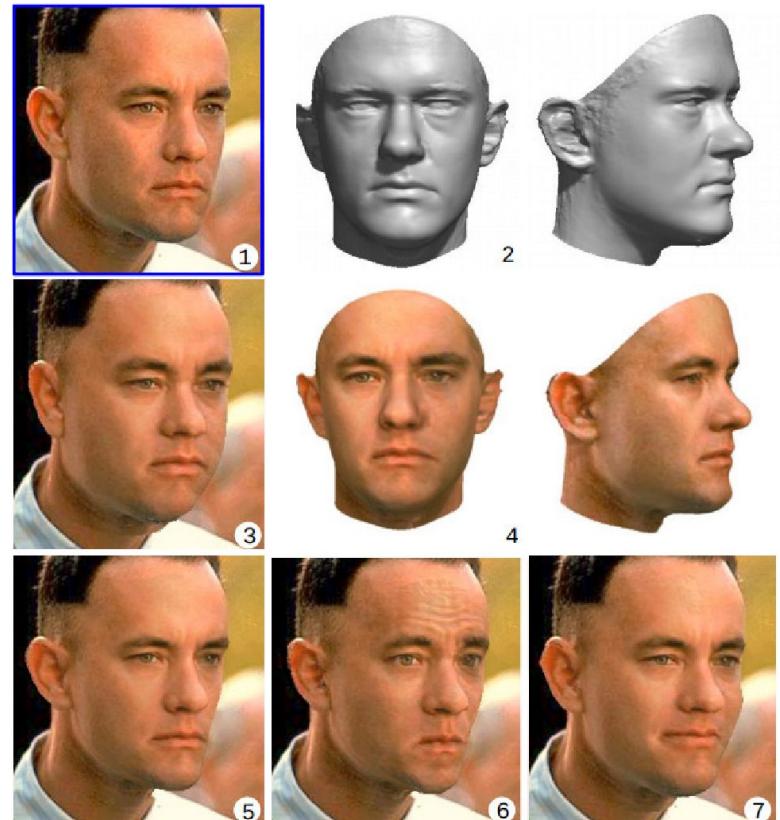


LeCun, Bottou, Bengio and Haffner: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.

A brief history of CV

1999: Morphable Models

- Single-view 3D face reconstruction
- Linear combination of 200 laser scans of faces
- Stunning results at the time

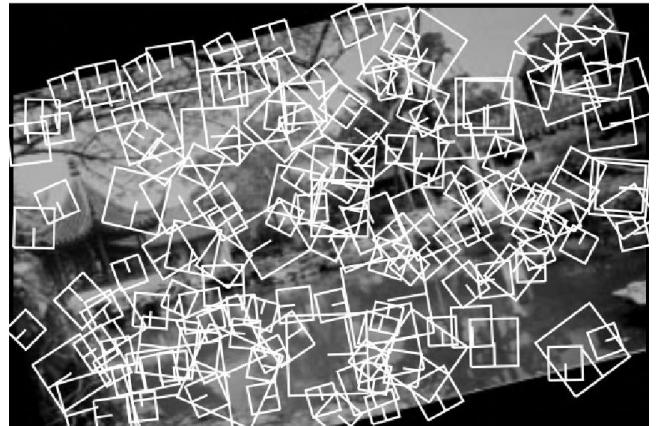
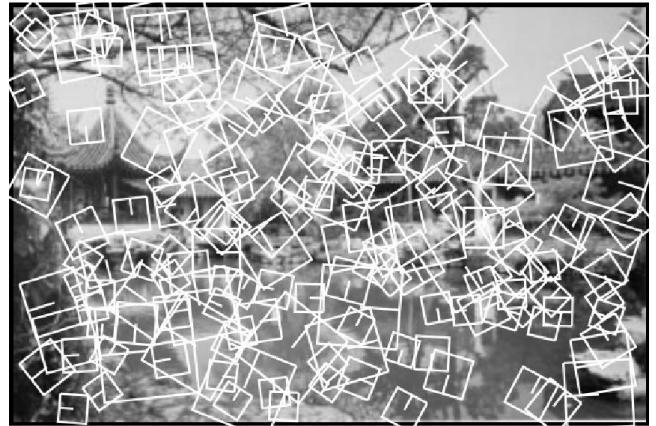


Blanz and Vetter: A Morphable Model for the Synthesis of 3D Faces. SIGGRAPH, 1999.

A brief history of CV

1999: SIFT

- Scale Invariant Feature Transform
- Detection and description of salient local features in an image
- Enabled many applications (e.g., image stitching, reconstruction, motion estimation, . . .)

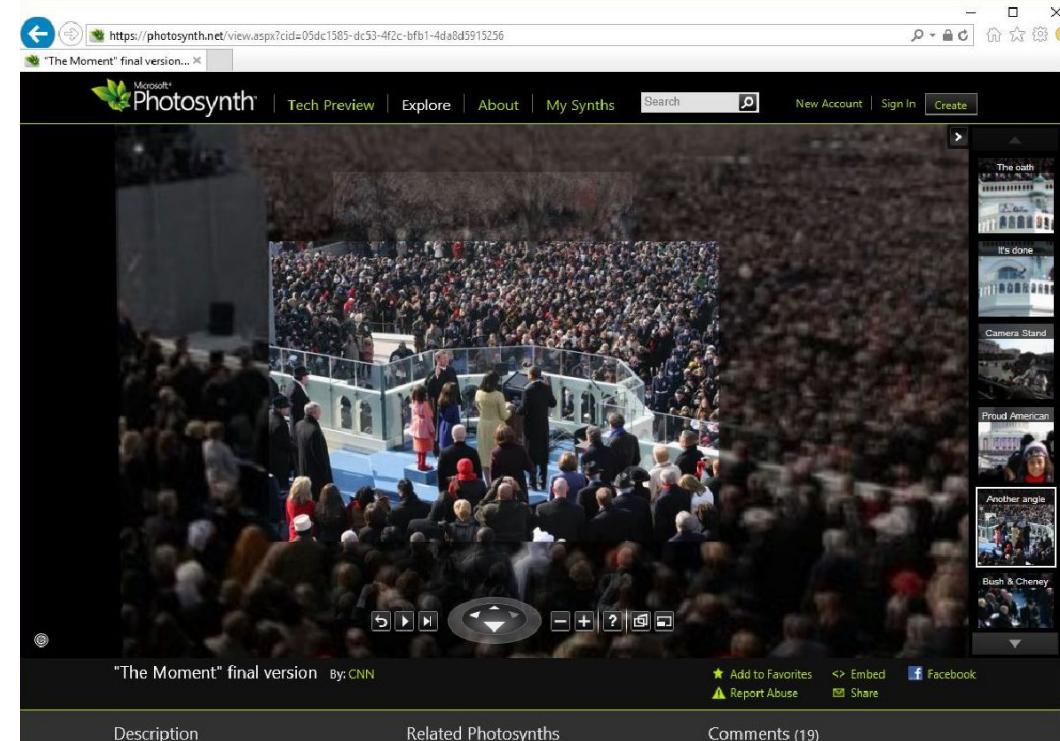


Lowe: Object Recognition from Local Scale-Invariant Features. ICCV, 1999

A brief history of CV

2006: Photo Tourism

- Large-scale 3D reconstruction from internet photos
- Key ingredients: SIFT feature matching, bundle adjustment
- Microsoft Photosynth (discont.)



Snavely, Seitz and Szeliski: Photo tourism: exploring photo collections in 3D. SIGGRAPH, 2006.

