香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)
信息枢纽
INFORMATION HUB
人工智能学域
ARTIFICIAL INTELLIGENCE

Introduction to Computer Vision

Course AIAA 4220

Week 2 - Lecture 3

Changhao Chen

Assistant Professor

HKUST (GZ)

Recap: What is Embodied AI?

- Embodied AI is a paradigm in AI where an agent (software or hardware) learns to perform tasks by sensing, interpreting, and acting within a physical or simulated environment.
- Unlike traditional AI models that process static data (like images or text), embodied agents have a "body" (virtual or real) that allows them to interact with and change their surroundings to achieve a goal.
- It‘s about moving from “seeing ” "thinking" to "doing."

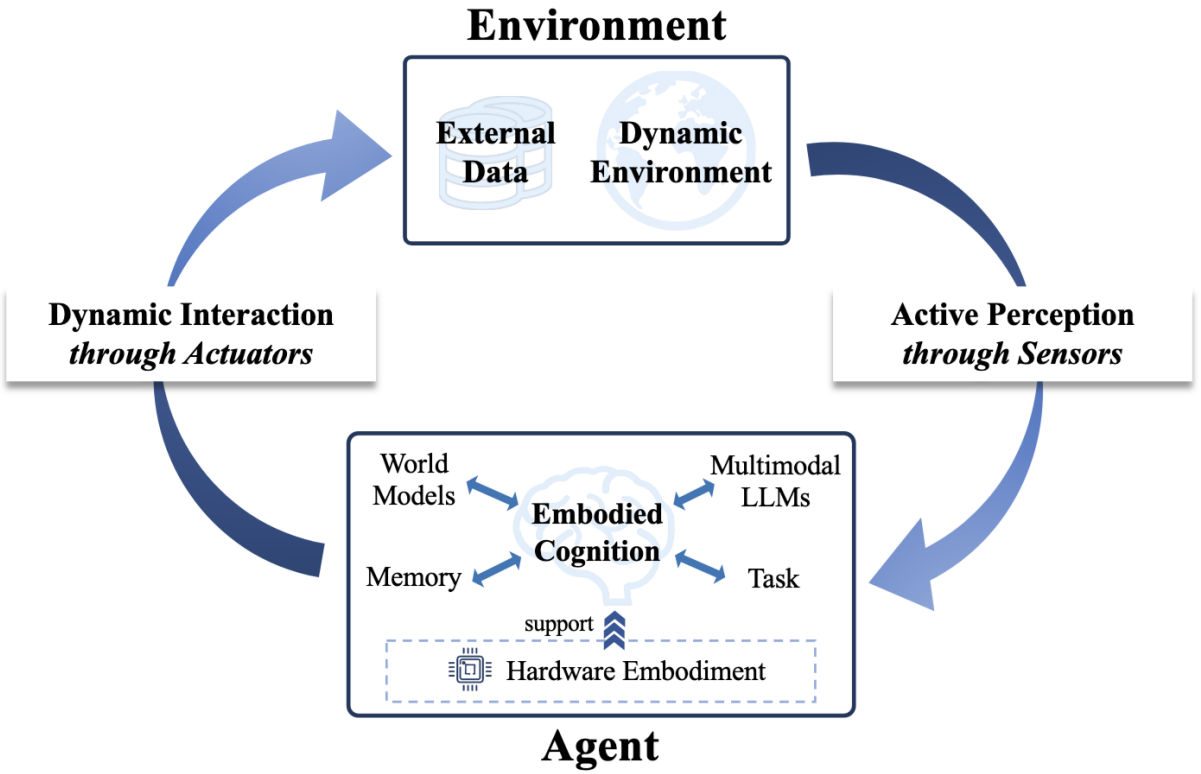
Recap: What is "Embodiment"?

Embodiment is the two-way relationship between an agent (an AI, a robot, an animal) and its environment, mediated through a physical form.

It's not just having a body. It's about how the body's characteristics (its shape, sensors, actuators, materials) shape:

- How it perceives the world.
- What actions it can take.
- How it learns and thinks.

The body is not just a vessel for the brain; it is a fundamental part of the intelligence system.



Recap: Cameras

Monocular Camera:

Single lens, low-cost

Rich information

Limitation: lacks depth perception

Stereo Camera:

Mimics human eyes → depth from disparity

Useful for near-field obstacle detection

More expensive & calibration-sensitive

Sensor Placement:

Front-facing cameras - lane keeping, forward vision

Surround-view (4–8 cameras) - 360° coverage

Image Analysis

Common tasks for image analysis

- Classification (Visual recognition)
- Object Detection
- Segmentation

Image Classification

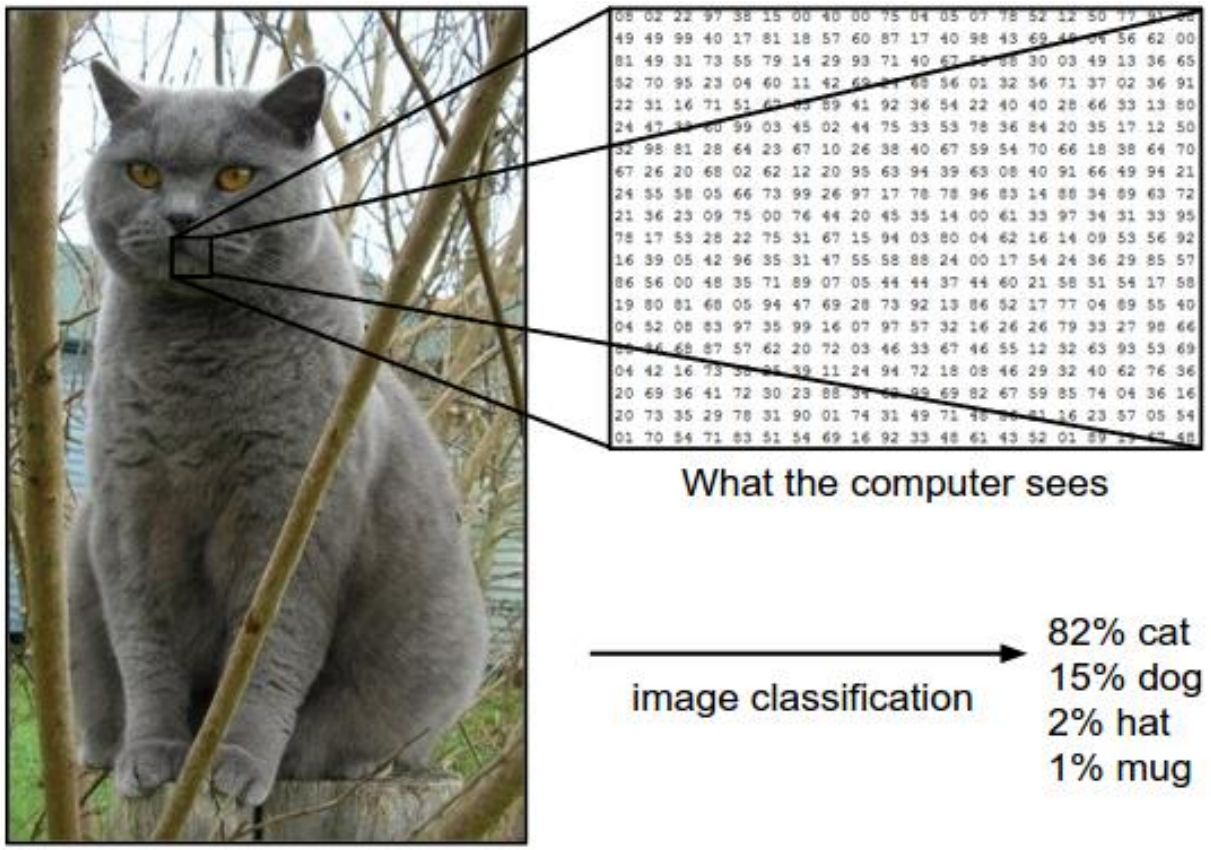


Image data

- Each pixel [0, 255]
- 3 channel - RGB

Given an image tensor $\mathbf{x}_i$ of shape [H, W, 3]

