Computer Vision

CS308

Feng Zheng

SUSTech CS Vision Intelligence and Perception

Week 1

南方科技大学

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

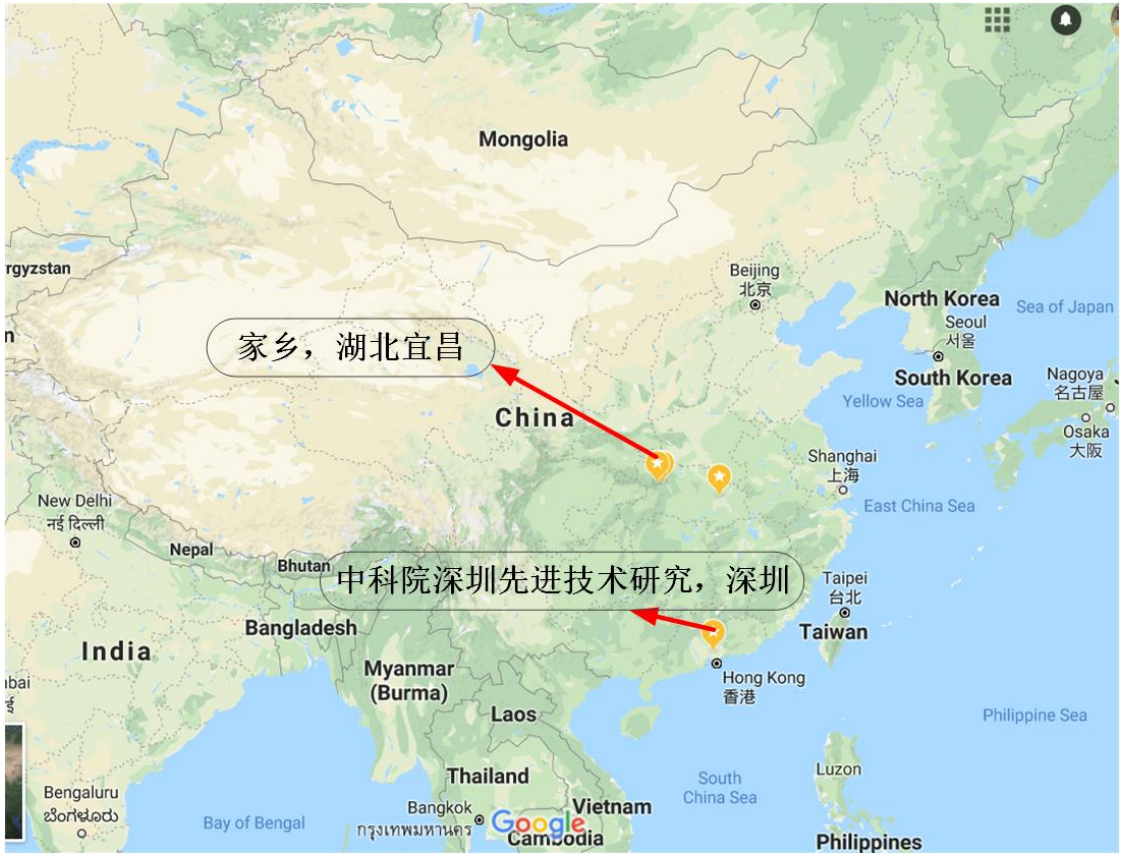
Content

- Introduction
- The state-of-the-arts
- Conclusions

Brief Biography

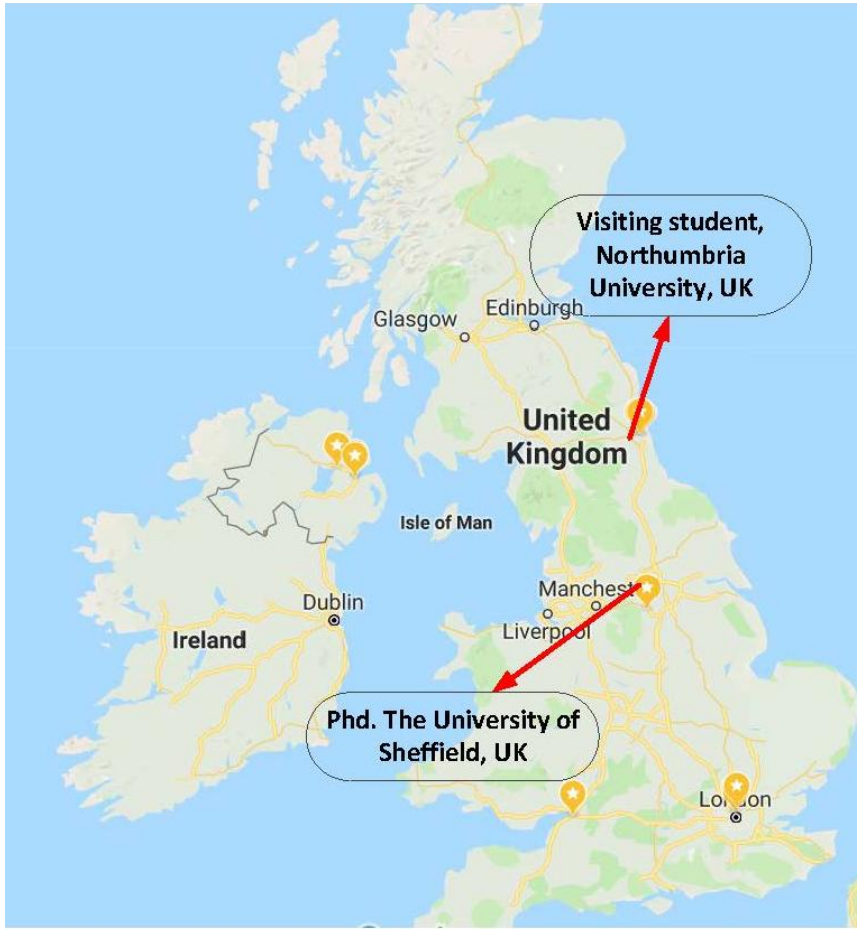
China

- Shenzhen Institutes of Advanced Technology (SIAT), CAS, Jul. 2009 - Sep. 2012



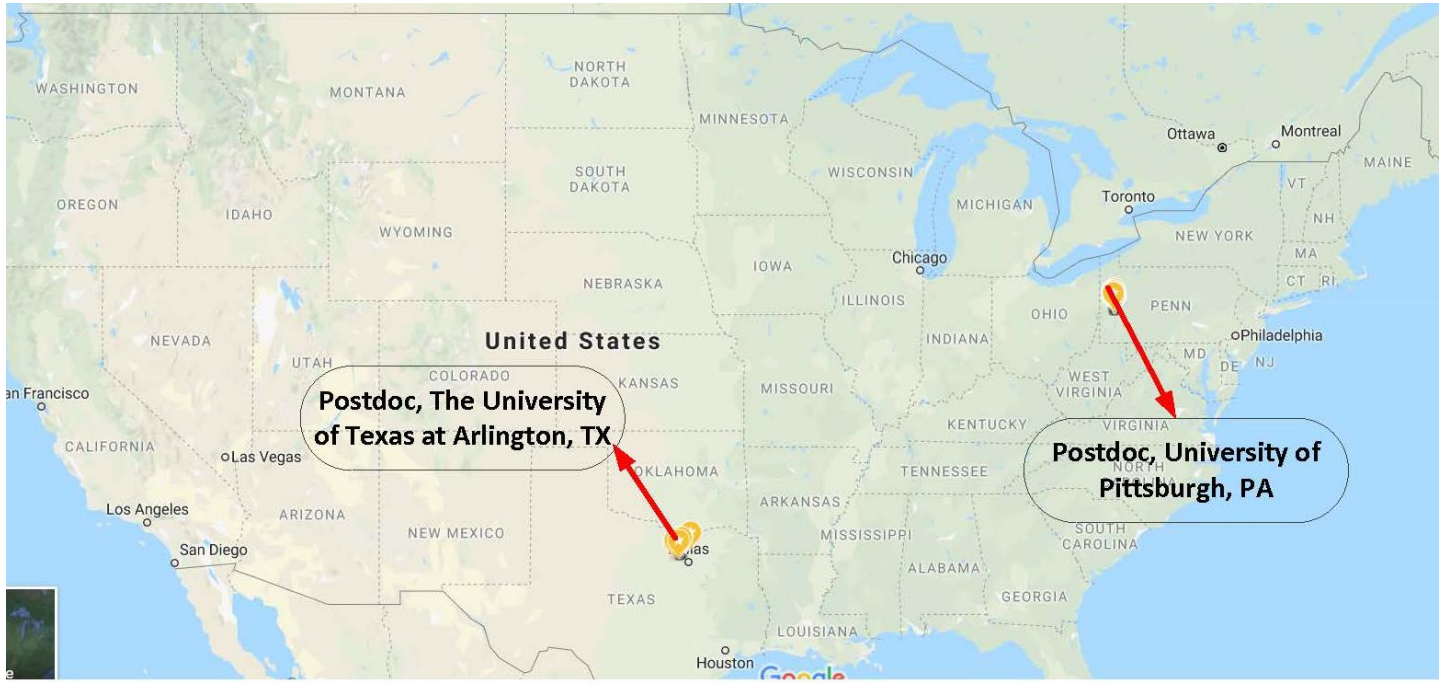
United Kingdom

- The University of Sheffield, UK, Oct. 2012 - Otc. 2016



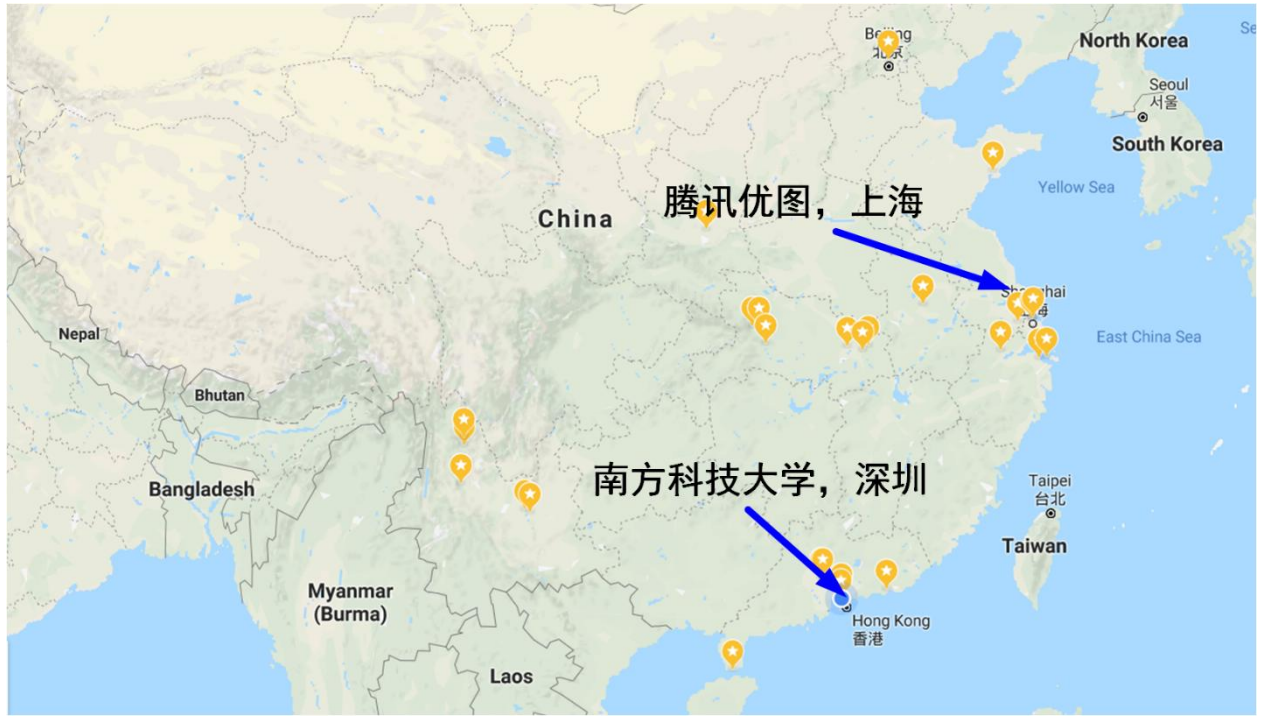
United States

- The University of Texas, at Arlington, Texas, USA, Dec. 2016 - Aug. 2017
- University of Pittsburgh, Pittsburgh, USA, Sep. 2017 - July 2018



China

- Southern University of Science and Technology, Shenzhen, China, Nov. 2018 - present
- Youtu Lab, Tencent, Shanghai, China, Aug. 2018 - Oct. 2018



个人简介

郑锋，国家优青、副教授（研究员）、博导

工作经历

2018.11-至今	南方科技大学，计算机科学与工程系	副教授/助理教授
2018.08-2018.10	腾讯优图实验室，上海	高级研究员
2017.08-2018.07	匹兹堡大学，美国	博后研究员
2016.12-2017.08	德克萨斯大学，美国	博后研究员
2009.07-2012.09	中国科学院深圳技术研究院，深圳	助理研究员/研究助理

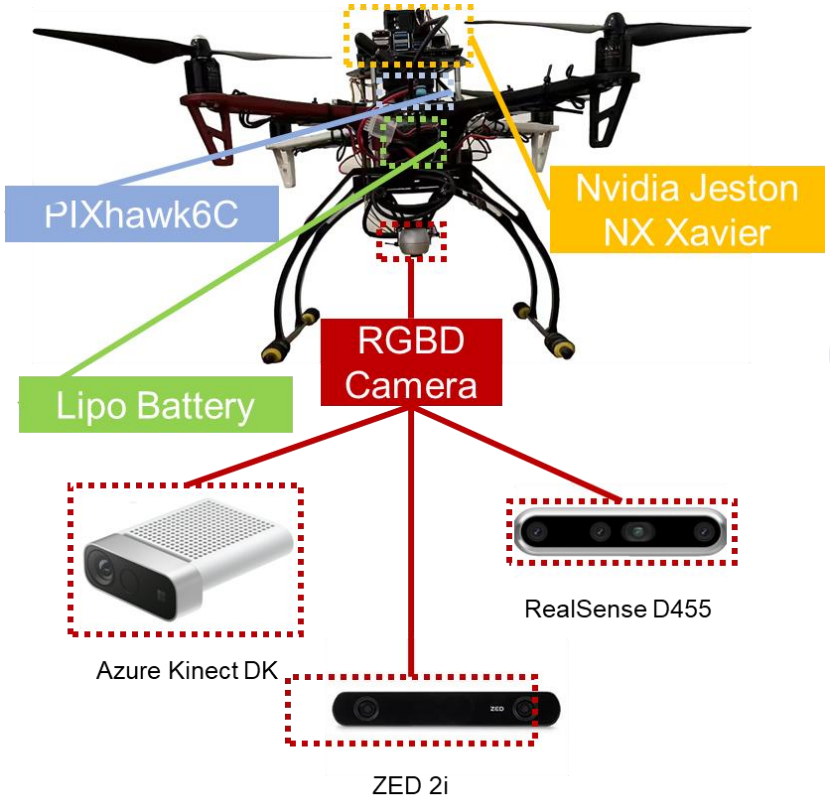
学习经历

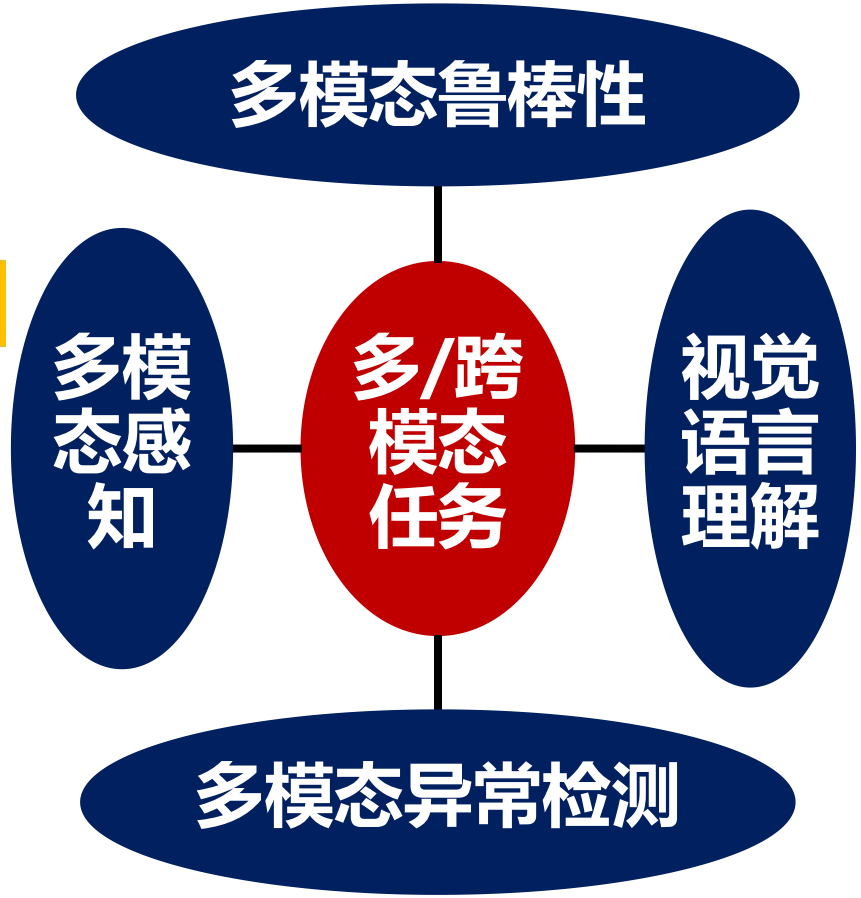
2012.10-2016.10	谢菲尔德大学，英国	博士（英国全额奖学金）

研究方向

人类获取知识的方式，从单一媒体，迅速演化到融合异地异时产生的大规模多媒体信息

研究方向





An amatuer wrestling event, two wrestlers begin their match as the announcer introduces them, guys wrestling each other in a competition, two people are fighting, wrestling champions are fighting on the ground