A brief history of CV

2007: PMVS

- Patch-based Multi View Stereo
- Robust reconstruction of various small and large objects
- Performance of 3D reconstruction techniques continues to increase

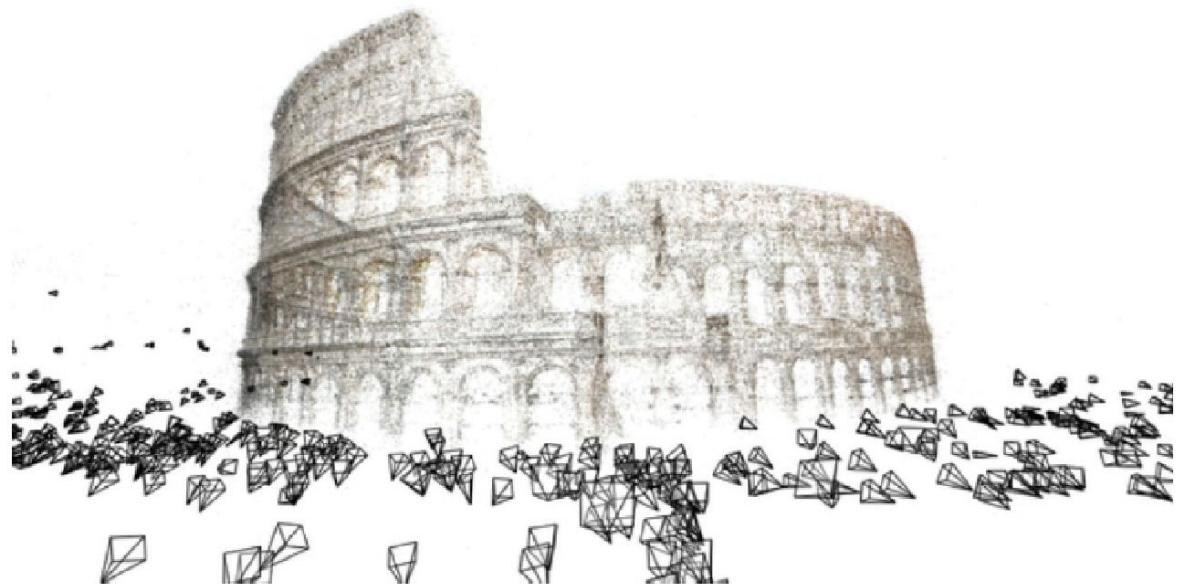


Furukawa and Ponce: Accurate, Dense, and Robust Multi-View Stereopsis. CVPR 2007.

A brief history of CV

2009: Building Rome in a Day

- 3D reconstruction of landmarks and cities from unstructured Internet photo-collections
- Follow-up: Rome on a Cloudless Day

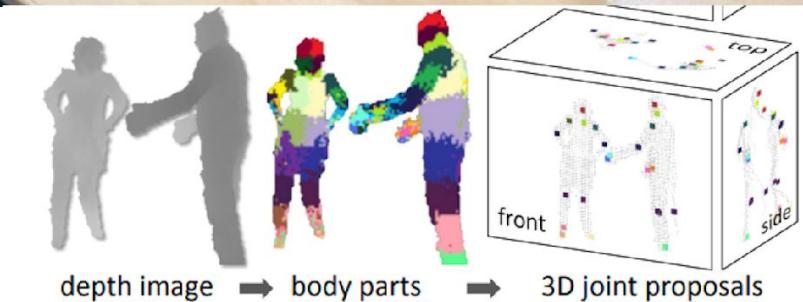


Agarwal, Snavely, Simon, Seitz and Szeliski: Building Rome in a day. ICCV, 2009.

A brief history of CV

2011: Kinect

- Active light 3D sensing
- ML for 3D pose estimation
- Multiple hardware generations
- Early versions failed to commercialize but heavily used for robotics and vision research



A brief history of CV

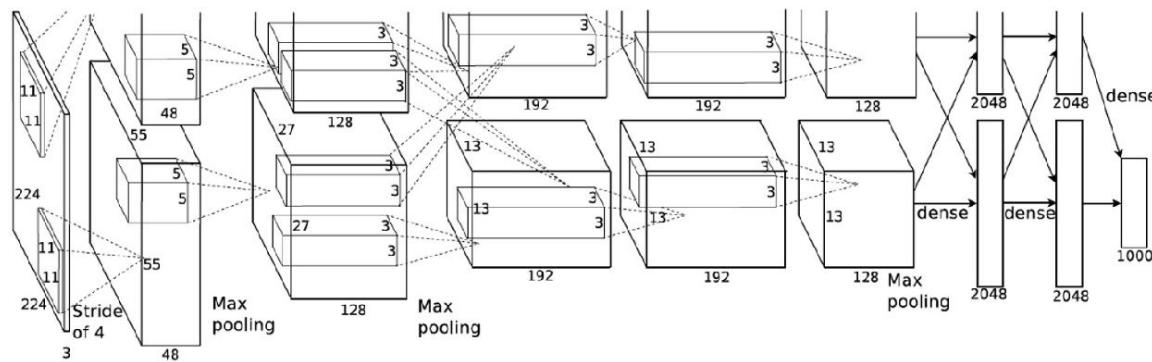
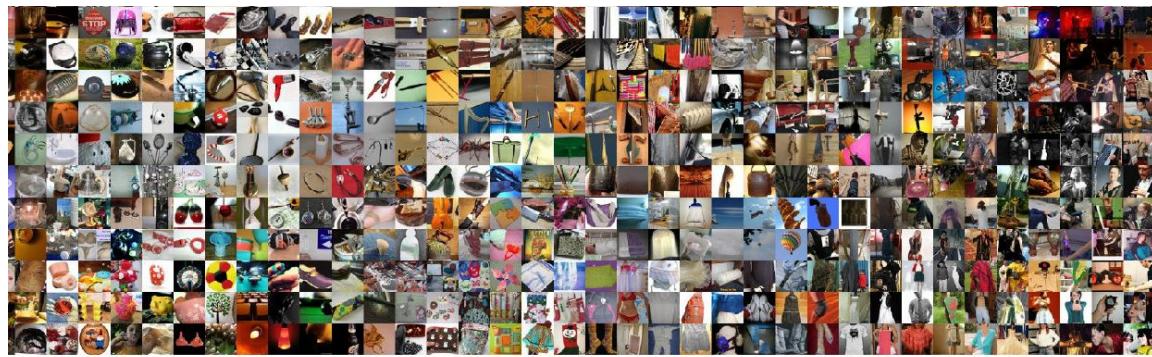
2009-2012: ImageNet and AlexNet

ImageNet <https://www.image-net.org/>

- Recognition benchmark (ILSVRC)
- 10 million annotated images
- 1000 categories

AlexNet

- First neural network to win ILSVRC via **GPU training, deep models, data**



Krizhevsky, Sutskever and Hinton: ImageNet classification with deep convolutional neural networks. NIPS, 2012.

