Introduction to Computer Vision

Lecture 1 - Overview

Prof. He Wang

About Me

- 王鹤
- Assistant Professor in Center on Frontiers of Computing Studies (CFCS)
- Joined PKU in September, 2021
- Received Ph.D. from Stanford in 2021
- Received Bachelor from Tsinghua in 2014

- Our lab: *Embodied Perception and InteraCtion (EPIC) Lab*
- Research interest: 3D vision, Robotics
- Homepage: https://hughw19.github.io/

北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

Stanford University

清華大學
Tsinghua University

Course Logistics

Objective: A Great Course on Computer Vision

- A self-included beginning course on computer vision
- A broad coverage of classic and deep vision from a modern perspective, to distinguish from online available vision courses
- To lay a solid foundation for applying and researching in computer vision

Logistics

- Instructor
  - He Wang (hewang@pku.edu.cn)
  - Office Hour: Friday 5:00PM - 6:00PM or under appoint.
  - Office location: Room 106-1, Courtyard No.5, Jingyuan
- TAs:
  - Mi Yan (dorisyan@pku.edu.cn)
  - Hao Shen (2301112029@pku.edu.cn)
  - Yuxing Chen (yuxingc_20@stu.pku.edu.cn)
- Class Time & Location
  - Wednesday 3:10PM - 6:00PM
  - Room 507, Teaching Building 2, Peking University

Prerequisite

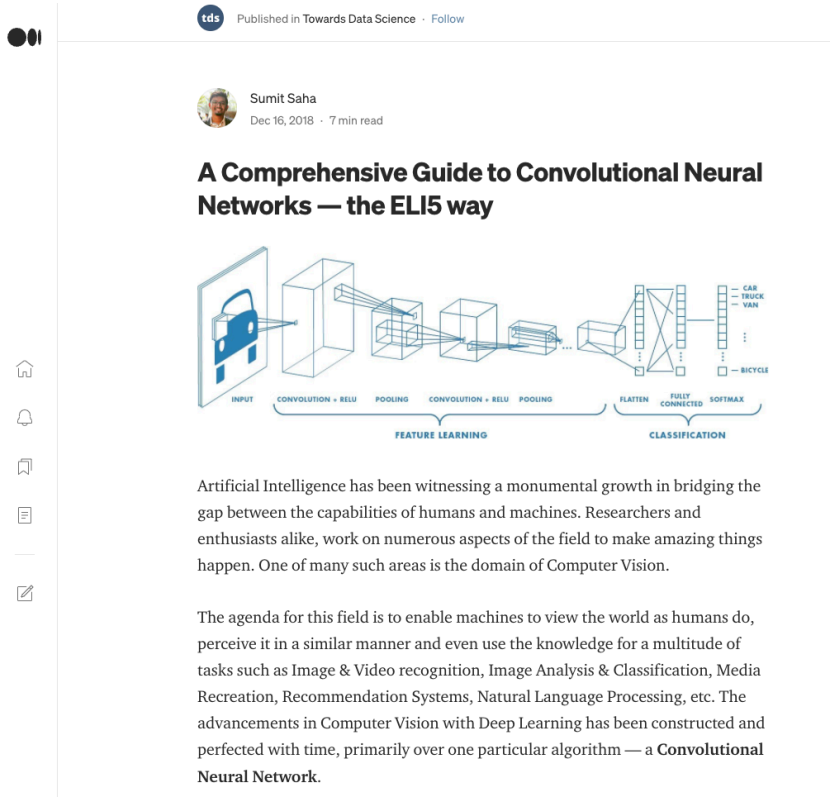
- Math:
  - Calculus
  - Linear Algebra
  - Basics of probability and statistics
- Proficiency in Python
- Optionally, have taken *Introduction to AI* or know some basic knowledge of machine learning and neural networks.

Books and References

- No required textbooks.
- Books for references:
  - On Deep Learning:
    - Ian Goodfellow and Yoshua Bengio and Aaron Courville. *Deep Learning.* MIT Press, 2016.
  - On Classic Computer Vision:
    - R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer, 2011.
    - D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach (2nd Edition).* Prentice Hall, 2011.

Books and References

- The most effective way to learn and check things:
  - Just google it!
  - Search in wikipedia!



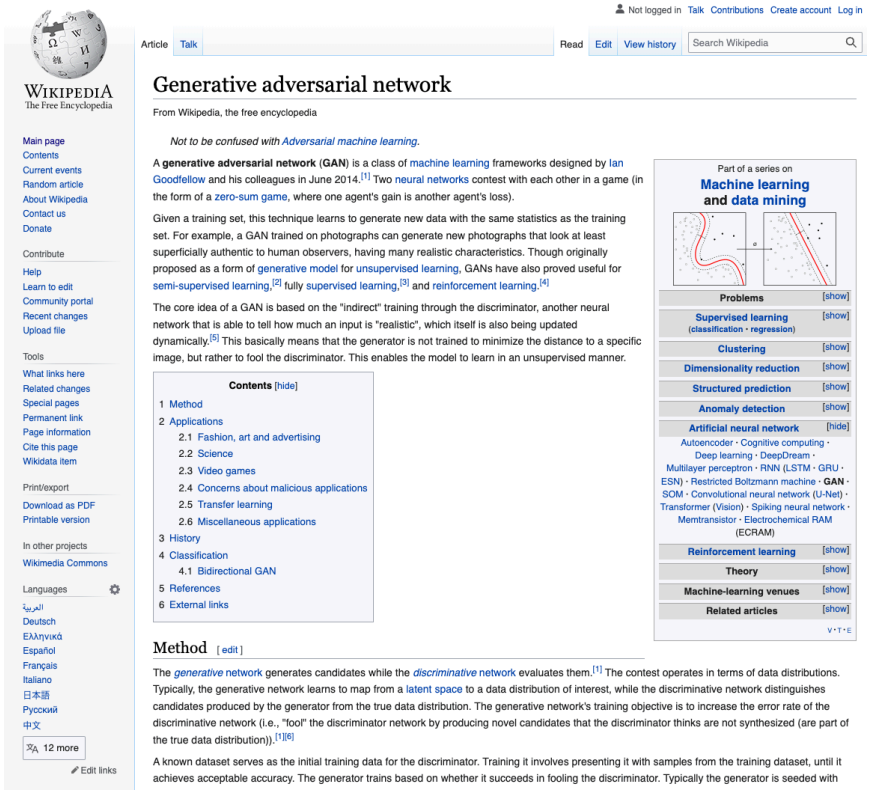
Published in Towards Data Science · Follow

Sumit Saha
Dec 16, 2018 · 7 min read

A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way

Artificial Intelligence has been witnessing a monumental growth in bridging the gap between the capabilities of humans and machines. Researchers and enthusiasts alike, work on numerous aspects of the field to make amazing things happen. One of many such areas is the domain of Computer Vision.

The agenda for this field is to enable machines to view the world as humans do, perceive it in a similar manner and even use the knowledge for a multitude of tasks such as Image & Video recognition, Image Analysis & Classification, Media Recreation, Recommendation Systems, Natural Language Processing, etc. The advancements in Computer Vision with Deep Learning has been constructed and perfected with time, primarily over one particular algorithm — a **Convolutional Neural Network**.



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate
Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file
Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item
Print/export
Download as PDF
Printable version
In other projects
Wikimedia Commons
Languages
العربية
Deutsch
Ελληνικά
Español
Français
Italiano
日本語
Русский
中文
12 more
Edit links

Generative adversarial network

From Wikipedia, the free encyclopedia

Not to be confused with Adversarial machine learning.

A **generative adversarial network** (**GAN**) is a class of machine learning frameworks designed by Ian Goodfellow and his colleagues in June 2014.[1] Two neural networks contest with each other in a game (in the form of a zero-sum game, where one agent's gain is another agent's loss).

Given a training set, this technique learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics. Though originally proposed as a form of generative model for unsupervised learning, GANs have also proved useful for semi-supervised learning,[2] fully supervised learning,[3] and reinforcement learning.[4]

The core idea of a GAN is based on the "indirect" training through the discriminator, another neural network that is able to tell how much an input is "realistic", which itself is also being updated dynamically.[5] This basically means that the generator is not trained to minimize the distance to a specific image, but rather to fool the discriminator. This enables the model to learn in an unsupervised manner.

Contents [hide]
1 Method
2 Applications
2.1 Fashion, art and advertising
2.2 Science
2.3 Video games
2.4 Concerns about malicious applications
2.5 Transfer learning
2.6 Miscellaneous applications
3 History
4 Classification
4.1 Bidirectional GAN
5 References
6 External links

Part of a series on
Machine learning and data mining

Problems [show]
Supervised learning (classification · regression) [show]
Clustering [show]
Dimensionality reduction [show]
Structured prediction [show]
Anomaly detection [show]
Artificial neural network [hide]
Autoencoder · Cognitive computing · Deep learning · DeepDream · Multilayer perceptron · RNN (LSTM · GRU · ESN) · Restricted Boltzmann machine · GAN · SOM · Convolutional neural network (U-Net) · Transformer (Vision) · Spiking neural network · Memtransistor · Electrochemical RAM (ECRAM)
Reinforcement learning [show]
Theory [show]
Machine-learning venues [show]
Related articles [show]
v · t · e

Method [edit]

The *generative* network generates candidates while the *discriminative* network evaluates them.[1] The contest operates in terms of data distributions. Typically, the generative network learns to map from a latent space to a data distribution of interest, while the discriminative network distinguishes candidates produced by the generator from the true data distribution. The generative network's training objective is to increase the error rate of the discriminative network (i.e., "fool" the discriminator network by producing novel candidates that the discriminator thinks are not synthesized (are part of the true data distribution)).[1][6]

A known dataset serves as the initial training data for the discriminator. Training it involves presenting it with samples from the training dataset, until it achieves acceptable accuracy. The generator trains based on whether it succeeds in fooling the discriminator. Typically the generator is seeded with

Courseworks and Grading Policy

- 4 assignments: each 10%, in total 40%
- 1 midterm exam: 30%
- 1 final exam: 30%
- Class/discussion board participation: up to 5% bonus

- Look up class schedule (https://pku-epic.github.io/Intro2CV_2024/schedule/) for release and due dates.

Assignments

- Late policy for assignments
  - If 1 day (0 - 24 hours) past the deadline, 15% off
  - If 2 day (24 - 48 hours) past the deadline, 30% off
  - Zero credit if more than 2 days.

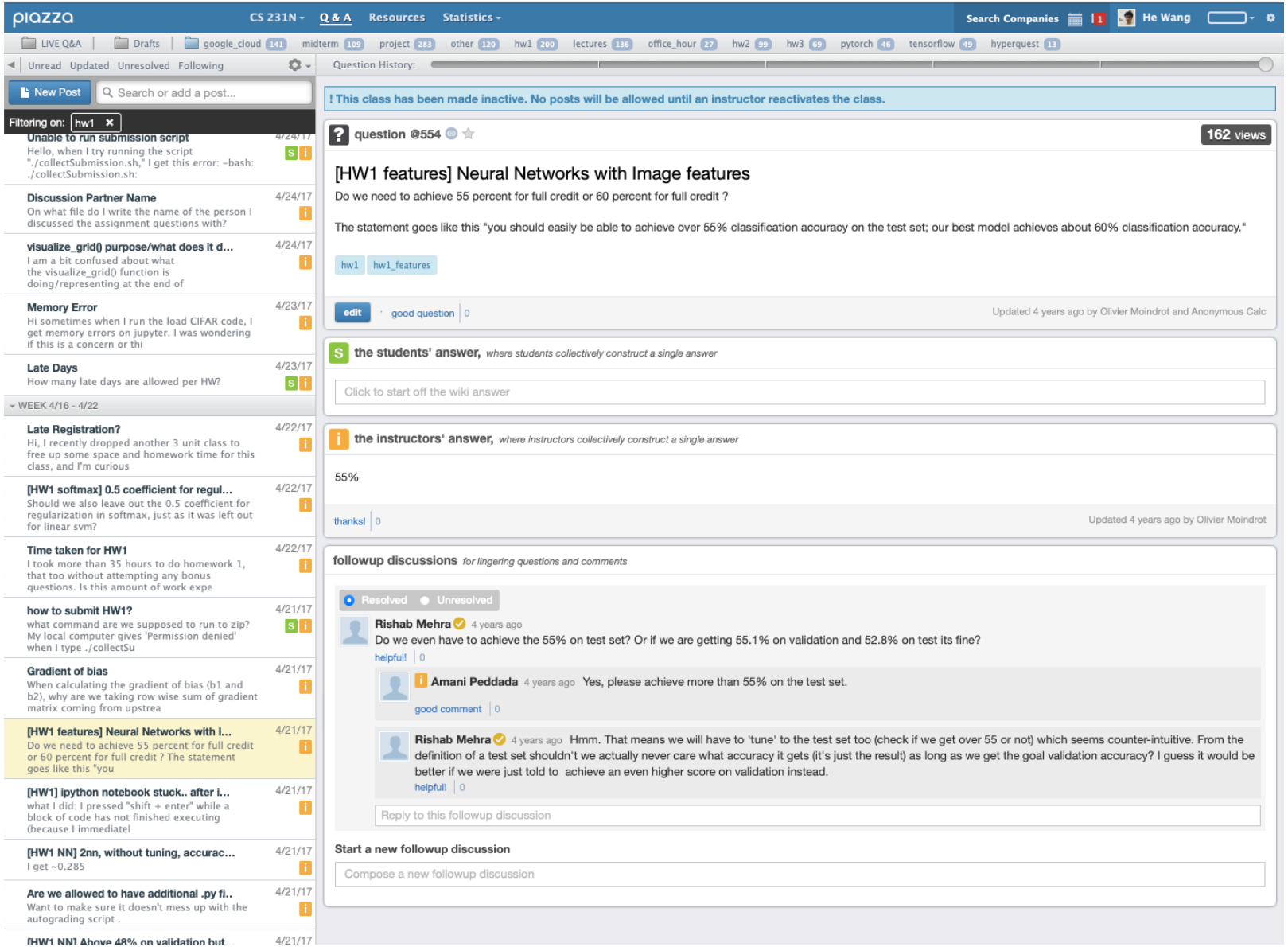
Midterm and Final Exams

- Midterm exam will be held in class.
- Final exam will be held in the afternoon of June 21.
- 1-page cheat sheet is allowed for both.

Course Website

- Website accessible to everyone: https://pku-epic.github.io/Intro2CV_2024/schedule/
- Internal course website in https://course.pku.edu.cn/
  - Slides/videos download
  - Assignment submission
  - Grades
  - Discussion board

My Experience back to Stanford

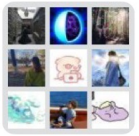


Discussion Board

- Course discussion board: share your questions!
- Will have a constant forum for discussing course material and three assignment forums each for one assignment.
- Your active participation in classes, discussions and my office hours will all count towards bonus.

WeChat Group

- WeChat group:
  - Use for notifications and announcements
  - Not an ideal place for tracking each individual questions.
  - Not recommended to WeChat me or TAs in person to ask questions that may also be interesting the other students. Use discussion board!



群聊: 2024春计算机视觉导论

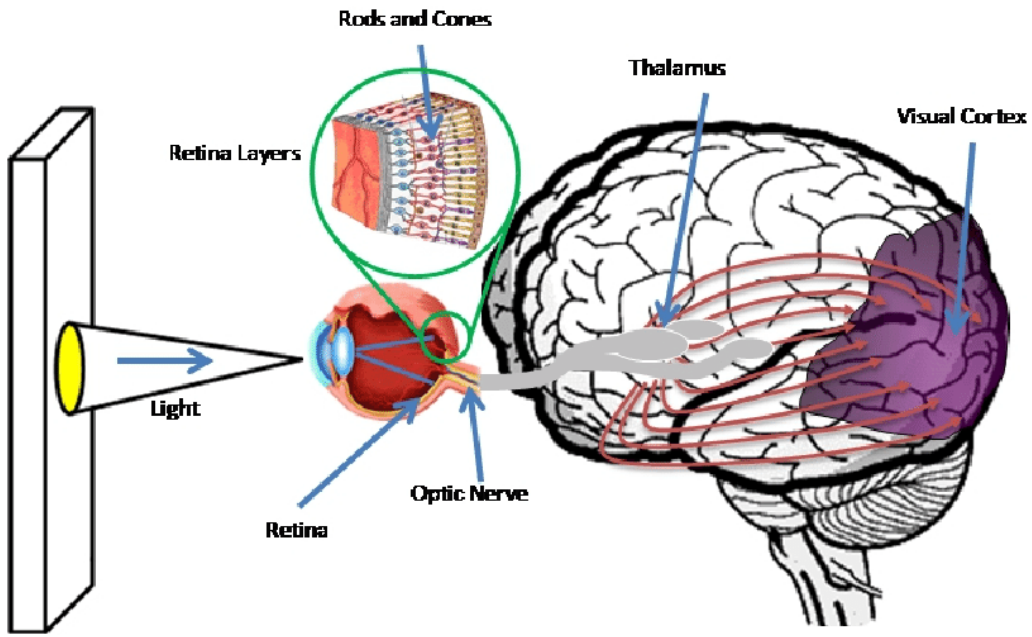
该二维码7天内(2月28日前)有效，重新进入将更新

What is Vision?

What is Vision?

Human Visual System

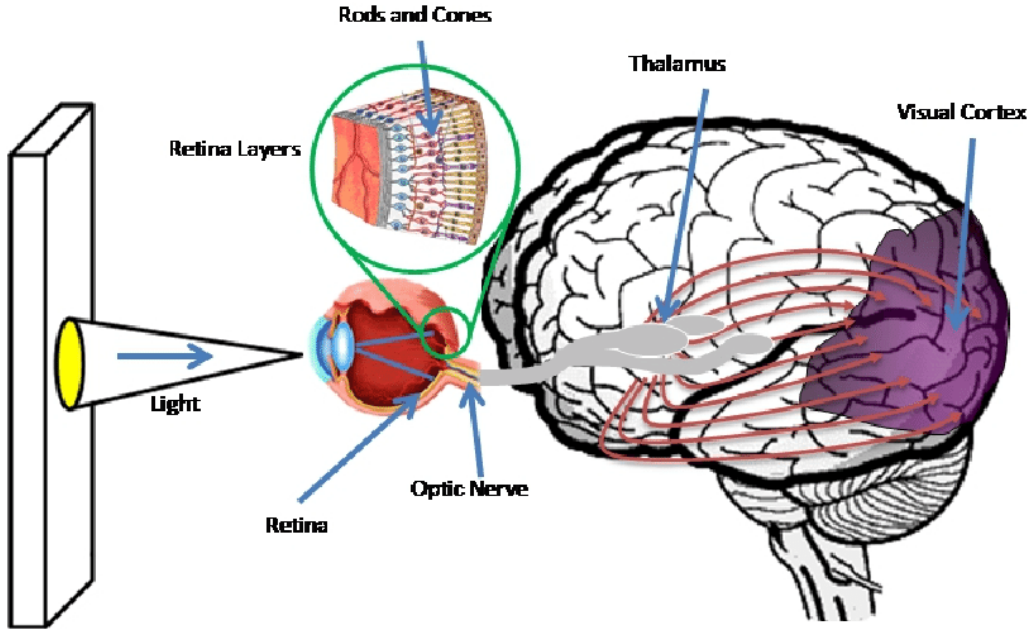
- The visual system comprises
  - Eyes (sensory organ)
  - Parts of the central nervous system
    - Retina layers
    - Optic nerve
    - Optic tract
    - Visual cortex



Maghami, Mohammad Hossein & Sodagar, Amir & Zabihian, Alireza & Asgarian, Farzad. (2012). Implantable Biomedical Devices. 10.5772/50336.