四个关于多/跨模态的研究方向

课题组介绍

研究助理教授	1人	硕士生	10人
博后研究员	1人	访问学生	8人
博士生	10人	研究助理	6人

课题组成员一共30余人

科研项目

项目负责人：共8项，金额991万				
2023.11-2025.01	中兴通讯股份有限公司	基于行业场景的视觉大模型技术研究	125万	PI
2022.11-2025.10	国家重点研发计划课题	智能生物眼的信息交互设计与实现	435万	PI
2022.01-2024.12	国家优秀青年科学基金	高效能跨媒体关联与理解	200万	PI
2021.12-2024.12	腾讯优图	工业场景下多模态视觉感知	100万	PI
2021.06-2022.05	腾讯数平	广告场景中的视频剪辑	15万	PI
2020.01-2023.12	国自然基金面上项目	面向多态异构大数据的 高效能检索算法研究	59万	PI
2020.10-2021.09	国防科技创新特区	空间环境事件关联与预报的 深度模型研究	50万	PI
2019.10-2020.09	国防科技创新特区	空间环境关联事件的 相似推荐算法研究	35万	PI
2019.10-2020.03	中科院空间中心	空间环境事件推荐算法研究	12万	PI

主要学术任职

- CCF-A类国际会议领域主席(Area Chair)
  - ACM MM 2020 & 2021; NeurIPS 2023
- 国际会议区域主席(Local Co-Chair)
  - IEEE ICME 2021 (CCF-B类) & IJCB 2021 (CCF-C类)
- CCF C类中国学术会议(Organization Chair)
  - PRCV 2022
- 国际会议程序委员会(PC)：CCF-A类
  - CVPR, ICCV, ICLR, AAAI, ICML, NeuIPS等30次以上
- 领域内重要杂志副编(Associate Editor)
  - IET Image Processing (CCF-C类)
- AI领域内主流杂志(>20个)审稿：
  - IEEE TPAMI/TNNLS/TCSVT/TCYB/TMM, PR, IS等

竞赛获奖

- 2020年，天池大数据竞赛-基于**多模态学习**的太阳风暴预测“第一名”
- Rank 1 in Make-up Temporal Video **Grounding** Track of PIC challenge at ACMMM 2022
- Rank 1 in Make-up Dense Video **Captioning** Track of PIC challenge at ACMMM 2022
- Rank 2 in Generic Event Boundary **Captioning** Track of LOVEU Challenge at CVPR 2022
- Rank 2 in Event **Dense-Captioning** Track of ActivityNet Challenge at CVPR 2022
- Rank 2 in Event **Dense-Captioning** Track of ActivityNet Challenge at CVPR 2021
- Rank 3 in **TinyAction** Challenge at CVPR 2021

学术研究

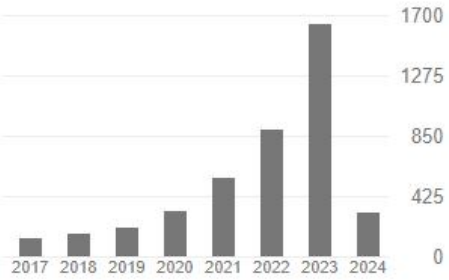
发表论文：100余篇

- **CCF-A类：76篇**
- **ACM/IEEE Transactions：27篇**
- **谷歌学术引用：4700余次，高被引2篇**

Google 学术搜索

引用次数 查看全部

	总计	2019 年至今
引用	4787	3948
h 指数	32	31
i10 指数	74	71



CSRankings: Computer Science Rankings

14	▼ SUSTech	2.0	33
	Faculty	*# Pubs*	*Adj. #*
	Feng Zheng VISION,AI	53	9.5
	Yinqian Zhang SECURITY	27	6.6
	Yu Zhang 0006 AI	20	4.7

CSRanking2018-2022指标：在清华、北大、以及美国前五大计算机专业高校CMU, Berkeley, MIT, UIUC, Stanford教研系列教授中，均能排进前十

开源贡献

Caption Anything

https://github.com/ttengwang/Caption-Anything

★ **1.6k** stars

Track Anything

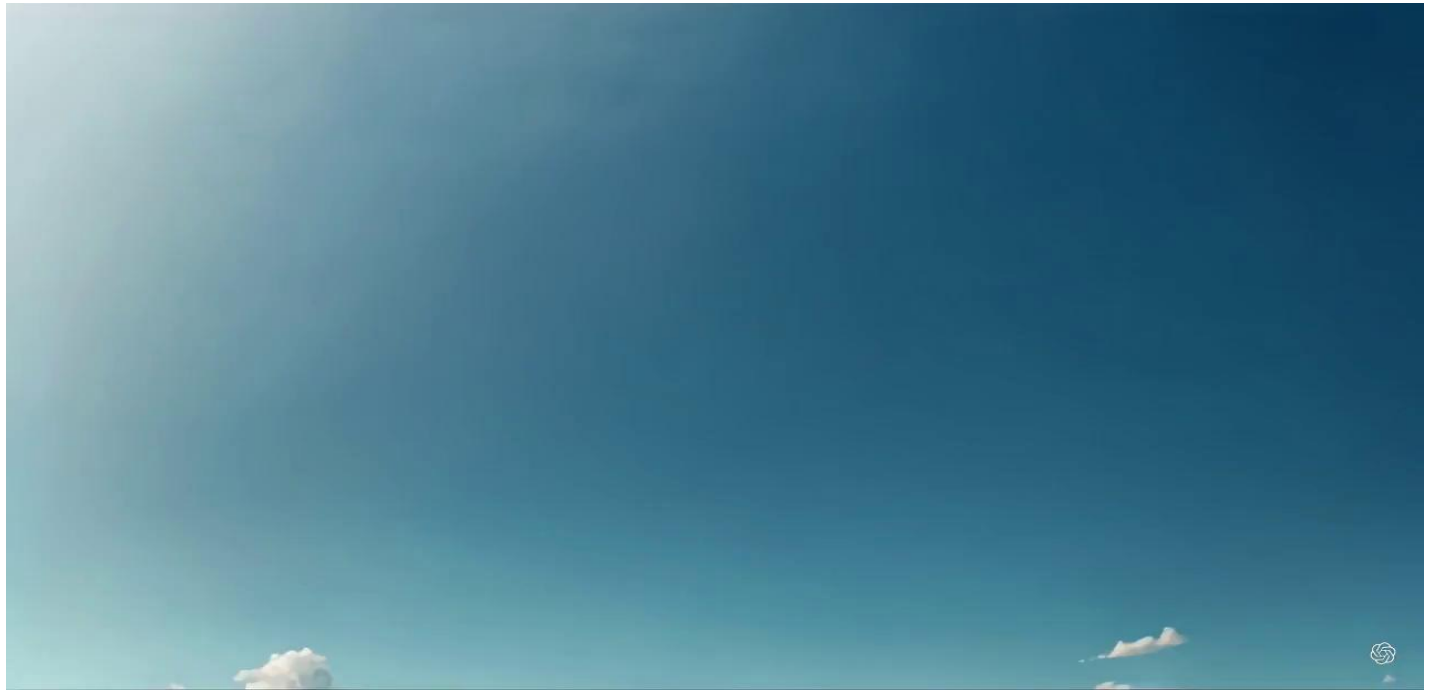
Code: https://github.com/gaomingqi/Track-Anything

Demo: https://huggingface.co/spaces/watchtowerss/Track-Anything

Report: https://arxiv.org/abs/2304.11968

★ **5.7k** stars

Video generation models as world simulators

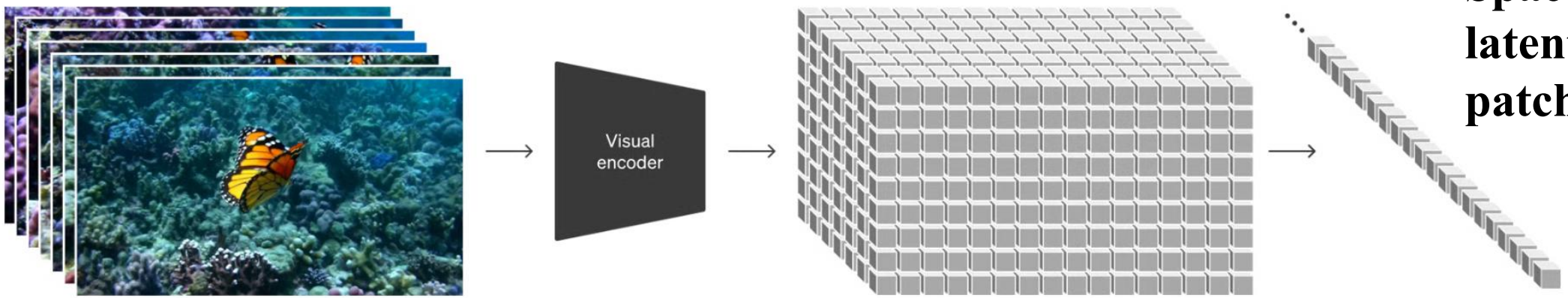


An image of a realistic cloud that spells “SORA”.

https://openai.com/research/video-generation-models-as-world-simulators#fn-19

SORA

Video compression network



Spacetime latent patches

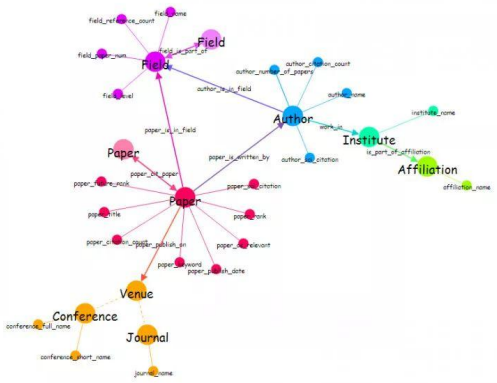
Scaling transformers for video generation

Introduction of CS308

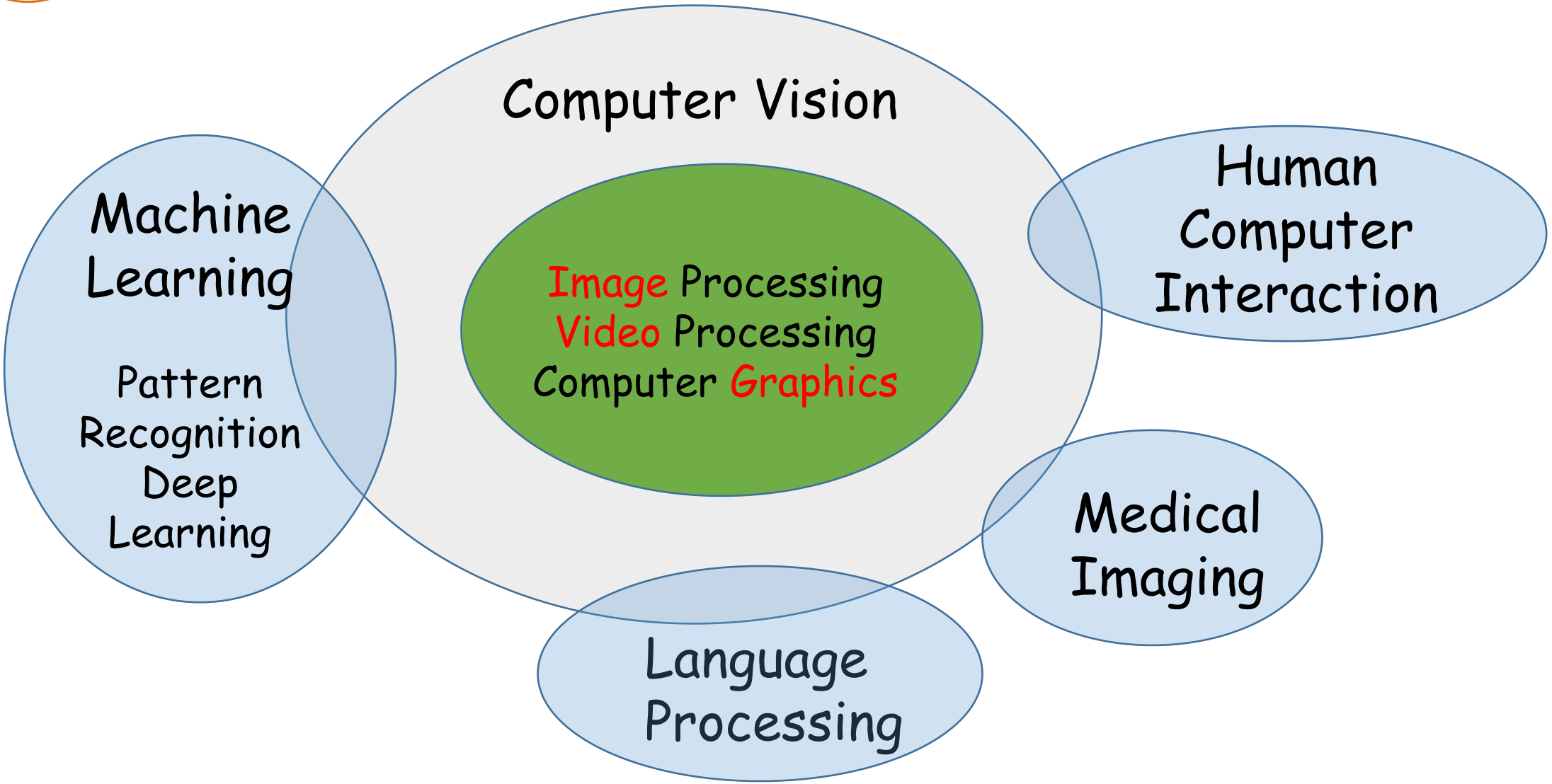
What Is Computer Vision?

- Visual computing is the **science** and **technology** of machines that **see** (capturing, understanding and prediction).
  - ➢ Come up with computational models of the human visual system
  - ➢ Build autonomous systems which could perform some of the tasks