A brief history of CV

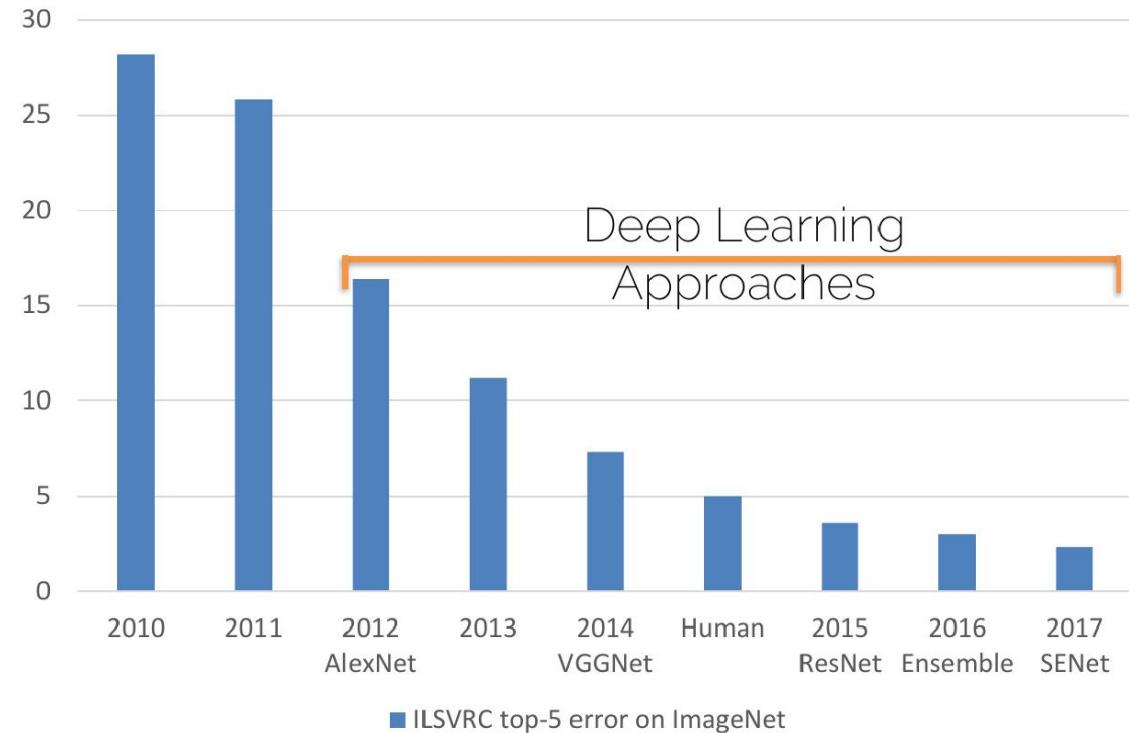
2009-2012: ImageNet and AlexNet

ImageNet <https://www.image-net.org/>

- Recognition benchmark (ILSVRC)
- 10 million annotated images
- 1000 categories

AlexNet

- First neural network to win ILSVRC via **GPU training, deep models, data**
- **Deep learning revolution from now on**

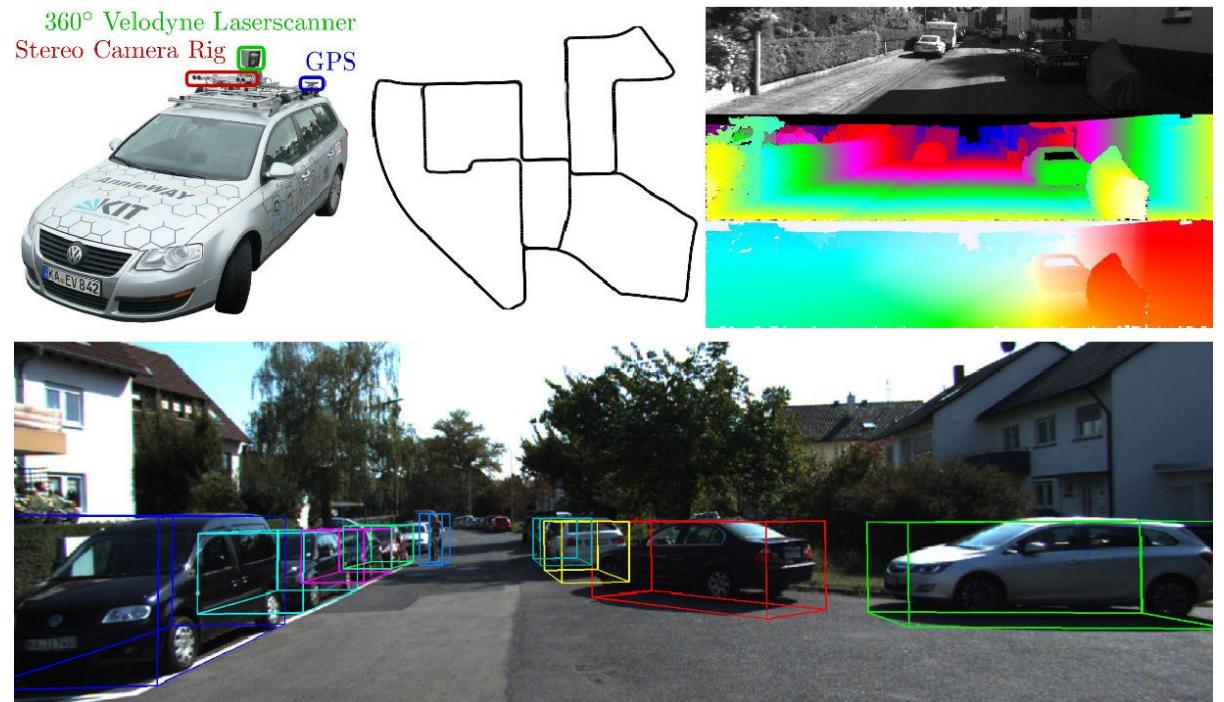


Krizhevsky, Sutskever and Hinton: ImageNet classification with deep convolutional neural networks. NIPS, 2012.

A brief history of CV

2002-now: Golden Age of Datasets

- Middlebury Stereo and Flow
- KITTI, Cityscapes: Self-driving
- PASCAL, MS COCO: Recognition
- ShapeNet, ScanNet: 3D DL
- Visual Genome: Vision/Language
- MITOS: Breast cancer

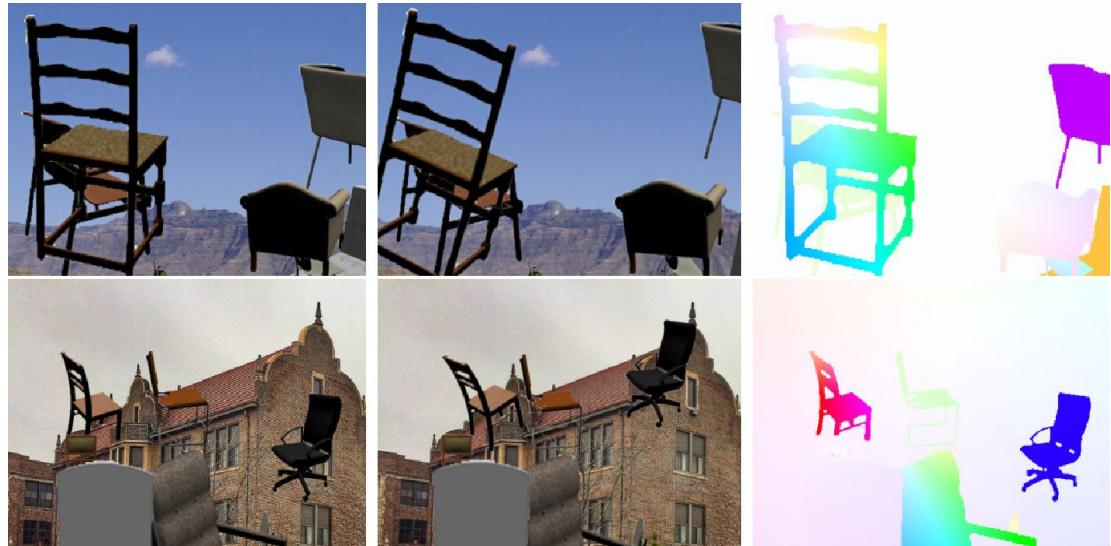


Geiger, Lenz and Urtasun: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. CVPR, 2012.

A brief history of CV

2012-now: Synthetic Data

- Annotating real data is expensive
- Led to surge of synthetic datasets
- Creating 3D assets is also costly
- But even very simple 3D datasets proved tremendously useful for pre-training (e.g., in optical flow)

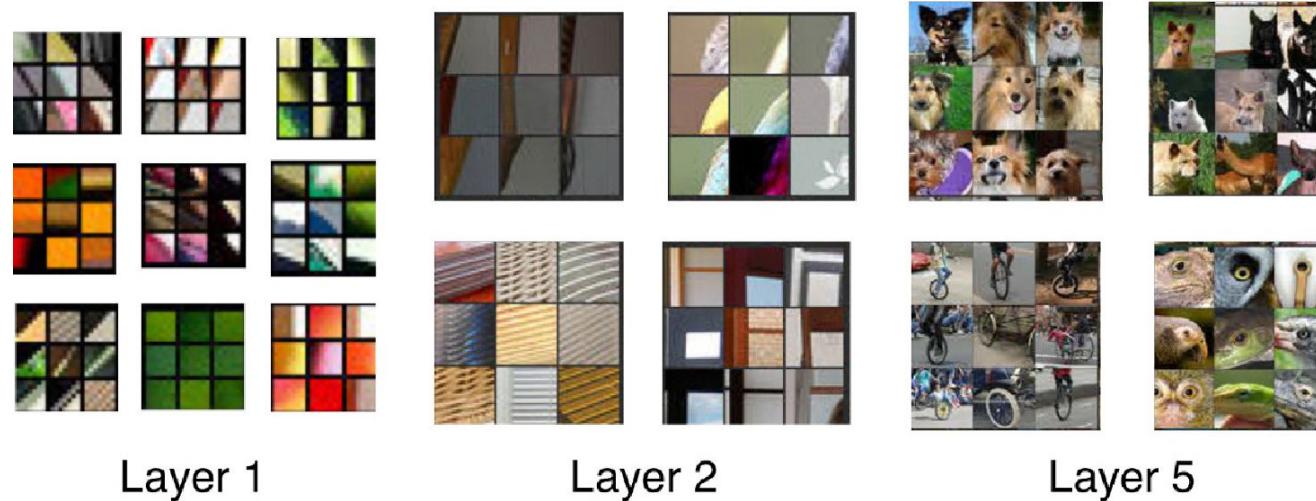


Dosovitskiy et al.: FlowNet: Learning Optical Flow with Convolutional Networks. ICCV, 2015.