Human Visual System

- The visual system comprises
  - Eyes (sensory organ)
  - Parts of the central nervous system
    - Retina layers
    - Optic nerve
    - Optic tract
    - Visual cortex



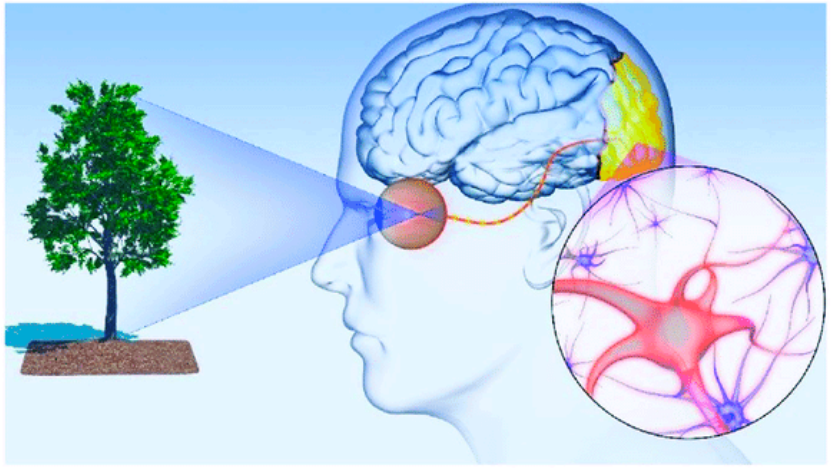
- Visual pathway: visual field→retina→optic nerve →…→optic tract→… →visual cortex

Maghami, Mohammad Hossein & Sodagar, Amir & Zabihian, Alireza & Asgarian, Farzad. (2012). Implantable Biomedical Devices. 10.5772/50336.

Human Visual System

- 83% information comes from vision (11% from audio, the others from smell, touch and taste).

- Carry outs a number of complex tasks:
  - visual sensation
  - visual perception
  - and visual motor coordination.



Wikipedia contributors. "Visual system." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 18 Jan. 2022. Web. 22 Feb. 2022.
Deng, Wei & Zhang, Xiujuan & Jia, Ruofei & Huang, Liming & Zhang, Xiaohong & Jie, Jiansheng. (2019). Organic molecular crystal-based photosynaptic devices for an artificial visual-perception system. NPG Asia Materials. 11. 77. 10.1038/s41427-019-0182-2.

Our Visual System Needs to Do

- the reception of light
- the formation of monocular neural representations
- color vision
- stereopsis and assessment of distances to and between objects

} Visual sensation

- pattern recognition
- the identification of particular object of interest
- motion perceptions
- the analysis and integration of visual information
- accurate motor coordination under visual guidance
- …

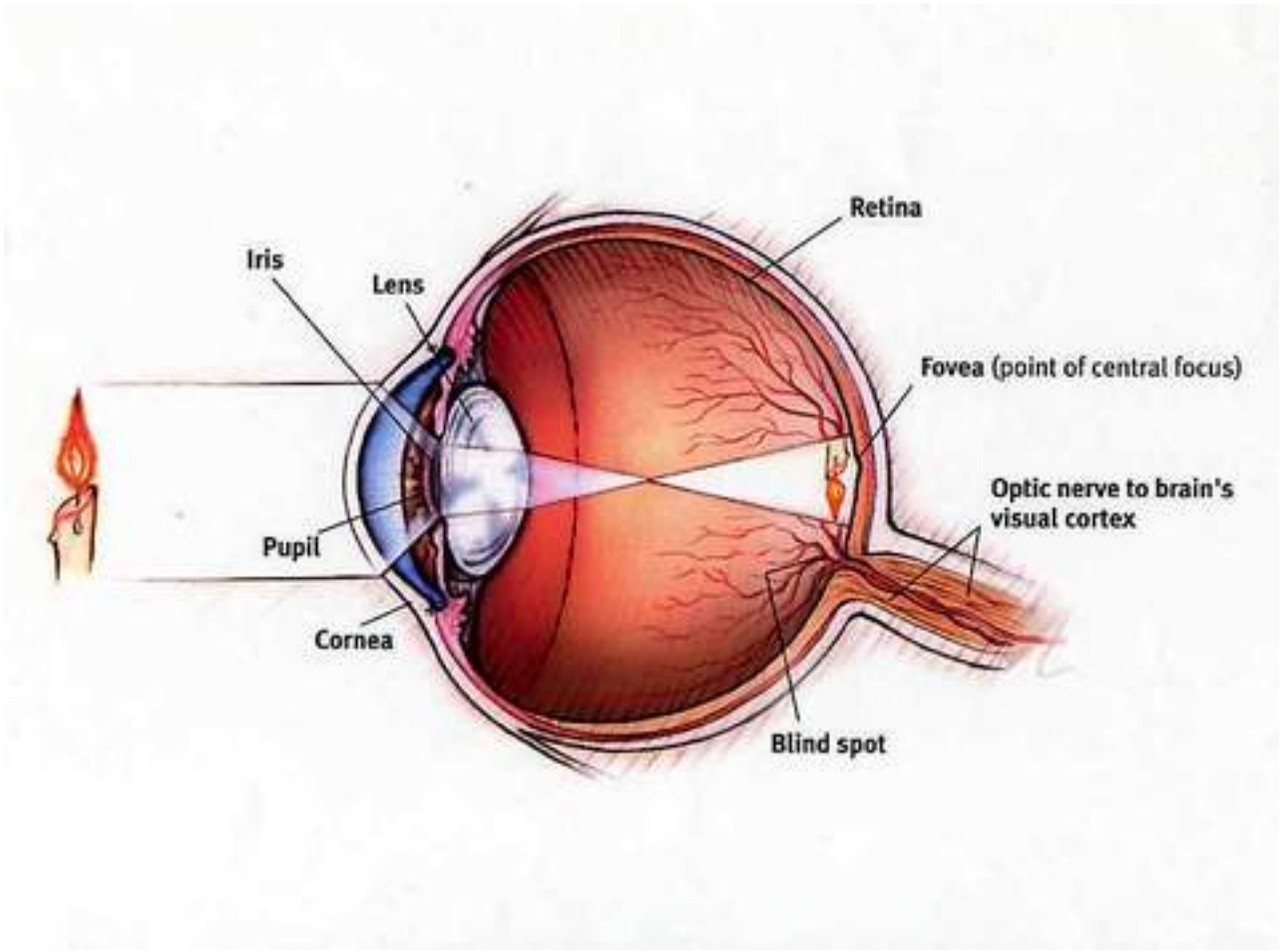
} Visual perception

- Integrating proprioception and vision signals (eye sensory feedbacks)
- Moving body parts as required to accomplish intended actions, *e.g.* eye-hand coordination in writing, eye-muscle coordination in sports
- …

} Visual motor coordination

Wikipedia contributors. "Visual system." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 18 Jan. 2022. Web. 22 Feb. 2022.

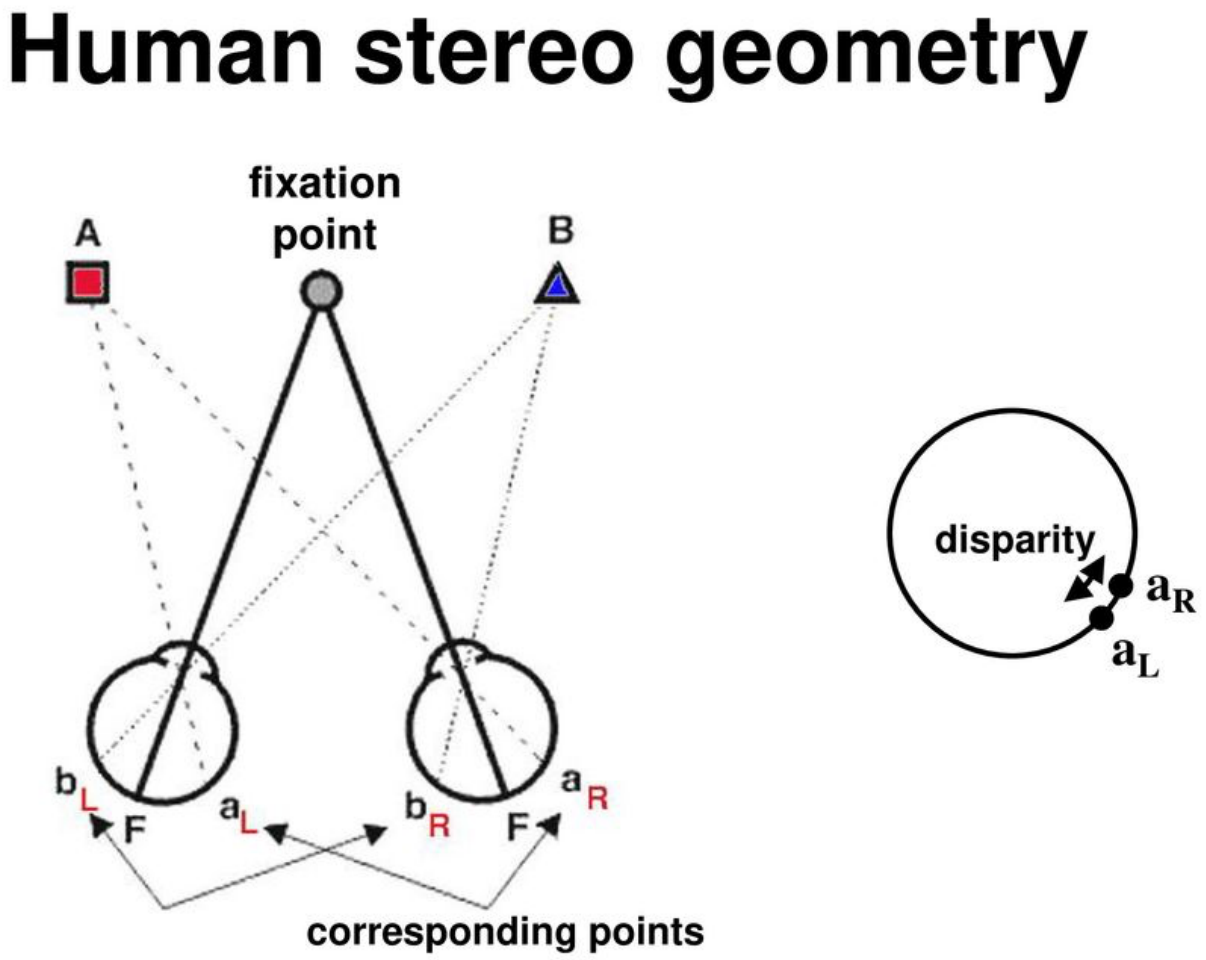
Visual Sensation: Monocular Vision



http://thebrucezone.weebly.com/3-image-formation.html

Visual Sensation: Binocular Vision and Stereopsis

- Human eyes are binocular.
- Senses distances through stereopsis



http://webvision.med.utah.edu/space_perception.html

S. Birchfield, Clemson Univ., ECE 847, http://www.ces.clemson.edu/~stb/ece847

Visual Perception

- Definition of visual perception in *Vision Science*:
  - the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect.

Palmer, Stephen E. *Vision science: Photons to phenomenology*. MIT press, 1999.

Visual Perception

- Concerns the acquisition of knowledge.
  - Fundamentally a cognitive activity.
  - Distinct from purely optical processes such as photographic ones.
  - Vision = eyes = camera? No, cameras have no perceptual capabilities at all.

Palmer, Stephen E. *Vision science: Photons to phenomenology*. MIT press, 1999.

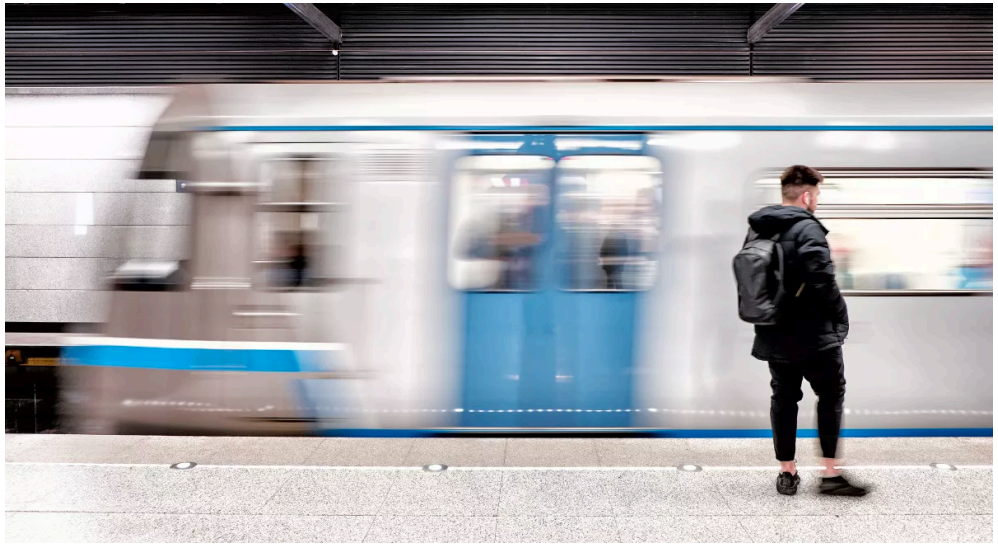
Visual Perception

- Concerns the acquisition of knowledge.
- The knowledge achieved by visual perception concerns objects and events in the environment.

Palmer, Stephen E. *Vision science: Photons to phenomenology*. MIT press, 1999.

Examples of Perception: Motion Perception

- Motion perception:the ability of the nervous system to discern the distance and speed of a moving object in relation to the **eye** that is seeing the object.

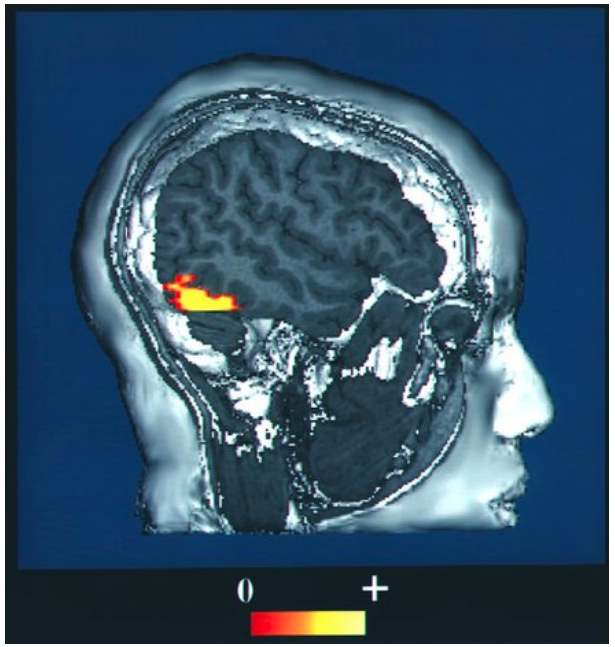


Wikipedia contributors. "Visual system." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 18 Jan. 2022. Web. 22 Feb. 2022.
Hülsdünker, Thorben, Martin Ostermann, and Andreas Mierau. "The speed of neural visual motion perception and processing determines the visuomotor reaction time of young elite table tennis athletes." *Frontiers in behavioral neuroscience* 13 (2019): 165.

Examples of Perception: Pattern Recognition

- One great example of pattern recognition is facial recognition.

Facial recognition: detect faces, perceive emotion, distinguish similar faces
(Dimitri Otis | Getty Images)



Corresponding area in the brain

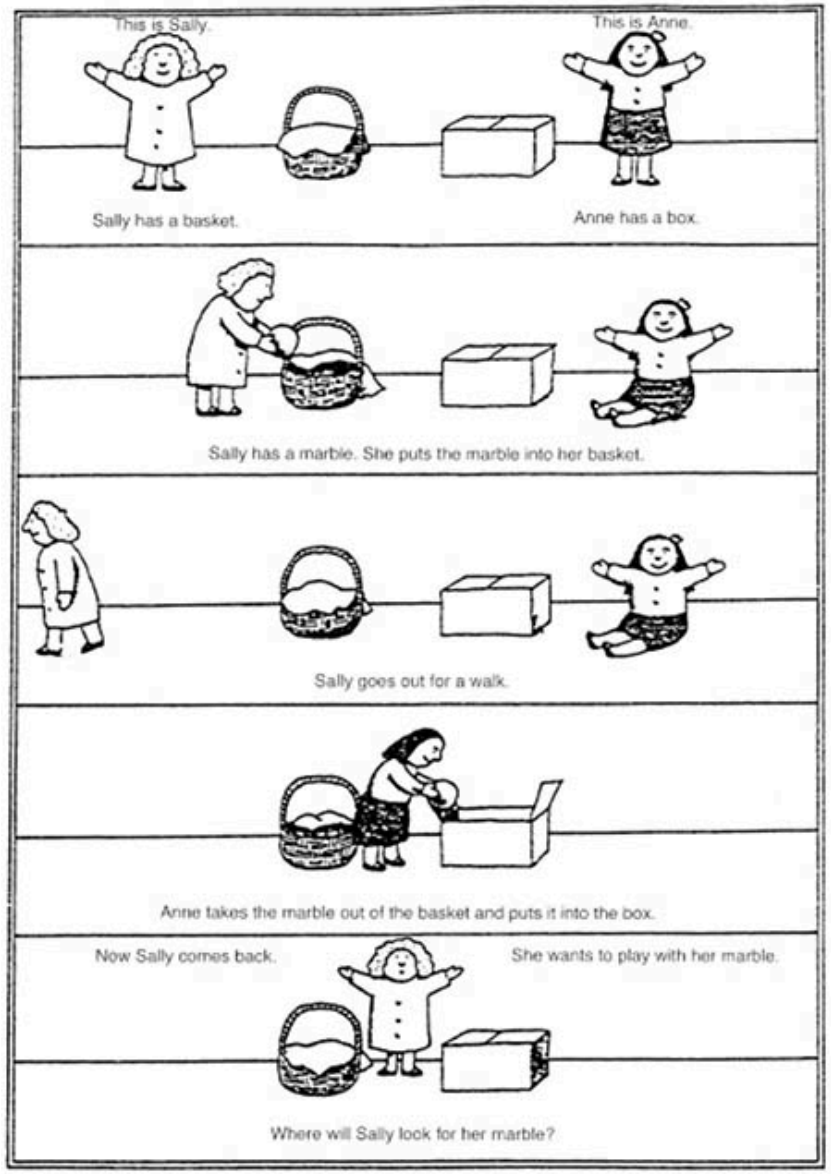
Wikipedia contributors. "Face perception." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 28 Jan. 2022. Web. 22 Feb. 2022.

Examples of Perception: Visual Cognition

- One example of visual cognition is false belief.

- Sally-Anne Test

  Sally takes a marble and hides it in her basket. She then "leaves" the room and goes for a walk. While she is away, Anne takes the marble out of Sally's basket and puts it in her own box. Sally is then reintroduced and the child is asked the key question, the *Belief Question*: "Where will Sally look for her marble?"



Wikipedia contributors. "Sally–Anne test." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 9 Feb. 2022. Web. 22 Feb. 2022.

Kosinski, Michal. "Theory of Mind May Have Spontaneously Emerged in Large Language Models." *arXiv preprint arXiv:2302.02083* (2023).