The classification model outputs a category label

$$\hat{y}_i \in \{1, \ldots, K\}$$

Source: Stanford CS231n

Image Classification

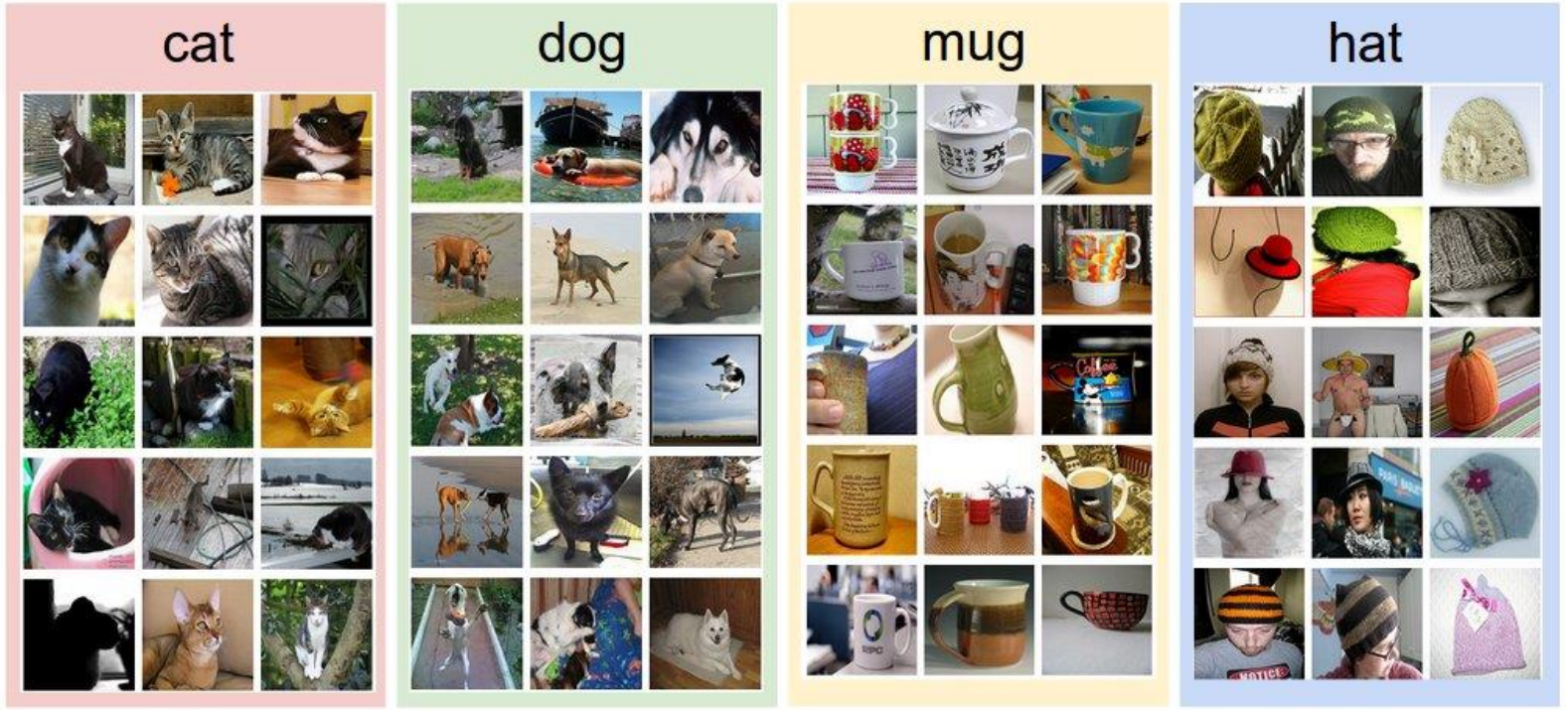


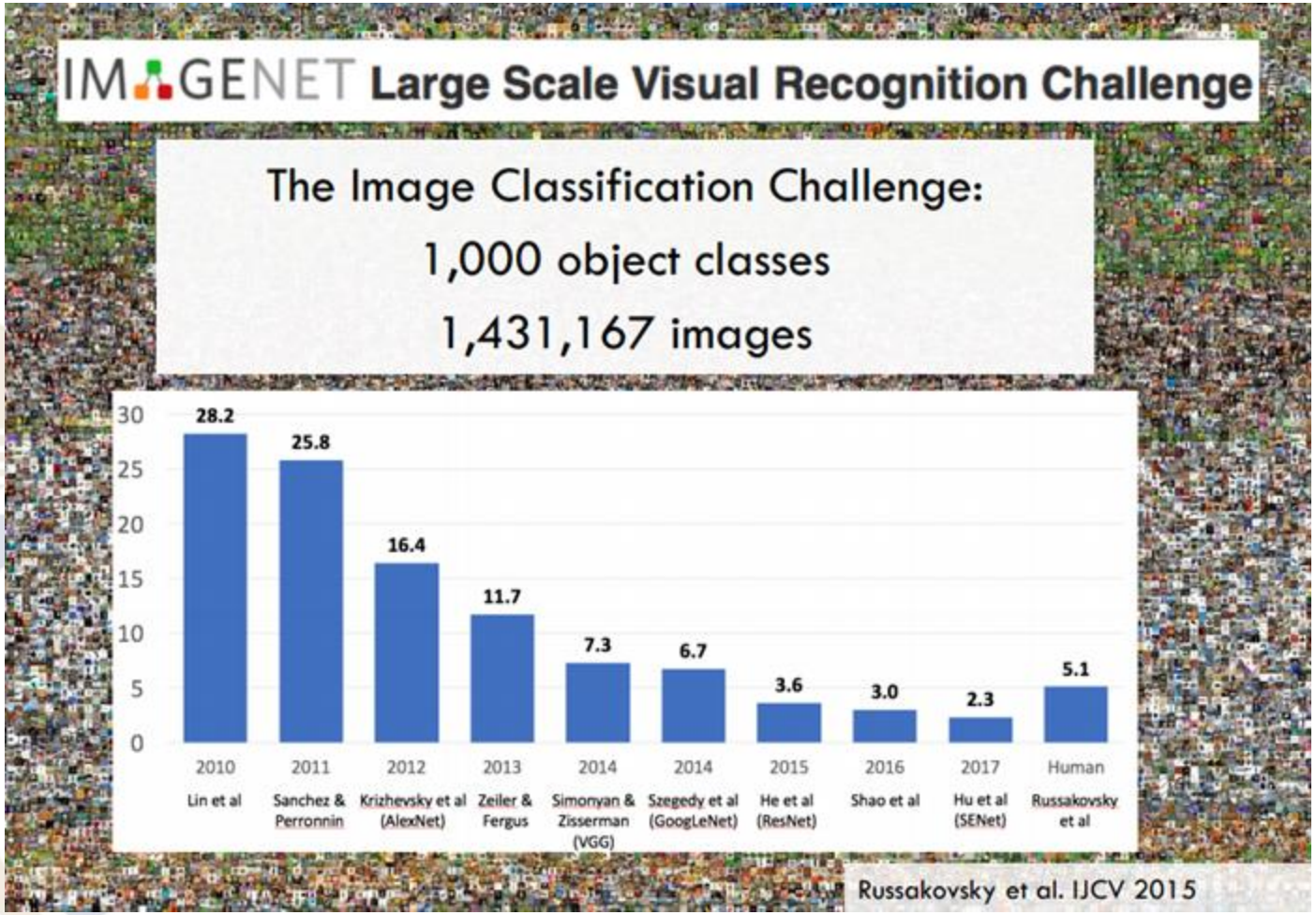
Image classification model:

$$\hat{y} = f(x; W)$$

W is the model weights, x is input image.

Source: Stanford CS231n

ImageNet Classification Challenge

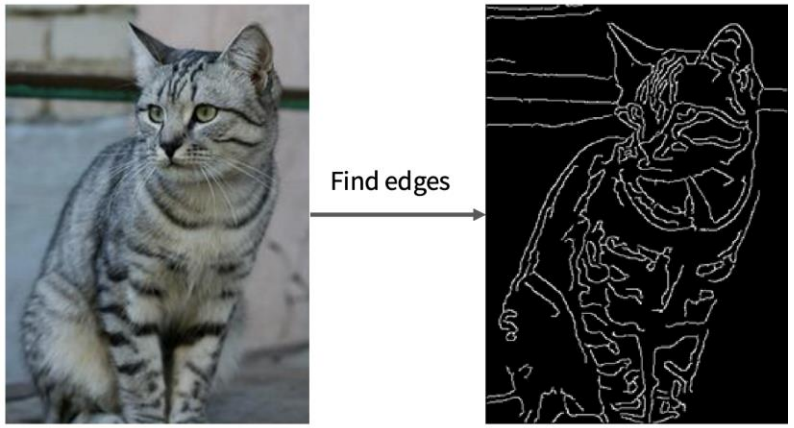


Challenge started in 2009

One of the most important image datasets in computer vision

ImageNet Classification Challenge

Local features (HOG, SIFT, etc.) -> Bag-of-Words -> SVM



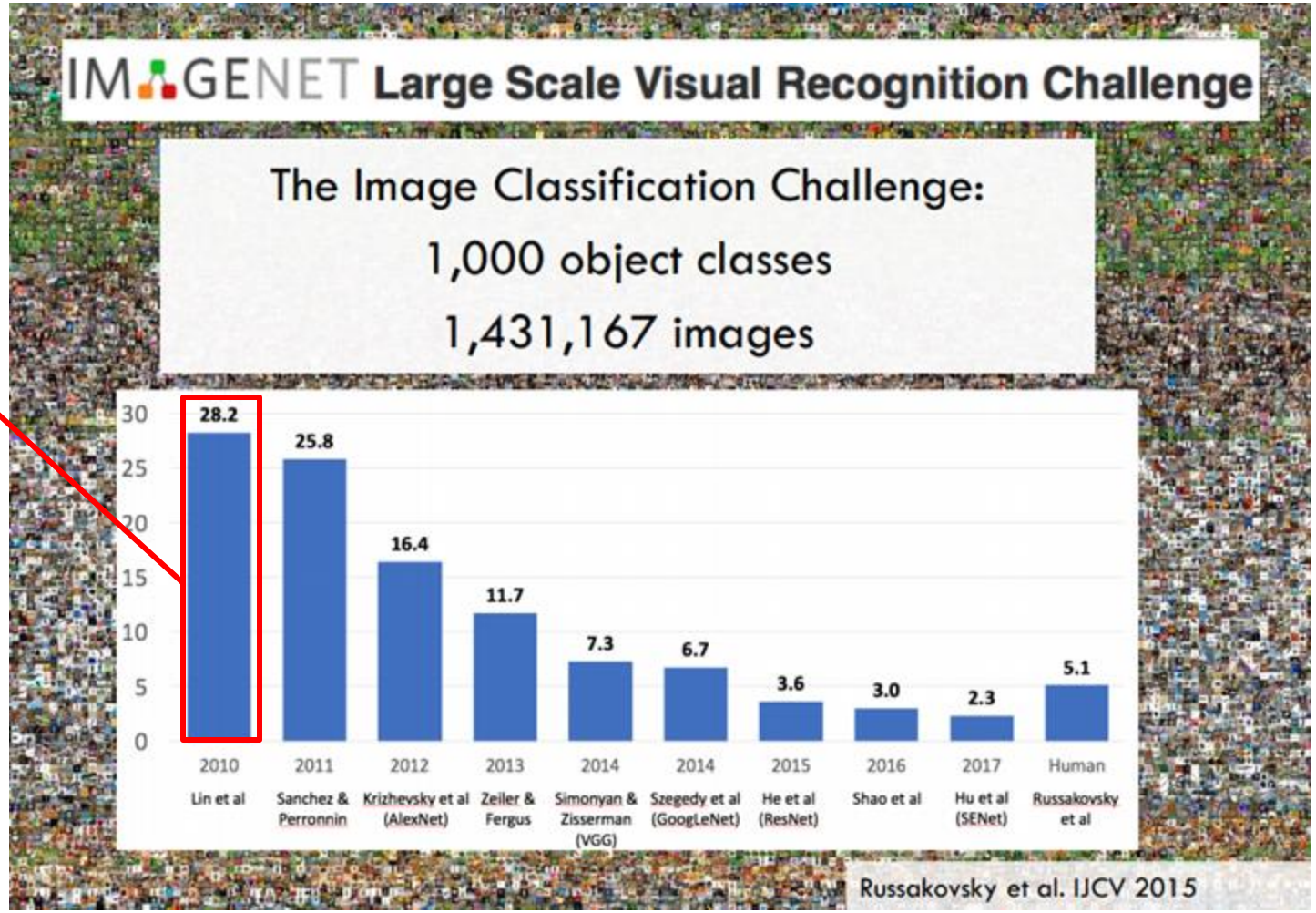


Image Analysis

Convolutional Neural Networks (CNNs)