+

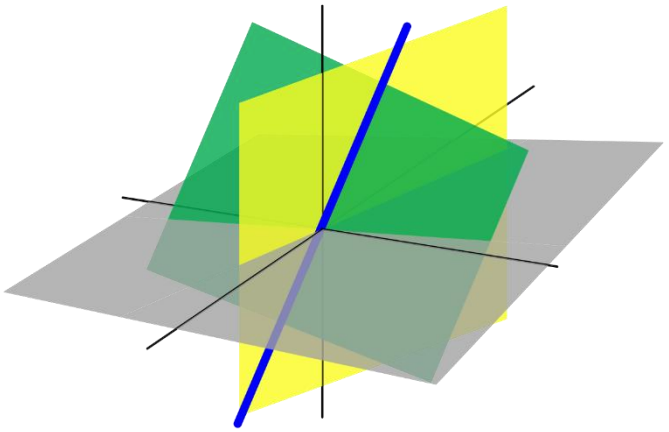


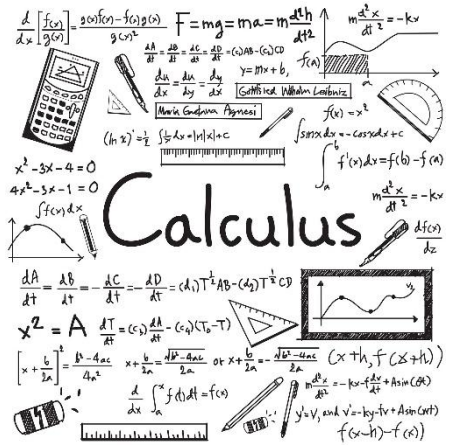
CS308-Computer Vision

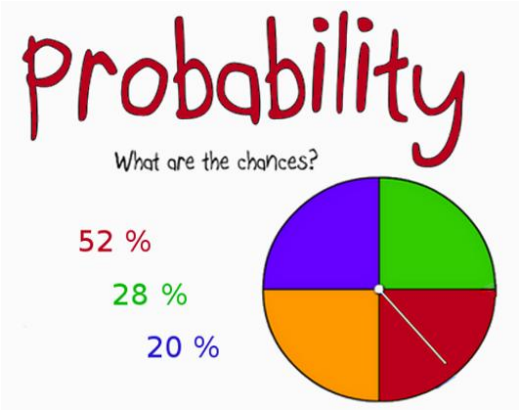


CS308-Computer Vision

- Linear algebra
- Basic calculus
- Probability



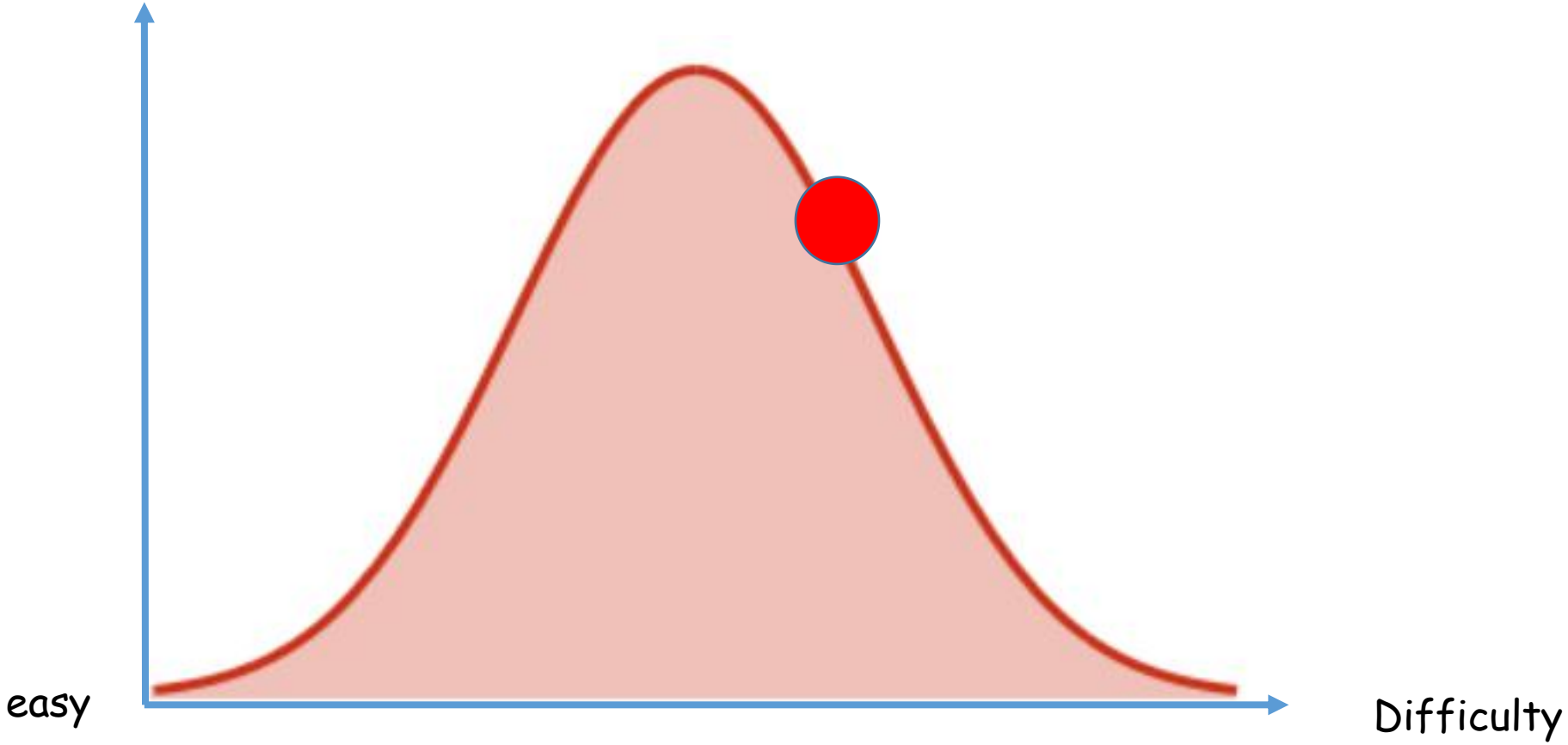




- Experience with image processing will help but is not necessary
- Experience with Python or Python-like languages will help

CS308-Target Student

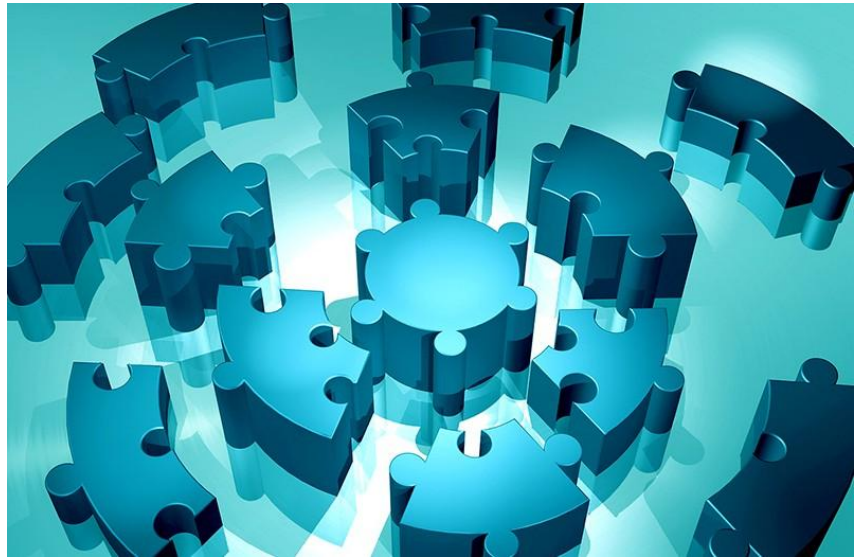
- A little higher than average ability



Expectations

- Understand the basics of computer vision (old)
- Know research trends (new)

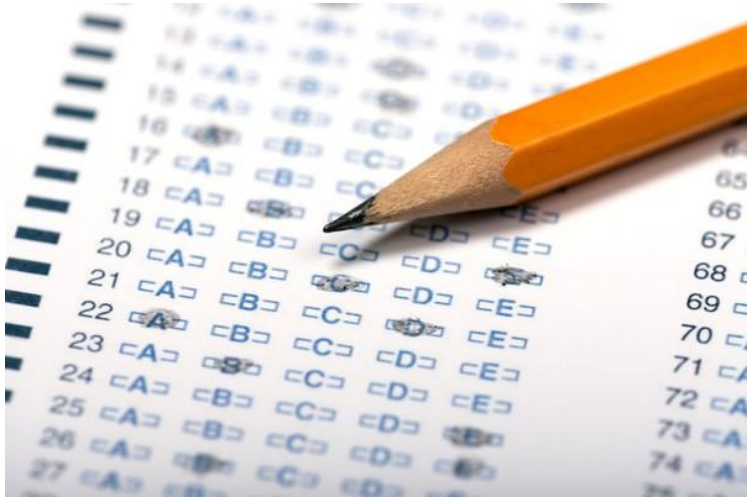
- Ability to model visual tasks
- Ability to implement the models



Exams test you on

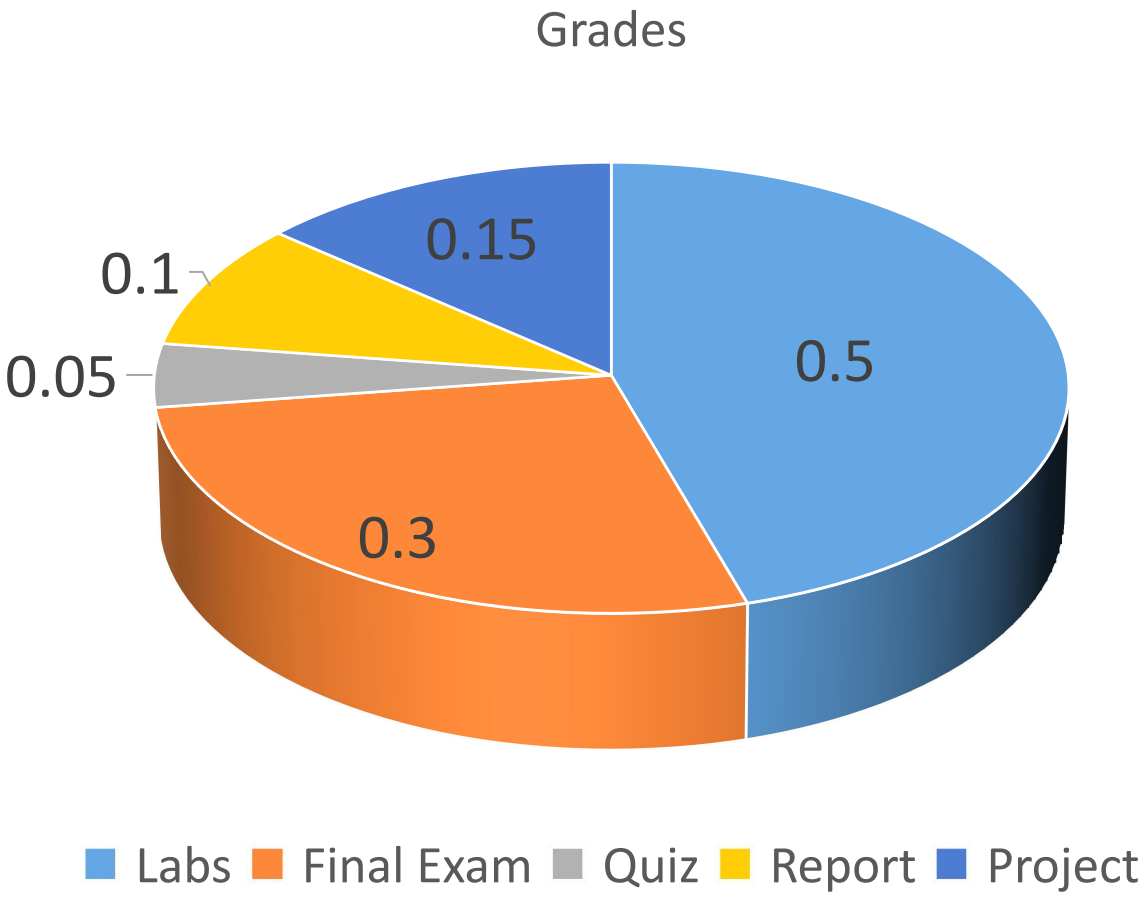
- General knowledge of visual technology

- Ability to model simple tasks
- Understand the general flow of the visual system (implement)



Grade Component

- Middle-Term Report 10%
- Final Exam 30%
- Labs 40%
- Project 15%
- Quiz: 5%



- LABS are VERY IMPORTANT

Honesty

- Get code from the internet for labs/assignments is perfectly **OK**
  - When you borrow, just say it.
  - You don't need to reinvent the wheel

~~**Plagiarism**~~

- **DON'T** pretend that you are the author of something that you didn't write.

QQ Group for CS308 Autumn

- Quick Response Code

- Blackboard: Computer Vision Spring 2024

群名称：CS308-CV-2024 Spring
群　号：747288356

Introduction of CV

Visual data: image

- The first photograph

Nicéphore Niépce. *View from the Window at Le Gras.* ca. 1826.

Robert Cornelius, *self-portrait, October or November 1839.*

Walden Kirsch as scanned into the SEAC computer by Russell A. Kirsch in 1957.

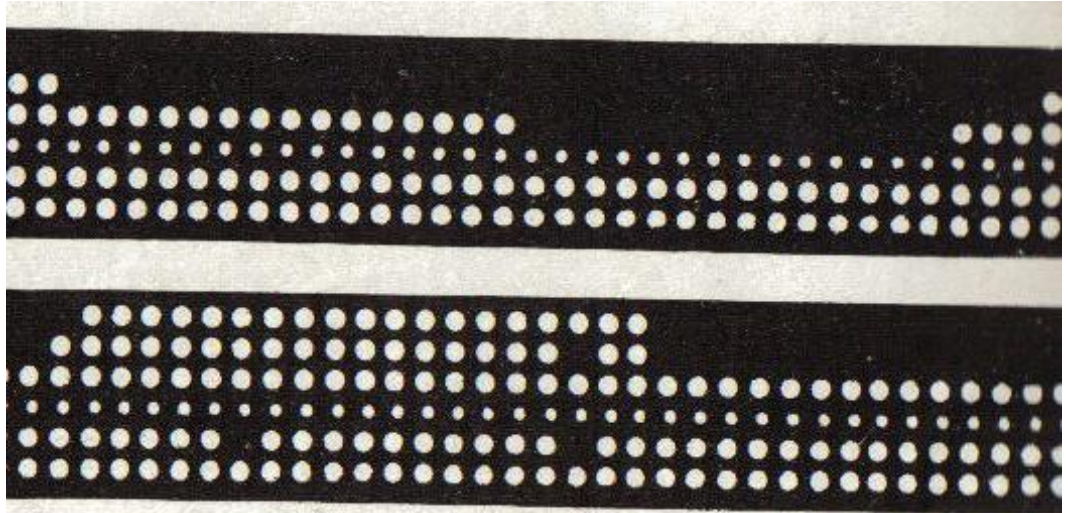
Visual data: video

- The first video (1888)

Visual data: digital imaging

- The first **digital** image was produced in 1920
  - Bartlane cable picture transmission system
  - Harry G. Bartholomew and Maynard D. McFarlane
  - London and New York

The Bartlane Transmitter.

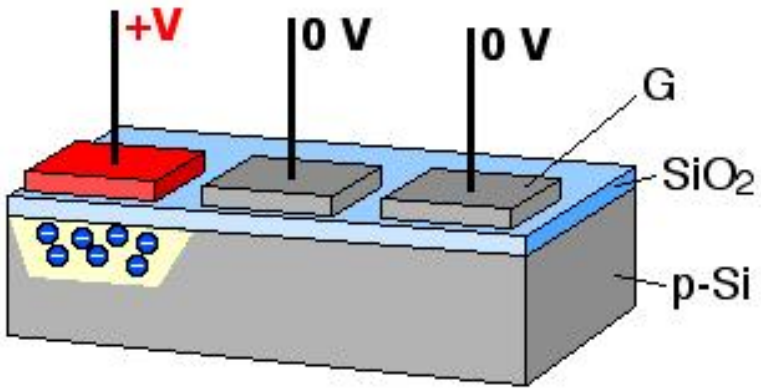


Samples of the Baudot Tape Used to Transmit Pictures by the Bartlane System.

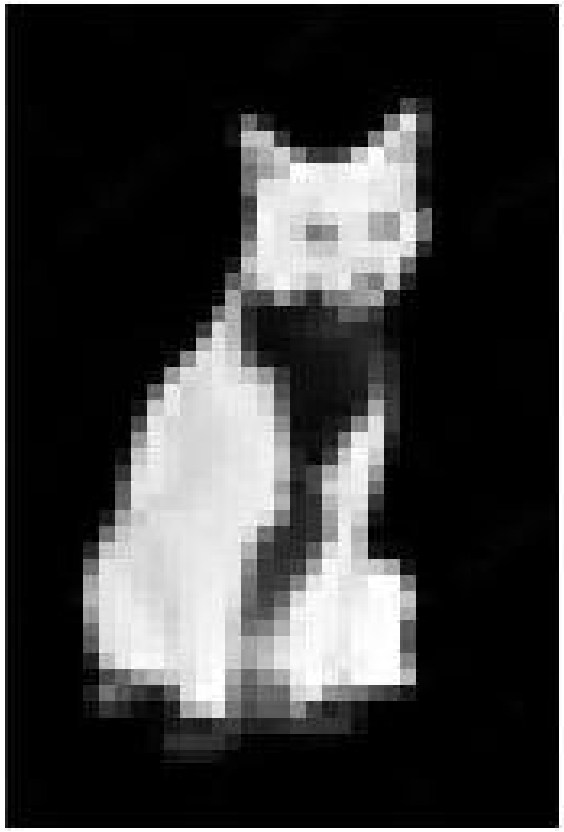
http://www.digicamhistory.com/1906_1920.html

Visual data: digital imaging

- Charge-coupled device(CCD)
  - AT&T Bell Labs(1969) by Willard Boyle and George E. Smith
  - A piece of lens
  - A Capacitor array (the photoactive region)
  - A control circuit