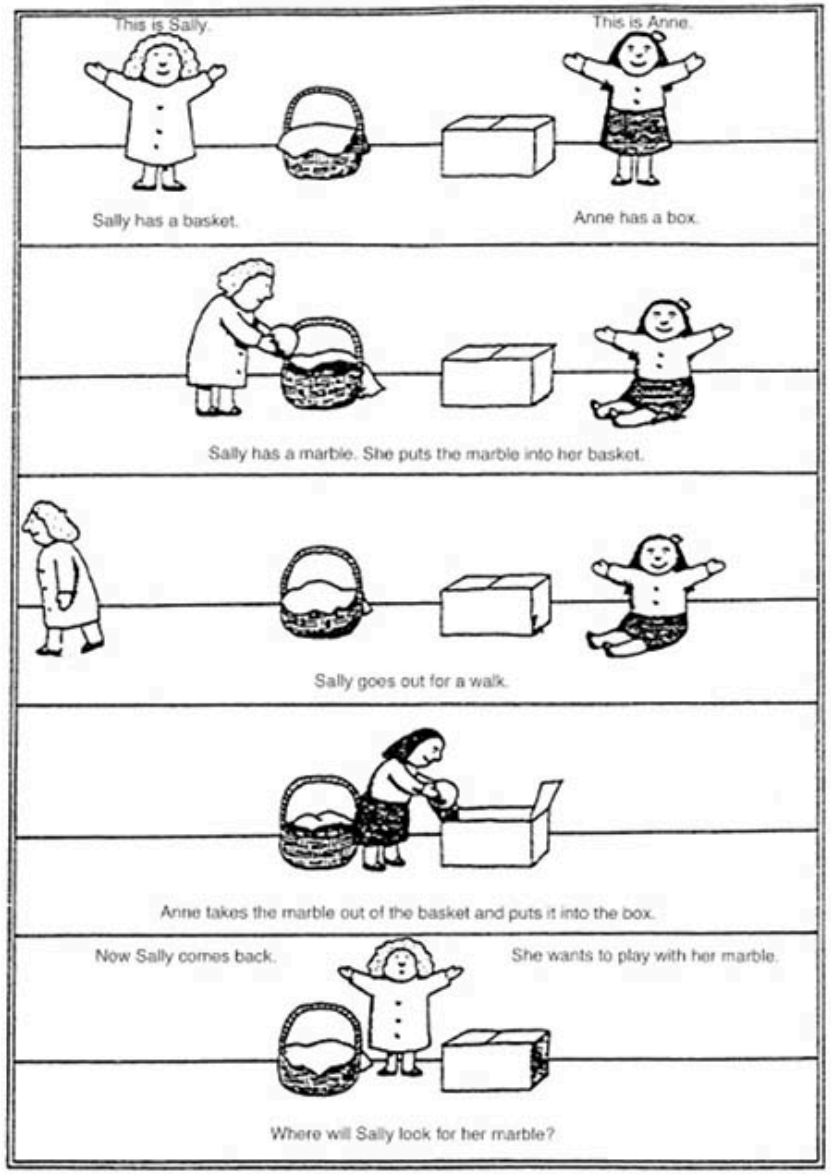
Examples of Perception: Visual Cognition

- One example of visual cognition is false belief.

- Sally-Anne Test

  Sally takes a marble and hides it in her basket. She then "leaves" the room and goes for a walk. While she is away, Anne takes the marble out of Sally's basket and puts it in her own box. Sally is then reintroduced and the child is asked the key question, the *Belief Question*: "Where will Sally look for her marble?"

- **Does ChatGPT have theory of mind?**



Wikipedia contributors. "Sally–Anne test." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 9 Feb. 2022. Web. 22 Feb. 2022.

Kosinski, Michal. "Theory of Mind May Have Spontaneously Emerged in Large Language Models." *arXiv preprint arXiv:2302.02083* (2023).

Visual Motor Coordination/Integration

- Visual motor control is the ability to coordinate visual information with motor output, where the eyes provide sensory feedback to adjust body motion.

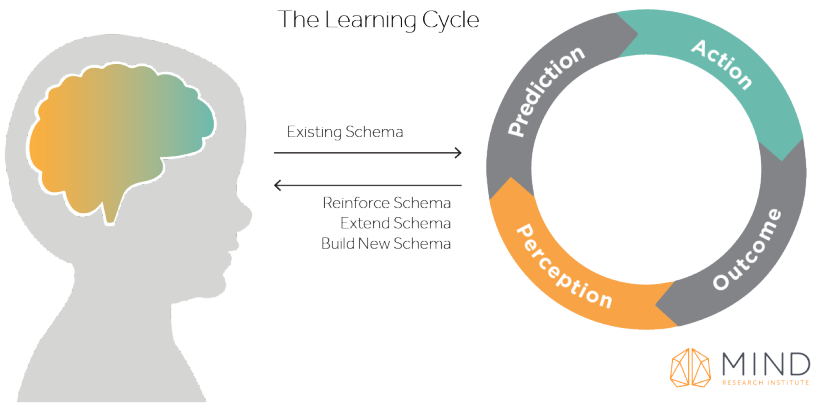
- It is crucial for coordinating the hands, legs, and the rest of the body's movements with what the eyes perceive.

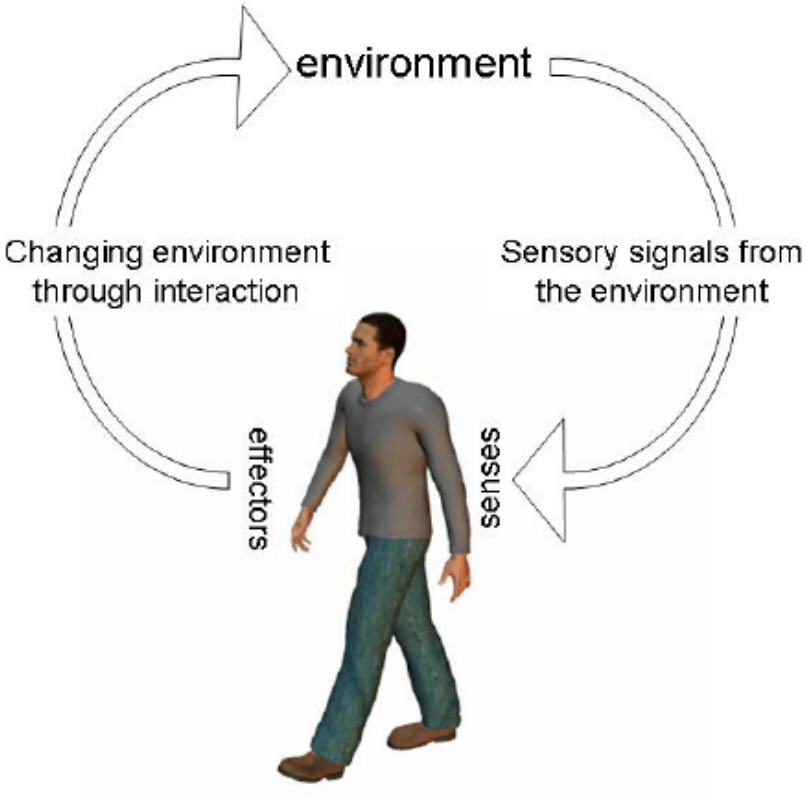
Examples of Visuomotor coordination: Eye-Hand Coordination

More Examples: Running, Balancing, etc.

How Humans Learn: Perception-Action Loop

- Perceive, forms hypotheses, and then take action to examine.
- Our brain makes sense of the world around us by creating and testing hypotheses about the way the world works.

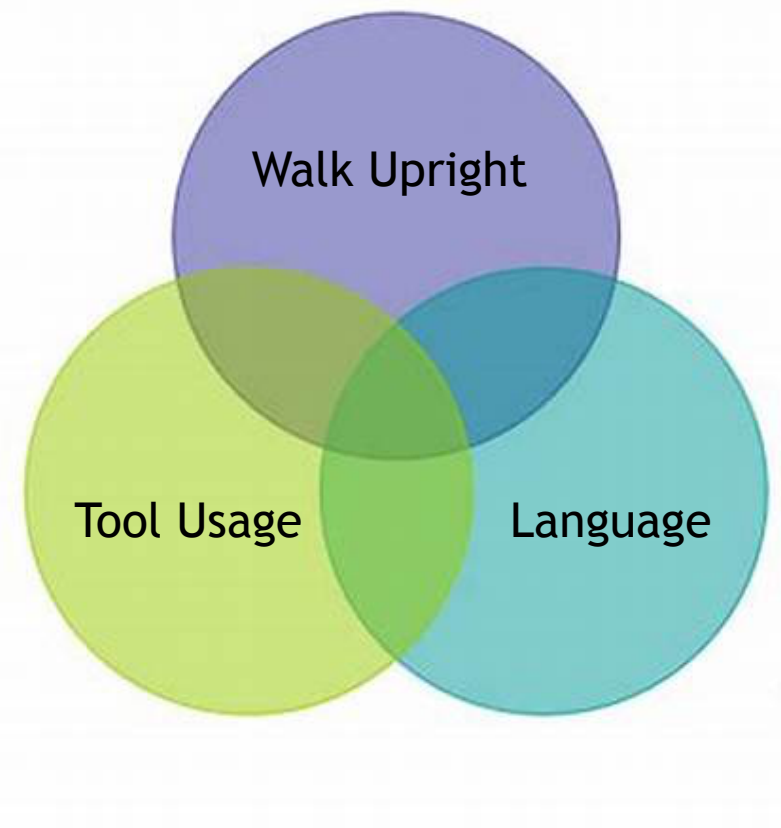




Perception-Action Loop

Vision and Langauge

The keys to evolution of human intelligence:



The interactions between vision and language:

Talk about what they see

Grounding according to language description

Summary of Human Visual System

- The visual system comprises eyes as sensor organ, which are connected to visual cortex in the brain.

- Visual tasks:
  - sensation
  - processing
  - perception
  - cognition
  - visuomotor coordination
  - language grounding
  - and more.

Computer Vision

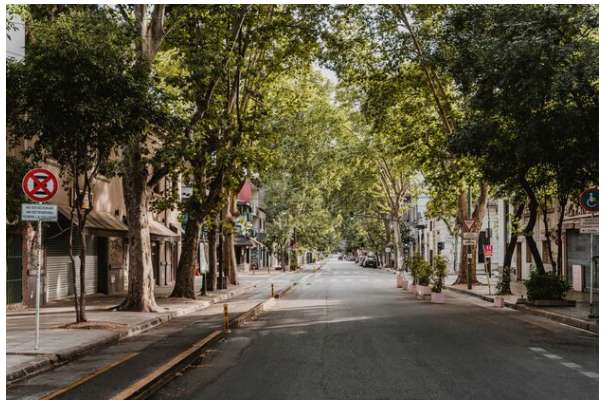
What is Computer Vision?

- Computer vision deals with
  - acquiring
  - processing and analyzing
  - understanding
  - generating or imagining