- AlexNet
- VGGNet
- ResNet (2016)

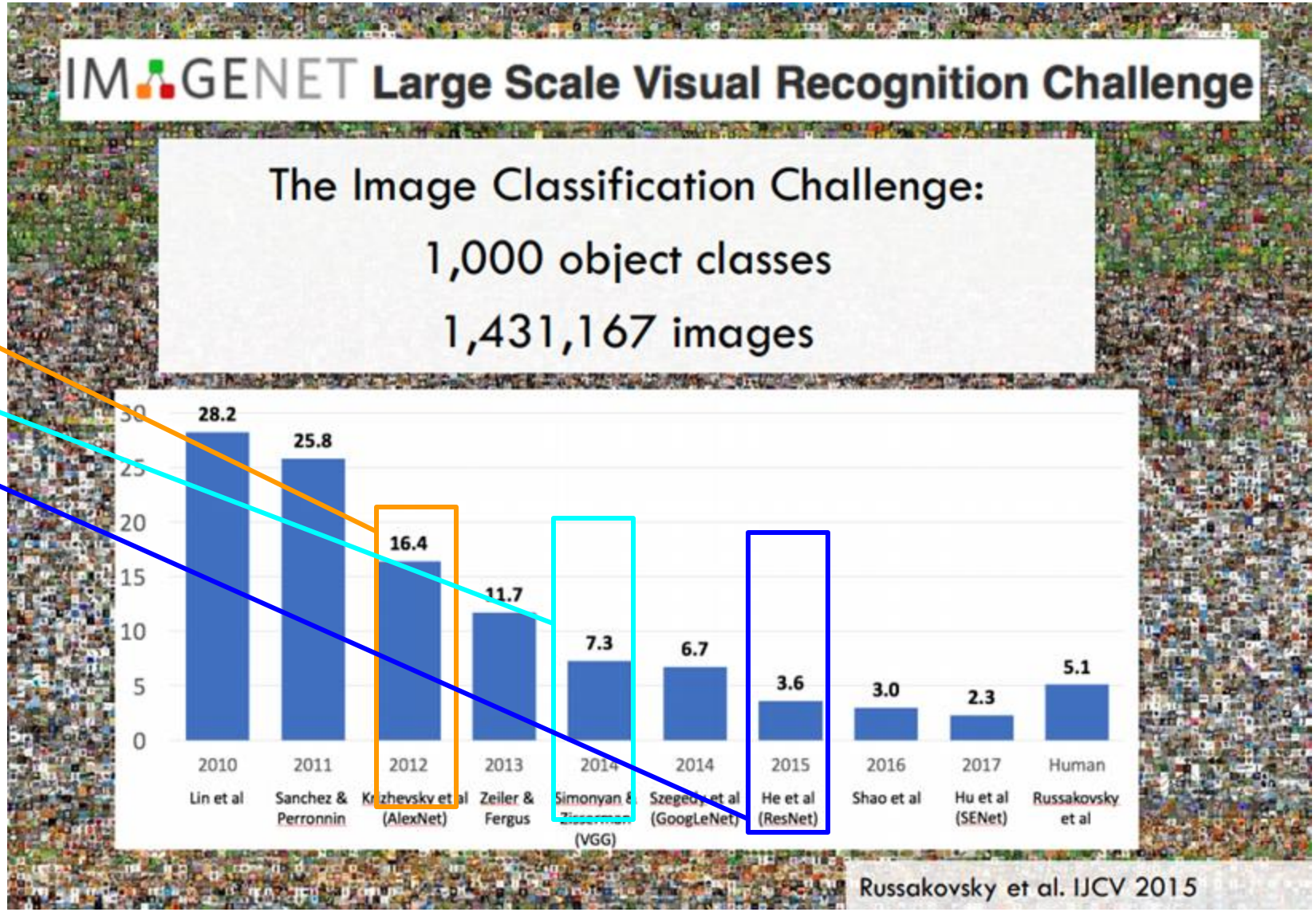


Image Analysis

CIFAR-10
32x32 color images
Balanced 50k
training 10k test

Image variable sizes
Usually cropped to 224x224
SOTA models use 448x448
1.4M images / 14M images

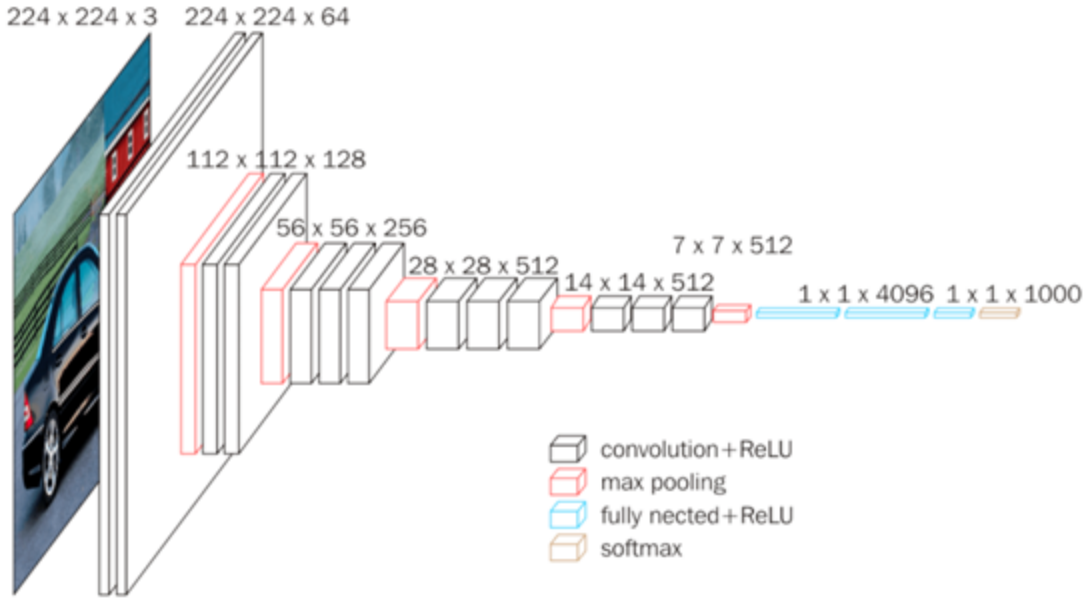
Classification (Visual recognition)

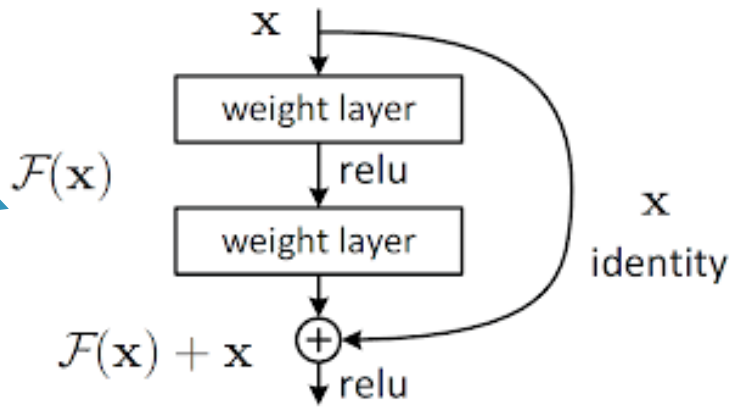
- Given an image, output probability over all N classes
- Key datasets: MNIST, CIFAR-10/100, ImageNet-1K, ImageNet-21K, JFT-300M/3B
- Other special dataset: CUB-200 (Bird Recognition), Flowers-102, iNaturalist (5k species categories), Stanford Cars (196 classes)

Image Analysis

CNN-based Solutions

- AlexNet
- VGGNet
- Inception
- ResNet
- MobileNet
- EfficientNet



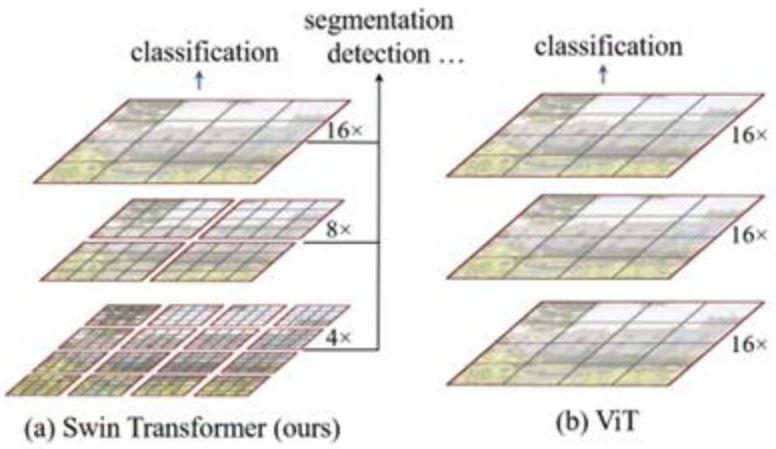


Residual connection: helps gradient flow -> solve gradient vanishing problem for deep CNNs

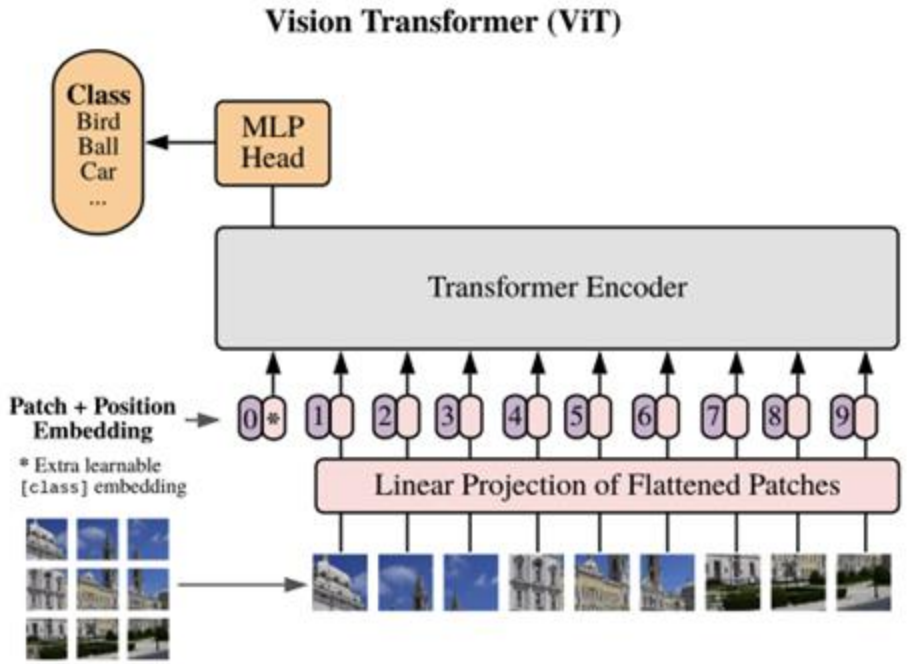
Image Analysis

Transformer-based Solutions

- ViT (20k citation from 2020)
- Swin Transformer (10k citation from 2021)



Hierarchical attention to save computation!