A brief history of CV

2014: Visualization

- Goal: provide insights into what the network (black box) has learned
- Visualized image regions that most strongly activate various neurons at different layers of the network
- Found that higher levels capture more abstract semantic information

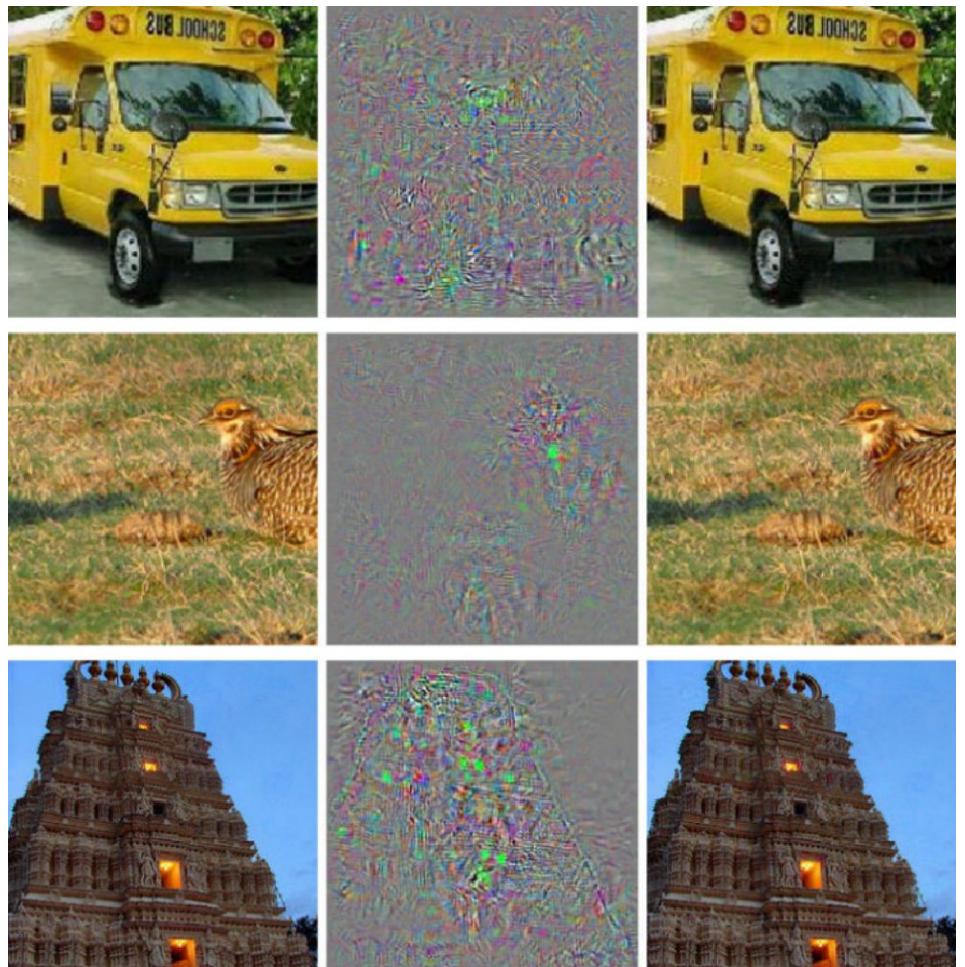


Zeiler and Fergus: CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. CVPR Workshops, 2014.

A brief history of CV

2014: Adversarial Examples

- Accurate image classifiers can be fooled by imperceptible changes (here magnified for visibility)
- All images in the right column are classified as “ostrich”



Szegedy et al.: Intriguing properties of neural networks. ICLR, 2014.

A brief history of CV

2014: Generative Adversarial Networks

- Deep generative models (VAEs, GANs) produce compelling images
- StyleGAN2 is state-of-the-art
- Results on faces hard to distinguish from real images
- Active research on image translation, domain adaptation, content and scene generation and 3D GANs

Moore's Law of AI
4.5 years of progress on faces

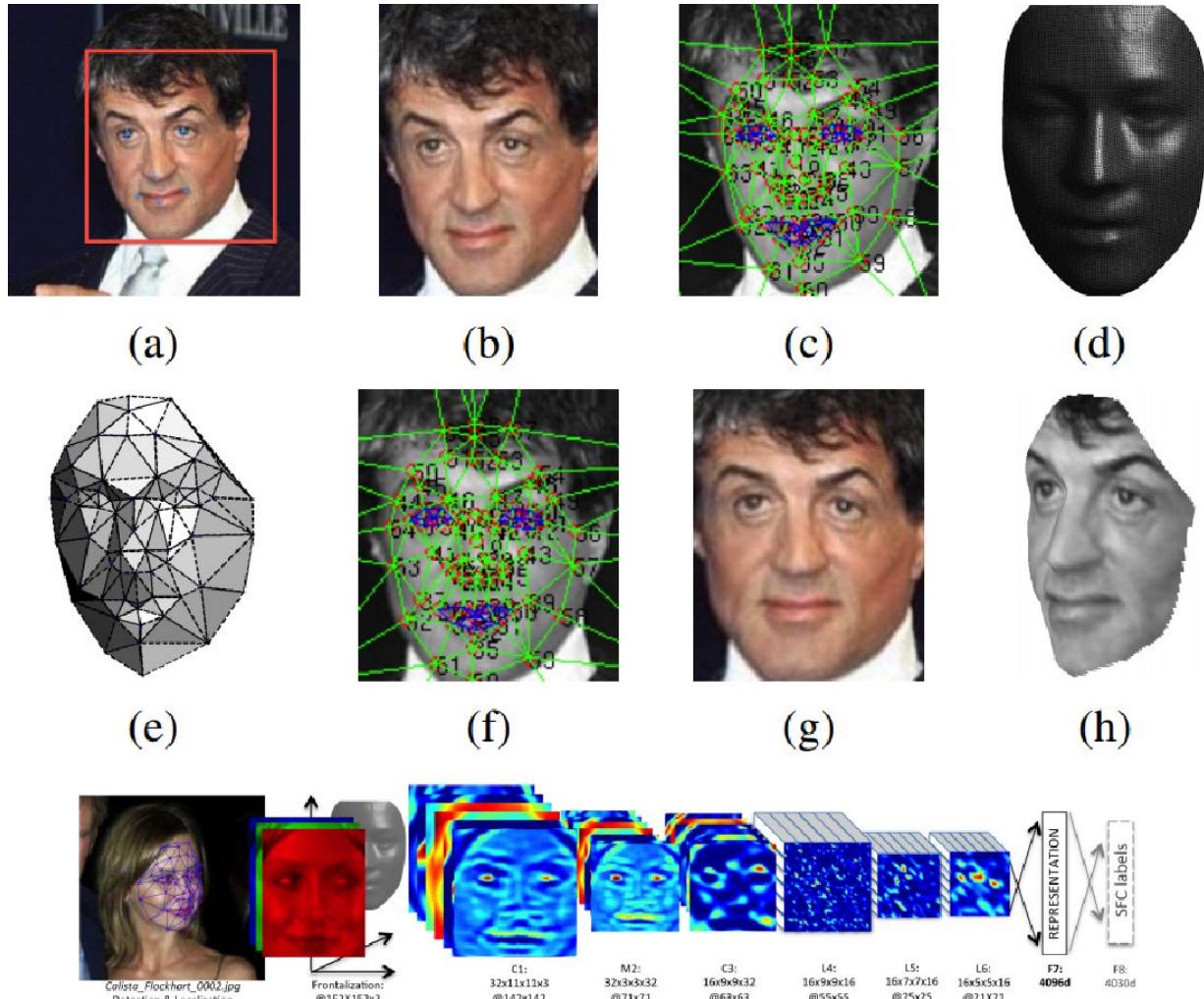


Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio: Generative Adversarial Networks. NIPS, 2014.