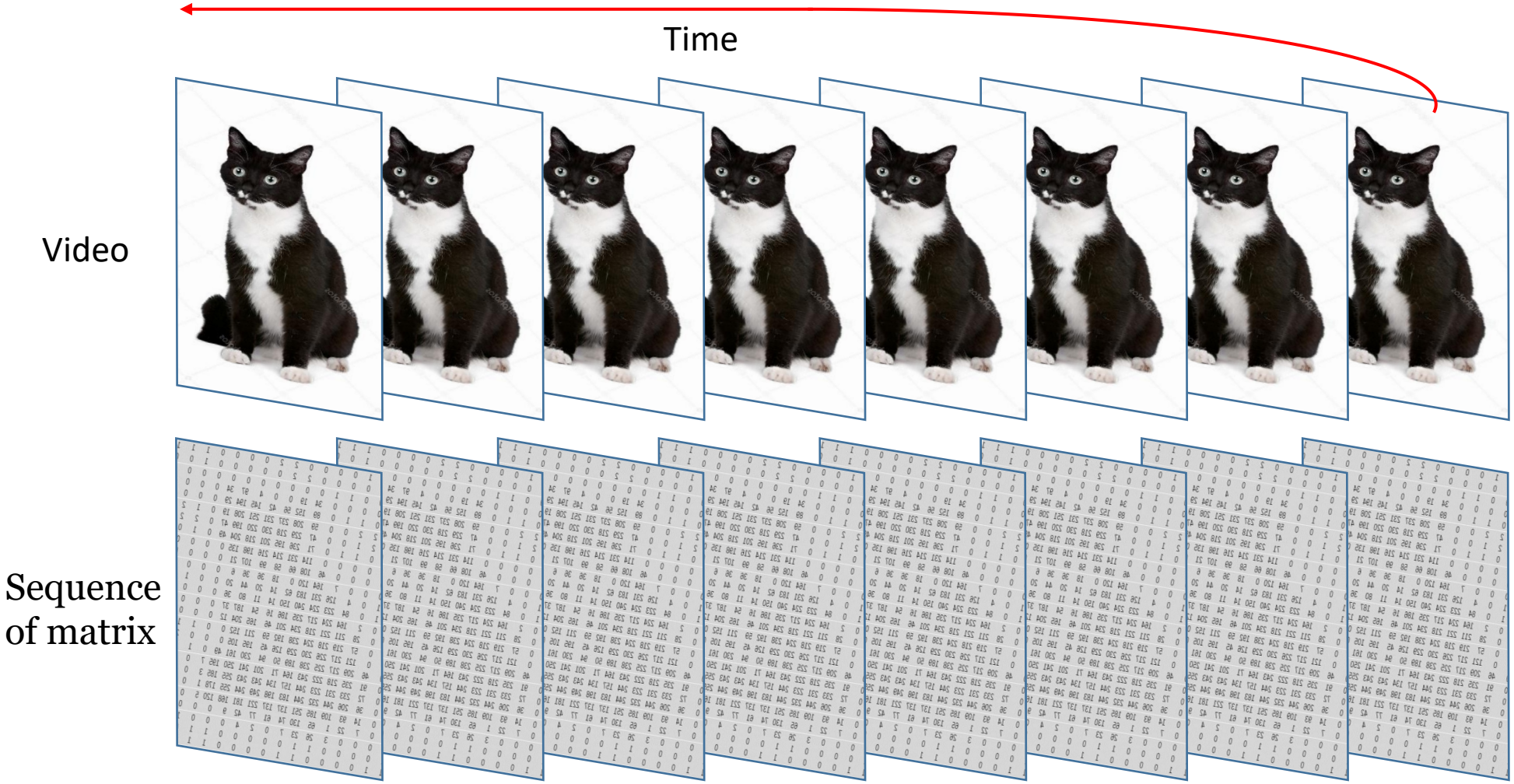


Visual data: matrix



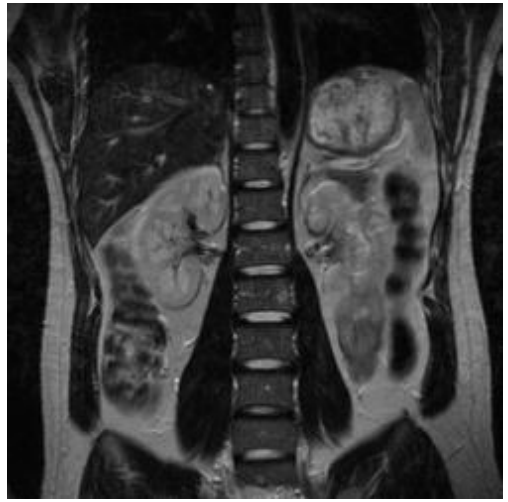
0	0	0	1	1	0	0	0	0	2	2	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	1	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	34	19	0	0	4	97	34	0	0	0	0
0	0	0	0	1	1	0	89	152	56	42	145	194	29	0	0	0	0
0	0	0	2	2	0	0	59	208	237	231	251	208	19	0	1	2	3
0	0	2	2	1	0	0	47	229	218	230	220	199	47	0	2	2	1
0	1	1	0	1	1	0	71	236	195	201	218	204	49	0	1	0	0
1	1	0	0	0	0	0	114	231	214	216	198	135	0	0	0	0	0
1	1	0	0	0	0	46	108	66	58	99	107	21	0	0	0	0	0
0	0	0	0	0	7	164	120	0	18	36	36	6	0	0	0	1	1
0	0	1	0	4	126	231	183	62	14	20	44	20	0	0	1	1	1
0	1	1	0	84	223	224	240	150	14	11	80	36	0	0	0	0	0
1	1	0	2	164	224	217	235	186	16	54	187	37	0	0	0	0	0
1	1	0	28	211	221	218	234	201	46	165	204	12	0	0	0	0	0
0	0	0	57	219	218	224	238	192	59	211	152	0	0	1	1	0	1
0	0	0	121	217	226	230	229	126	45	195	105	0	0	0	2	3	3
0	0	46	209	217	225	238	189	50	94	230	161	49	0	1	3	1	0
0	0	91	235	218	222	243	164	71	201	241	250	195	7	0	0	0	0
1	0	72	233	231	223	244	157	134	243	243	255	185	3	0	0	0	0
0	0	36	206	244	232	244	183	198	249	244	255	178	1	0	0	0	0
0	0	14	93	109	185	251	137	137	221	181	168	120	5	0	1	1	0
0	0	7	22	1	65	130	74	61	77	42	9	0	0	0	1	1	1
0	0	0	0	3	26	23	7	0	2	4	0	0	0	1	1	0	0
1	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0
1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0

Visual data: sequence of matrix



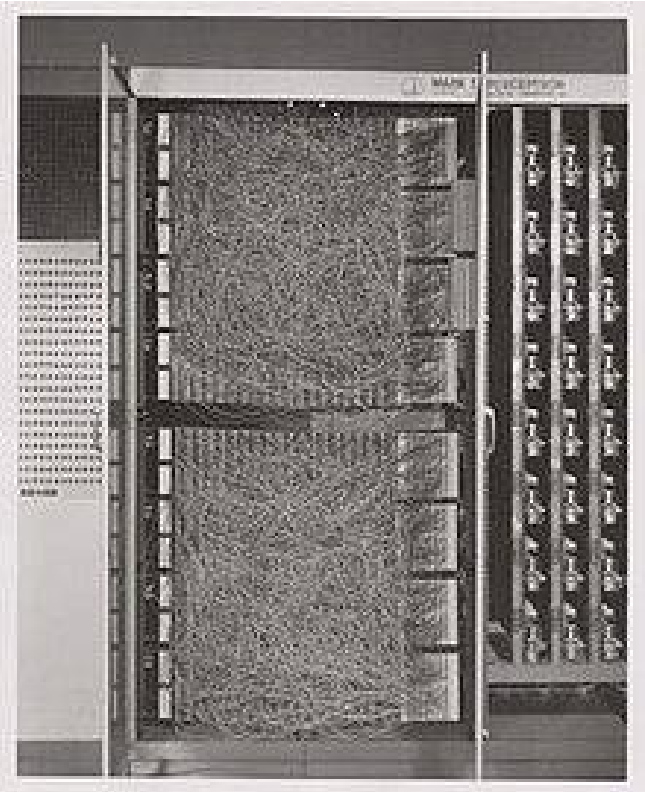
Visual data: more

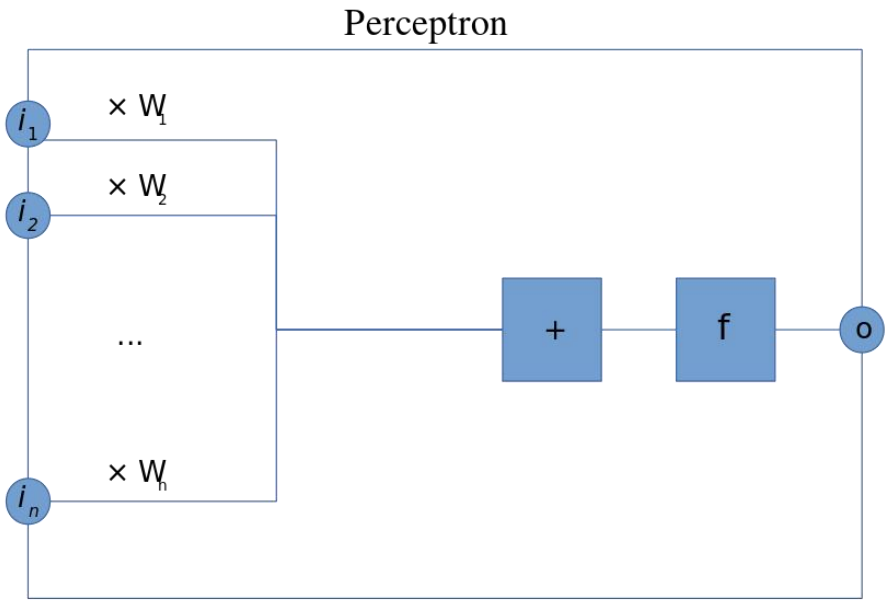
- Depth image (Time of flight, structured-light)
- Ultrasound imaging
- Magnetic resonance imaging



Early work

- Frank Rosenblatt (1957): using "Perceptron" machine to sort images into very simple categories like **triangle** and **square**

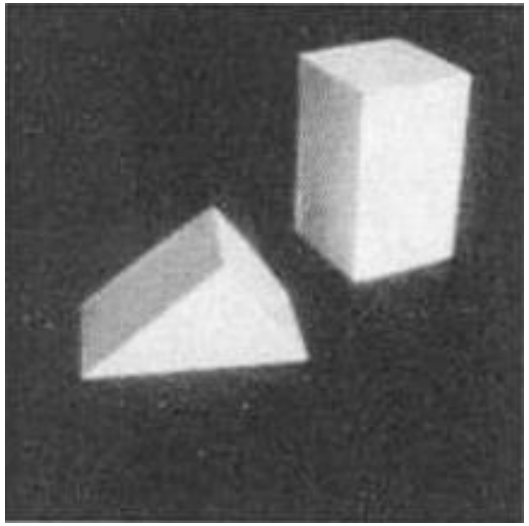




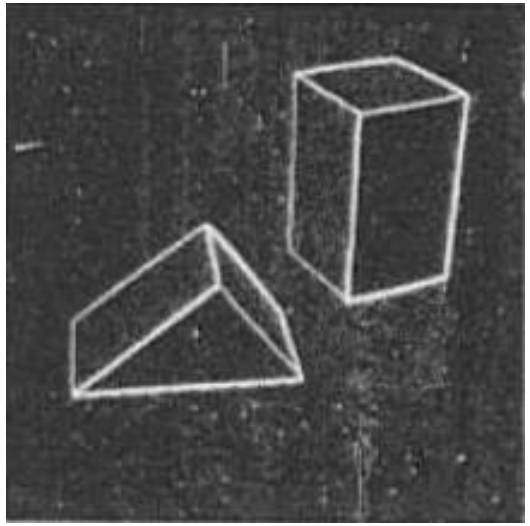
$$o = f\left(\sum_{k=1}^{n} i_k \cdot W_k\right)$$

Early work

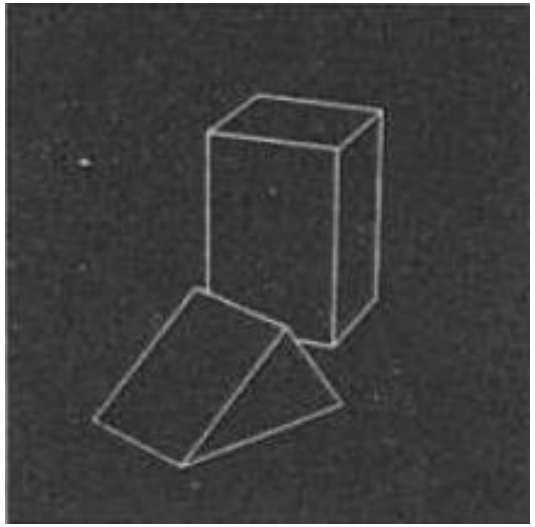
- Larry Roberts at MIT(1960): extracting 3D geometrical information from 2D perspective views of blocks
  - Machine perception of three-dimensional solids
  - Father of computer vision



Input image



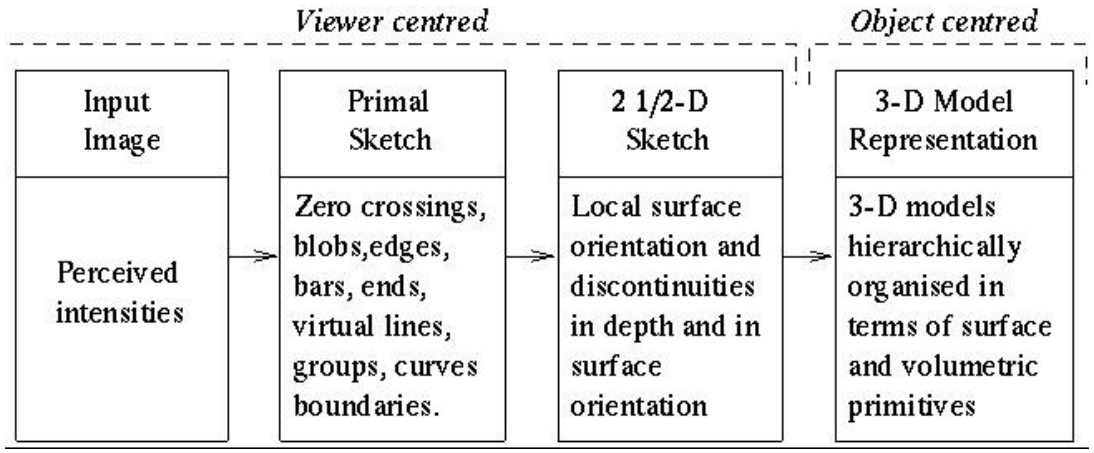
2x2 gradient operator

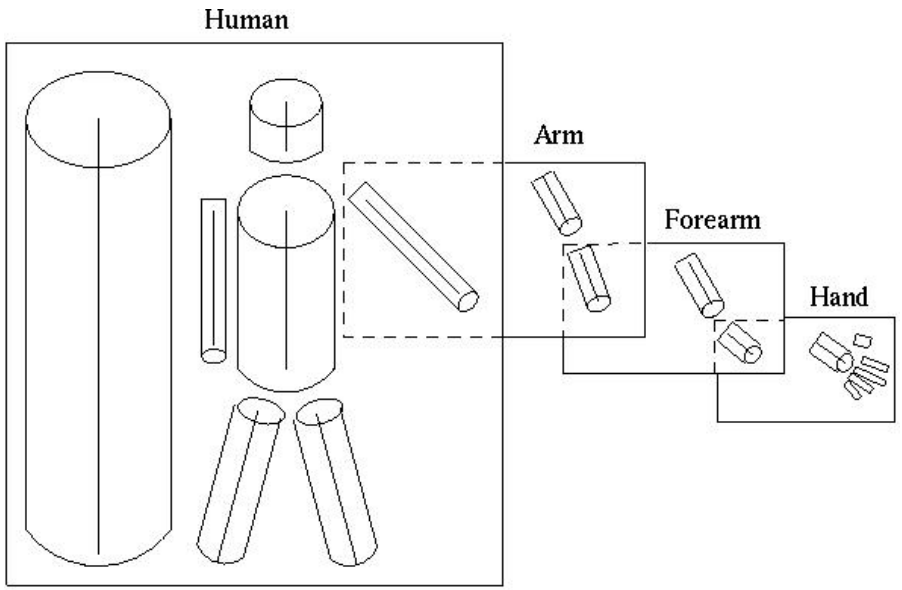


Computed 3D model

Early work

- Marvin Minsky at MIT (1966): connecting a camera to a computer
- David Marr at MIT (1978): proposing a bottom-up approach to scene understanding





Developing

Basics

Advanced

1980s

2012

https://papers.nips.cc › paper › 4824-imagenet-cla... PDF

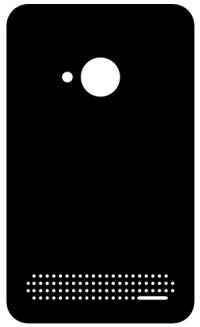
ImageNet Classification with Deep Convolutional Neural Nets

by A Krizhevsky · Cited by 86442 — **ImageNet Classification with Deep Convolutional Neural Networks**. Alex Krizhevsky. University of Toronto kriz@cs.utoronto.ca. Ilya Sutskever.

Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NeuIPS, 2012.

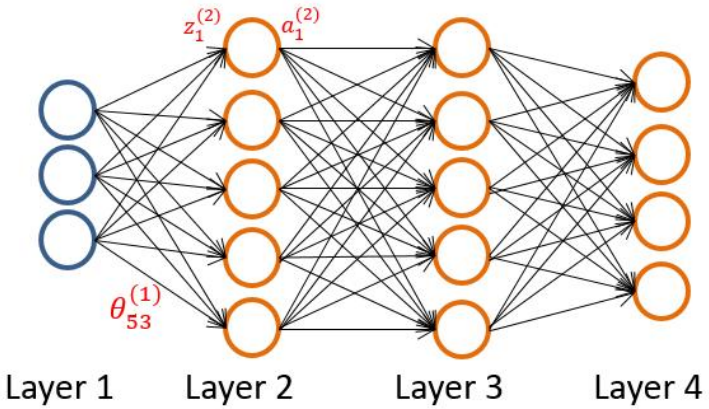
Why it bring about a renaissance?