  visual data, …

Visual Data Acquisition

RGB camera



RGB image

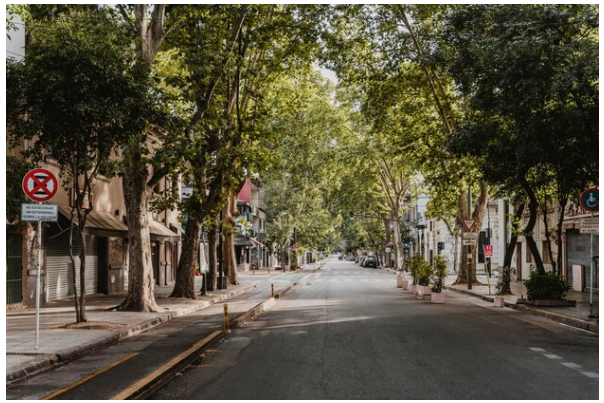
Visual Data Acquisition

- Different types of sensors and visual data

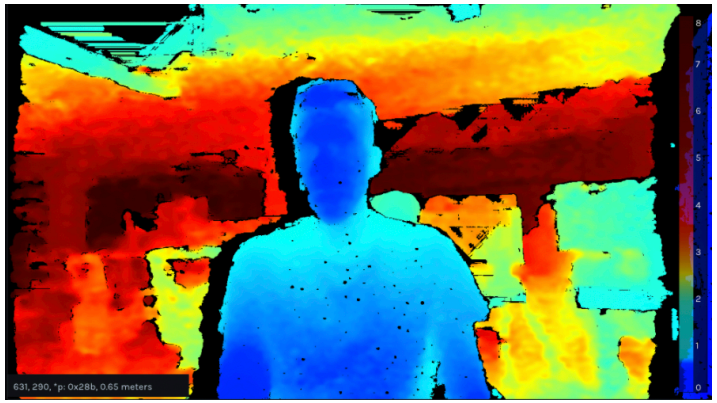
RGB camera

Depth camera

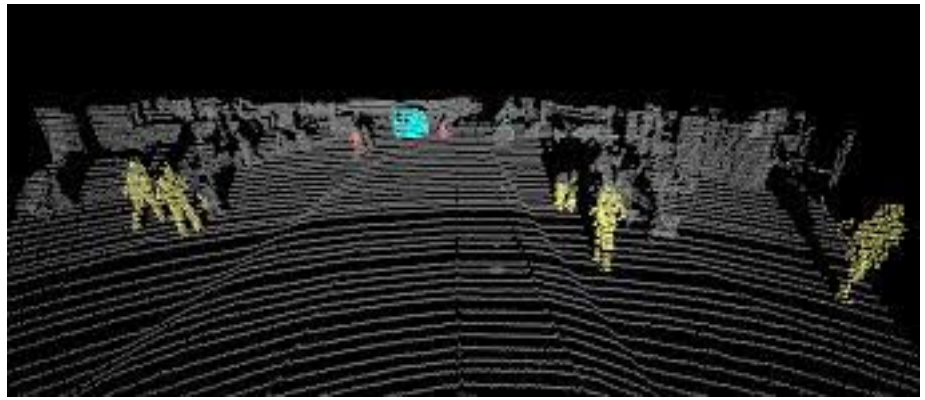
LiDAR



RGB image



Depth image



LiDAR point cloud

Beyond Single Frame and Single View

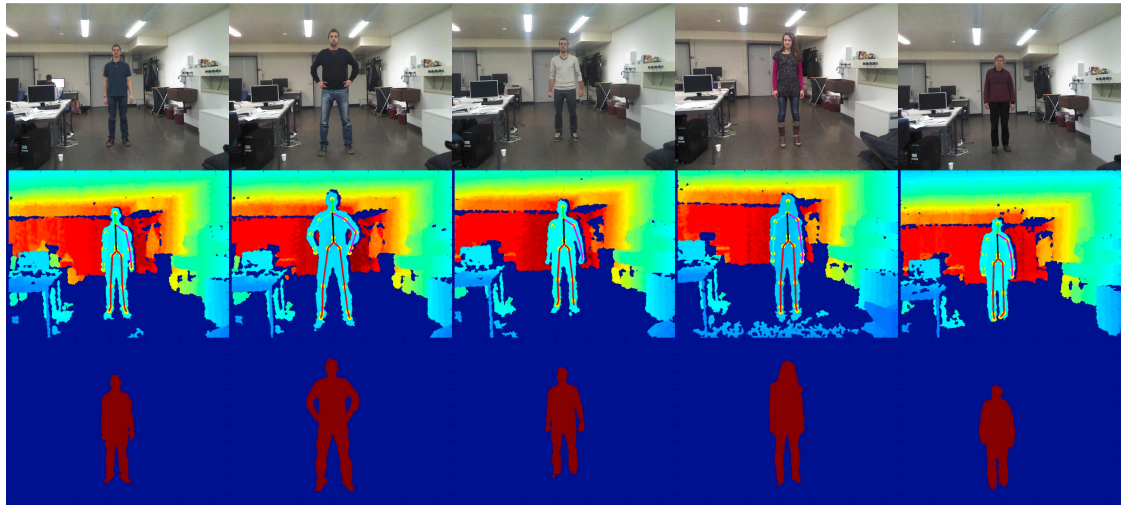
Stereo images

Multiview images

Panoramic images

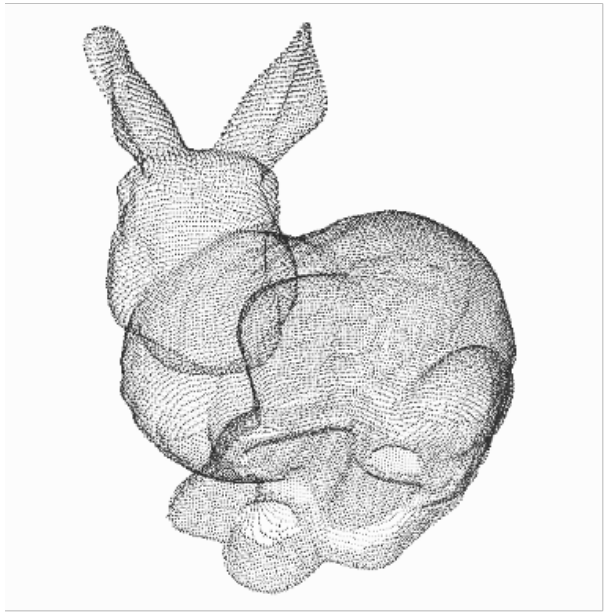
Beyond Single Frame and Single View

- RGB video

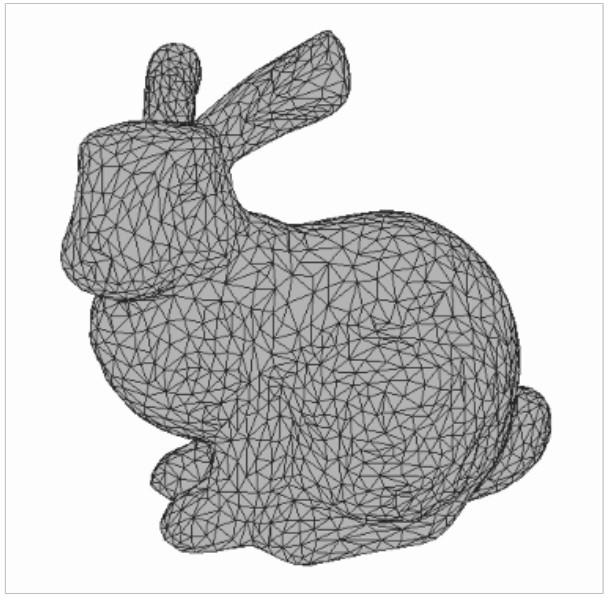


- RGBD video

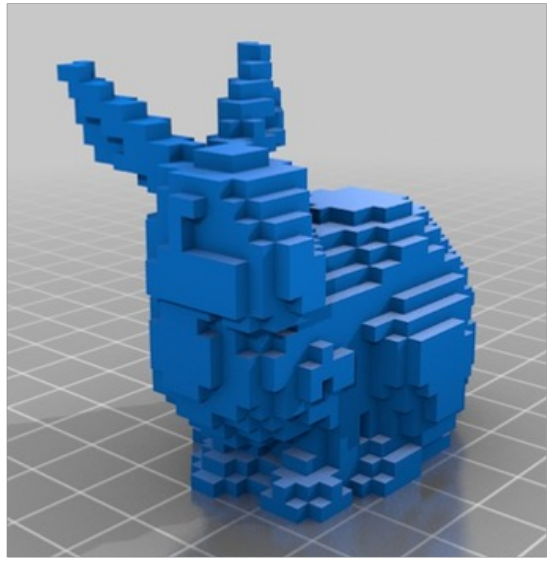
True 3D Visual Data



Point Cloud



Surface Mesh

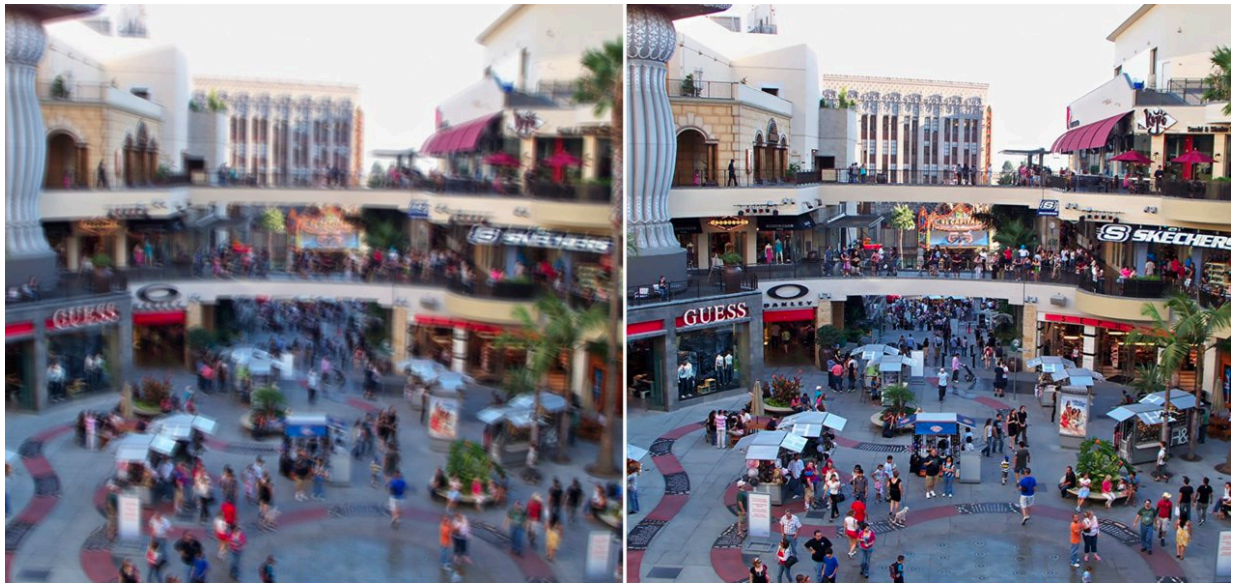


Volumetric

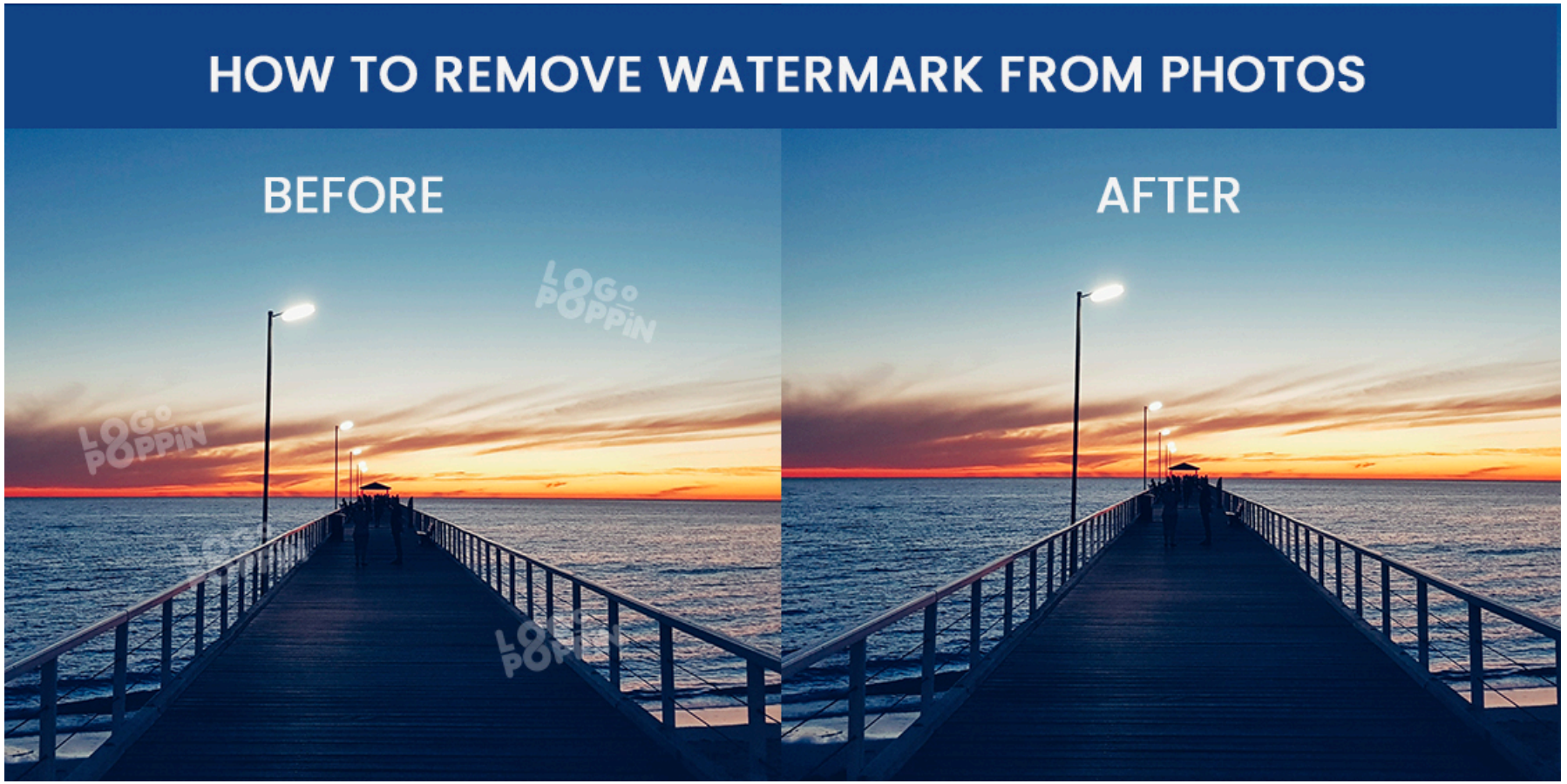
Low-Level Vision: Processing and Feature Extaction

- Low-level vision deals with
  - Image processing
    - image denoising/deblur
    - contrast enhancement
    - …
  - Feature extractions
    - edge/corner detection
    - optical flow/correspondence

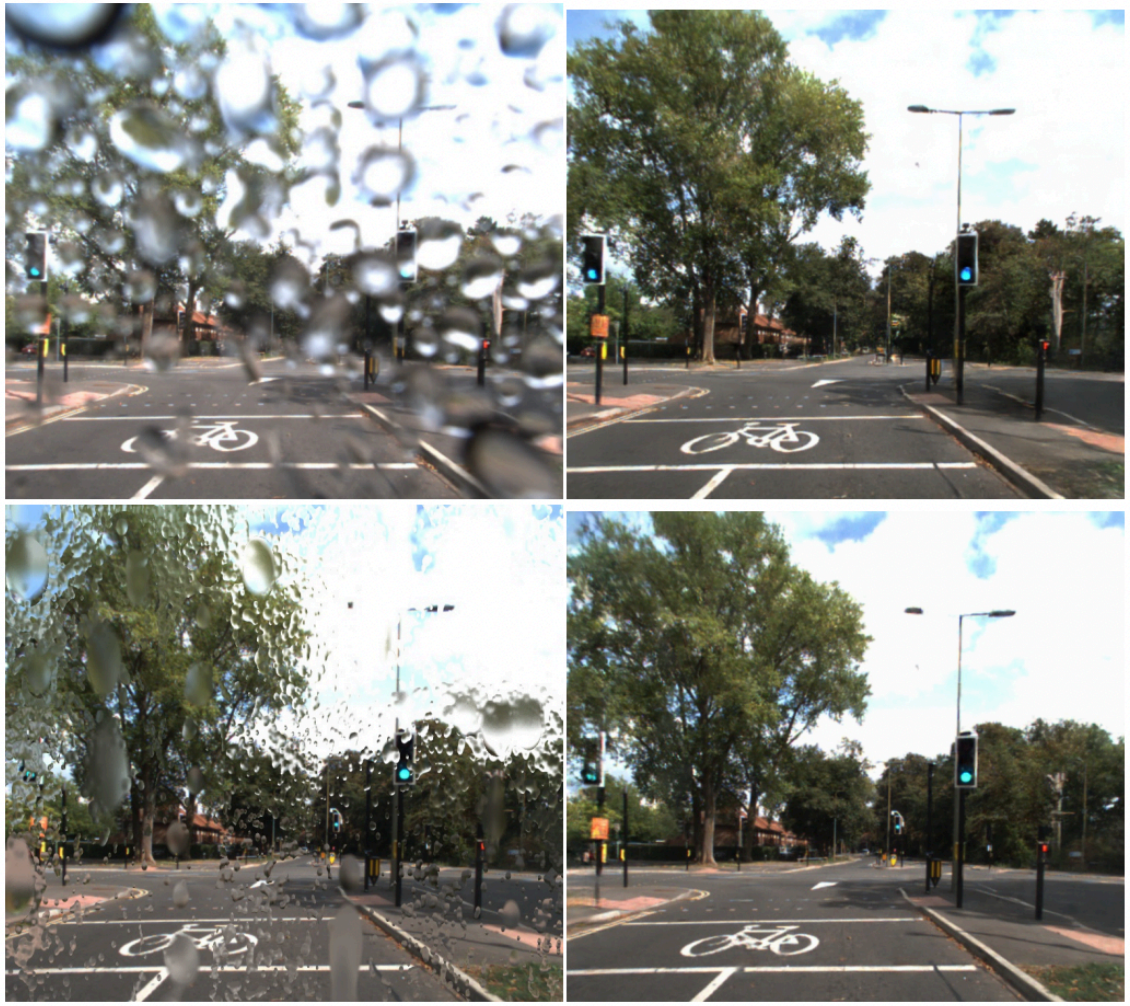
Applications: Motion Deblurring



Applications: Watermark Removal



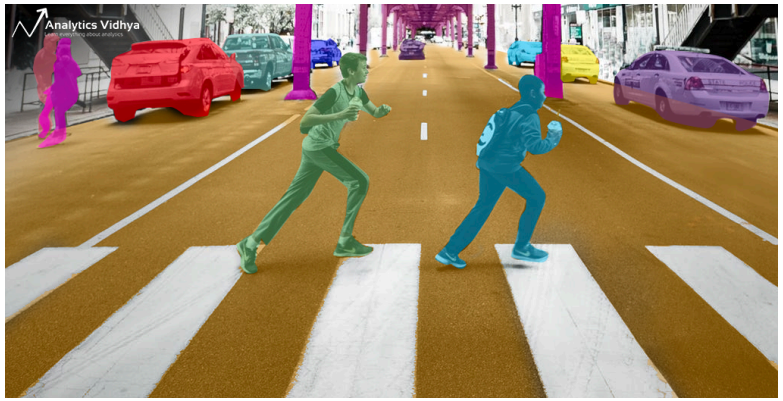
Applications: De-Raining

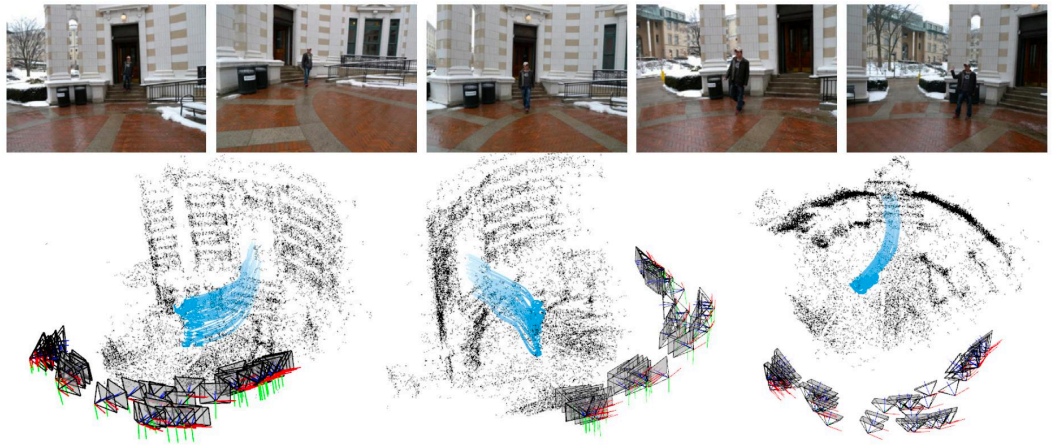


Porav, Horia, Tom Bruls, and Paul Newman. "I can see clearly now: Image restoration via de-raining." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

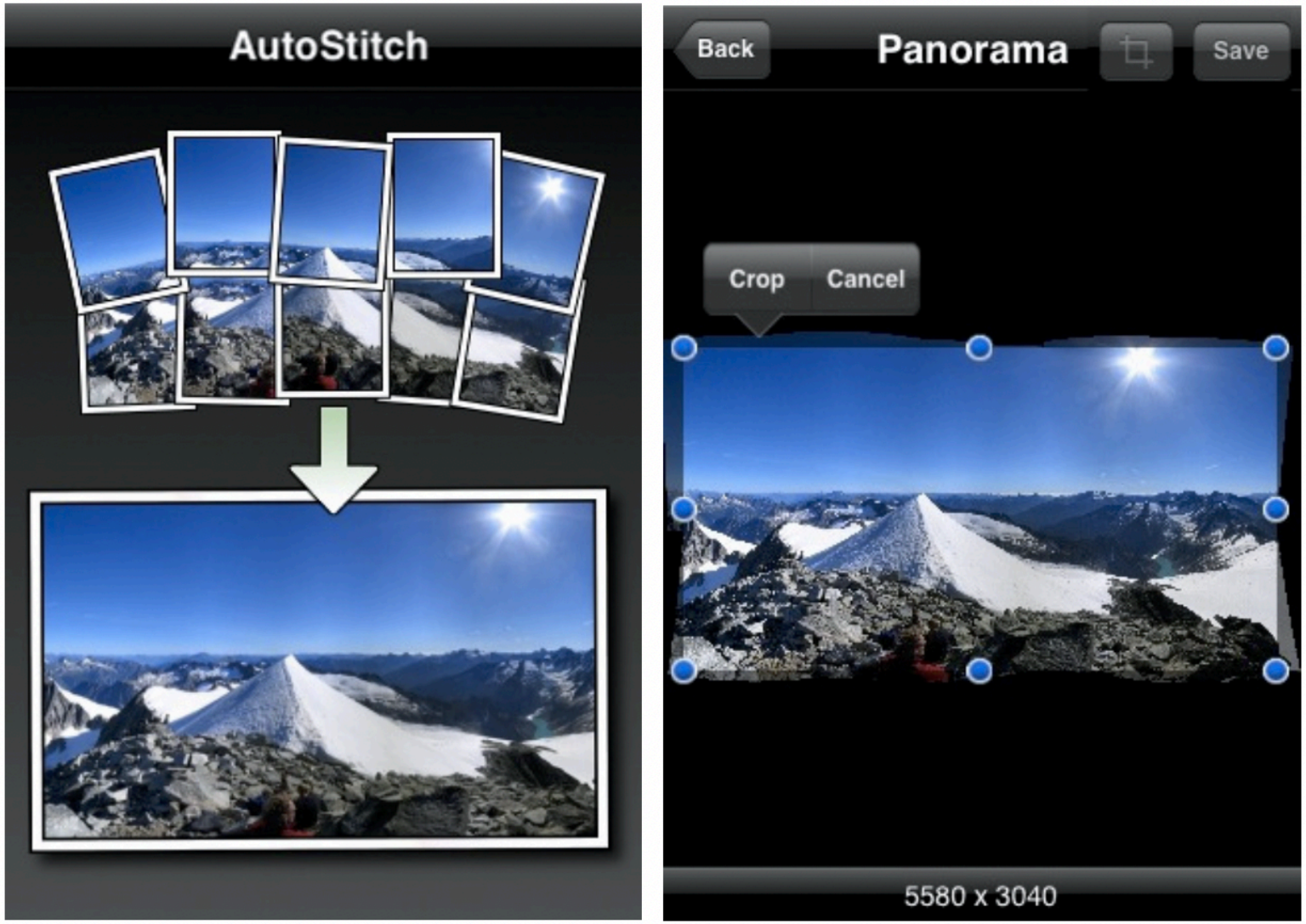
Mid-Level Vision

- Mid-level vision begins to make inferences about the world based on those measurements.
  - Analyzing local structures (grouping based segmentation, motion analysis, etc.)
  - 3D reconstruction using features obtained from the low-level vision.



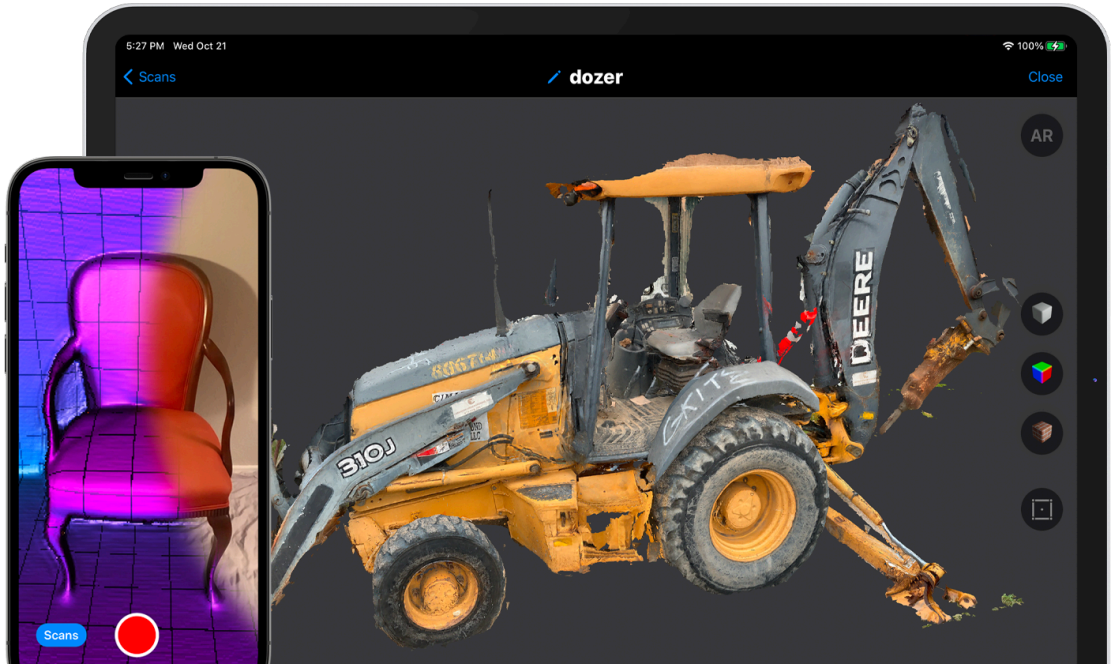


Applications: Panoramic Photography



Applications: 3D Object Scanner

Traditional line scanner



iPad Pro LiDar Scanner App

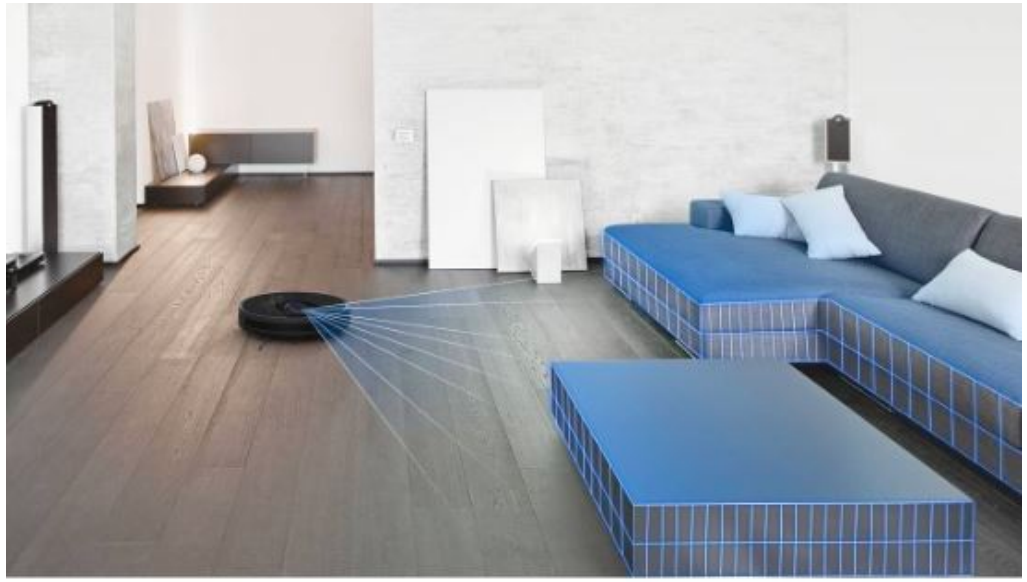
Applications: 3D Modeling of Landmarks

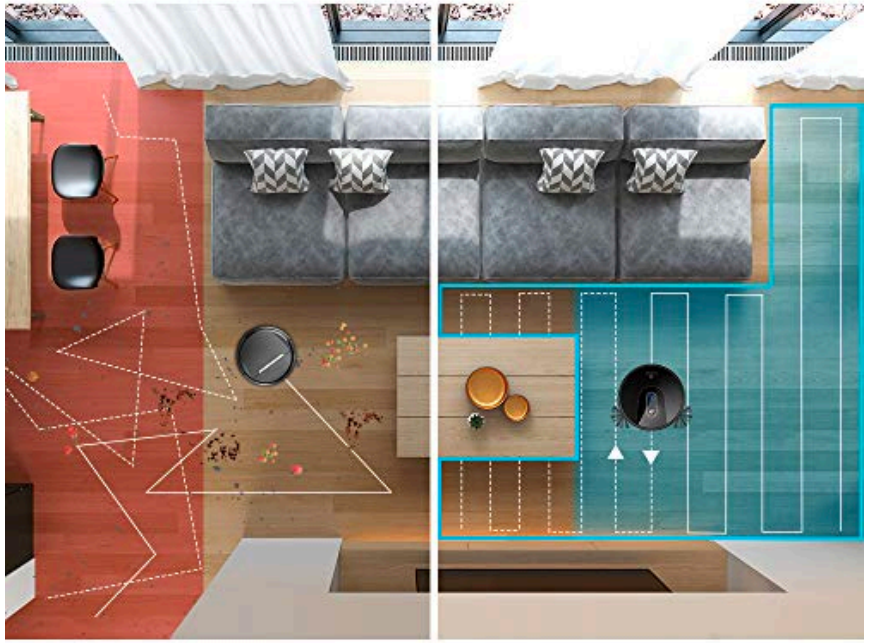
Building Rome in a day.