A brief history of CV

2014: DeepFace

- Combination of model-based alignment with deep learning for face recognition
- First model to reach human-level face recognition performance

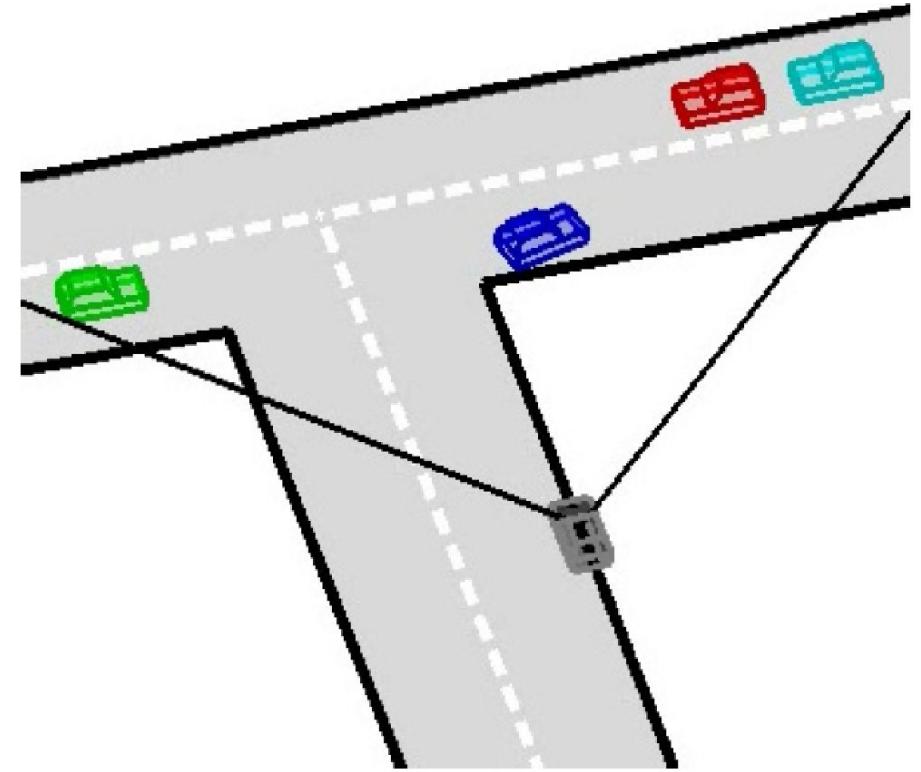


Taigman, Yang, Ranzato and Wolf: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR, 2014.

A brief history of CV

2014: 3D Scene Understanding

- Parsing RGB and RGB-D images into holistic 3D scene representations
- Methods for indoors and outdoors



Geiger, Lauer, Wojek, Stiller and Urtasun: 3D Traffic Scene Understanding From Movable Platforms. PAMI, 2014.

A brief history of CV

2014: 3D Scanning

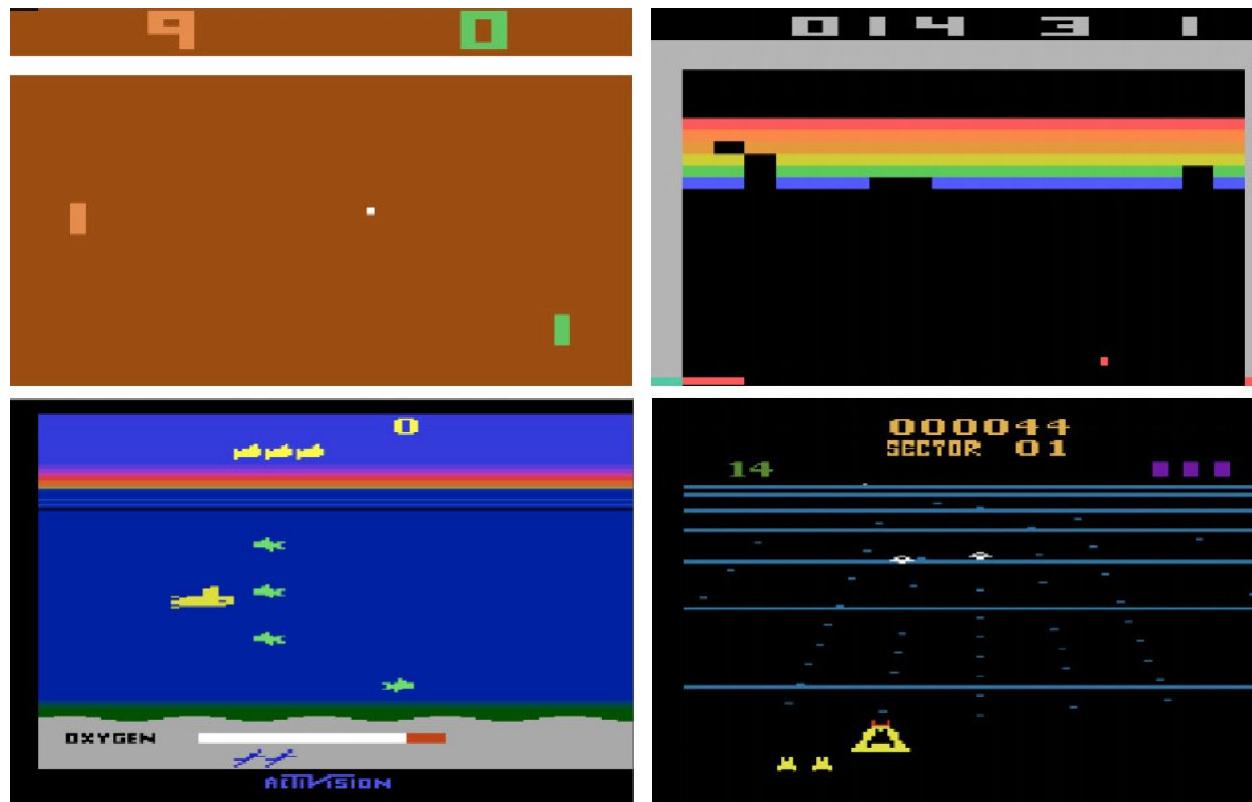
- 3D scanning techniques allow for creating accurate replicas
- Debevec's team scans Obama
- Exhibition in Smithsonian
<https://dpo.si.edu/blog/>



A brief history of CV

2015: Deep Reinforcement Learning

- Learning a policy (state → action) through random exploration and reward signals (e.g., game score)
- No other supervision
- Success on many Atari games
- But some games remain hard



Mnih et al.: Human-level control through deep reinforcement learning. Nature, 2015.