- Mobile technology with built-in cameras (data)



- Computing power (devices)
- Mass storage

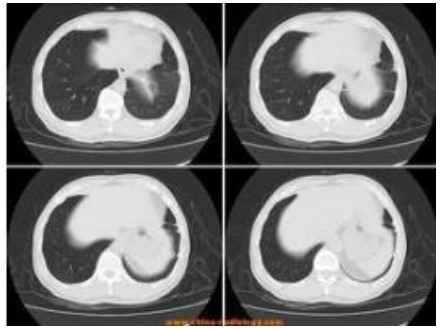
- New algorithms (models)
  - Support vector machine
  - Convolutional neural networks
  - Transformer



Why it matters

- Applications of visual computing

Safety



Health

Security

Facility

Fun

Industry

The state-of-the-arts

Areas of visual computing

- Image classification
- Object detection
- Image segmentation
- Pose estimation
- Visual language
  - Image captioning
  - VQA…
- Object tracking
- Object identification
- View synthesis (3D)

……

Image classification

- What is image classification?
  - Building a model (function) from hypothesis space
  - Image to label
  - Matrix to number

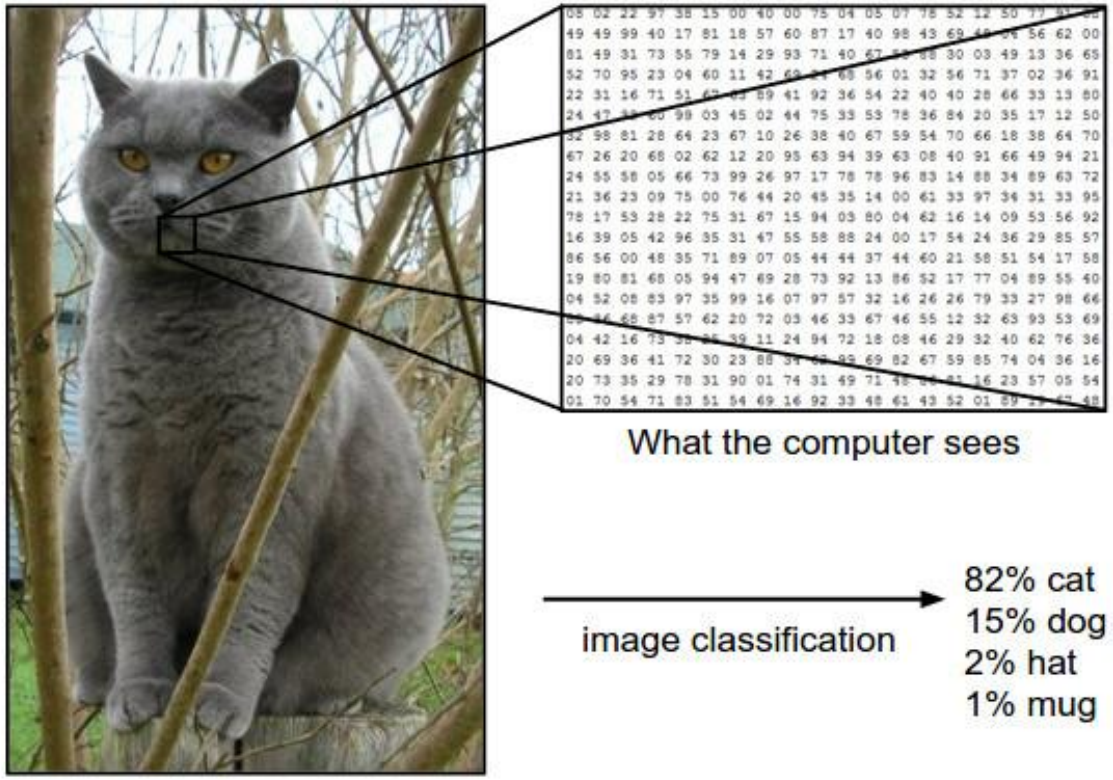


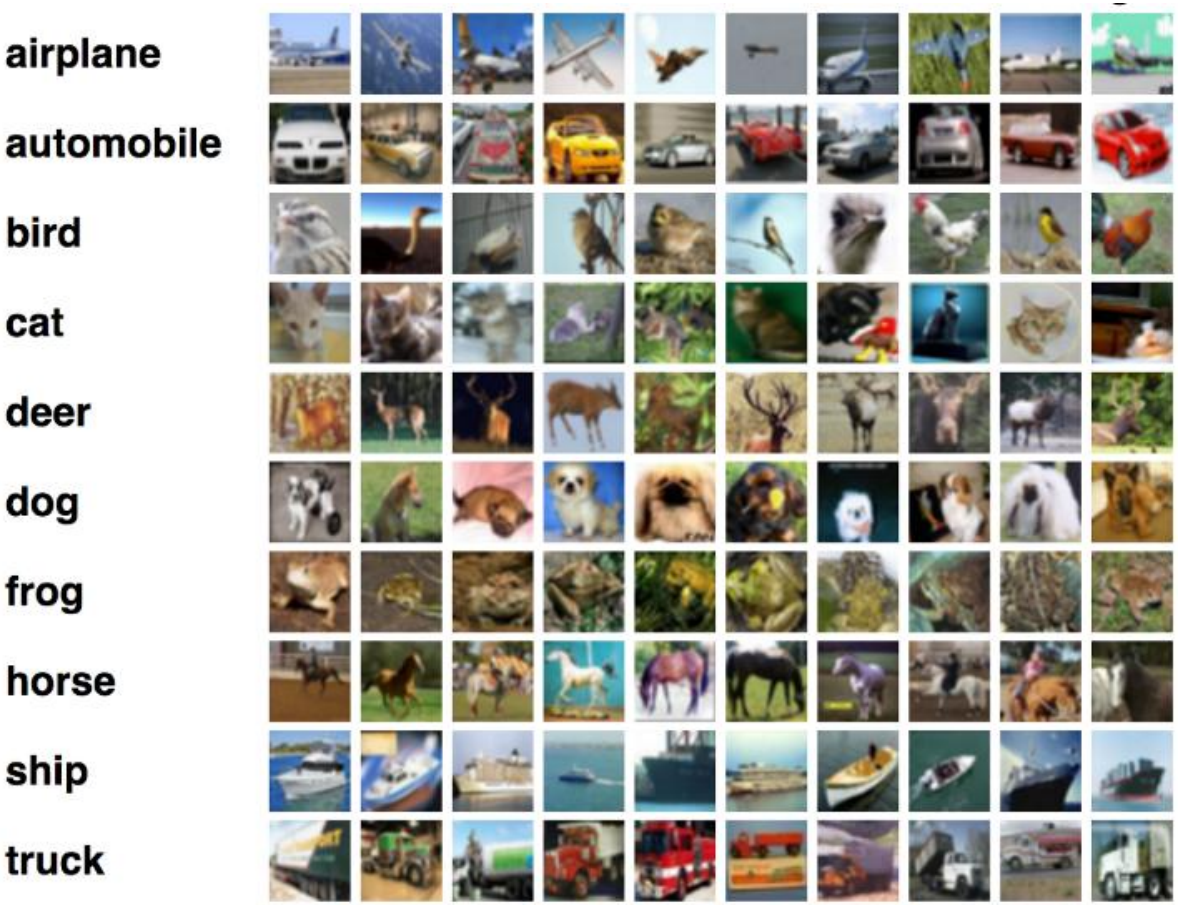
Image classification

- Task 1: MNIST database of handwritten digits
  - ICML 2013: 99.79%
  - It has a training set of 60,000 examples, and a test set of 10,000 examples

http://yann.lecun.com/exdb/mnist/

Image classification

- The CIFAR-10 dataset
  - arXiv 2015: 96.53%
  - 60000 32x32 colour images
  - 10 classes,
  - 6000 images per class.
  - 50000 training images
  - 10000 test images.



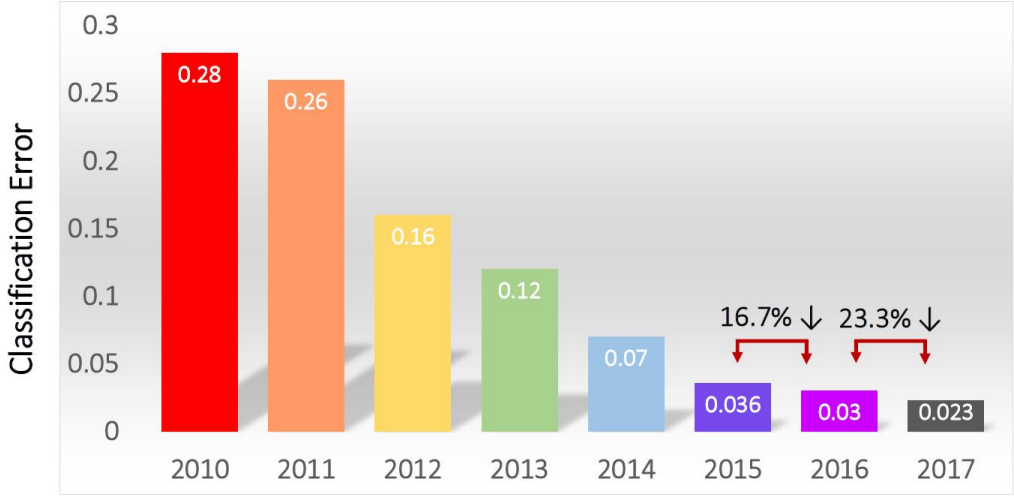
https://www.cs.toronto.edu/~kriz/cifar.html

Image classification

- IMAGENET: Large Scale Visual Recognition Challenge
  - 1000 classes: 1M train images and 100K test images

http://www.image-net.org/

Classification Results (CLS)



ImageNet: A large-scale hierarchical image database - IEEE ...

by J Deng · 2009 · Cited by 31462 — We introduce here a new **database** called "**ImageNet**", a **large-scale** ontology of **images** built upon the backbone of the WordNet structure. **ImageNet**...

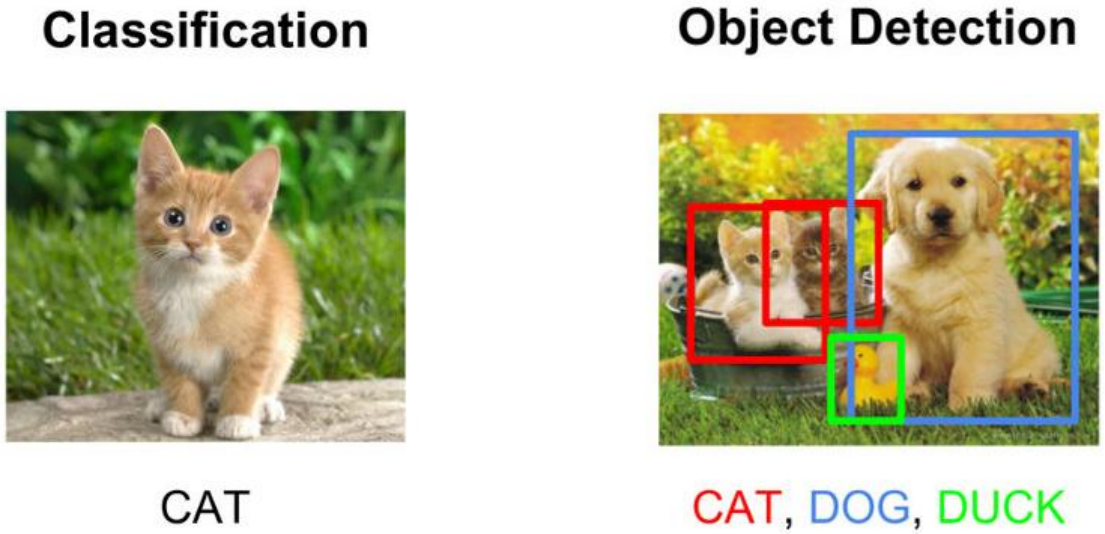
Date Added to IEEE Xplore: 18 August 2009 **DOI:** 10.1109/CVPR.2009.5206848

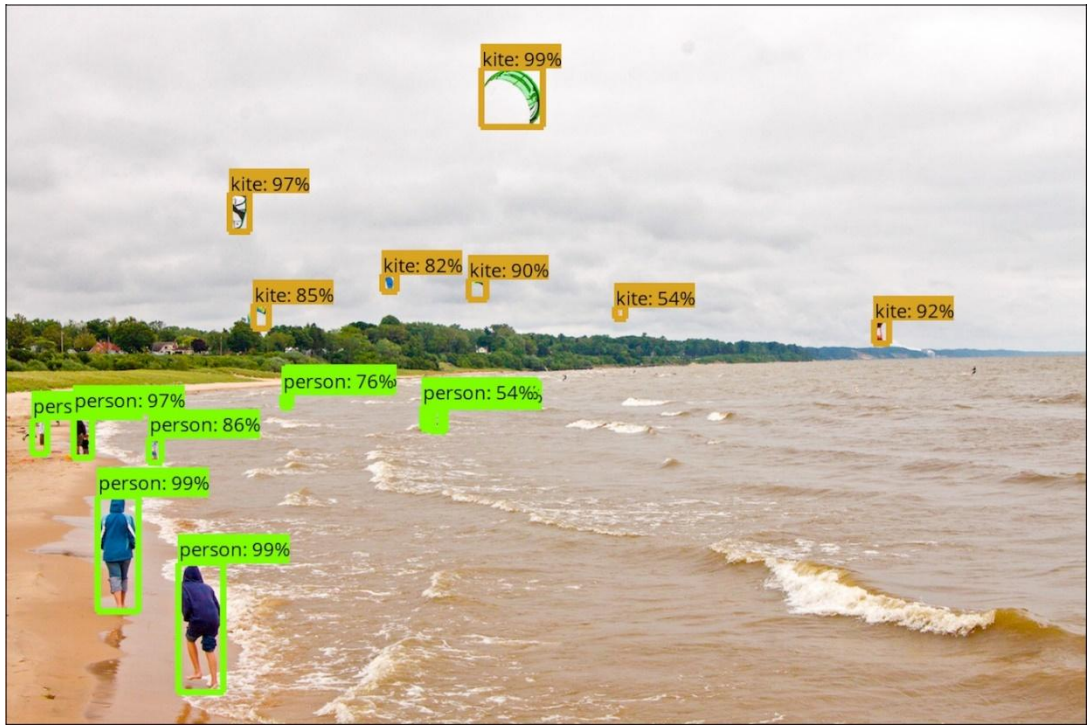
Date of Conference: 20-25 June 2009 **INSPEC Accession Number:** 10836047

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, **ImageNet: A Large-Scale Hierarchical Image Database.** *IEEE Computer Vision and Pattern Recognition (CVPR), 2009.*

Object detection

- What is object detection?
  - Input: image
  - Output: locations of objects





Object detection

- The frameworks

Faster R-CNN

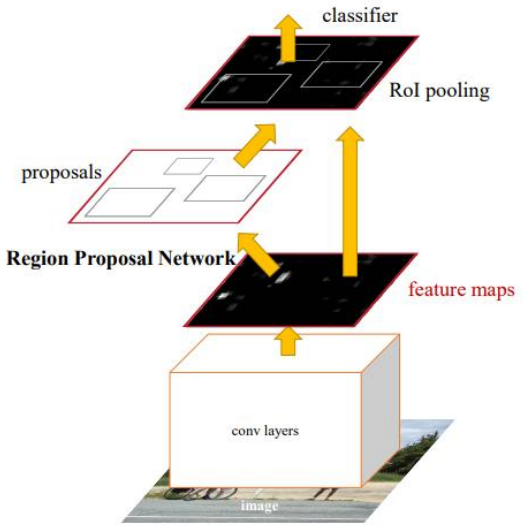
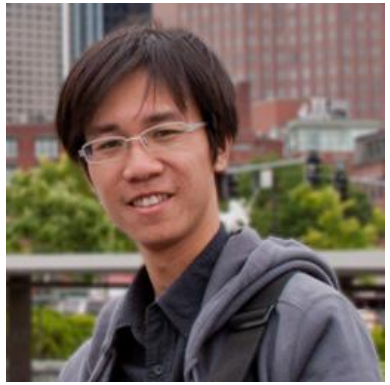


Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.



http://kaiminghe.com/

Cited by

	All	Since 2015
Citations	145231	142905
h-index	53	53
i10-index	61	61

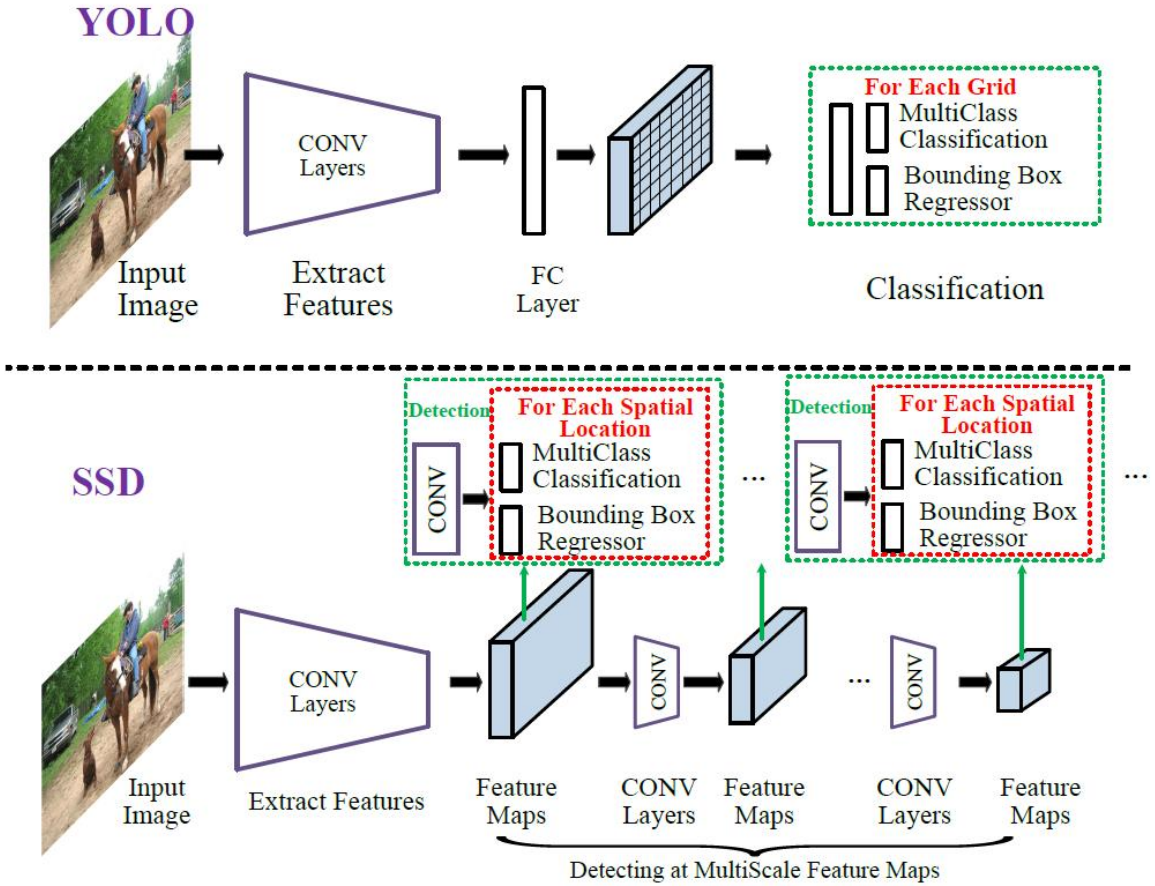
https://arxiv.org › cs

Faster R-CNN: Towards Real-Time Object Detection with ...

by S Ren · 2015 · Cited by 47205 — An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end...