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., & Szeliski, R. (2011). Building rome in a day. *Communications of the ACM*, *54*(10), 105-112.

Applications: SLAM

- SLAM = Simultaneous localization and mapping

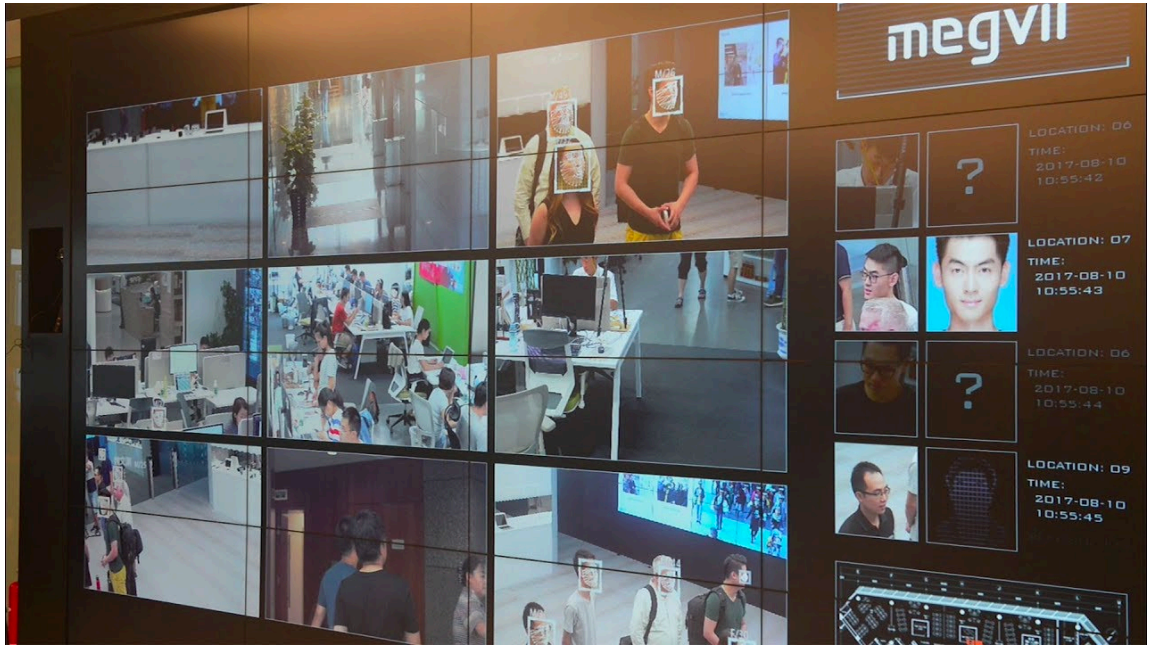


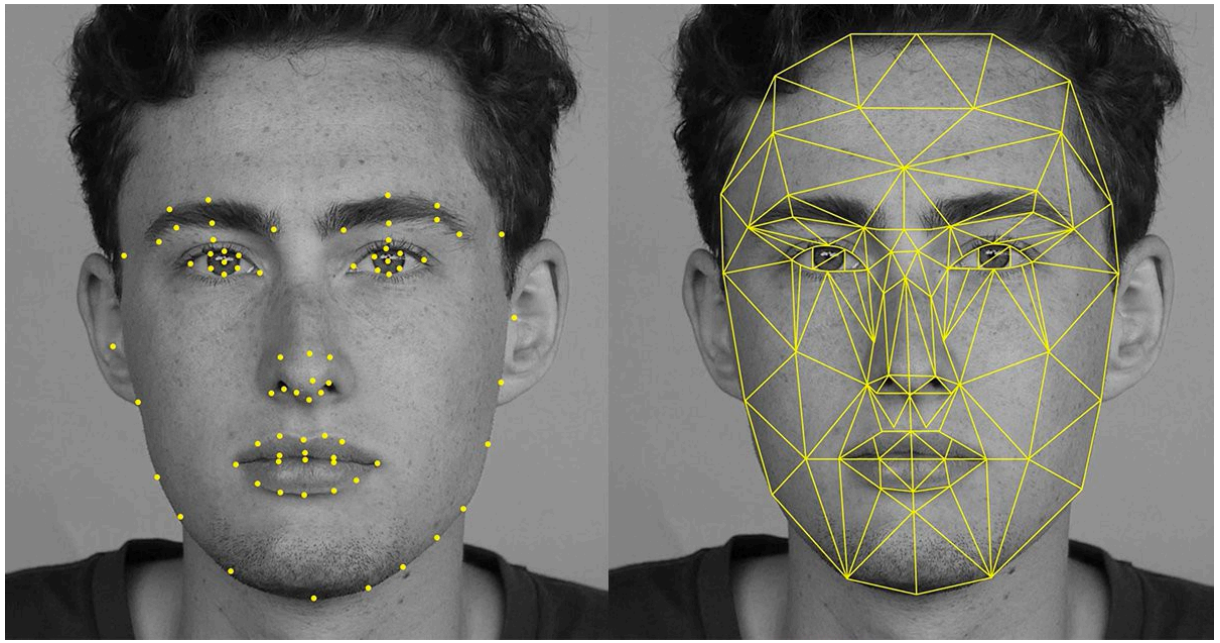


High-Level Vision: Understanding

- High-level vision analyzes the structure of the external world that produced those images and generates **semantic representation/interpretations**, including
  - object recognition and detection
  - scene understanding
  - activity understanding
  - etc.

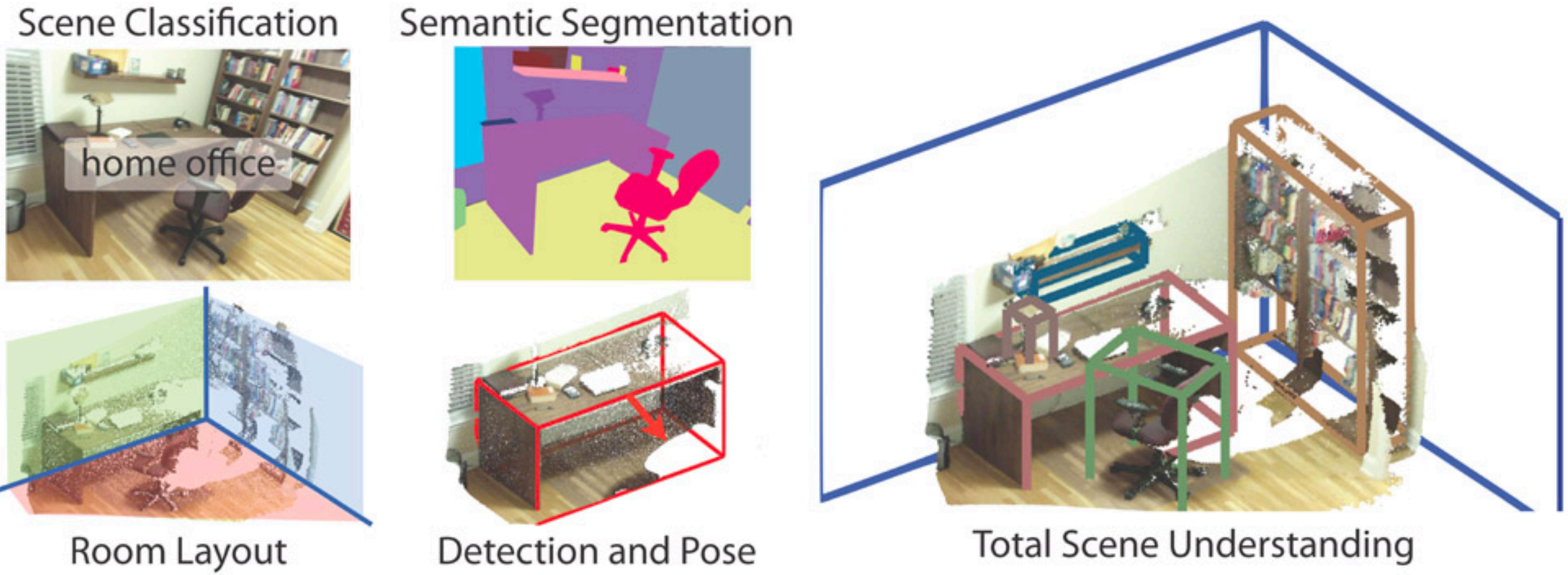
Applications: Facial Recognition





https://github.com/3ba2ii/Facial-Keypoints-Detector-CVND

Task: Scene Understanding



Song, Shuran, Samuel P. Lichtenberg, and Jianxiong Xiao. "Sun rgb-d: A rgb-d scene understanding benchmark suite." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Applications: Augmented Reality

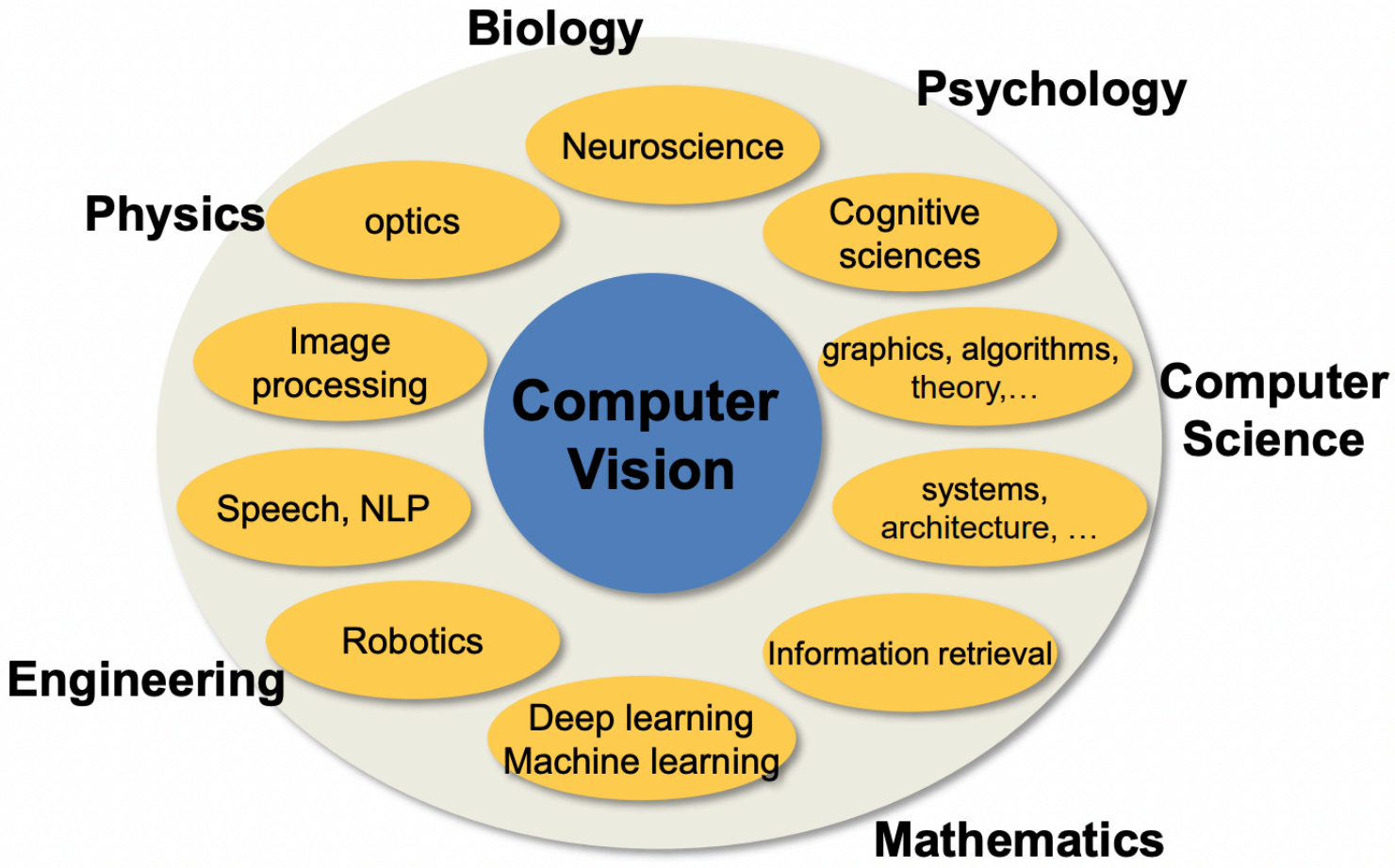
Assisting furniture layout

E-Learning

Applications: Amazon Go

- Cashier-free store. A very complicated vision system.

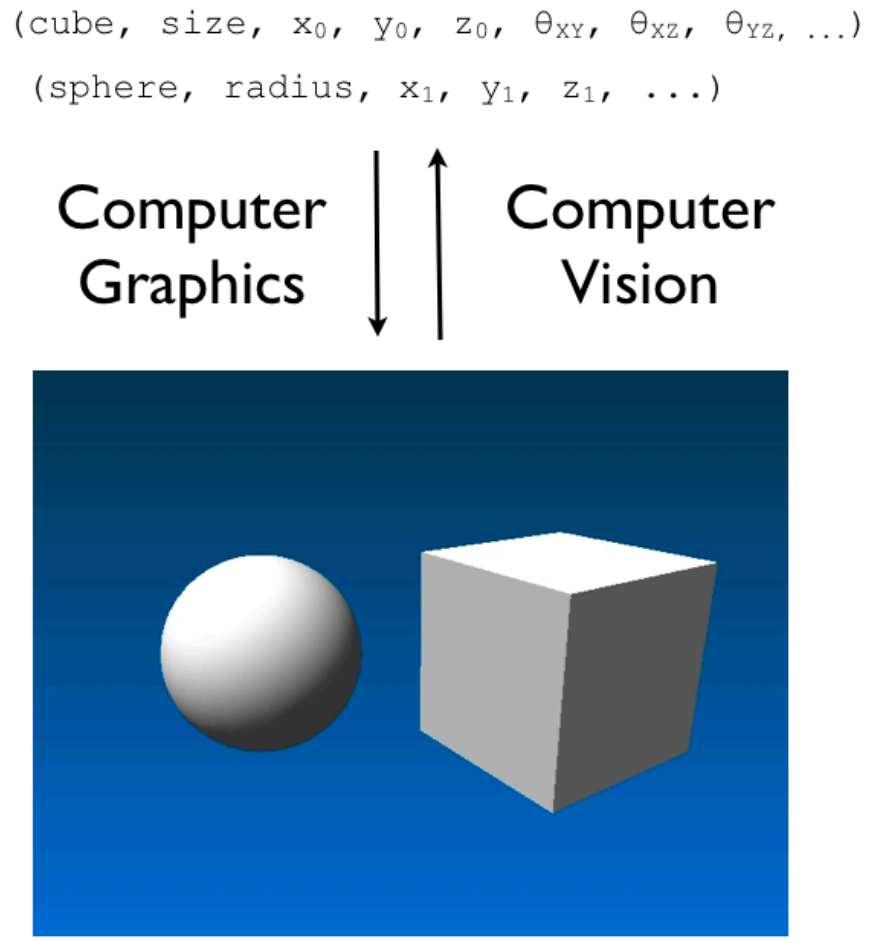
Interdisciplinary Nature of Computer Vision



Borrowed from Stanford CS231N Lecture 01.

Vision and Graphics

- Graphics:
  - go from the parameter space to the image space (rendering)
- Vision:
  - inverse graphics
  - more ill-posed
  - arguably harder



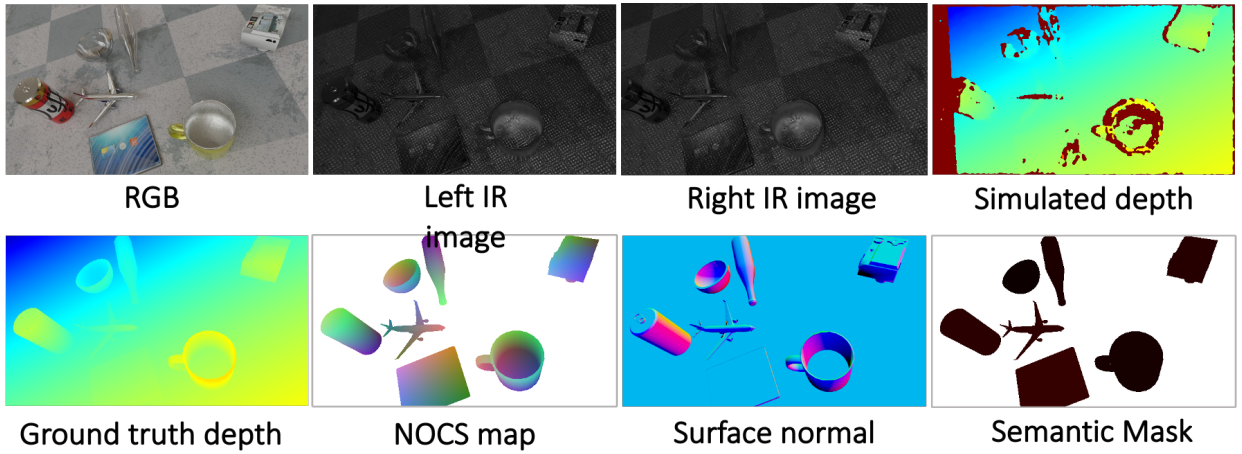
http://aboutintelligence.blogspot.com/2009/01/computer-vision-vs-computer-graphics.html

Leveraging Graphics for Vision

Synthetic data comes with free labels!