16x16 patch + position embedding to process original image & self-attention/multi-head attention block

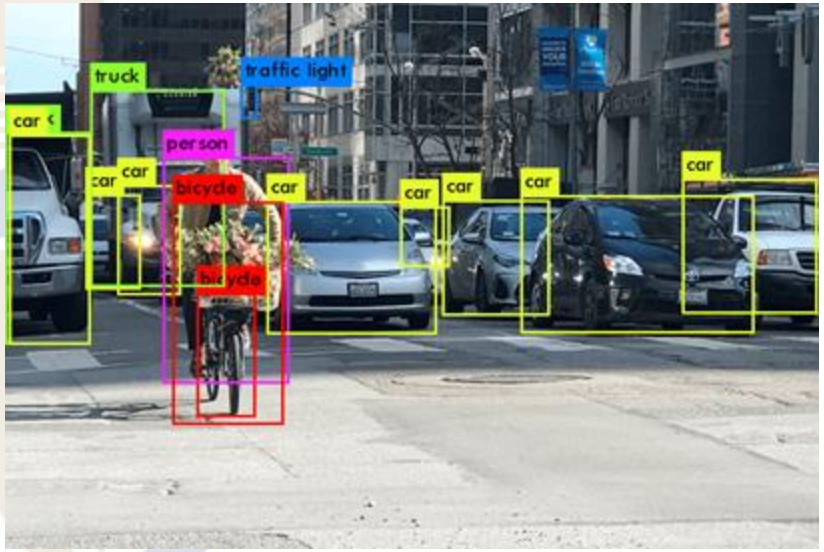
Image Analysis

Detection:

Given image, output bounding boxes/masks

Datasets: PASCAL VOC, MS COCO, Open Image

OpenImage - 9 million images (v6), partially labeled, 9600 object classes



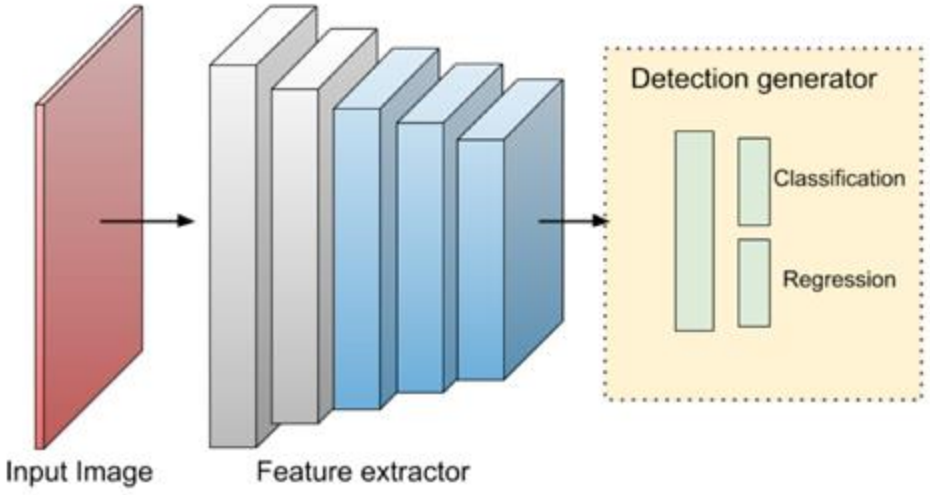
PASCAL VOC - 2012 dataset with 20 object classes.1.4K images for training

MS COCO - The go-to dataset for object detection 120K images with bounding boxes, masks and captions. 80 object classes