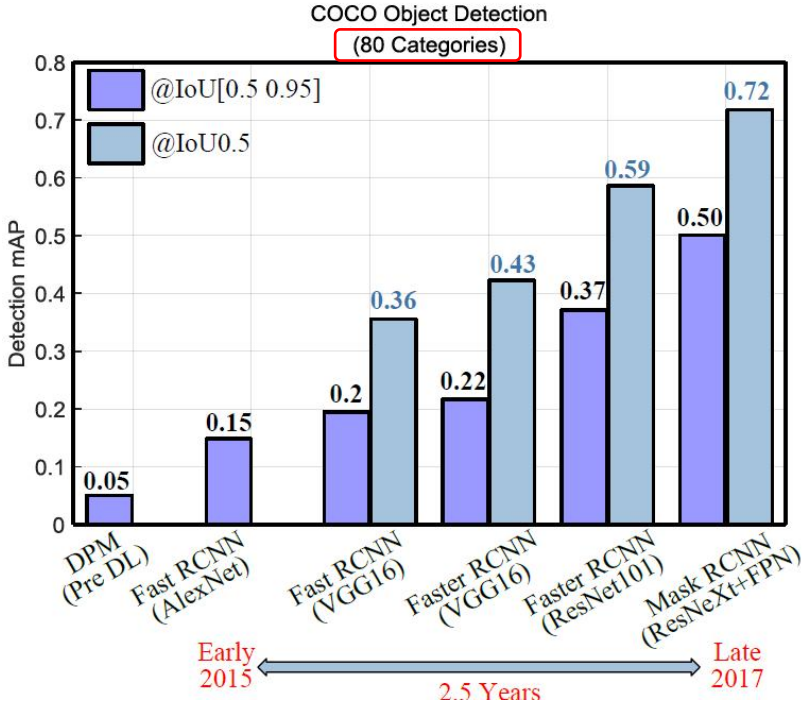
Cite as: arXiv:1506.01497

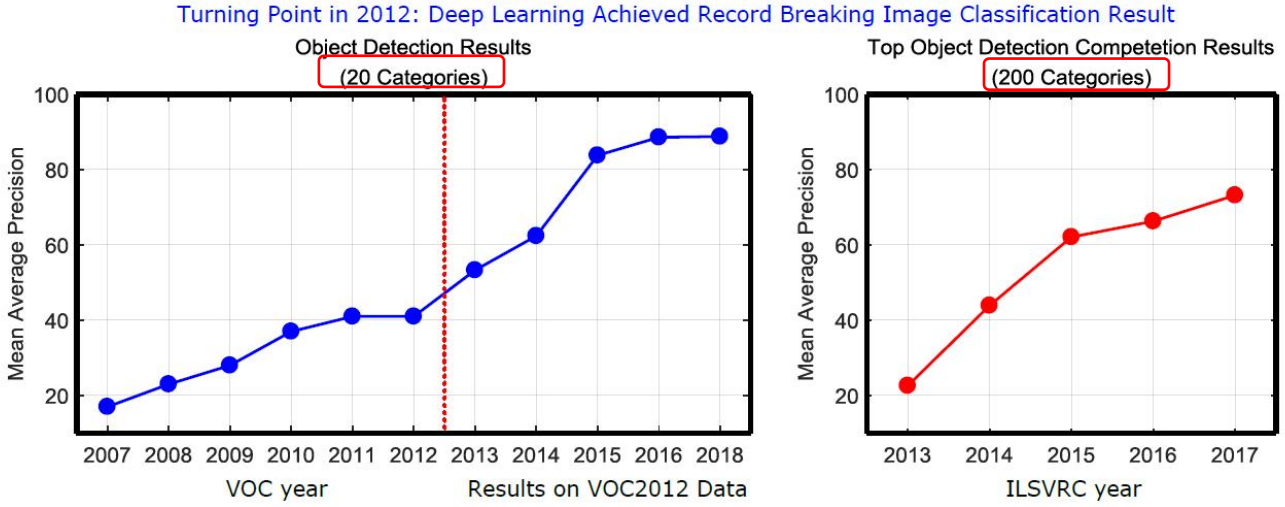
You've visited this page 2 times. Last visit: 22/12/20

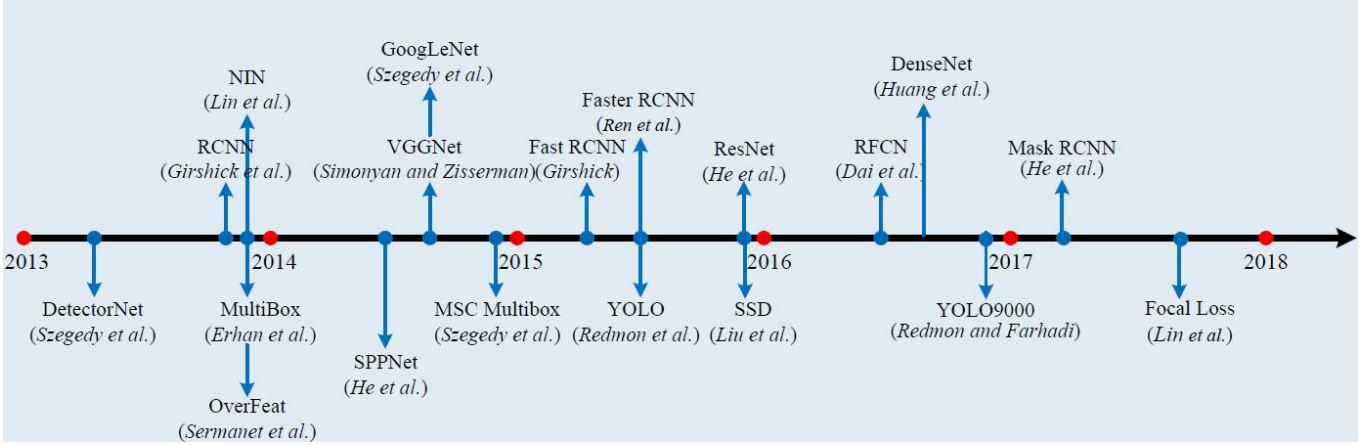


Object detection

- Milestones







https://www.youtube.com/watch?v=9DQGuD1sHBI

Visual Grounding

- What is visual grounding?
  - Input: image + language
  - Output: locations of objects

Query from GuessWhat?!:	Query from ReferCOCO:
is it a person? Yes is the person a man? No is it the left woman? No does she wear glasses? Yes	woman in black
	Query from ReferCOCO+: woman wearing glasses with smile
	Query from ReferCOCOg: a woman in black playing a game with her friends

Figure 1. In VG, the query can be a dialogue, a sentence or a phrase, and the target of the query is an instance in the image.

Image segmentation

- What is segmentation?
  - Input: image
  - Output: regions, structures

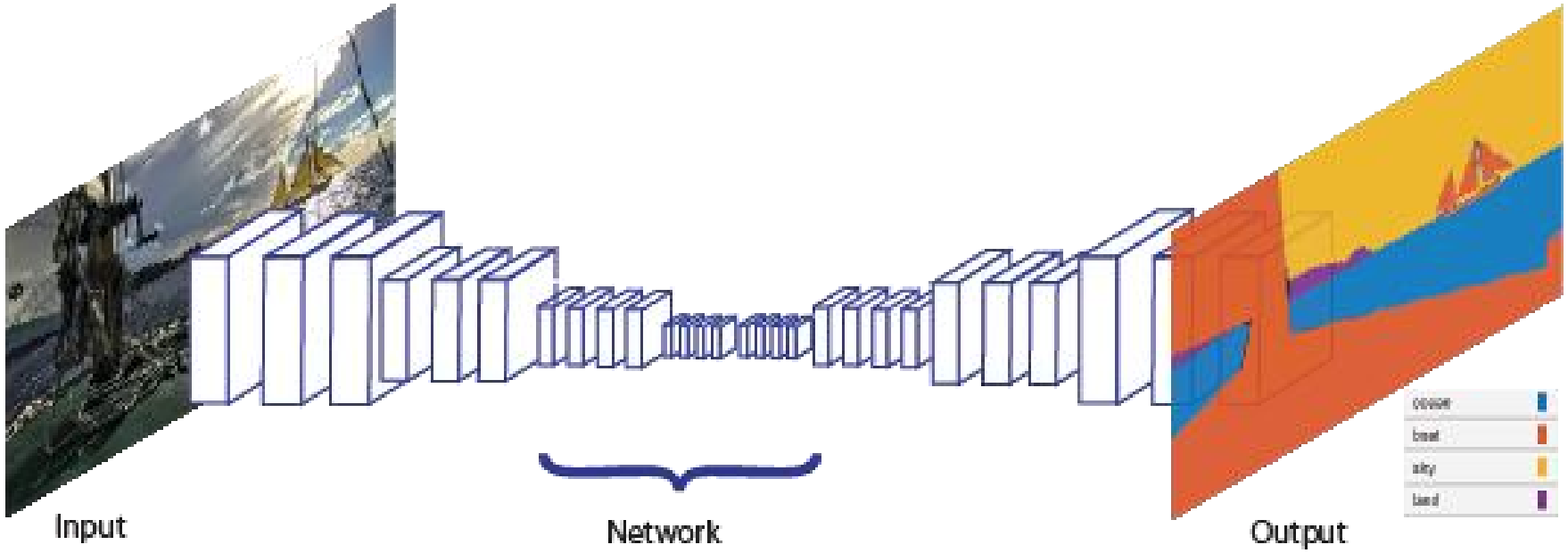


Image segmentation

- What is **semantic** segmentation?
  - Idea: recognizing, understanding what's in the image.
  - "Two men riding on a bike in front of a building on the road. And there is a car"

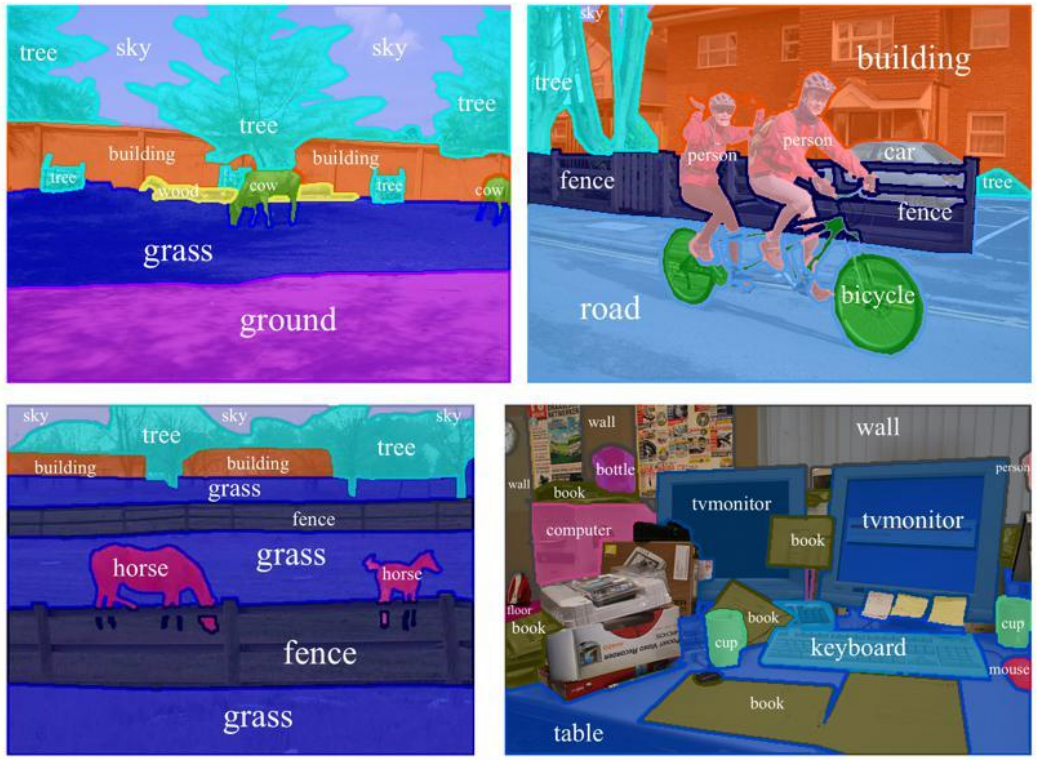


Image segmentation

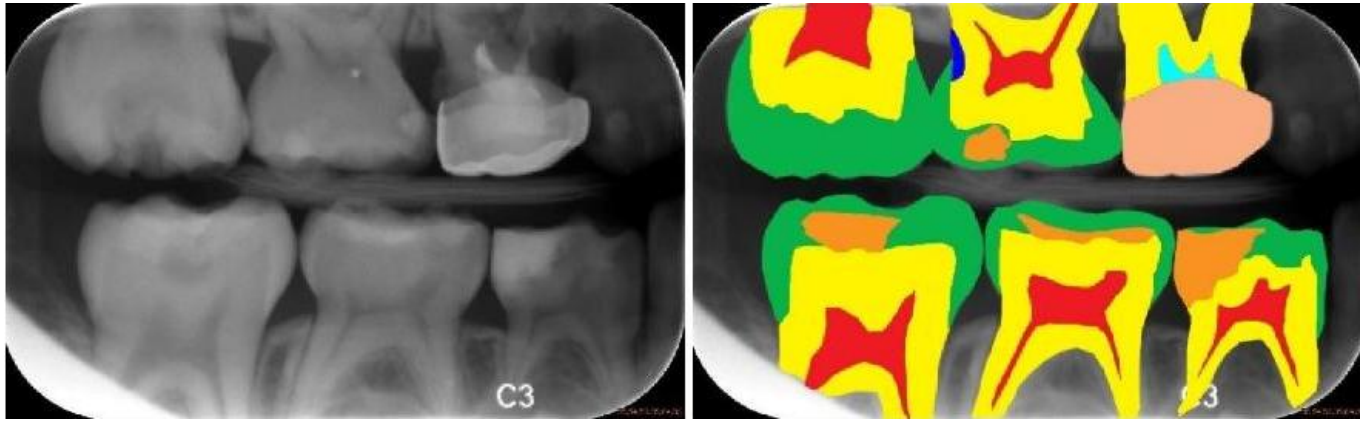
- Why semantic segmentation?
  - Robot vision and understanding
  - Autonomous driving
  - Medical purposes
- Mask R-CNN

https://arxiv.org › cs

[1703.06870] Mask R-CNN - arXiv

by K He · 2017 · Cited by 20599 — Abstract: We present a conceptually simple, flexible, and general framework for object instance segmentation.