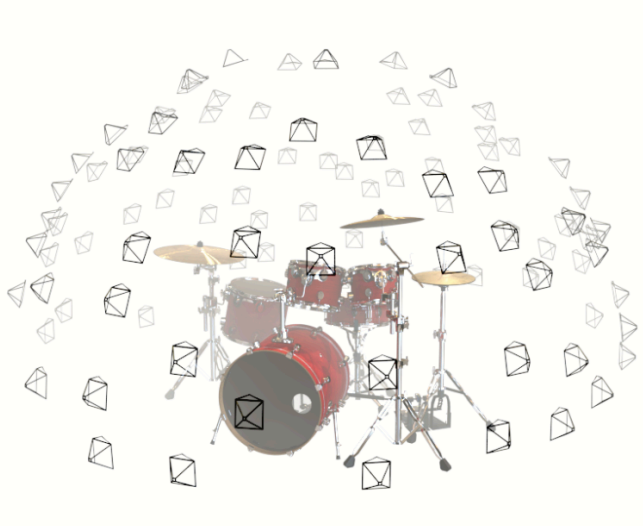
For outdoor semantic segmentation

For table-top depth/pose/surface normal prediction (ECCV 2022 from EPIC Lab)



Vision does Graphics Job

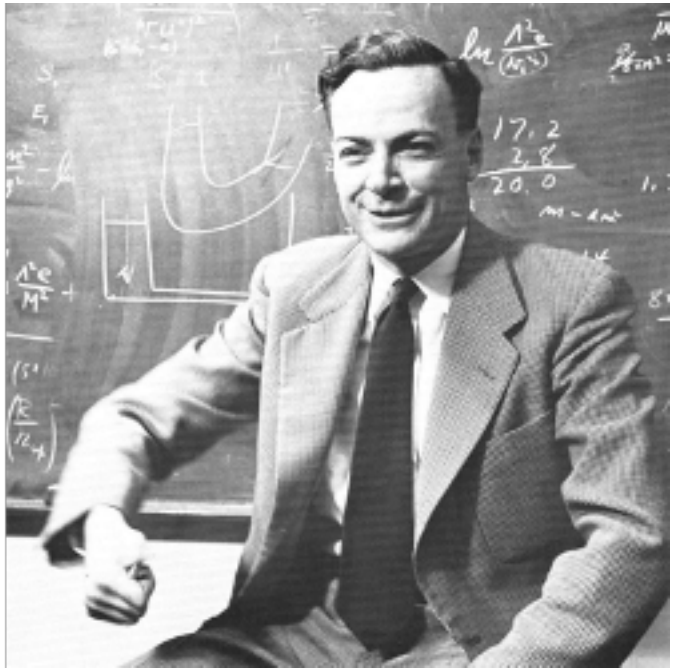
Training network to reproduce all input views of the scene

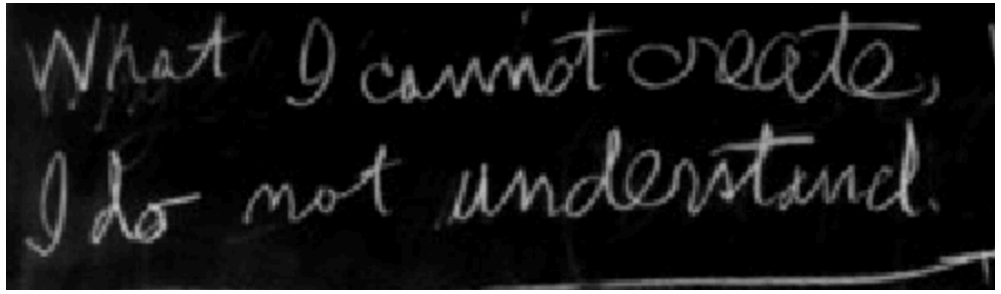


Neural Radiance Field (NeRF)

A brief intro to NeRF: https://www.youtube.com/watch?v=JuH79E8rdKc

Visual Content Generation





- Generation or imagination is not a core function of human visual system.
- Richard Feynman: “What I cannot create, I do not understand”
- Thus, computer vision also deals with generation.

Applications: Human Face Generation

StyleGAN for facial image generation

Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

Applications: Facial Reenactment

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. Advances in Neural Information Processing Systems, 32.

Applications: Style Transfer

Content target

+

Style reference

=

Combination image

Neural Style Transfer

https://gitee.com/happyjoejoe/deep-learning-with-python-notebooks/blob/master/8.3-neural-style-transfer.ipynb

Applications: Text-to-Image Diffusion Model

Imagen: https://imagen.research.google/

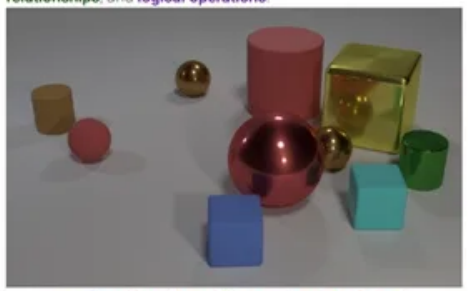
More Vision Language Tasks

- Referring Expression
  - ReferIt Game, *EMNLP 2014*
  - RefCOCO, *ECCV 2016*
  - GuessWhat?!, *CVPR 2017*
- Visual Dialog
  - VisDial, *CVPR 2017*
  - Image Grounded Conversation, *ACL 2017*
  - Dialog-based Image Retrieval, *NIPS 2018*
- Text 2 image/video

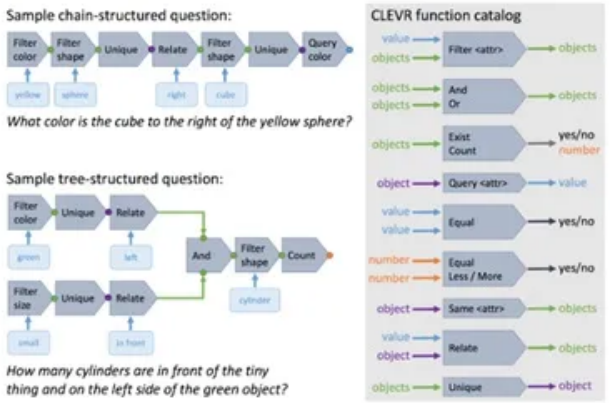
User1: My son is ahead and surprised!
User2: Did he end up winning the race?
User1: Yes he won, he can't believe it!

Intersection between computer vision and natural language processing (NLP)

Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or red things?



Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." In *CVPR*. 2017.

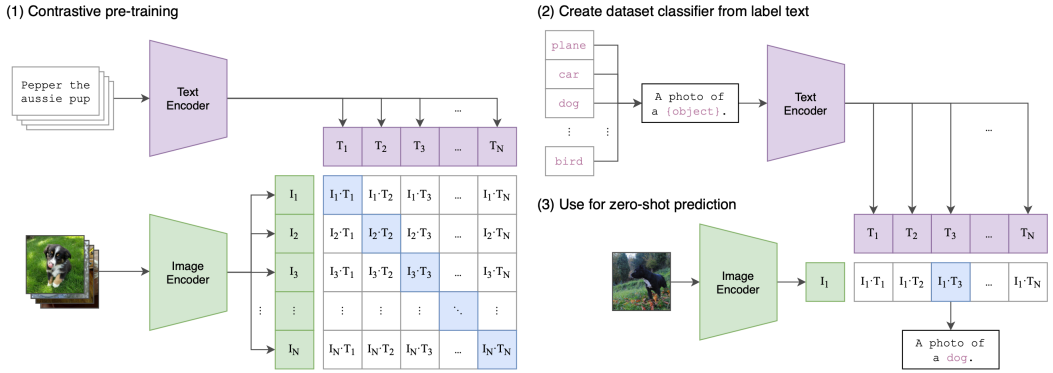


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

What Else in Computer Vision?

- Computer vision deals with
  - acquiring
  - processing and analyzing
  - understanding
  - generating or imagining

  visual data, and, what's more,
  - providing visual feedbacks for body motions
  - helping making decisions

  for embodied agents.

Embodied AI and Embodied Vision

👁️ **See**: perceive their environment through vision or other senses.

🎤 **Talk**: hold a natural language dialog grounded in their environment.

👂 **Listen**: understand and react to audio input anywhere in a scene.

🕹️ **Act**: navigate and interact with their environment to accomplish goals.

🤔 **Reason**: consider and plan for the long-term consequences of their actions.

Embodied AI is the field for solving AI problems for virtual robots that can move, see, speak, and interact in the virtual world and with other virtual robots — these simulated robot solutions are then transferred to real world robots.

--- Luis Bermudez, Overview of Embodied Artificial Intelligence

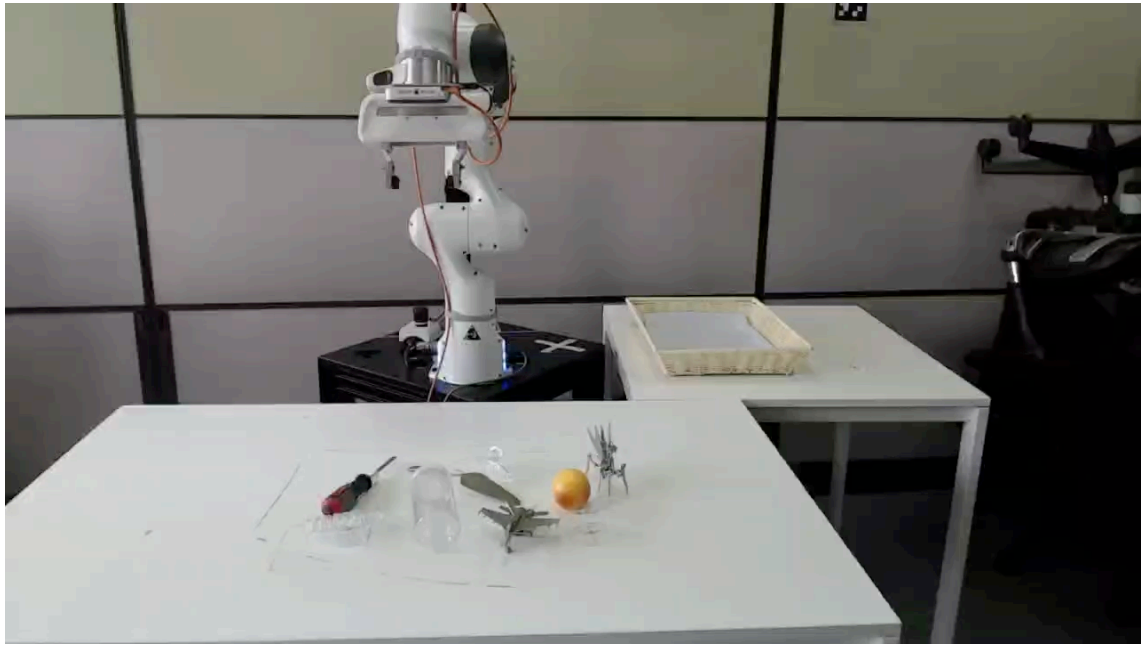
Home robots

Applications: Autonomous Driving

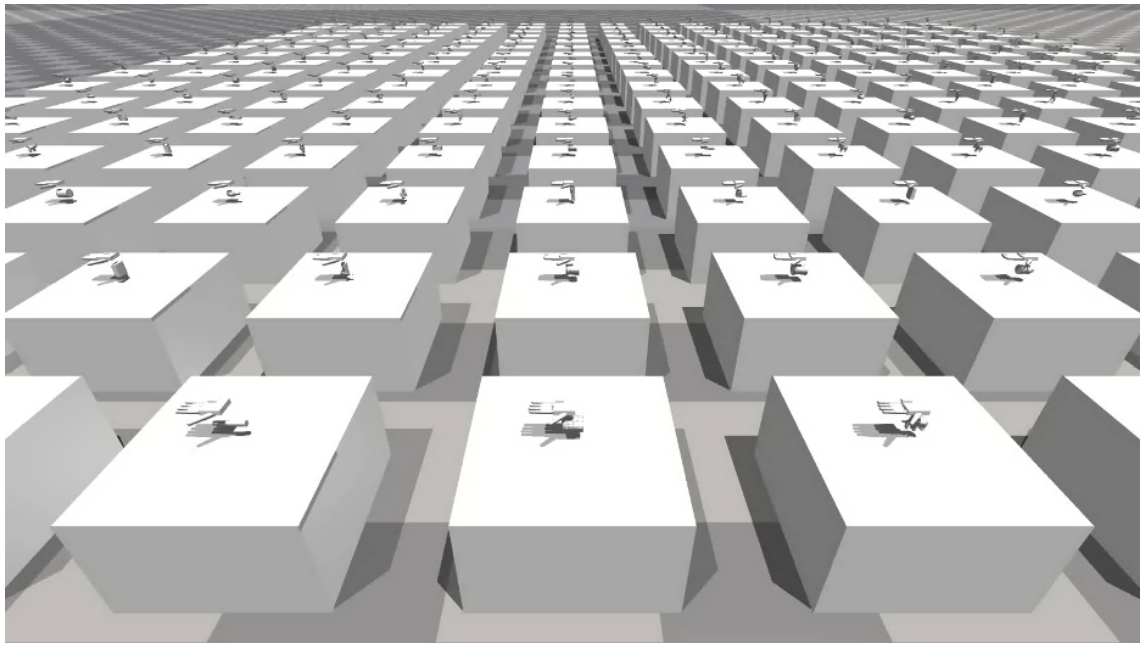
Tesla: pure vision solution
No LiDAR but multi-cameras

Waymo
LiDAR-based Solution

Applications: Embodied Vision

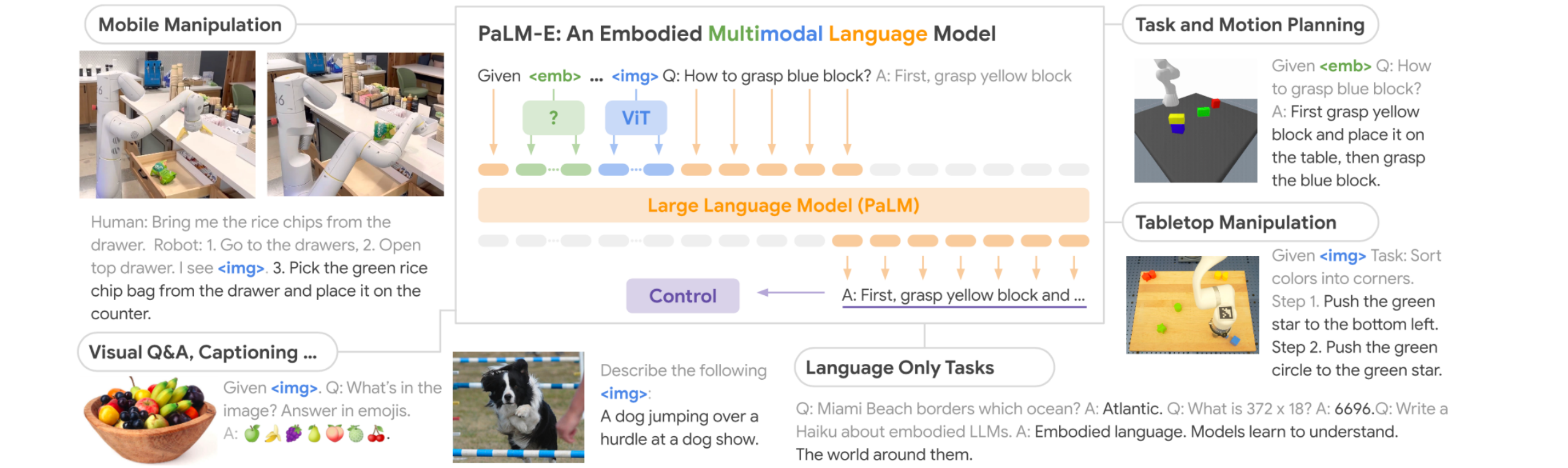


ASGrasp (ICRA 2024)

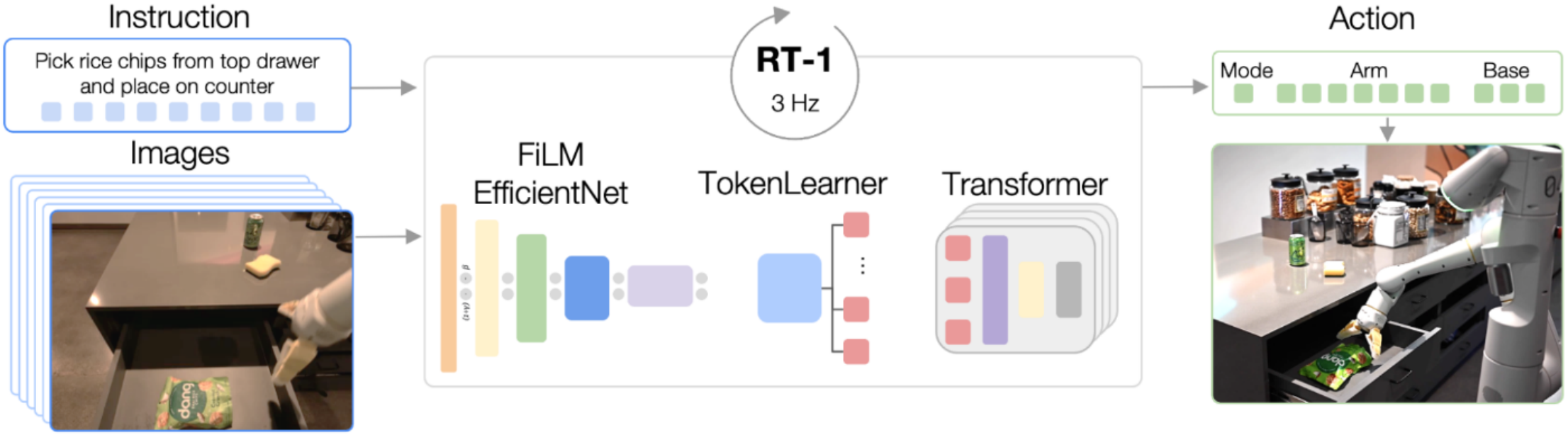


Dexterous hand grasping (ICCV 2023 Best paper finalist)

Large Vision-Language Model



Embodied Vision-Language Model: RT-1



RT-1's architecture: The model takes a text instruction and set of images as inputs, encodes them as tokens via a pre-trained FiLM EfficientNet model and compresses them via TokenLearner. These are then fed into the Transformer, which outputs action tokens.

Summary of Computer Vision

- Compared to human vision, computer vision deals with the following tasks:
  - visual data acquisition (similar to human eyes but comes with many more choices)
  - image processing and feature extraction (mostly low-level)
  - analyze local structures and then 3D reconstruct the original scene (from mid-level to high-level)
  - understanding (mostly high-level)
  - generation (beyond the scope of human vision system)
  - and further serving embodied agents to make decisions and take actions.