A brief history of CV

2016: Style Transfer

- Manipulate photograph to adopt style of another image (painting)
- Uses deep network pre-trained on ImageNet for disentangling content from style
- <https://deepart.io/>

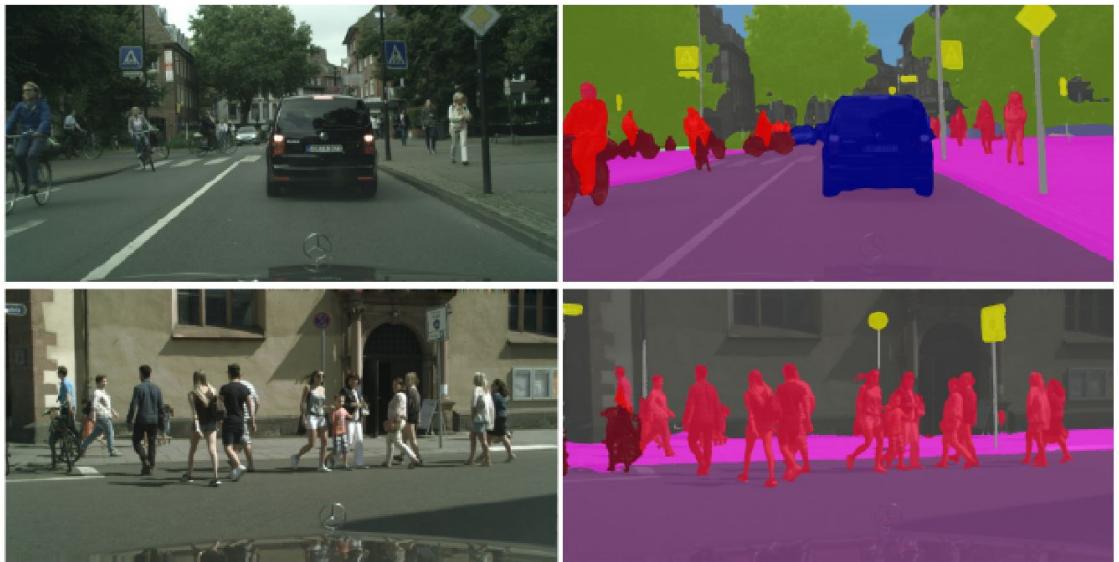


Gatys, Ecker and Bethge: Image Style Transfer Using Convolutional Neural Networks. CVPR, 2016.

A brief history of CV

2015-2017: Semantic Segmentation

- Assign semantic class to every pixel
- Semantic segmentation starts to work on challenging real-world datasets (e.g., CityScapes)
- 2015: FCN, SegNet
- 2016: DeepLab, FSO
- 2017: DeepLabv3

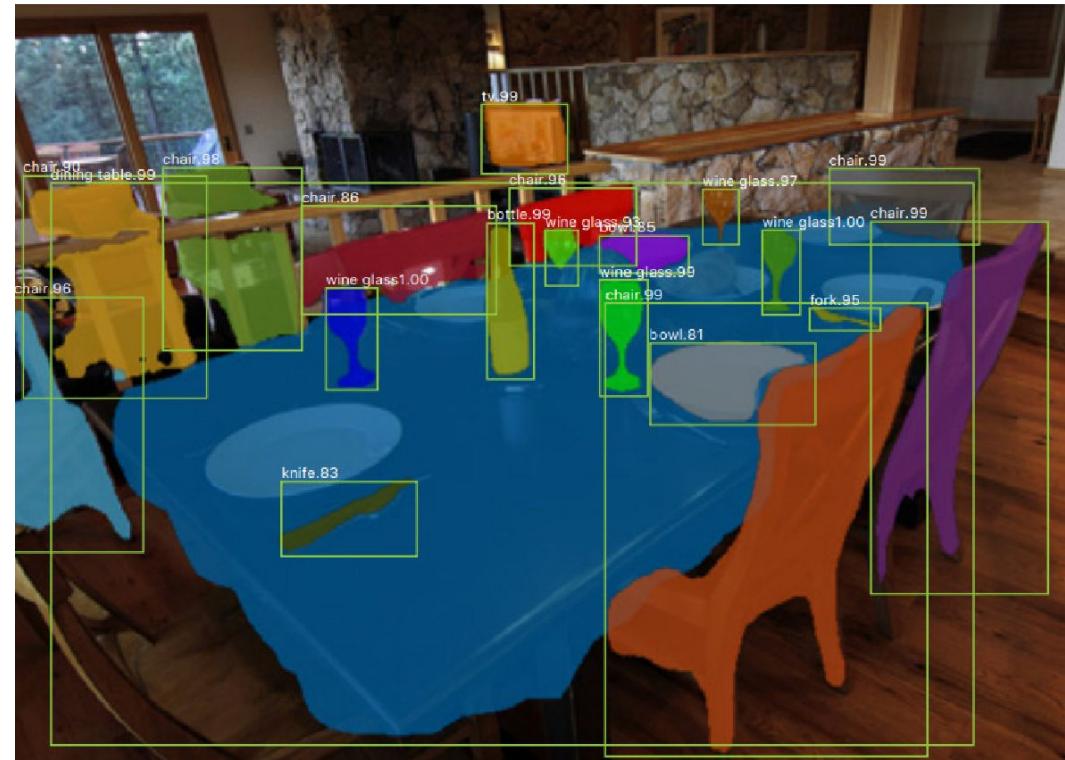


Kundu, Vineet and Koltun: Feature Space Optimization for Semantic Video Segmentation. CVPR, 2016.

A brief history of CV

2017: Mask R-CNN

- Deep neural network for joint object detection and instance segmentation
- Outputs “structured object”, not only a single number (class label)
- State-of-the-art on MS-COCO



A brief history of CV

2016-now: YOLO

- Real-Time Object Detector
- YOLO v12 (2025) object detection, classification, and segmentation tasks



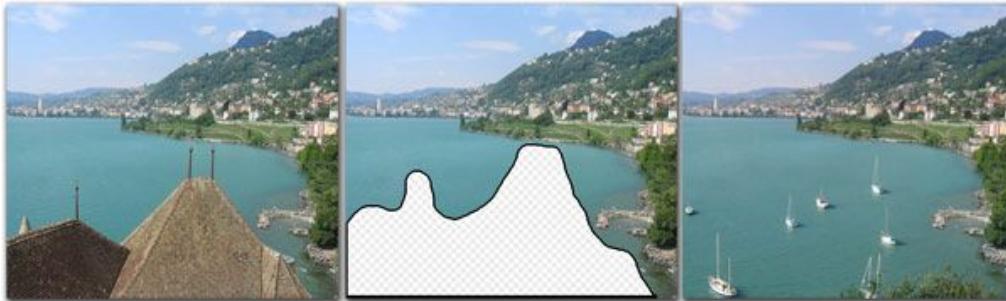
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In 2016 CVPR

A brief history of CV

- Improve photos (“Computational Photography”)



Super-resolution (source: 2d3)



Inpainting / image completion
(image credit: Hays and Efros)



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

A brief history of CV

2017: Image Captioning

- Growing interest in combining vision with language
- Several new tasks emerged including image captioning and visual question answering
- However, models still lack understanding / commonsense



"little girl is eating piece of cake."



"baseball player is throwing ball in game."



"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."



"a young boy is holding a baseball bat."



"a cat is sitting on a couch with a remote control."



"a woman holding a teddy bear in front of a mirror."



"a horse is standing in the middle of a road."