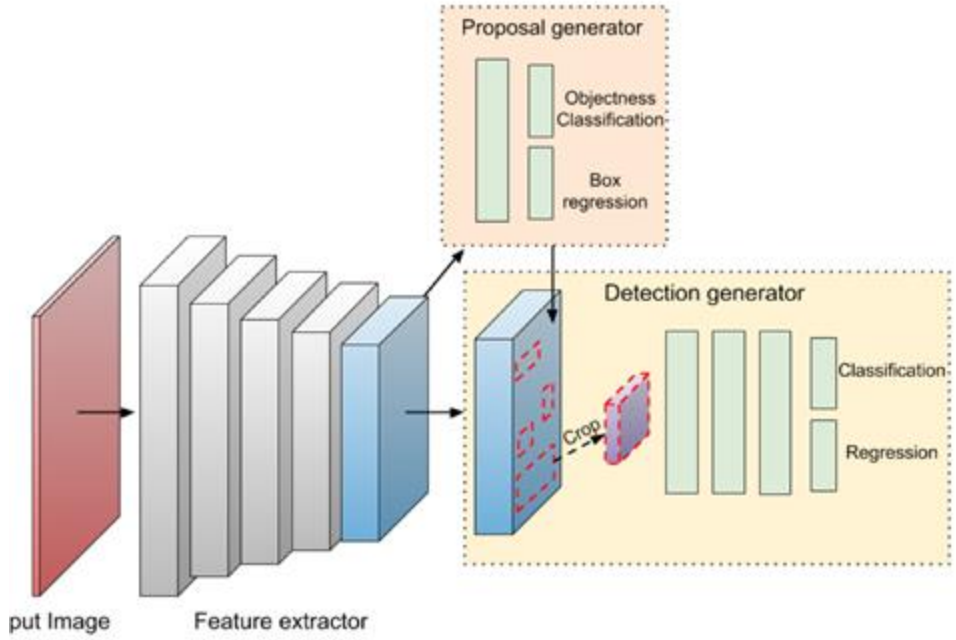
Image Analysis

Detection:
Given image, output bounding boxes/masks

One Stage Method, such as RetinaNet
Two Stages Method, such as Faster R-CNN



(b) One-stage RetinaNet

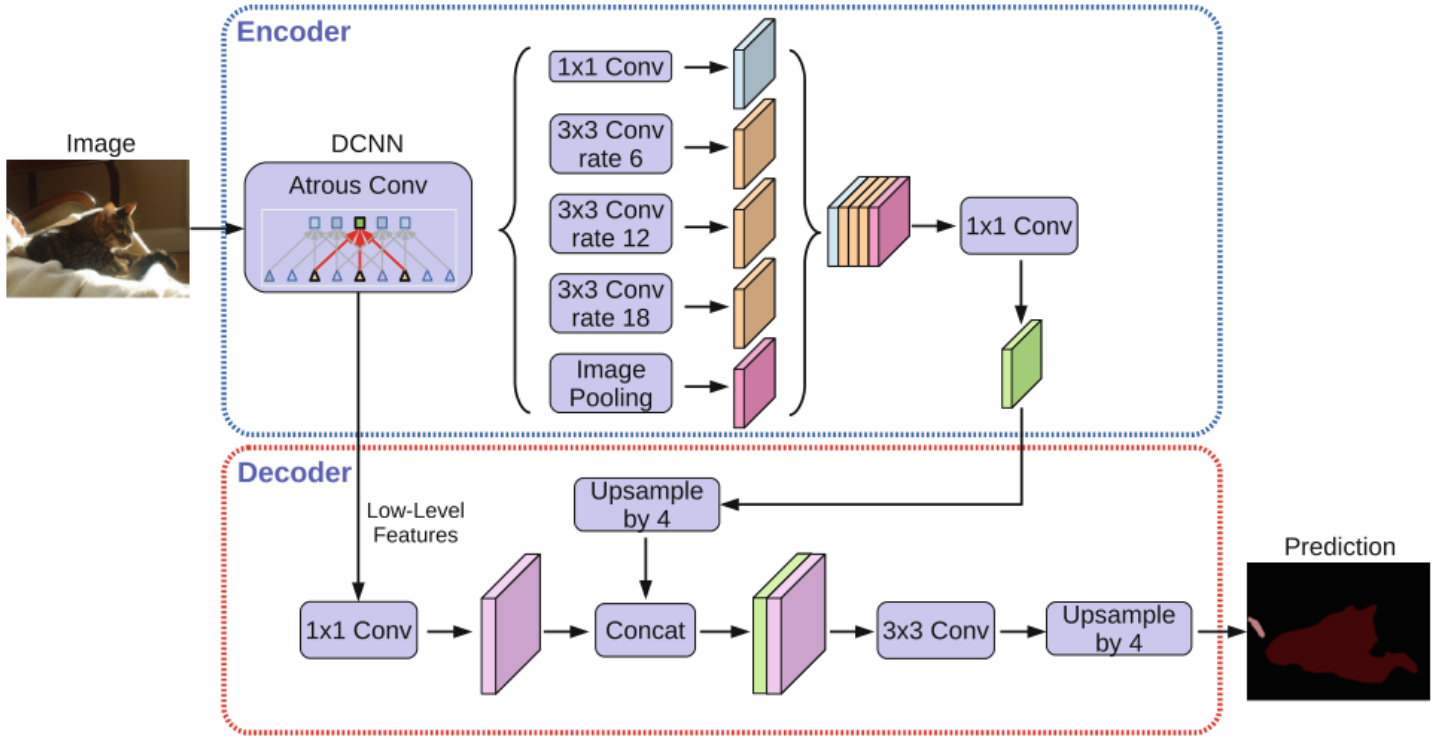


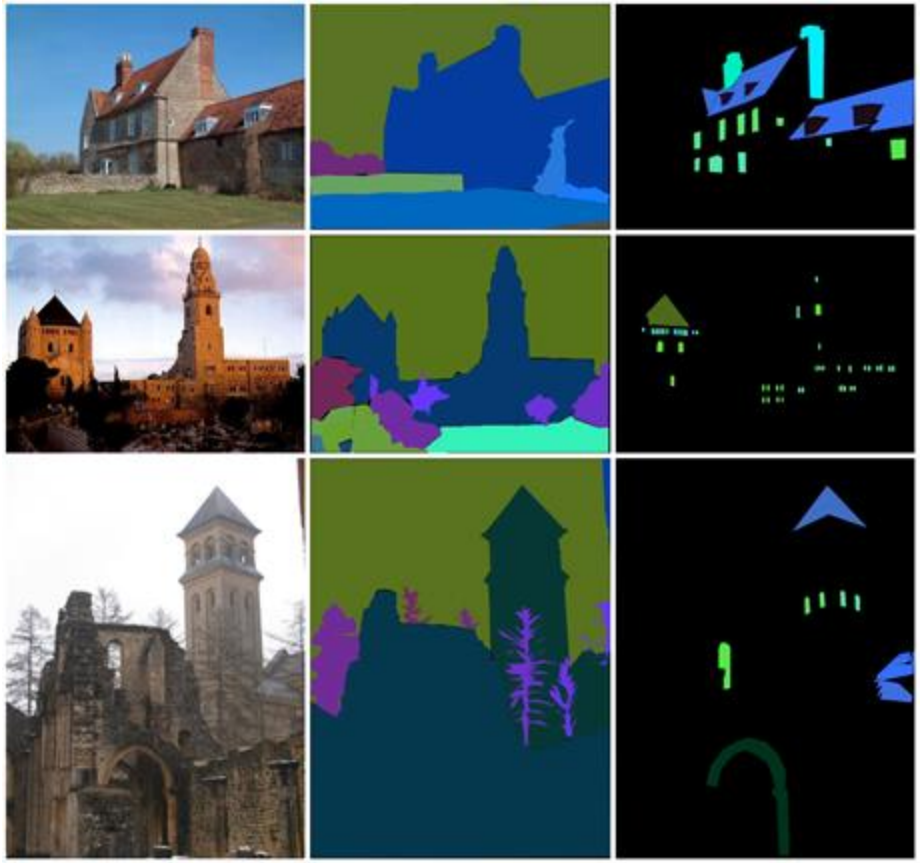
(a) Two-stage Faster R-CNN

Image Analysis

Segmentation: Given image, classify every pixels

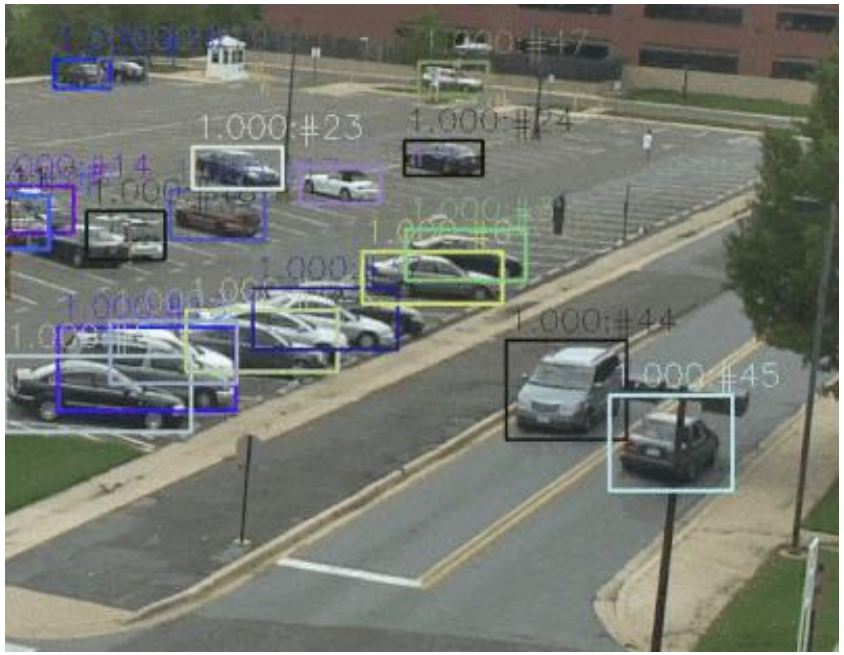
- Cityscape, ADE20K
- Notable methods: DeepLab





Video Understanding

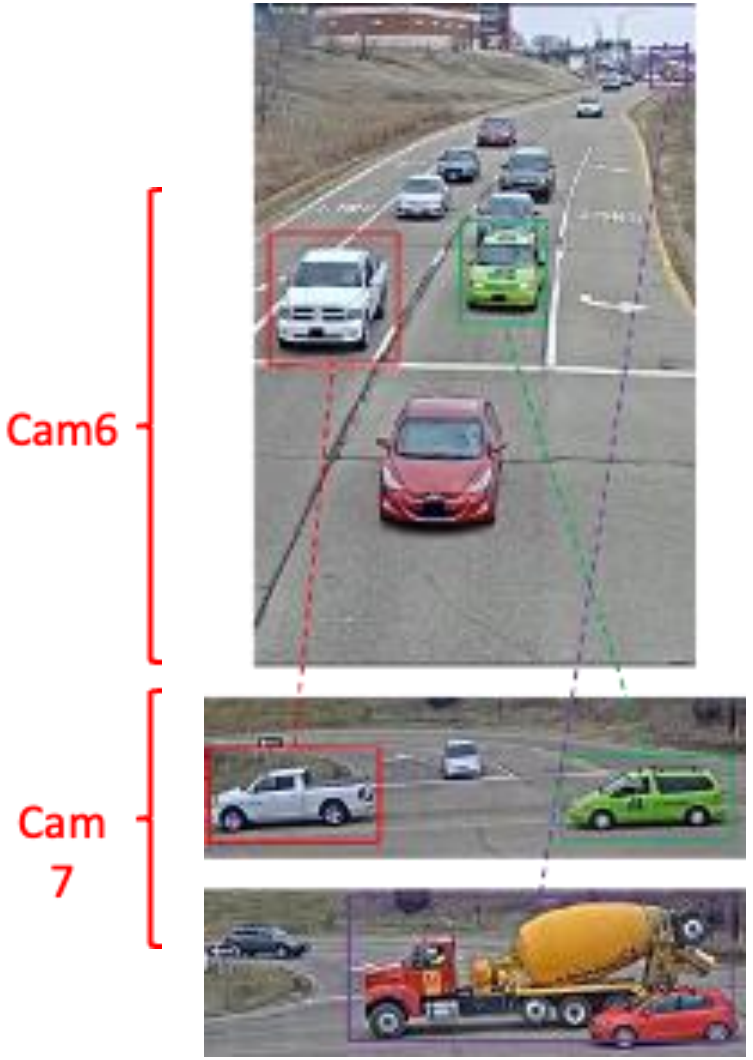
Object Tracking: Need to assign “ID” for each object



Video Understanding

City-scale Vehicle Tracking and ReID

- Multi-target multi-camera (MTMC) tracking aims to track the vehicles over large areas within multiple camera networks.
- Different from classical multiple object tracking (MOT) which only focuses on tracking objects within a single camera, MTMC needs to resort to multiple cameras.
- Moreover, the characteristics of moving vehicles bring unique challenges for multi-camera vehicle tracking.



Video Understanding

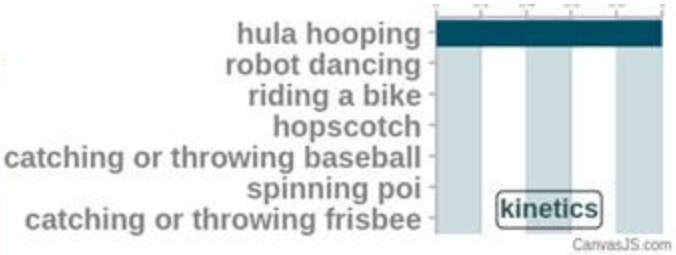
Trajectory Prediction

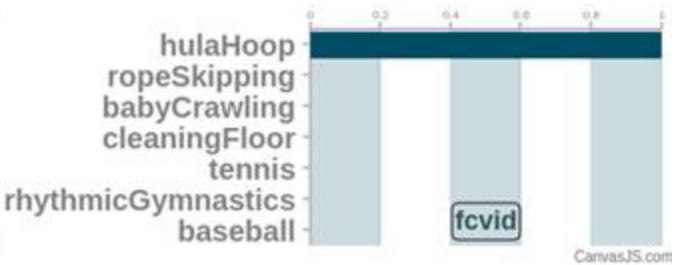
- Trajectory prediction / activity prediction
- Could be used to improve traffic safety and smart robot assistants