A Brief History to Computer Vision

The Birth of Artificial Intelligence

Alan Turing and Turing test

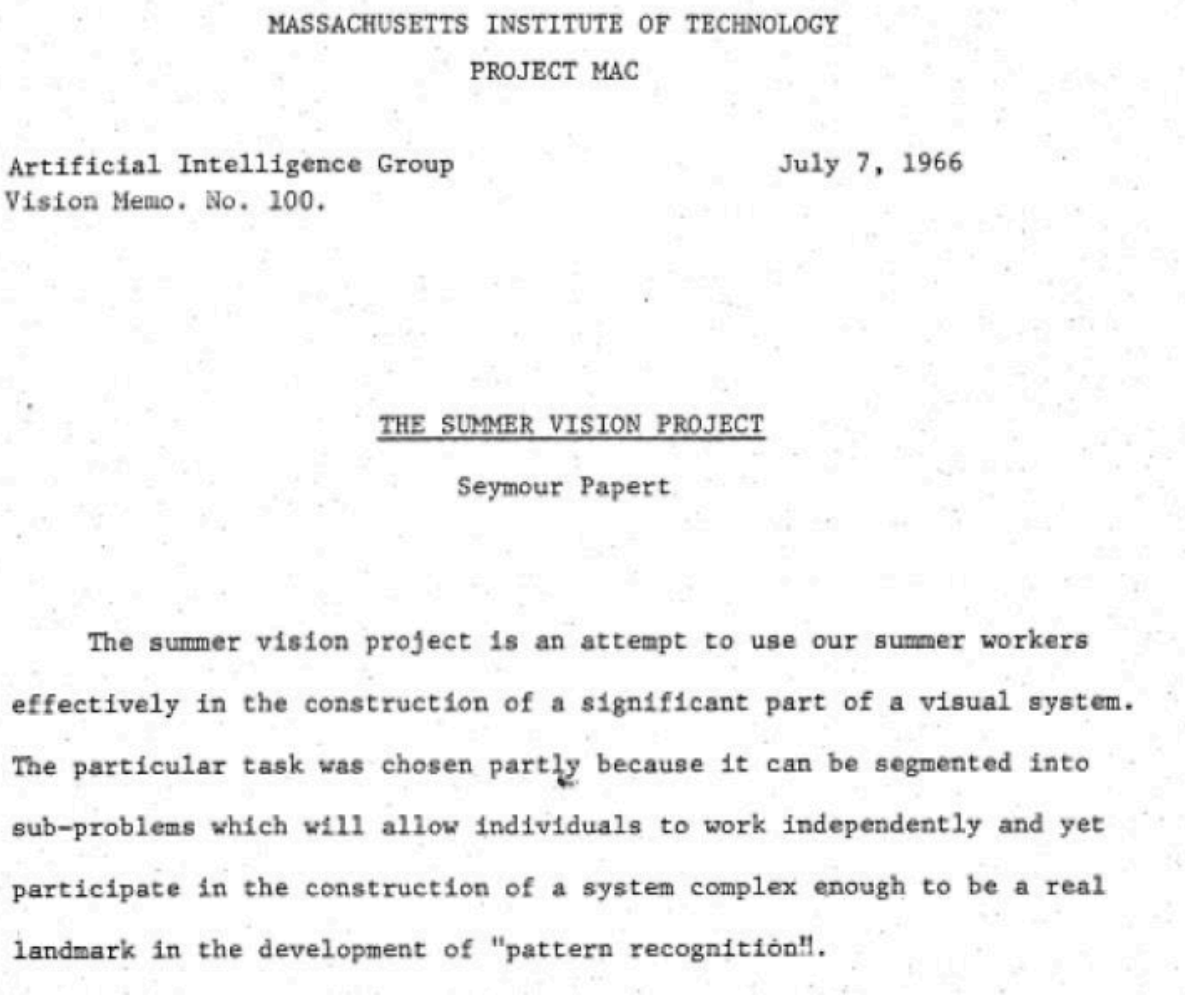
1950, Turing wrote the article ***"Computing machinery and intelligence"***, in which he described what would become known as the **"Turing Test"**.

The Dartmouth Conference

August 1956. *From left to right: Oliver Selfridge, Nathaniel Rochester, Ray Solomonoff, Marvin Minsky, Trenchard More, John McCarthy, Claude Shannon.*

Early in 1960s: CV as a Summer Project

- A visual perception component of an ambitious agenda to mimic human intelligence.
- AI pioneers believed that solving the “visual input” problem would be easier than solving higher-level reasoning and planning.
- Marvin Minsky at MIT asked his undergrad Gerald Jay Sussman to “spend the summer linking a camera to a computer and getting the computer to describe what it saw”. *However, we know this is not that easy.*



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

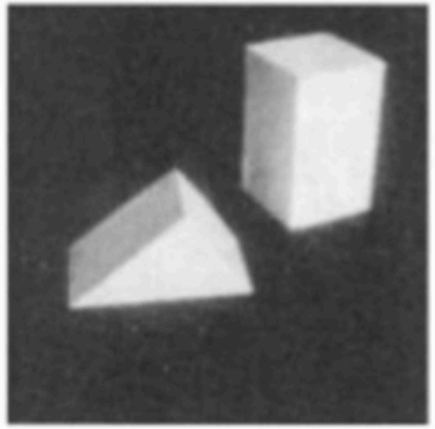
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

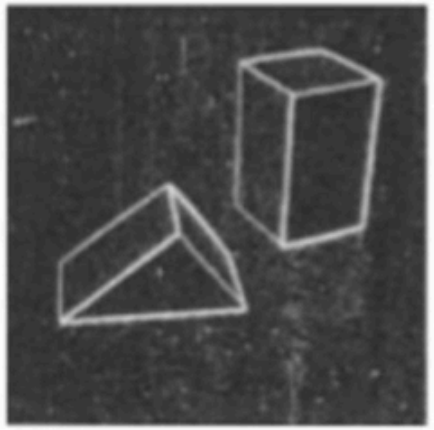
Early in 1960s: Interpretation of Synthetic Objects

Larry Roerts
1963, 1st thesis of Computer Vision

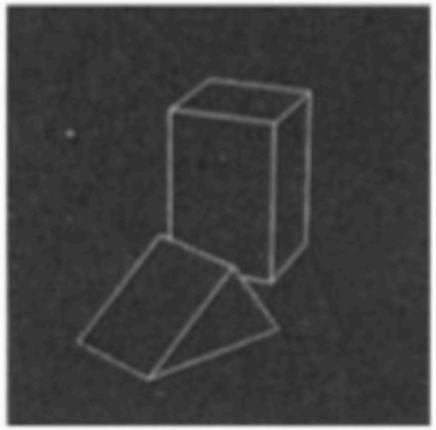
Ph.D. thesis "Machine Perception of Three-Dimensional Solids"



Input image



2x2 gradient operator



computed 3D model
rendered from new viewpoint

Borrowed from Stanford CS231N Lecture 01.

1970s/1980s: reconstruction as the first step

- What distinguished computer vision from the already existing field of digital image processing:
  - the desire to recover the three-dimensional structure of the world from images
  - And use this as a stepping stone to- wards full scene understanding

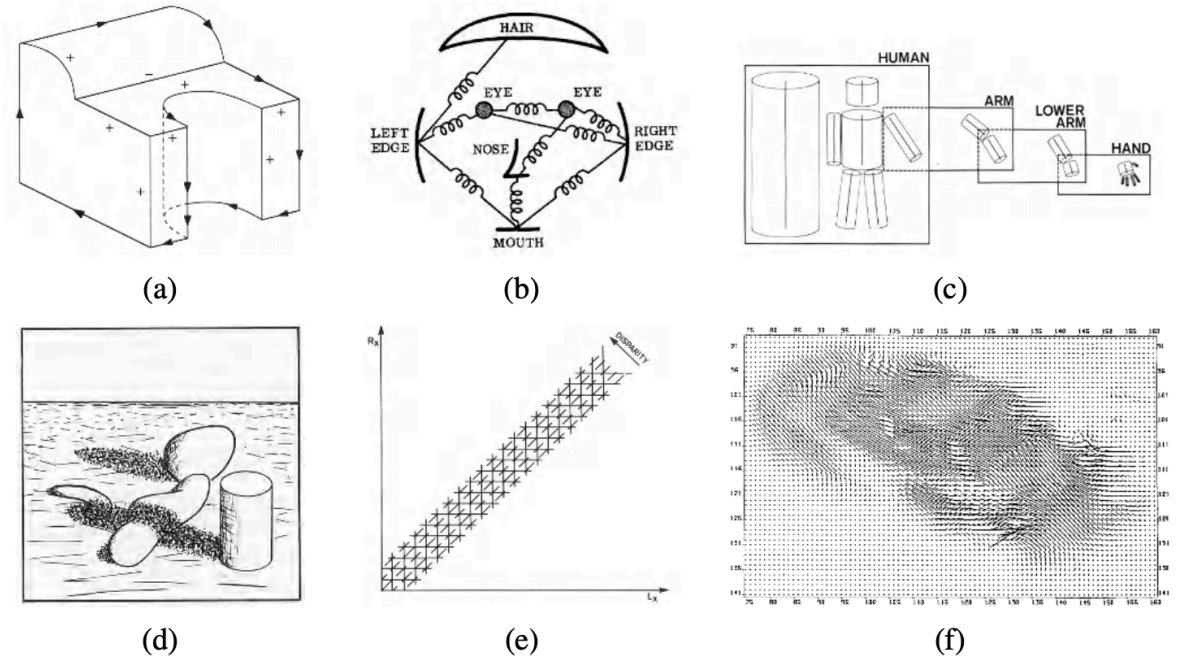
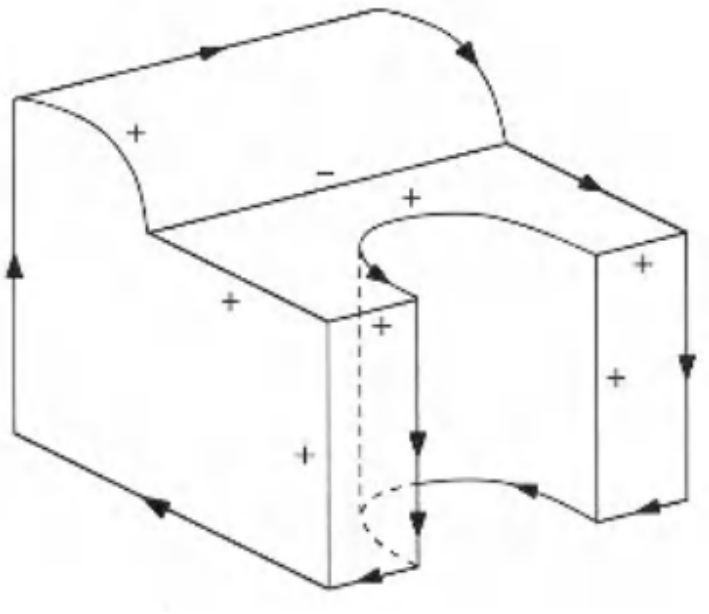


Figure 1.7 Some early (1970s) examples of computer vision algorithms: (a) line labeling (Nalwa 1993) © 1993 Addison-Wesley, (b) pictorial structures (Fischler and Elschlager 1973) © 1973 IEEE, (c) articulated body model (Marr 1982) © 1982 David Marr, (d) intrinsic images (Barrow and Tenenbaum 1981) © 1973 IEEE, (e) stereo correspondence (Marr 1982) © 1982 David Marr, (f) optical flow (Nagel and Enkelmann 1986) © 1986 IEEE.

Szeliski, Richard. Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

Basic Ideas

- Extracting edges and then inferring the 3D structure of an object or a “blocks world” from the topological structure of the 2D lines



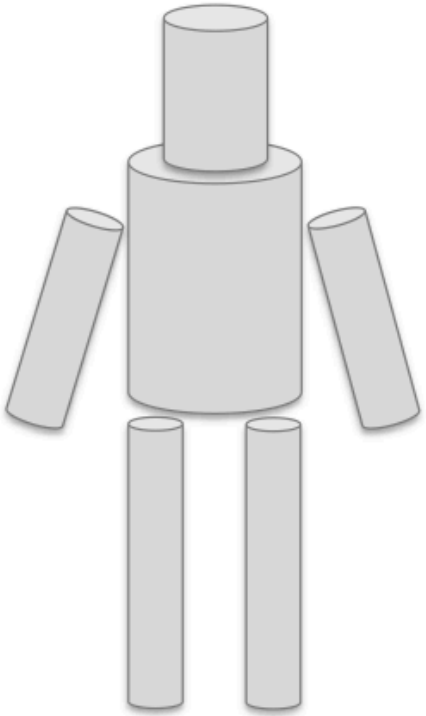
Line labeling (Nalwa 1993)

Szeliski, Richard. Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

Three-dimensional Modeling of Non-polyhedral Objects

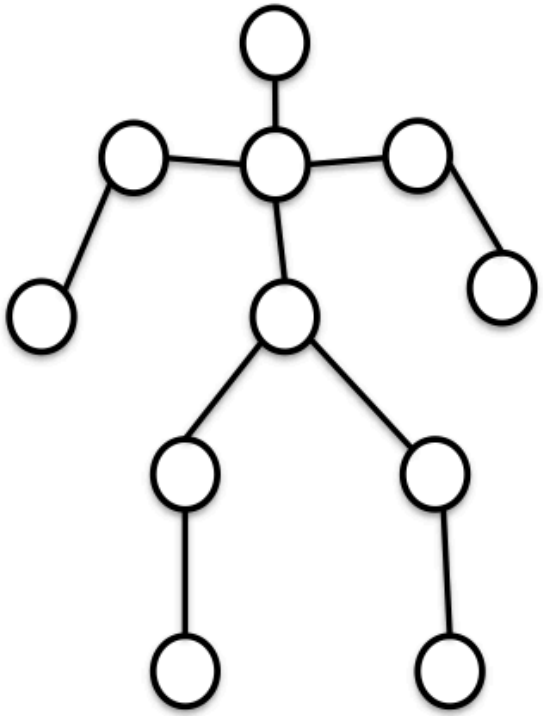
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

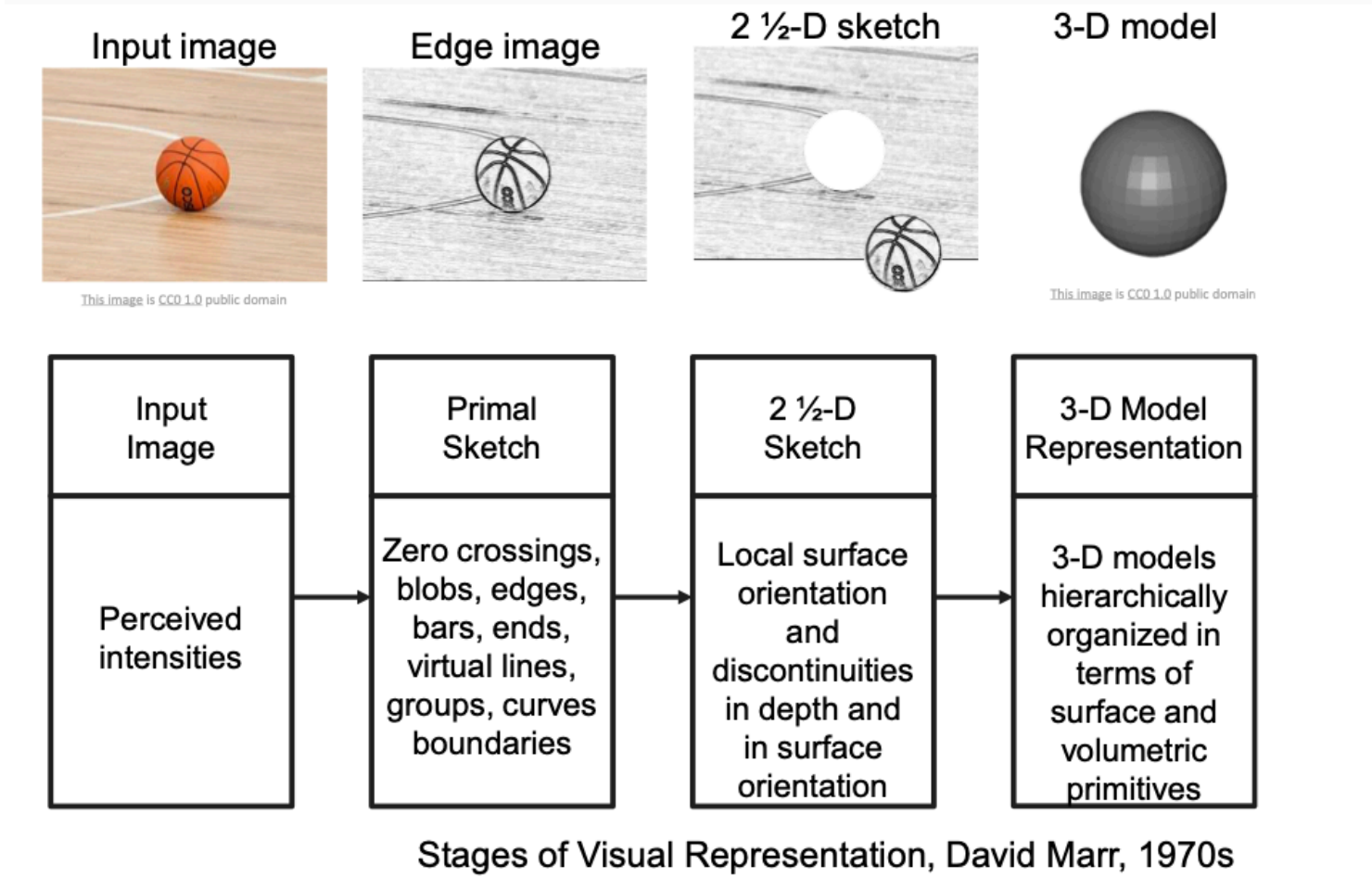
Fischler and Elschlager, 1973



Borrowed from Stanford CS231N Lecture 01.

David Marr's 2.5-D Sketch

- 2.5-D Sketch:
  - A surface based representation that bridges 2D and 3D
  - Depth-from-X: computed from a 2-D image-based representation (primal sketch) via extracting information about
    - surface orientation
    - depth from a variety of sources, such as shading, stereo, and motion.



Stages of Visual Representation, David Marr, 1970s

Palmer, Stephen E. Vision science: Photons to phenomenology. MIT press, 1999.
Stanford CS 231N Lecture 01.