A brief history of CV

2018: Human Shape and Pose

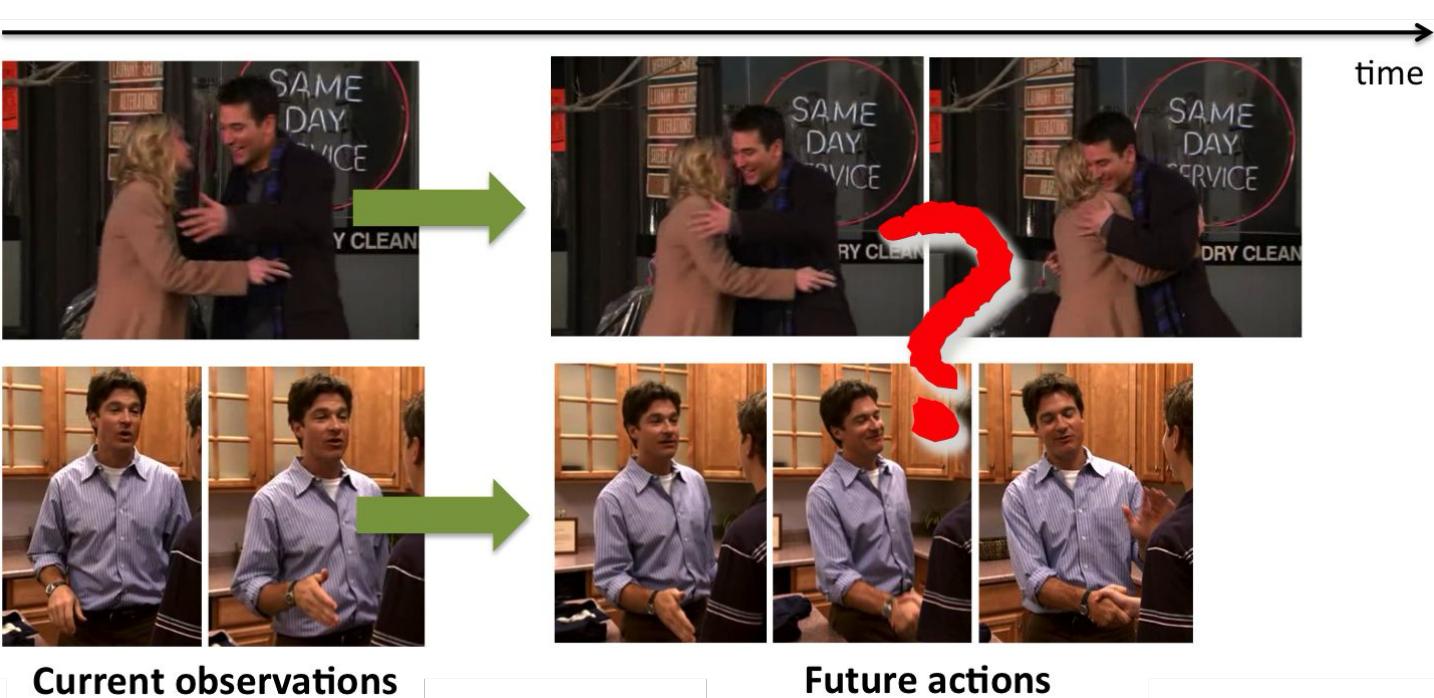
- Human pose/shape models mature
- Rich parametric models (e.g., SMPL, STAR)
- Regression from RGB images only
- Models of pose-dependent deformation and clothing



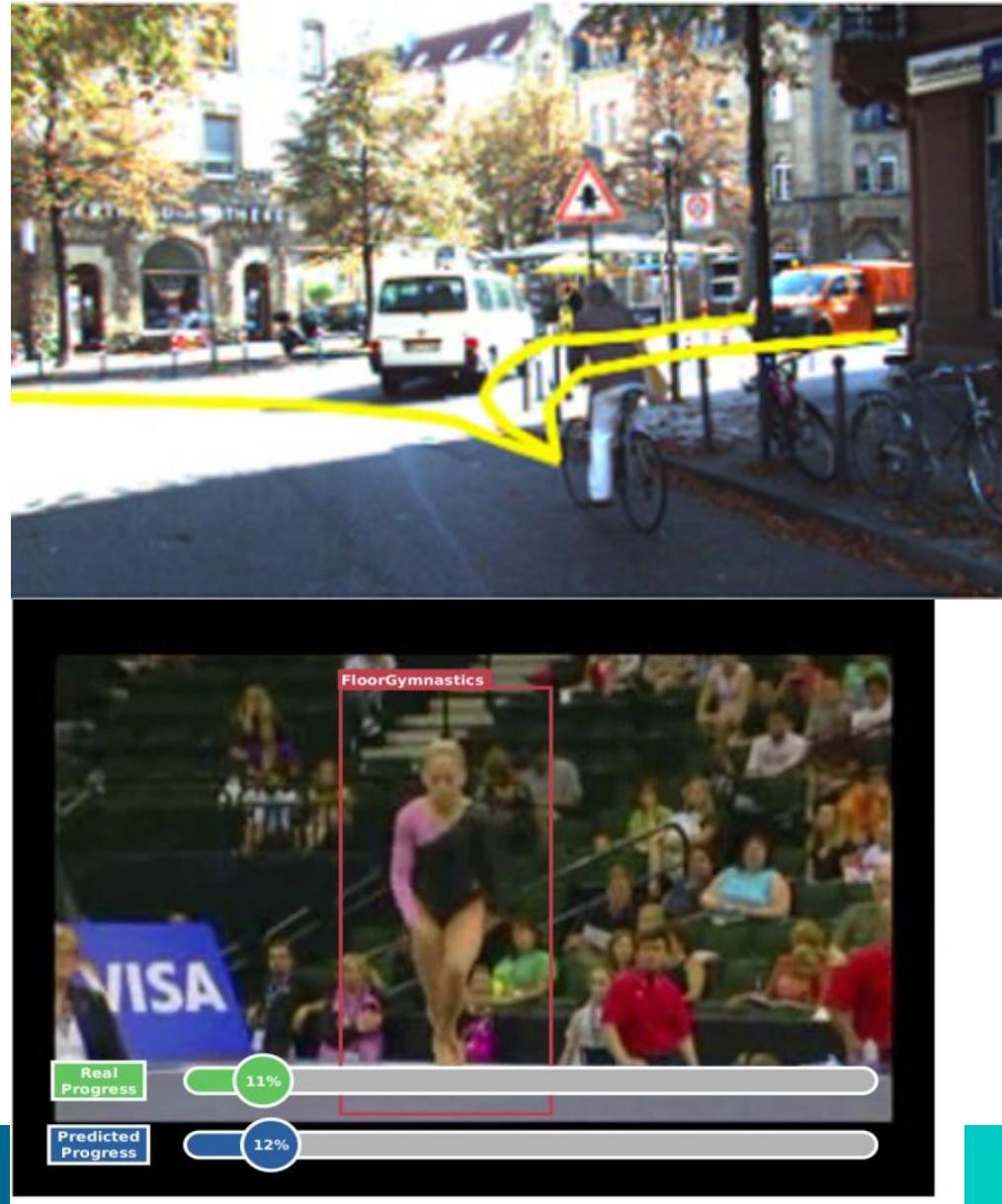
Kanazawa, Black, Jacobs and Malik: End-to-End Recovery of Human Shape and Pose. CVPR, 2018.

A brief history of CV

- Action recognition and forecasting



Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. CVPR 2018



A brief history of CV

2019: DeepFake and Deepfake detection

- Generation of one of the first deepfake dataset FaceForensics++
- First detector to distinguish fake/real images of faces

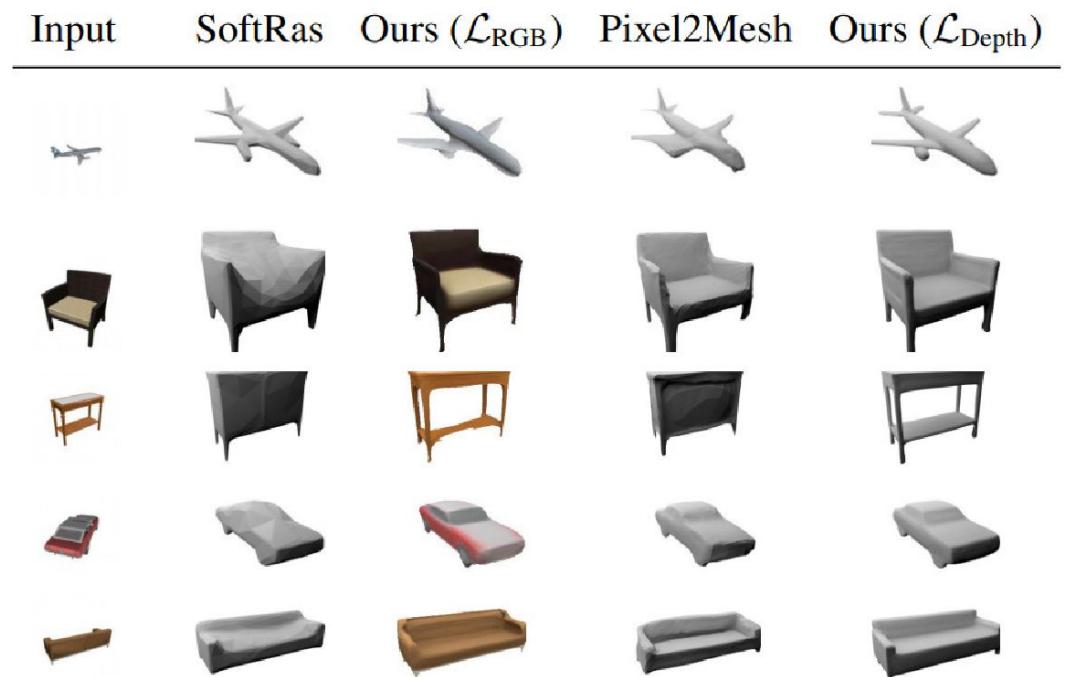


A. Rossler, et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in ICCV 2019

A brief history of CV

2016-2020: 3D Deep Learning

- First deep models to output 3D representations
- Voxels, point clouds, meshes, implicit representations
- Prediction of 3D models even from a single image
- Geometry, materials, light, motion