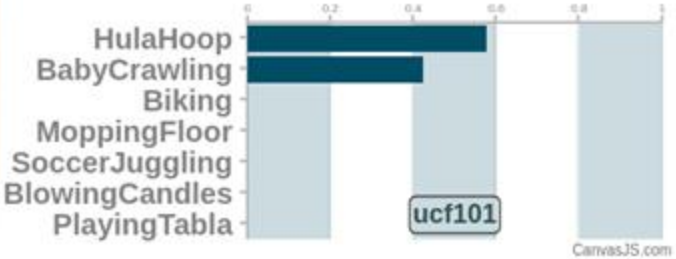
Video Understanding

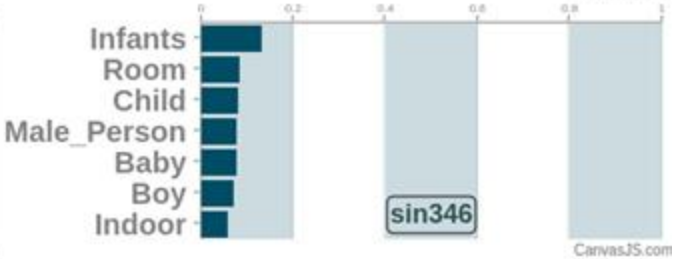
Action Recognition

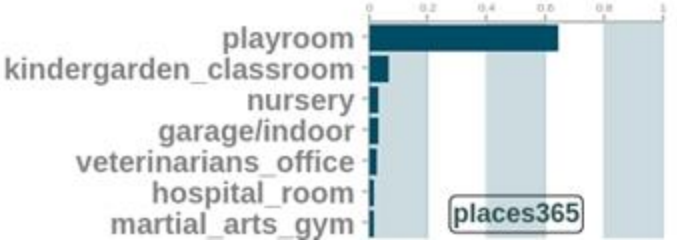
Given a video, output one class label

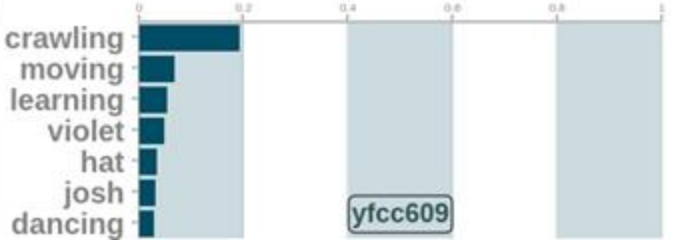








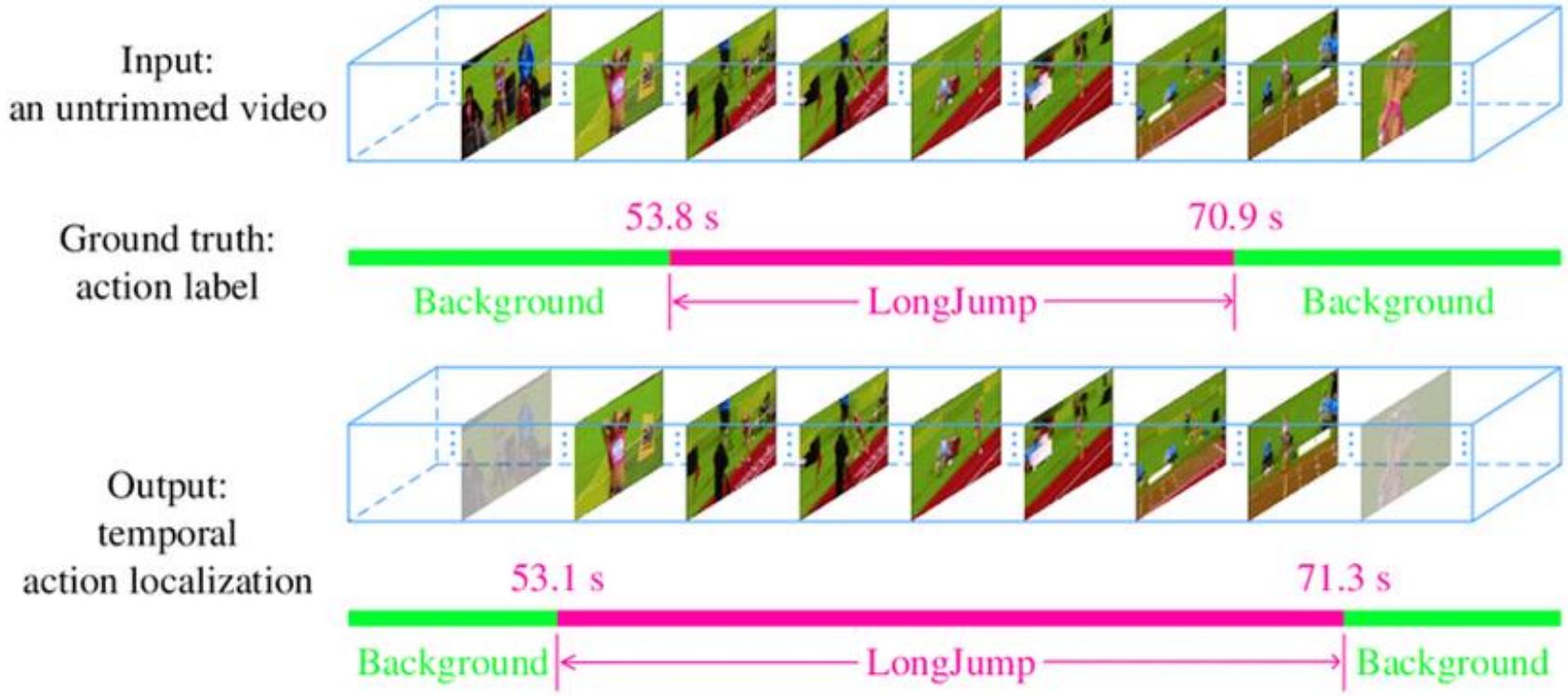




Video Understanding

Temporal Action Localization

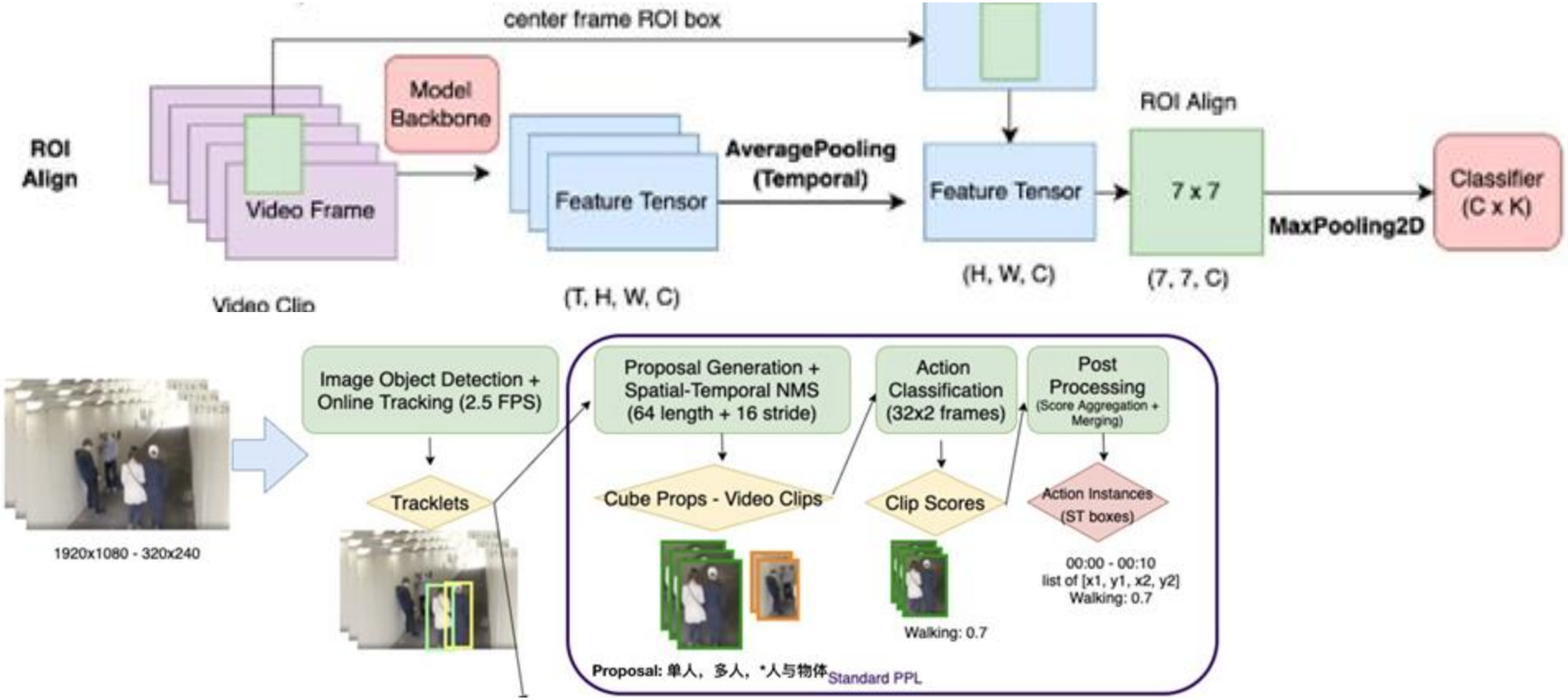
- Given (untrimmed) video, localize action's start and end time
- Datasets: THUMOS, ActivityNet



Video Understanding

Action Detection

- Given video, output bounding boxes
- Datasets: AVA, VIRAT, MEVA (surveillance videos)
- Methods: See NIST TRECVID ActEV challenges