3D Reconstruction

Structure from Motion
(Tomasi and Kanade 1992)

Dense stereo matching
(Boykov, Veksler, and Zabih 2001)

Multi-view reconstruction
(Seitz and Dyer 1999)

Szeliski, Richard. Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

Recognition and Segmentation

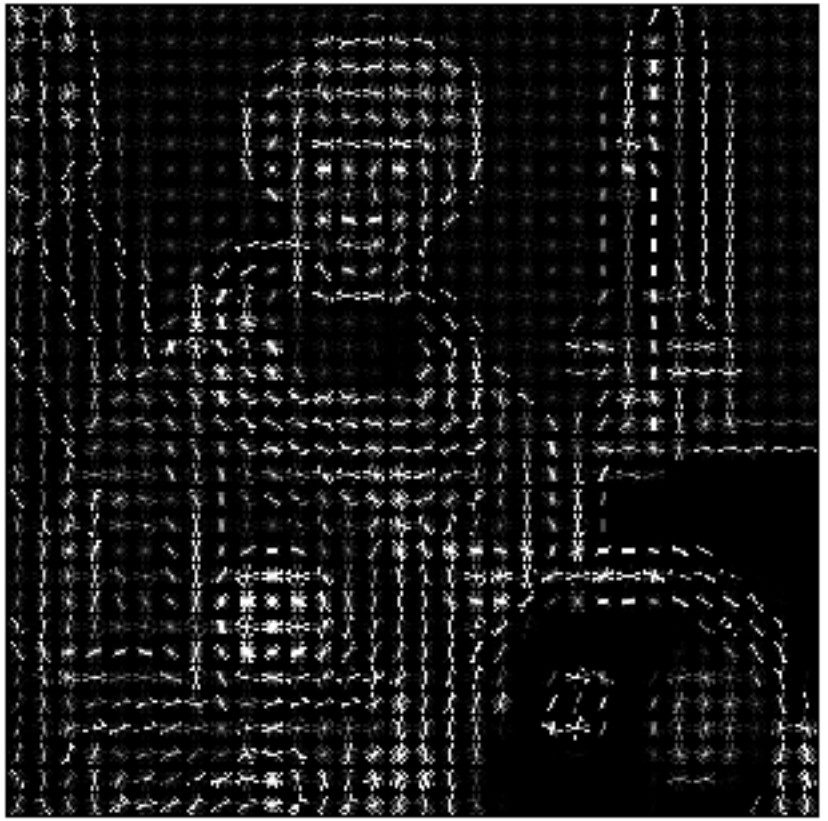
D. Lowe. IJCV, 1992

Normalized Cut (Shi & Malik, 1997)

Descriptors

Input image

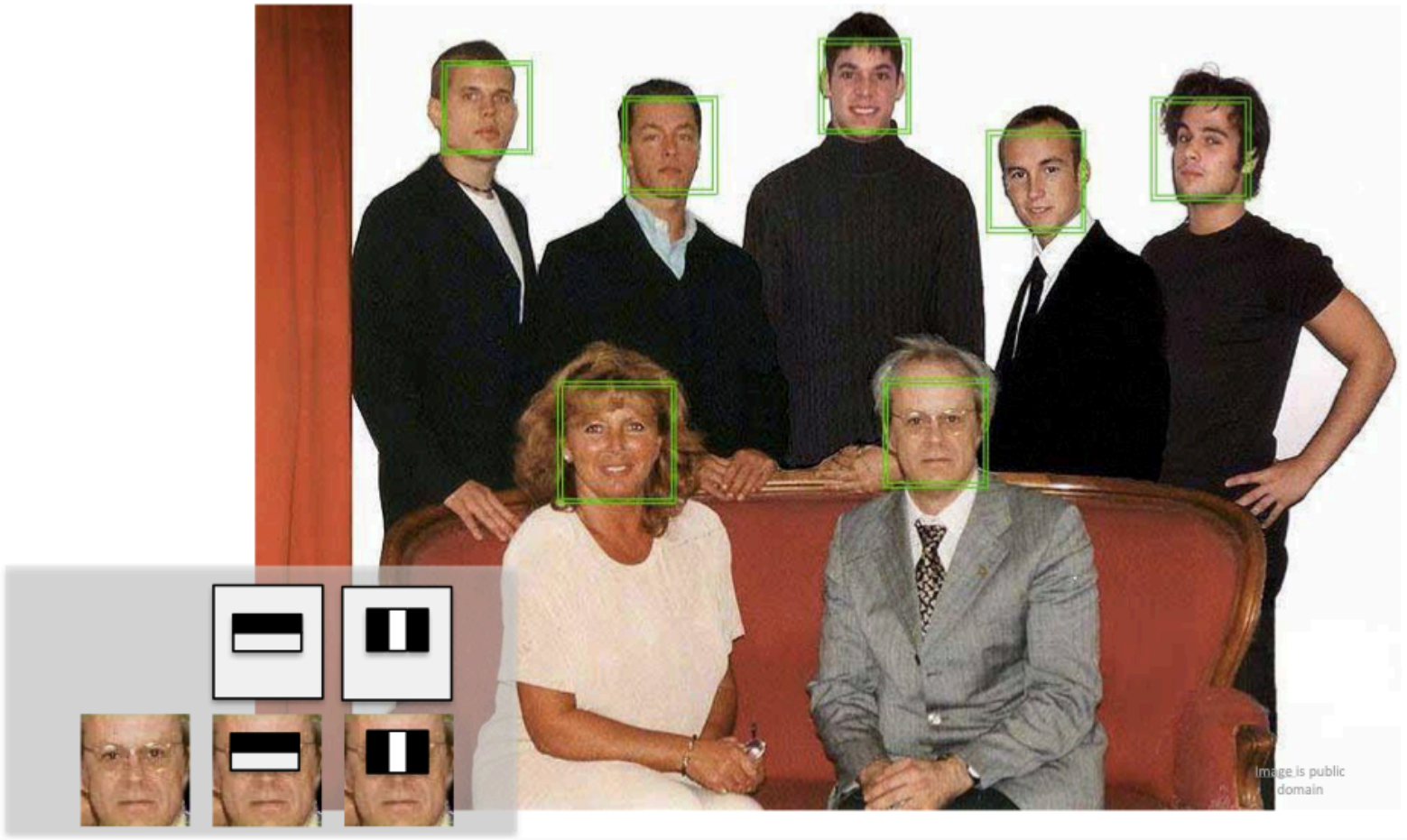
Histogram of Oriented Gradients



Histogram of Gradients (HoG) Dalal & Triggs, 2005

Credit: https://iq.opengenus.org/object-detection-with-histogram-of-oriented-gradients-hog/

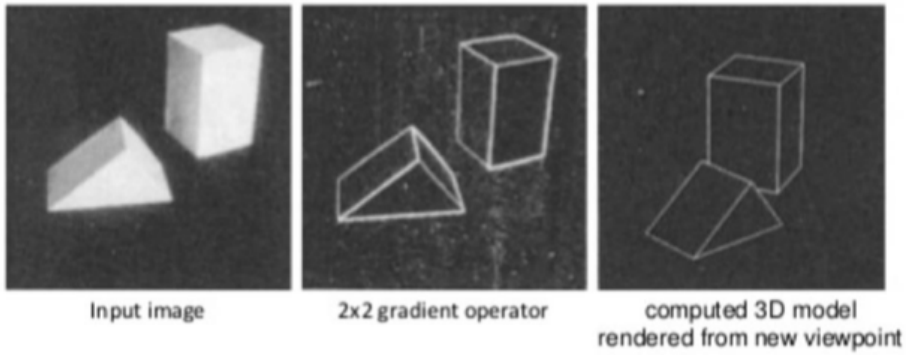
Detection

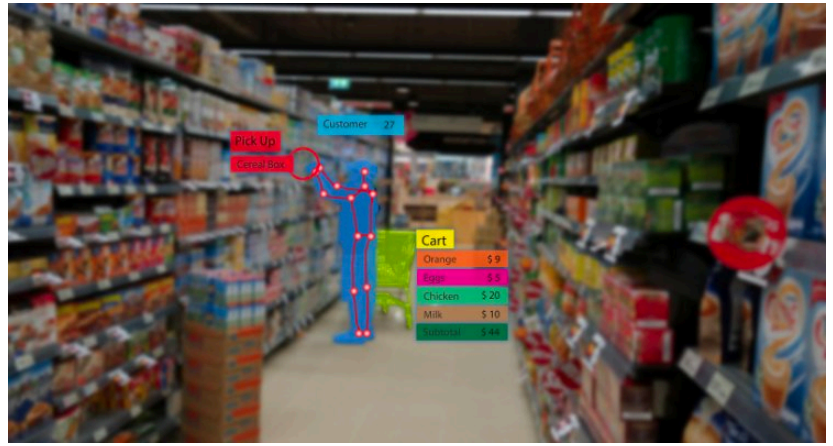


Face Detection, Viola & Jones, 2001

CV from the Classic Era to the Deep Learning Era

- Previous works don't leverage learning.
- However, many techniques and concepts proposed by them are still foundations for modern computer vision.
- Current trend:
  - From non-learning based method to **learning-based method**
  - Rely on **big data**
  - Requires more **computation resources**.





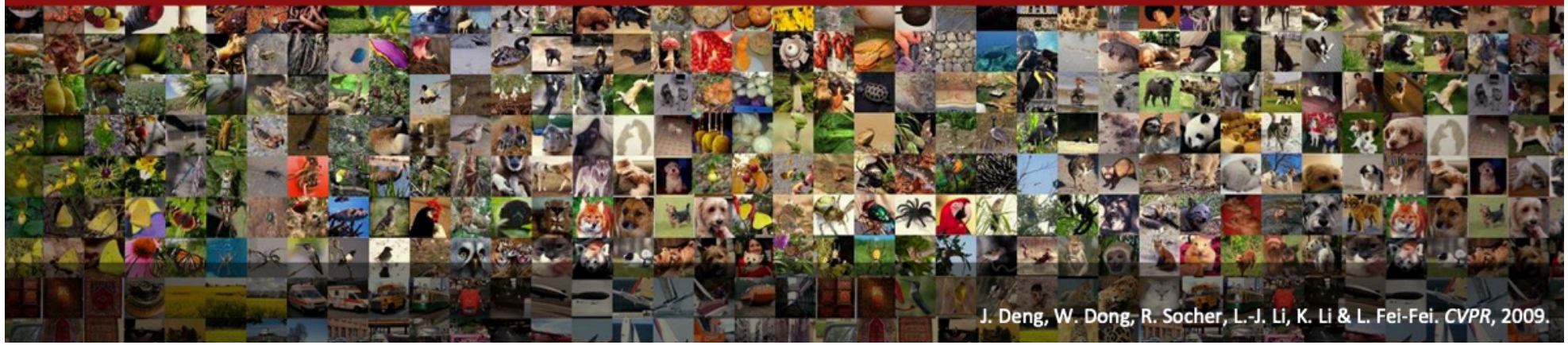
Algorithm: Deep Learning

2018 Turing Awards: Geoffrey Hinton, Yann LeCun, and Yoshua Bengio

Data: ImageNet and Its Benchmark

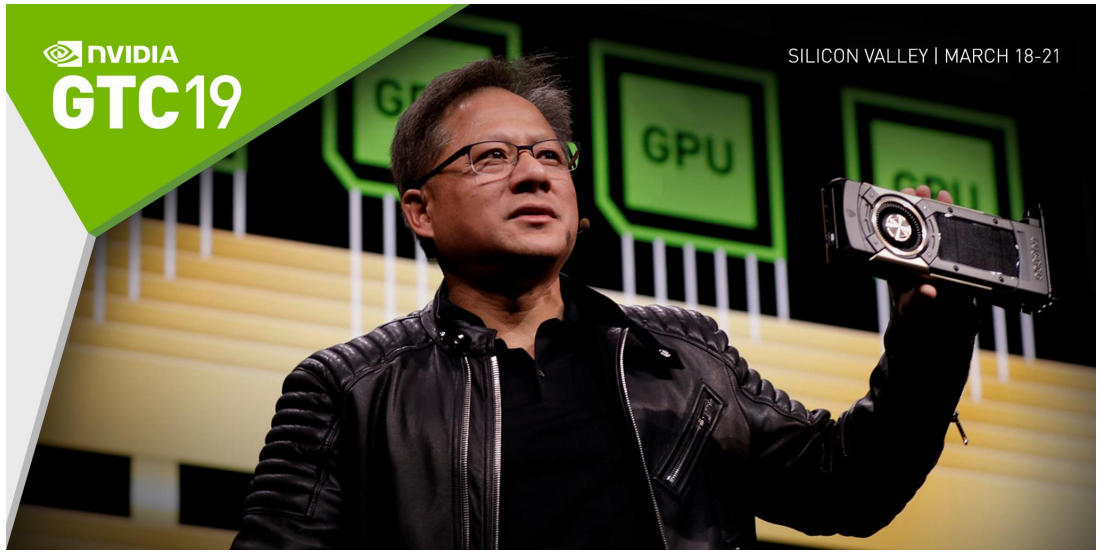
IMAGENET

22,000 categories ⋮ 15,000,000 images

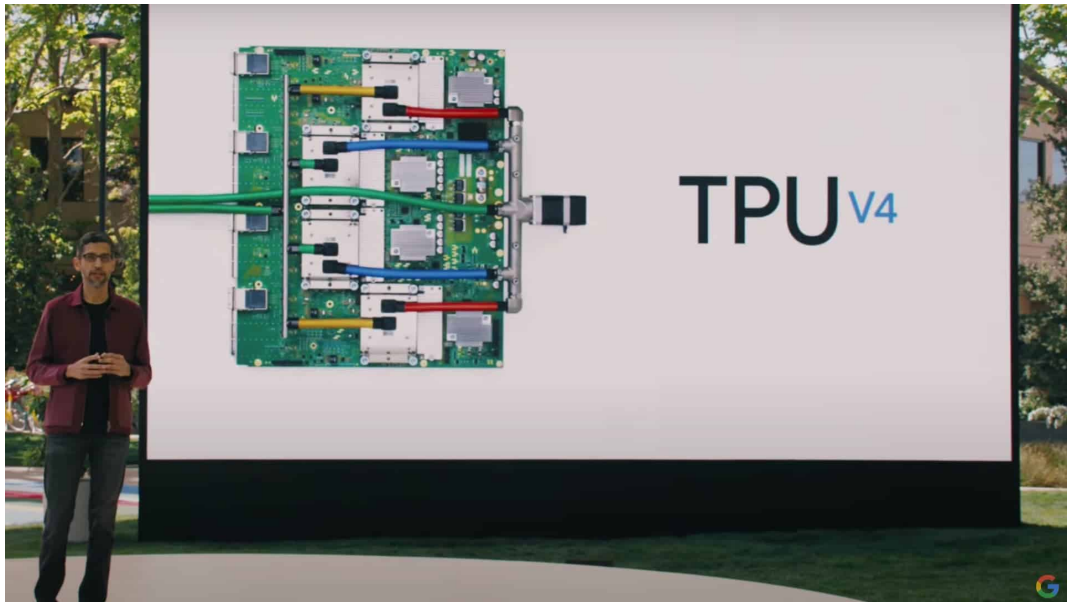


J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li & L. Fei-Fei. *CVPR*, 2009.

Computational Resources: GPU



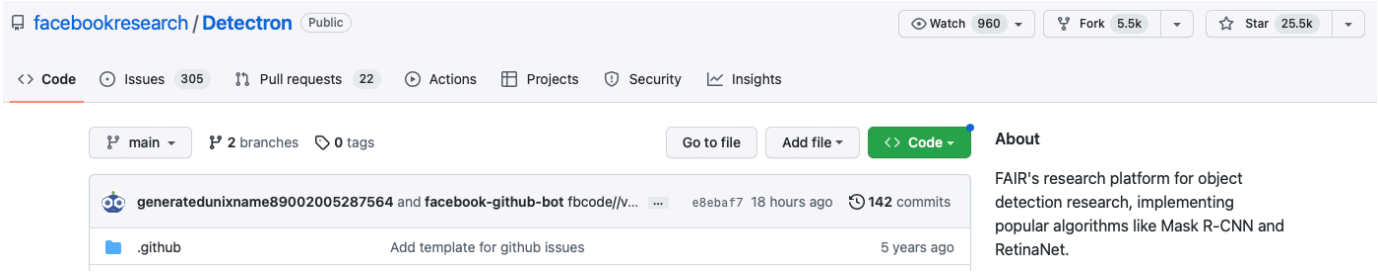
NVIDIA and its GPU

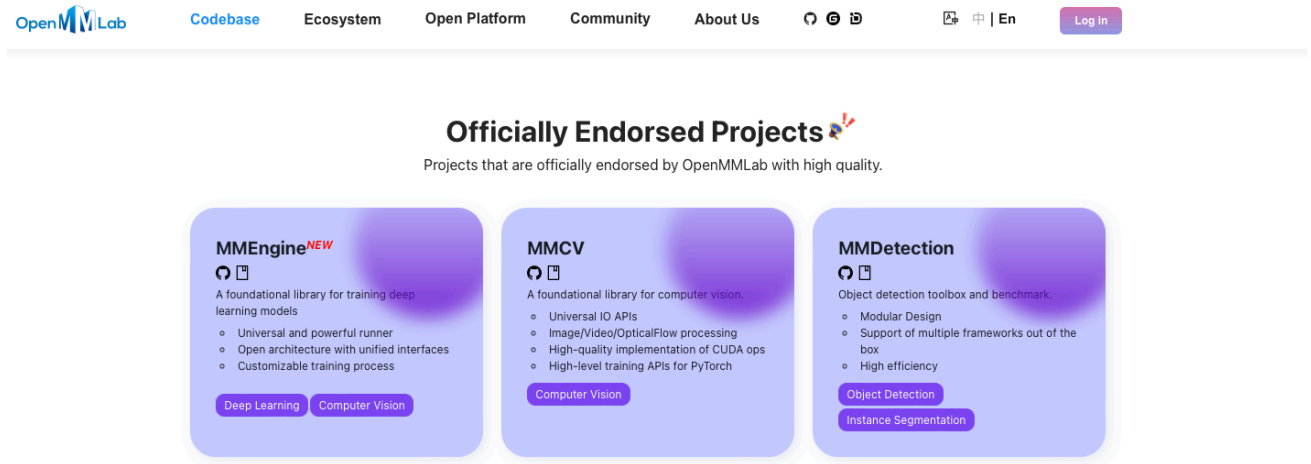


Google and its TPU

Deep Learning Frameworks and Opensource

PyTorch





Course Organization

Course Organization: Low-Level Vision and Deep Learning

Lecture	**03/01/2023 15:10 Wednesday**	Classic Vision Techniques	Feature extraction and classic descriptors; Fitting: Least Square, RANSAC, Hough Voting; Optimization: First-order/Second-order methods.
Lecture	**03/08/2023 15:10 Wednesday**	Deep Learning I	Supervised Learning; Linear classifier and Logistic Regression; MLP and Backpropagation; CNN;
Lecture	**03/15/2023 15:10 Wednesday**	Deep Learning II	Why CNN is better? Training Neural Networks: data preprocessing, weight initialization, optimizer, learning rate Improve CNN training: underfit (Batchnorm, ResNet)
Lecture	**03/22/2023 15:10 Wednesday**	Deep Learning III	CNN training: overfit; classification, segmentation
Lecture	**03/29/2023 15:10 Wednesday**	Deep Learning IV, 3D Vision I (Camera Model)	Attention and Transformer; Transformations; Camera Model: orthographic, weakly persepctive, perspective

(Tentative schedule)

For the complete and up-to-date schedule, see https://hughw19.github.io/Intro2CV_23Spring/schedule/.

Course Organization: Mid-Level and 3D Vision

Lecture	**03/29/2023 15:10 Wednesday**	Deep Learning IV, 3D Vision I (Camera Model)	Attention and Transformer; Transformations; Camera Model: orthographic, weakly persepctive, perspective
Lecture	**04/12/2023 15:10 Wednesday**	3D Vision II (From Single View and Epipolar Geometry to Stereo Vision)	Camera calibration; Epipolar Geometry: Vanishing lines/points, Essential matrix, Fundamental matrix; Passive and active stereo system;
Lecture	**04/19/2023 15:10 Wednesday**	3D Vision III (3D data)	Depth sensors; 3D data (voxel, mesh)
Lecture	**04/26/2023 15:10 Wednesday**	3D Vision IV (3D Deep Learning)	3D data (SDF, point cloud); 3D Deep Learning: Point, Mesh; 3D Deep Learning: Sparse Voxel Conv

(Tentative schedule)

For the complete and up-to-date schedule, see https://hughw19.github.io/Intro2CV_23Spring/schedule/.

Course Organization: High-Level Vision

Lecture	**03/22/2023 15:10 Wednesday**	Deep Learning III	CNN training: overfit; classification, segmentation
Lecture	**05/10/2023 15:10 Wednesday**	Object Detection and Instance Segmentation	2D Object detector (SSD, RCNN series, YOLO); Instance Segmentation, Panoptic Segmentation; 3D object detection and instance segmentation
Lecture	**05/17/2023 15:10 Wednesday**	Pose and Motion	6D pose; rotation representations: Euler angle, axis angle, quaternion; Instance-level 6D pose estimation: PoseCNN, rotation regression; category-level pose estimation: NOCS, orthogonal procrustes; two-frame motion, optical flow.
Lecture	**05/24/2023 15:10 Wednesday**	Temporal Data Analysis	RNN, LSTM, GRU; Video Analysis: 3D CNN, Two-stream network, ConvRNN
Lecture	**05/31/2023 15:10 Wednesday**	Generative Model	PixelRNN/CNN, VAE, Diffusion Model, GAN, conditional generative model
Lecture	**06/07/2023 15:10 Wednesday**	Advanced Topics in Computer Vision	Neural Rendering: Neural Radiance Field (NeRF); Embodied AI.

(Tentative schedule)

For the complete and up-to-date schedule, see https://hughw19.github.io/Intro2CV_23Spring/schedule/.

Next week: Lecture 2, Classic Vision Methods