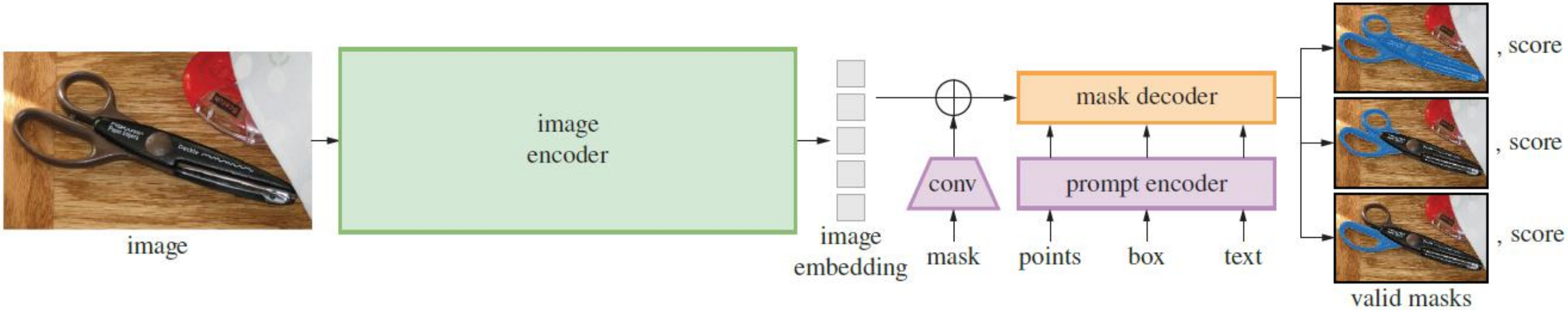
Cite as: arXiv:1703.06870



https://www.youtube.com/watch?v=OOT3UIXZztE

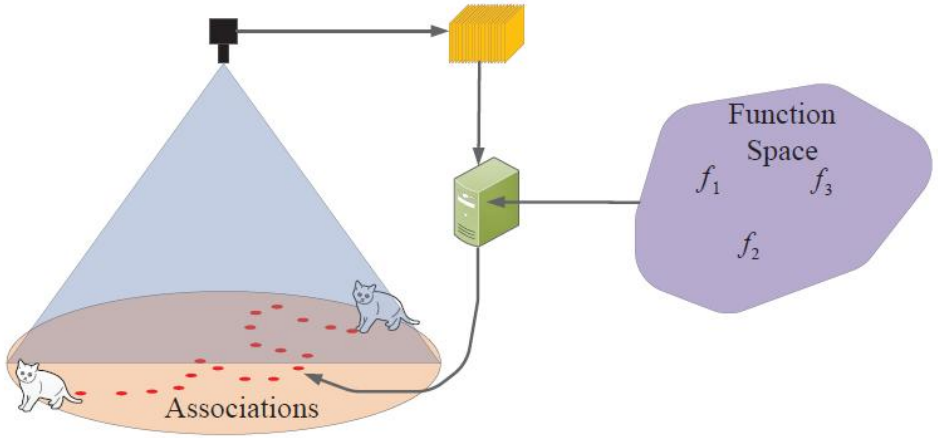
Segment Anything

- Input: image + prompt
- Output: masks



Object tracking

- What is object tracking?
  - Input: video
  - Output: trajectory



Referring Video Segmentation

- Input: language + video
- Opuput: a set of masks

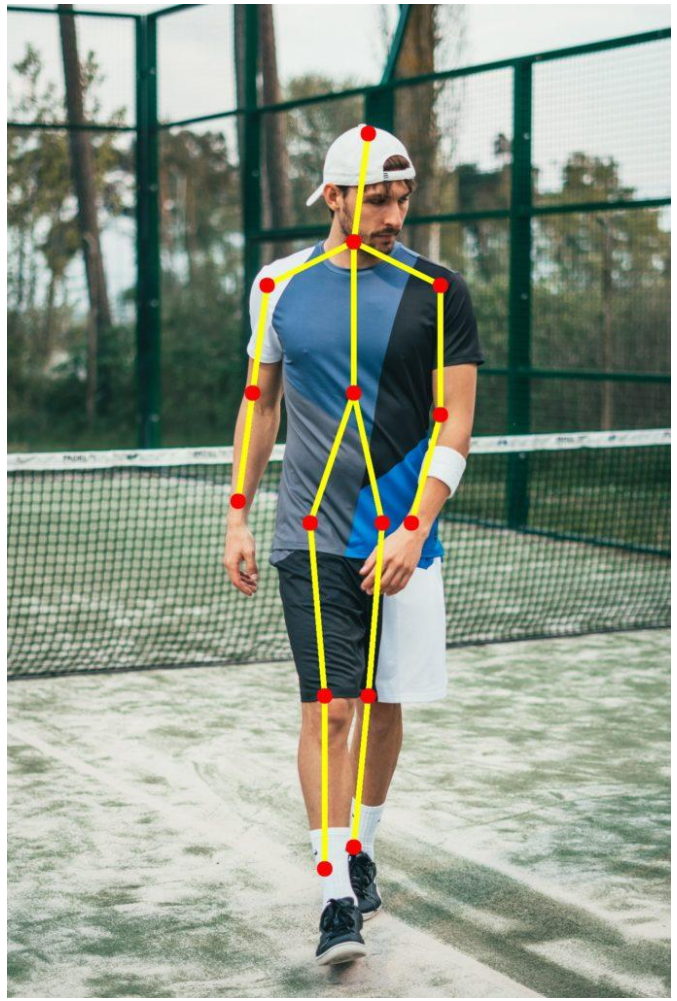
A man in a white T-shirt is surfing

Object tracking

- Human-computer interaction
- Human-mobile interaction

Pose estimation

- What is pose estimation?
  - Keypoint Detection
  - Input: image
  - Output: configuration



https://lmb.informatik.uni-freiburg.de/projects/hand3d/

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

SOTA-Visual and Language

Image captioning

- What is image captioning?
  - It is the process of generating textual description of an image
  - It uses both Natural Language Processing and Computer Vision to generate the captions

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Image captioning

- Why caption generation?
  - Generating summaries for YouTube videos
  - Captioning unlabeled images
  - Semantic search

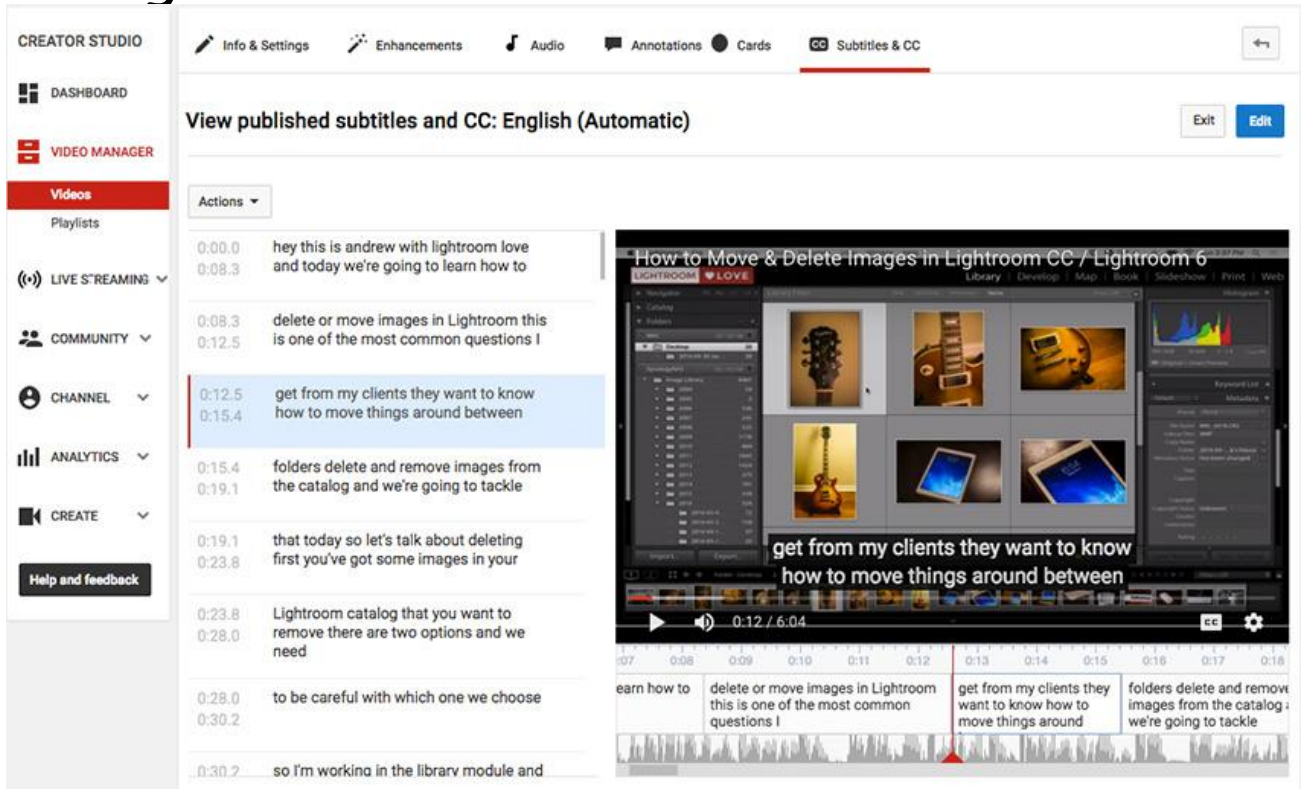
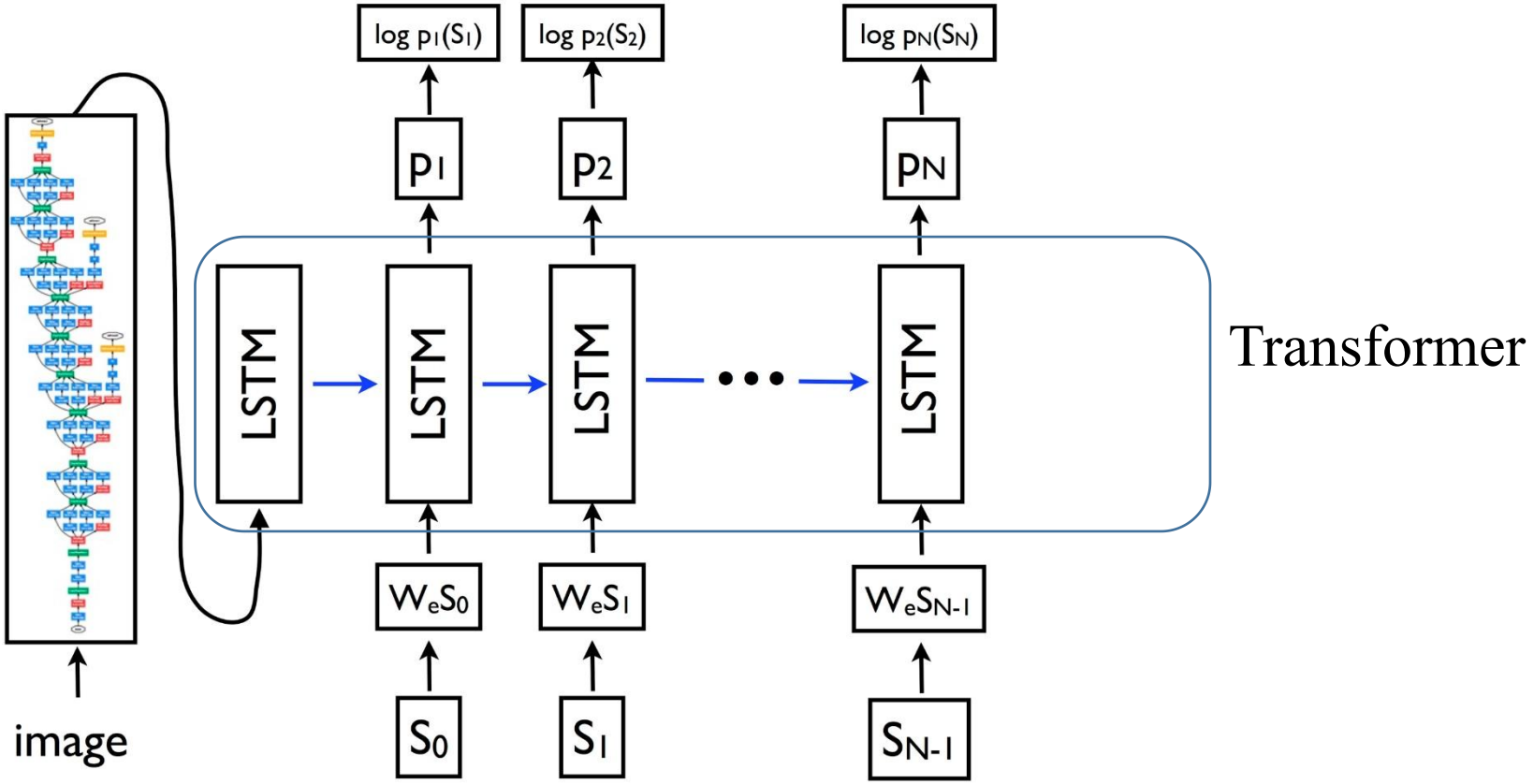


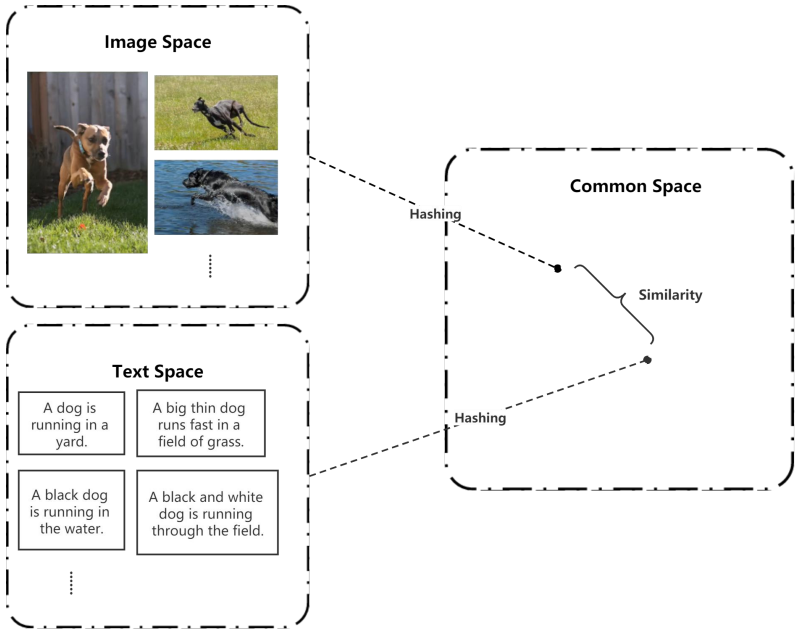
Image captioning

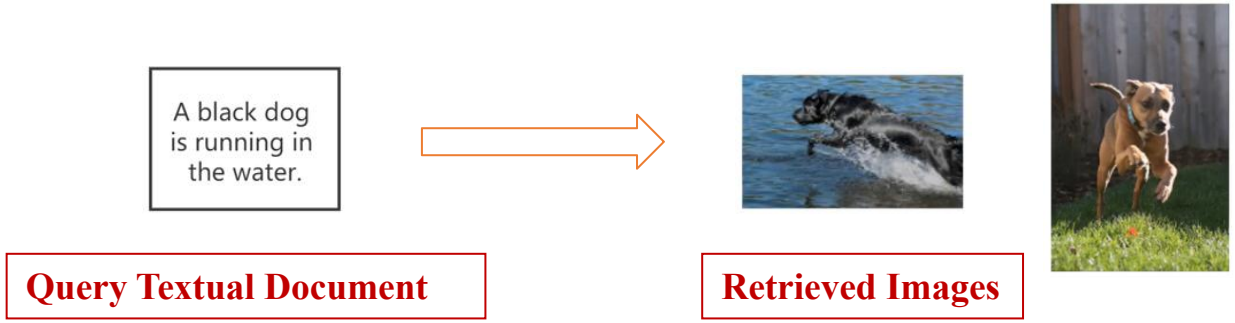
- Framework of image captioning



Cross-modal retrieval

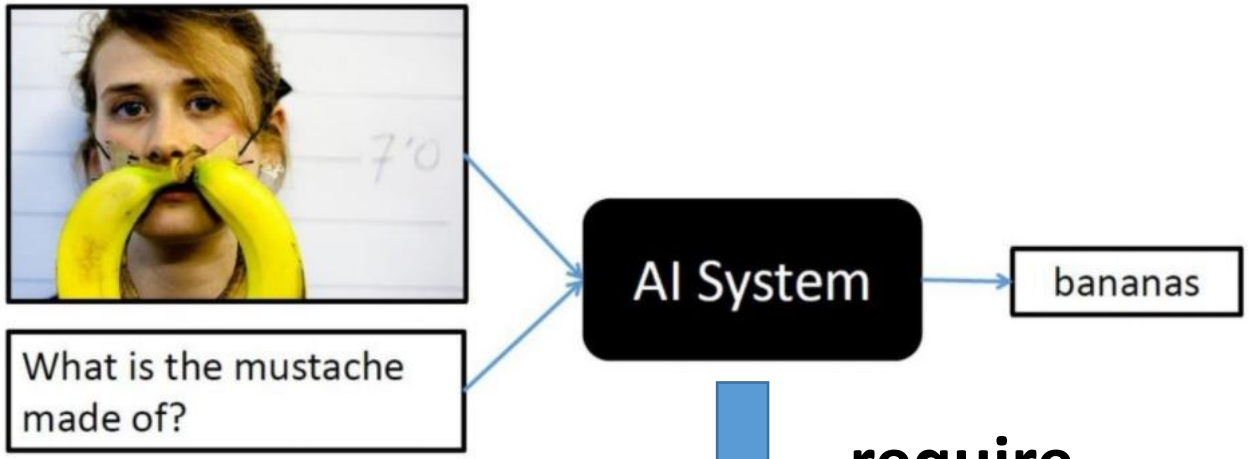
- Cross-modal Retrieval
  - ➢ Support similarity search for multi-modal data, e.g., the retrieval of images in response to a query textual document or vice versa.