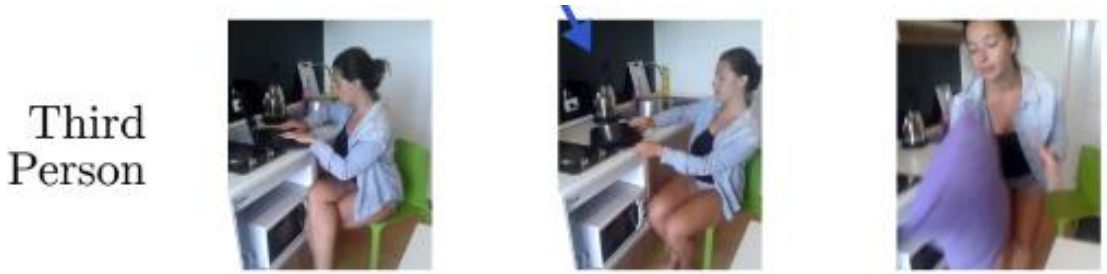


Egocentric Perception

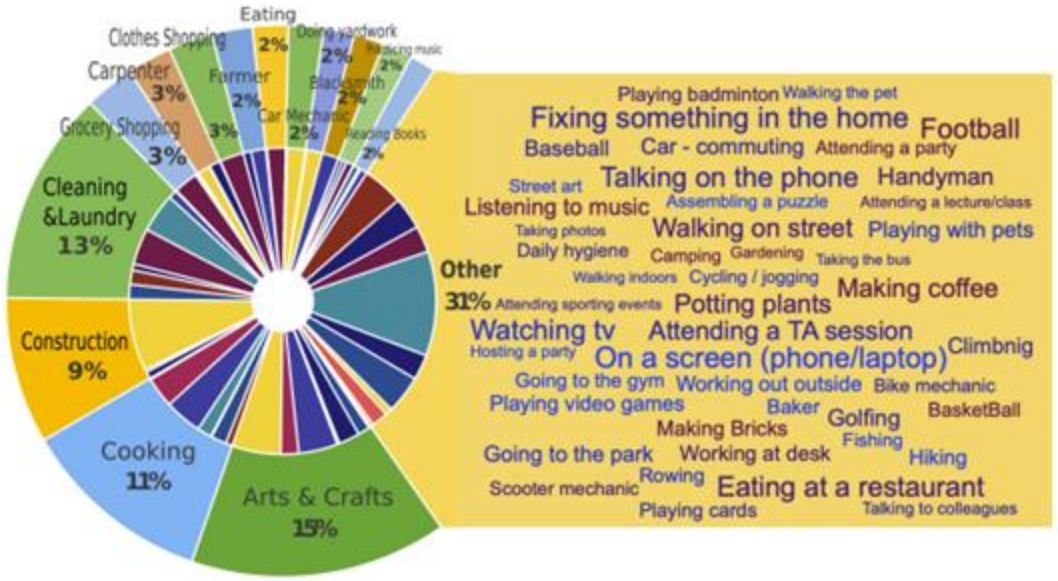
perspective view perception / First-person view perception

Used for robot vision and augmented reality

Vs.

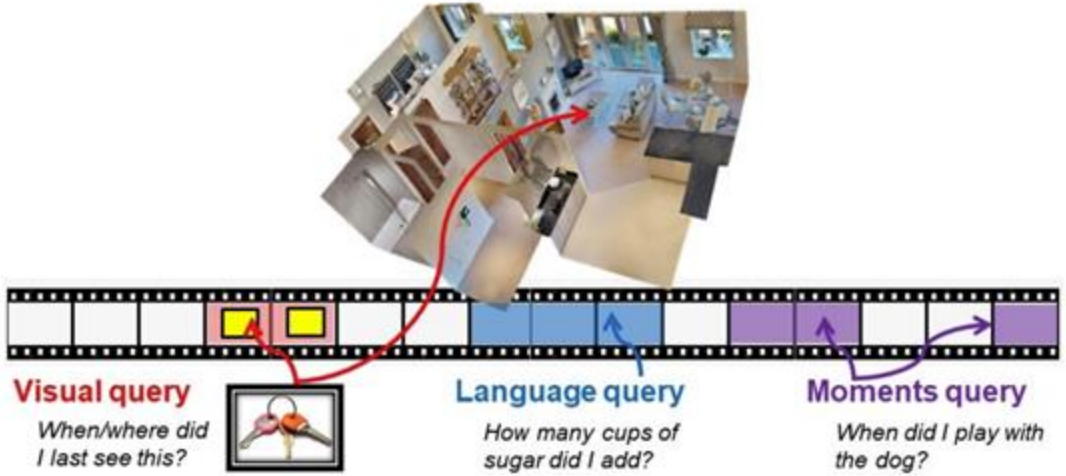


Egocentric Perception



Ego4D Dataset

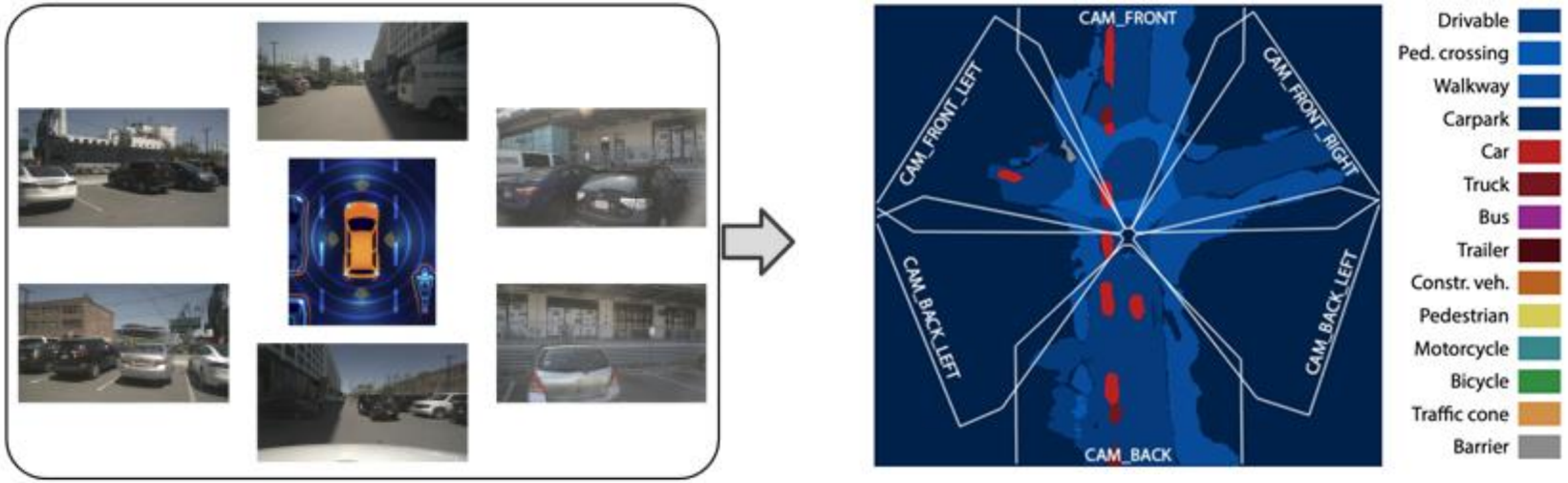
- 3000+ hours of ego-centric videos (head-mounted go-pros, etc.) over 74 locations in the world
- Daily activities



Bird-Eye-View Perception

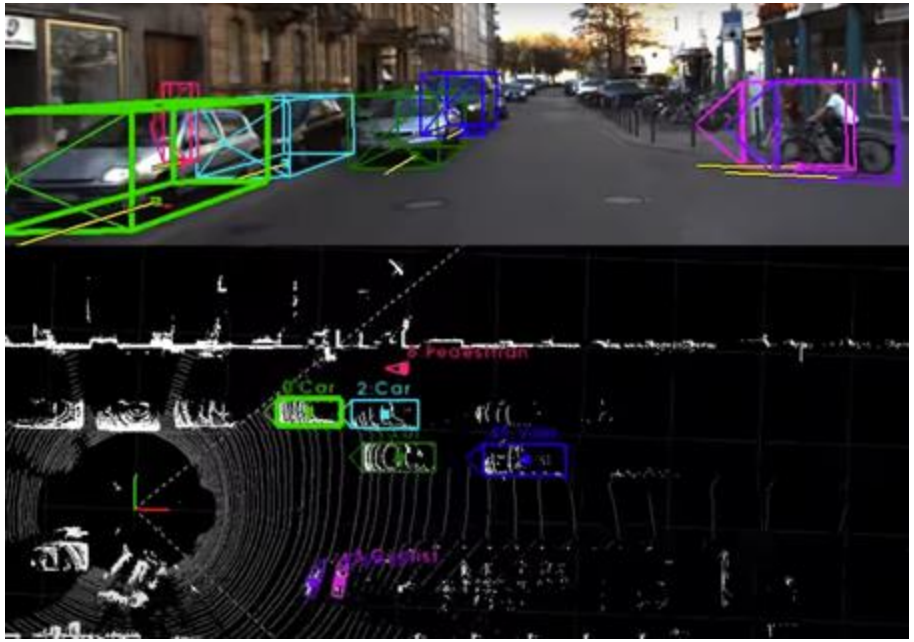
What is BEV perception?

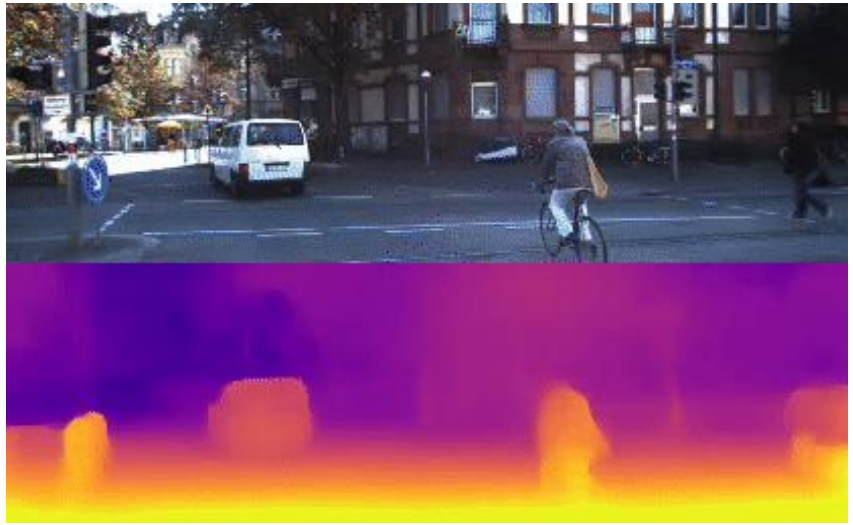
- Traditional perception tasks (detection, segmentation, tracking) are performed in the 2D image plane (front/view perspective). This view is inherently limited for spatial reasoning.
- Bird's-Eye View (BEV) perception transforms multi-camera and sensor data into a unified, top-down 2D representation.