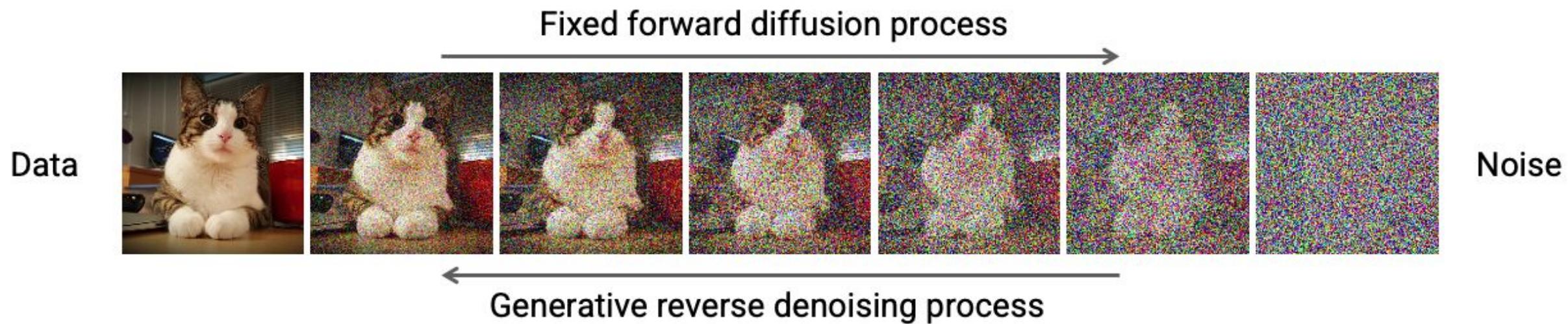


Niemeyer, Mescheder, Oechsle, Geiger: Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. CVPR, 2020

A brief history of CV

2020-now: Diffusion models

- Generative process for images and videos
 - Conditioned or not
 - They have gained popularity more and more (many commercial products DALL-E, Midjourney, RODIN, SORA etc..)



Jonathan Ho et al. Denoising diffusion probabilistic models. In NeurIPS, 2020.

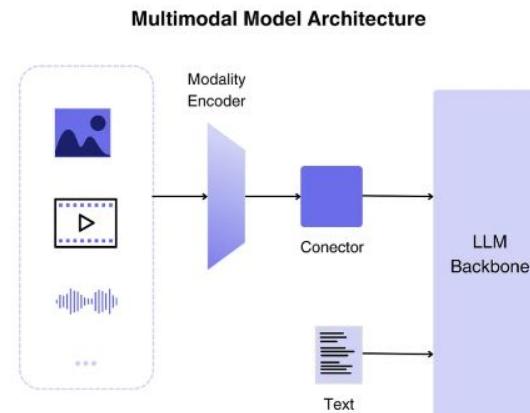
A brief history of CV

2024-now: Multimodal Large Language Models (MLLMs)

Evolution of Multimodal Large Language Models (MLLMs)



Figure 2. Some of the Multimodal Large Language Models (MLLMs) developed between 2022 and 2024



(1) Describe the scene

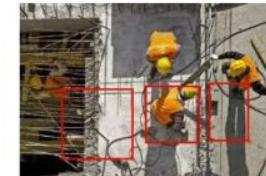
GPT-4o: The scene depicts a construction site viewed from above. The area appears to be a building under construction, with various construction materials and equipment present. There are four workers in the scene, all engaged in different construction activities, such as handling planks and wires. The workers are wearing orange safety vests and yellow hard hats.

(2) Number of workers with a hard hat

GPT-4o: All four workers are wearing hard hats.

(3) Bounding box coordinates

GPT-4o: Top left worker: $(x_1, y_1, x_2, y_2) = (200, 150, 230, 180)$. Top right worker: $(x_1, y_1, x_2, y_2) = (300, 150, 330, 180)$. Bottom left worker: $(x_1, y_1, x_2, y_2) = (200, 250, 230, 280)$. Bottom right worker: $(x_1, y_1, x_2, y_2) = (300, 250, 330, 280)$. Note: The exact coordinates above are estimated visually and may not be precise without actual image processing tools.



Construction site: detecting and counting hard hats

Ex 1. Question/Prompt: As an AI system working on the construction industry, you excel at analyzing and reporting safety issues that might occur in a construction site. Given the following image, please, (1) describe the scene, (2) provide the **number of workers** wearing a **hard hat**, (3) provide the **bounding box** coordinates of the hard hats identified in (2).



ml-ferret

LLaVA: The image shows a construction site with workers engaged in various activities. There are individuals wearing hard hats, which indicates that they are working on a construction project and are required to wear protective gear to comply with safety regulations.

ML-Ferret: The scene displays a construction site with a building under construction. A group of workers, dressed in orange jumpsuits, are actively engaged in various tasks related to the construction process.

A brief history of CV

- Applications and commercial products



Google Portrait Mode

Image credits Andreas Geiger



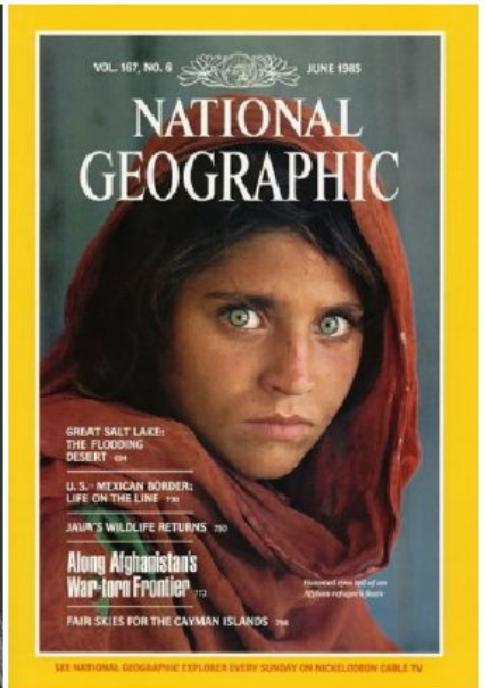
Skydio 2 Drone



Self-Driving Cars



Microsoft HoloLens



Iris Recognition

Current Challenges

- Un-/Self-Supervised Learning
- Interactive learning
- Accuracy (e.g., self-driving)
- Robustness and generalization
- Inductive biases
- Understanding and mathematics
- Memory and compute
- Ethics and legal questions

Acknowledgements: some slides and material from Bernt Schiele, Mario Fritz, Michael Black, Bill Freeman, Fei-Fei Li, Justin Johnson, Serena Yeung, R. Szeliski, Ioannis Gkioulekas, Roni Sengupta, Andreas Geiger

Contacts

Alcor Lab



WEBSITE
<https://alcorlab.diag.uniroma1.it/>



EMAIL
alcor@diag.uniroma1.it

Personal contact



PROF. Irene Amerini
amerini@diag.uniroma1.it



SAPIENZA
UNIVERSITÀ DI ROMA