Visual Question Answering (VQA)

- Given an image and a question (text) about the image
  - Aim to provide an accurate natural language answer



require

recognition | detection | classification | Commonsense reasoning | Relationship mining

Generative Images

Image generation (CycleGAN)

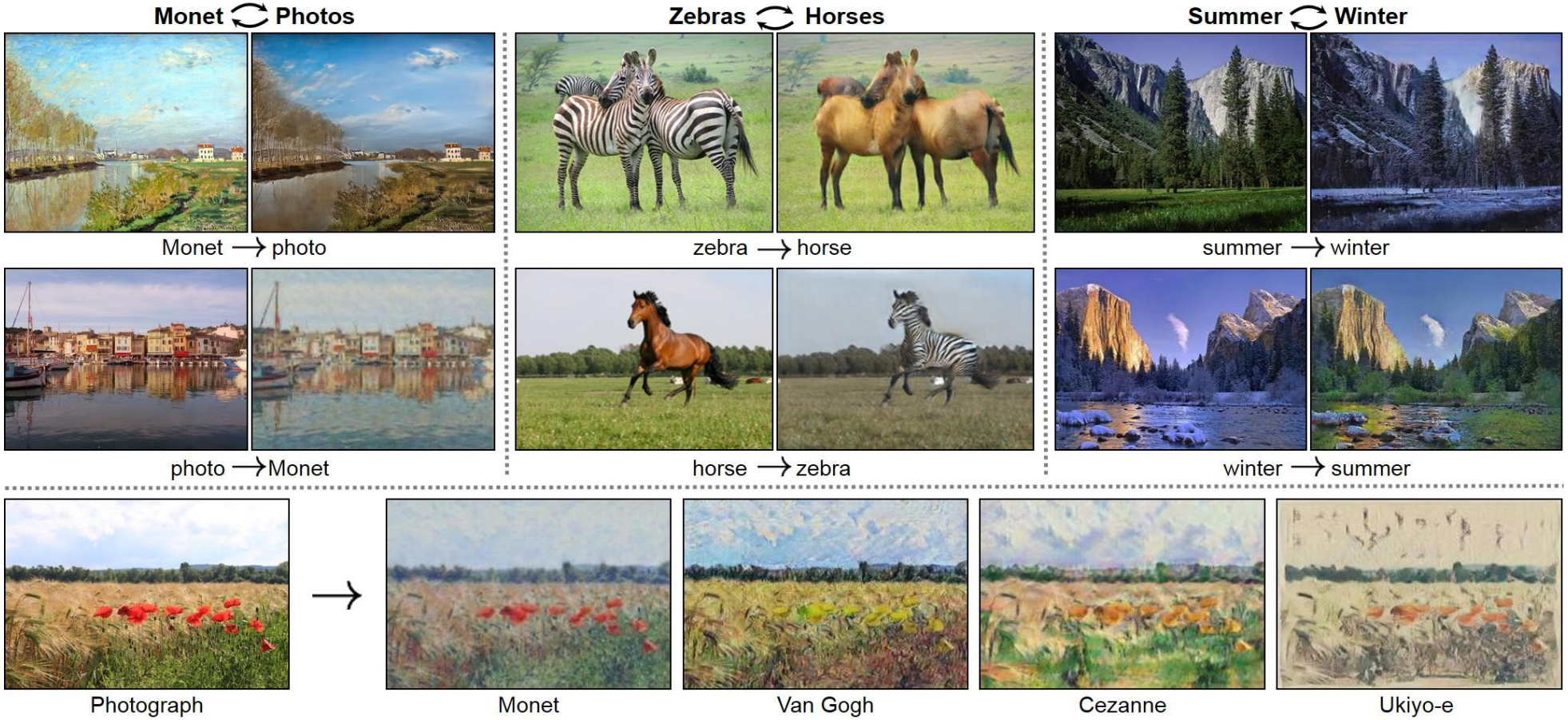
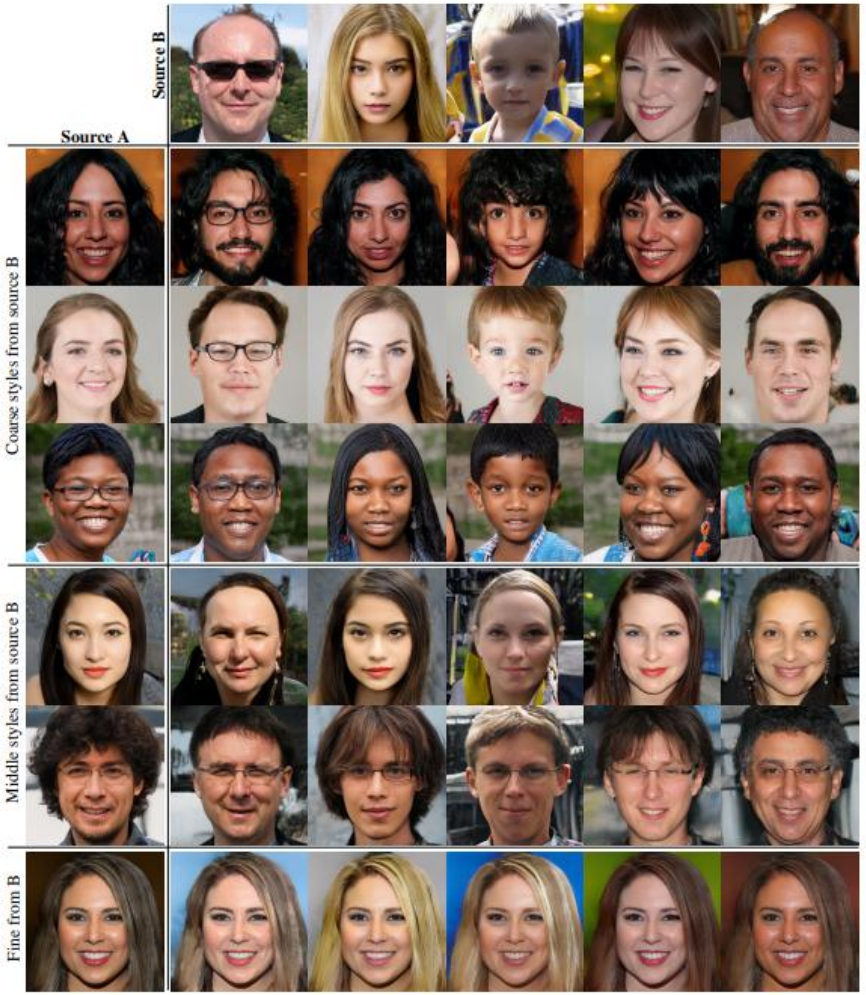


Image generation



朱茵→杨幂

A Style-Based Generator Architecture for Generative Adversarial Networks

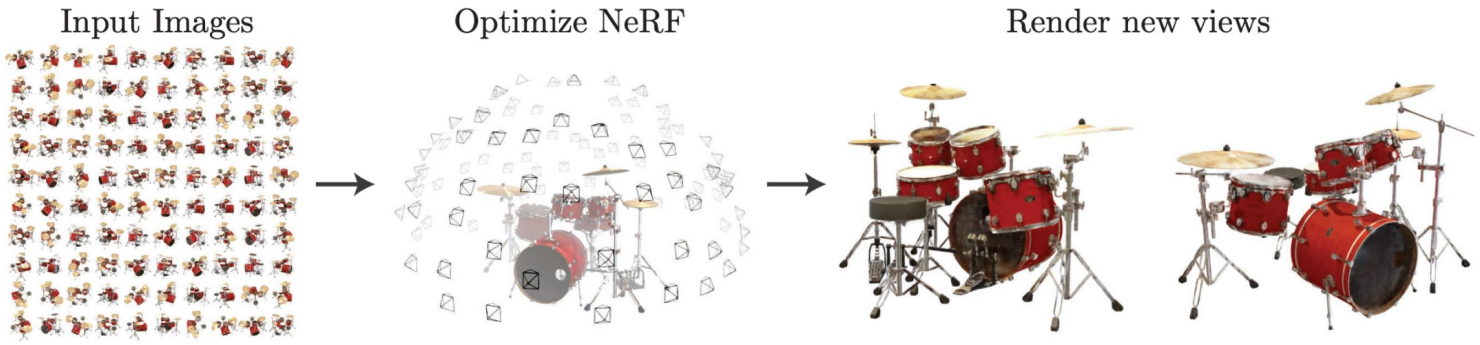
Tero Karras
NVIDIA
tkarras@nvidia.com

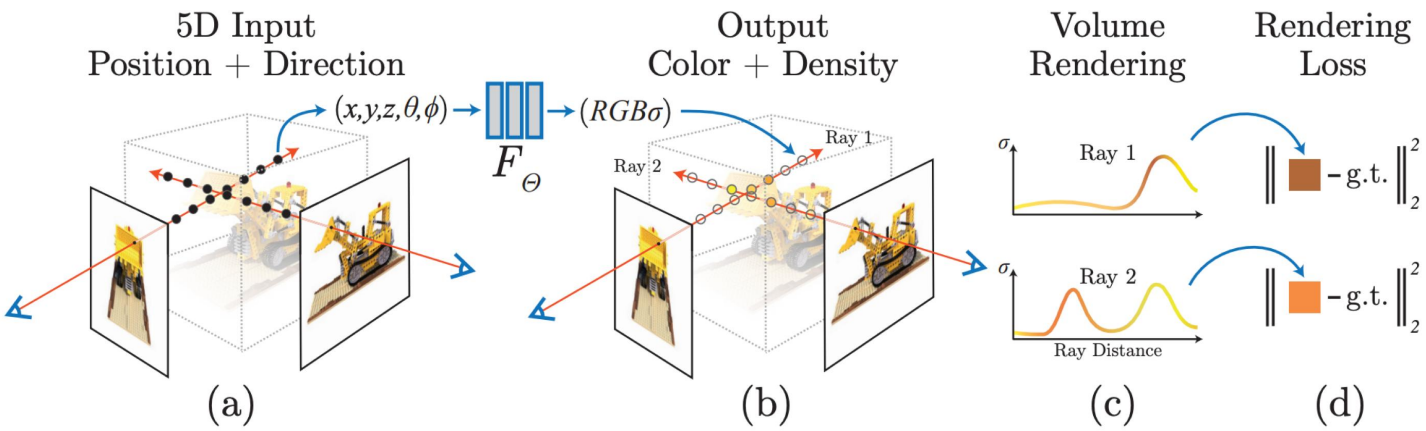
Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com

Neural Radiance Field (NeRF)

- Mildenhall et al. ECCV 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis





Neural Radiance Field (NeRF)

- Videos in another slide

Video generation

- input: language
- output: video

Applications

Code: https://github.com/gaomingqi/Track-Anything

Demo: https://huggingface.co/spaces/watchtowerss/Track-Anything

Report: https://arxiv.org/abs/2304.11968

★ **5.7k** stars

Demo: Video Annotation

Demo: Video Object Inpainting

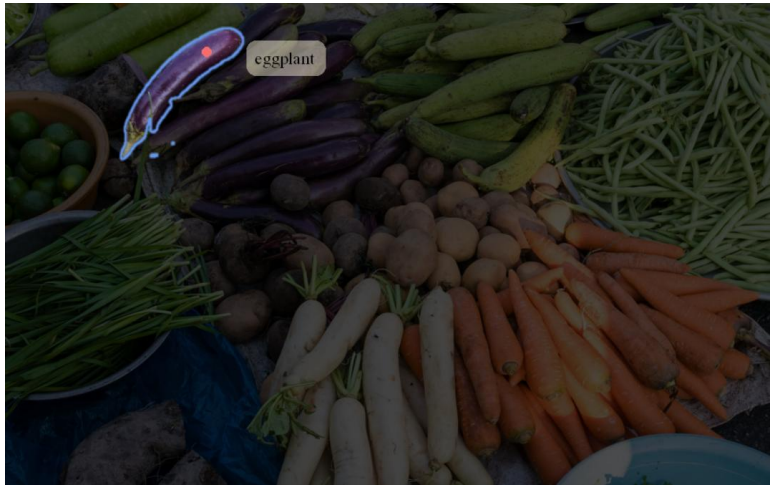
Demo: Synthetic Object Tracking

Caption Anything

https://github.com/ttengwang/Caption-Anything

★ **1.5k** stars

Demo: Visual Control



Demo: Language Control

English: a horse with a white mane and tail

French: Un cheval avec une crinière et une queue blanches

Chinese: 一匹白鬃白尾的马

Spanish: Un caballo con crin y cola blancas

Positive+Factual: The majestic horse boasts a stunning white mane and tail

Positive+Imagination: The majestic equine boasts a stunning white mane and tail

Negative+Factual: The horse's white mane and tail did not complement its appearance

Negative+Imagination: The horse galloping uncontrollably

Demo: Object-Centric Chatting

caption: A corgi dog.

What are the characteristics of the corgi?

The VQA tool shows that the corgi in the selected image has brown and white colors.

Can you see its tongue?

Yes, according to the VQA tool, the corgi's tongue is visible in the selected image

What's the color of the eye of the corgi?

According to the VQA tool, the color of the corgi's eye in the selected image is brown.

caption: A tiny bird with iridescent feathers hovers mid-air - a hummingbird.

What's the color of this hummingbird?

The hummingbird in the selected region is brown and white.

What's color of the eye of this hummingbird?

Based on the VQA tool, the color of the hummingbird's eye in the selected region is black.

What is the hummingbird doing?

Based on the VQA tool, the hummingbird in the selected region is flying.

caption: A stool adorned with yellow butterflies.

What's the shape of the stool?

The shape of the stool is square.

How many feet of the stool?

According to the VQA tool, the stool has 4 feet.

What's the stool made of?

According to the VQA tool, the stool is made of wood.

Demo: Caption Everything

In this image, we see a panda bear and her cub sitting on the ground. The mother panda is larger in size and has black and white fur, while her cub is smaller and also has black and white fur. They are both sitting on the ground, with the mother panda slightly behind her cub. The ground appears to be covered in grass and there are no other visible objects in the image. It seems like they are in a natural environment, perhaps in a forest or a zoo.

In this image, a raccoon can be seen sitting on a rock. The raccoon appears to be of average size and has a distinctive black and white fur pattern. The rock is large and appears to be made of a rough, textured material. In the background, there are two instances of scene text, one labeled " Jannes" and the other labeled "Pockele", both positioned towards the right side of the image. Overall, the image seems to depict a peaceful outdoor scene with the raccoon as the main focus.

In the center of the image, there is a large bed with a black headboard and footboard. The bed appears to be queen-sized and is covered in white sheets and a gray comforter. The headboard is rectangular in shape and has a smooth, glossy texture. The footboard is also rectangular and has a similar texture to the headboard. The bed is positioned in the middle of a room, but no other objects or details are visible in the image.

In this image, we see a row of violins hanging on a wall. The violins are all similar in size and shape, with their curved bodies and long necks. They appear to be made of wood, with a smooth and polished texture. The violins are positioned close together, with their strings facing outward. The wall behind them is a neutral color, providing a simple backdrop for the instruments. Overall, this image depicts a collection of well-crafted violins, likely displayed in a music-related setting.

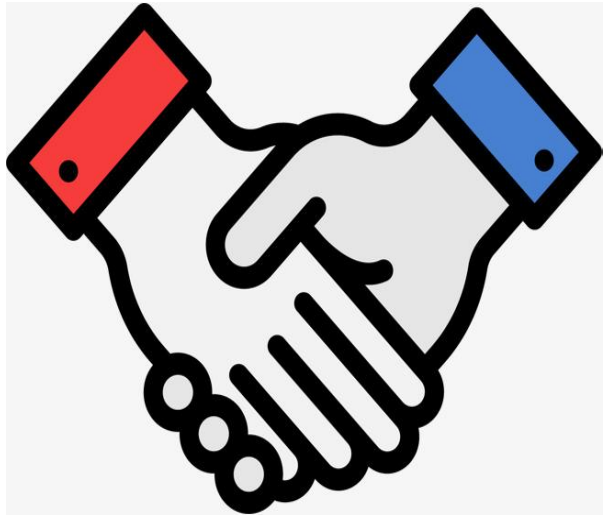
Conclusions

Conclusions

- Visual computing is very significant

- Visual computing is very interesting

- Visual computing is easy to study



Thanks

zhengf@sustc.edu.cn