Bird-Eye-View Perception

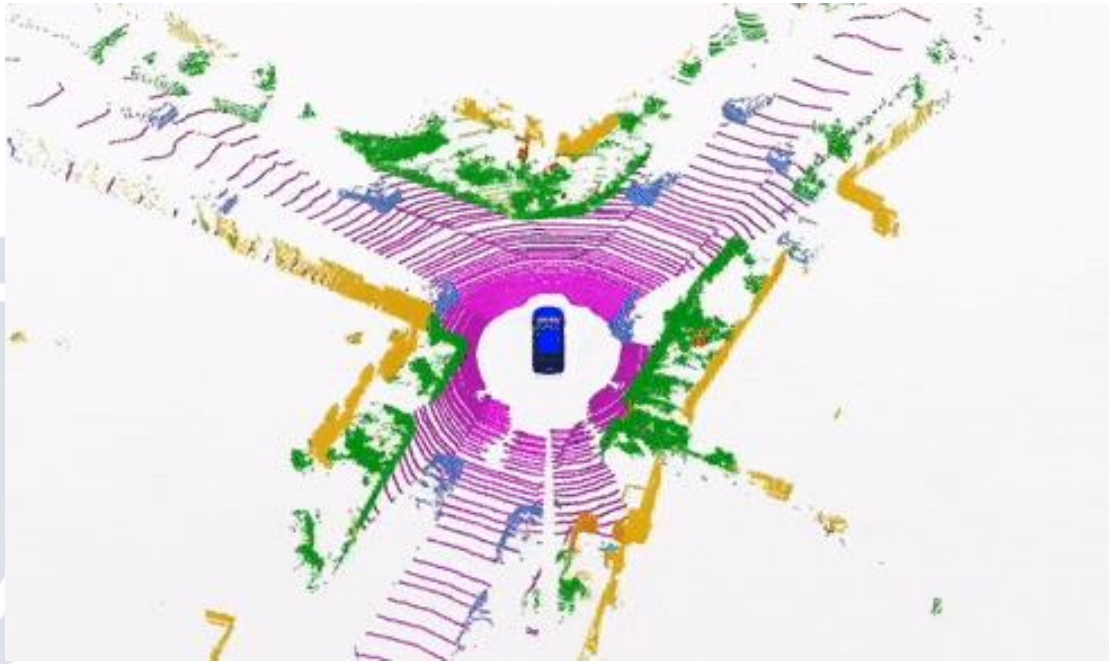
- It provides a spatially consistent and occlusion-aware environment model, which is crucial for downstream tasks like path planning and decision-making in autonomous systems.

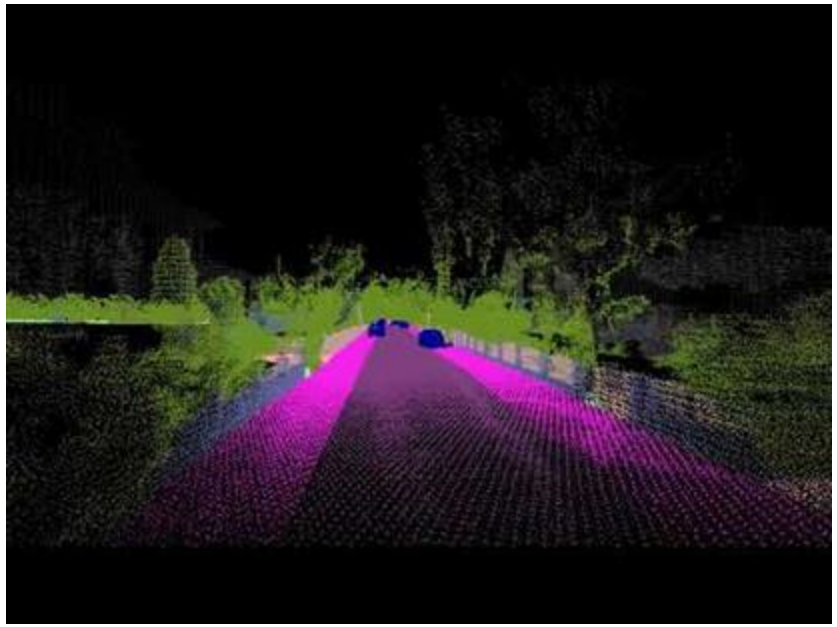




Bird-Eye-View Perception

Dataset: SemanticKITTI - RGB frames, 3D bbox, 3D segmentation





Information for Project 1

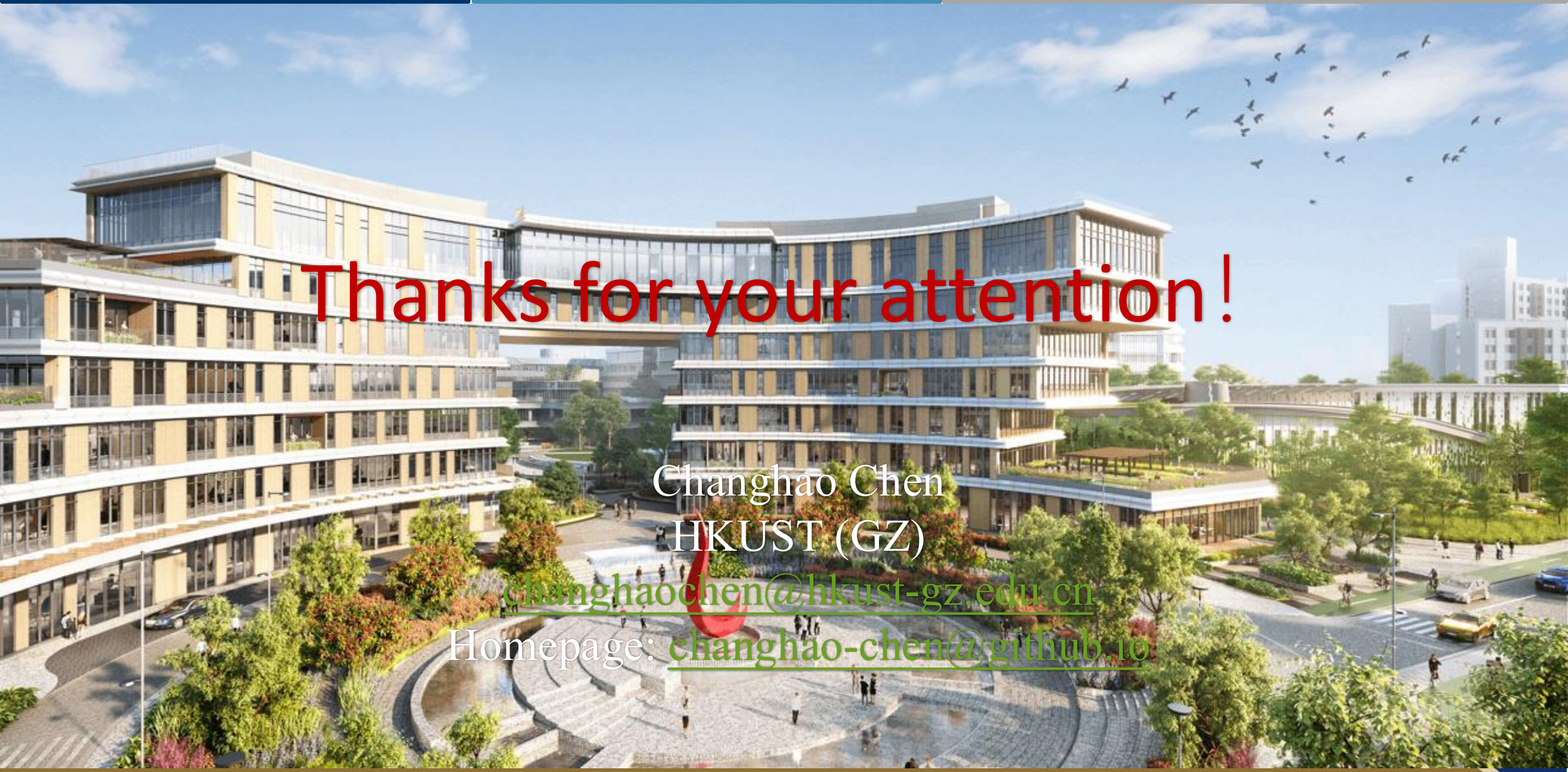
You will need to start forming groups (1-3 ppl) now.

Our TA will let you know how to use canvas to form a team, and submit your presentation ppt (ddl: Oct 14 2025).

- Project 1: Paper Presentation (20%)
  - You will select 1 paper from our list of Embodied AI research to present. This should show that you understand the paper, and you could present it to students who do not know the paper well. **You have 10 minutes to present and 5 minutes to handle at least one question (from TA/peers/teacher)**
    - Will be released on Sep 10
    - P1 presentation on Oct 15
  - Can be done in groups of 1-3 people. **Score multiplier: 1.1x for solo, 1.05x for two people and 1.0x for three people group.**
  - Project 1 is worth 20% credit.
    - You will be graded based on Understanding & Content (40%), Presentation & Communication (30%), Critical Analysis & Discussion (30%), by the teacher, TAs, and peers.

Paper List

Number	Theme	Paper Title	Conference/Journal
1	3D Vision	VGGT: Visual Geometry Grounded Transformer	CVPR 2025
2	3D Vision	BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers	ECCV 2022
3	3D Vision	Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data	CVPR 2024
4	3D Vision	MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection	ICCV 2023
5	3D Vision	Point Transformer V3: Simpler, Faster, Stronger	CVPR 2024
6	Navigation	AerialVLN :Vision-and-Language Navigation for UAVs	ICCV 2023
7	Navigation	SEEK: Semantic Reasoning for Object Goal Navigation in Real World Inspection Tasks	RSS 2024
8	Navigation	NavCoT: Boosting LLM-Based Vision-and-Languagee Navigation via Learning Disentangled Reasoning	TPAMI 2025
9	Navigation	HOP+: History-Enhanced and Order-Aware Pre-Training for Vision-and-Language Navigation	TPAMI 2023
10	Navigation	MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors	CVPR 2025
11	VLA	OpenVLA: An Open-Source Vision-Language-Action Model	PMLR 2024
12	VLA	RDT-1B: A DIFFUSION FOUNDATION MODEL FORBIMANUAL MANIPULATION	CoRR 2024
13	VLA	π0: A Vision-Language-Action Flow Model for General Robot Control	CoRR 2024
14	Manipulation	Diffusion policy: Visuomotor policy learning via action diffusion	IJRR 2025
15	Manipulation	Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware	RSS 2023



Thanks for your attention!

Changhao Chen
HKUST (GZ)
changhaochen@hkust-gz.edu.cn
Homepage: changhao-chen@github.io