

CS163: Deep Learning for Computer Vision

Lecture 19:
Trends and Open Problems in Computer
Vision and Tips for Doing Research

Announcement

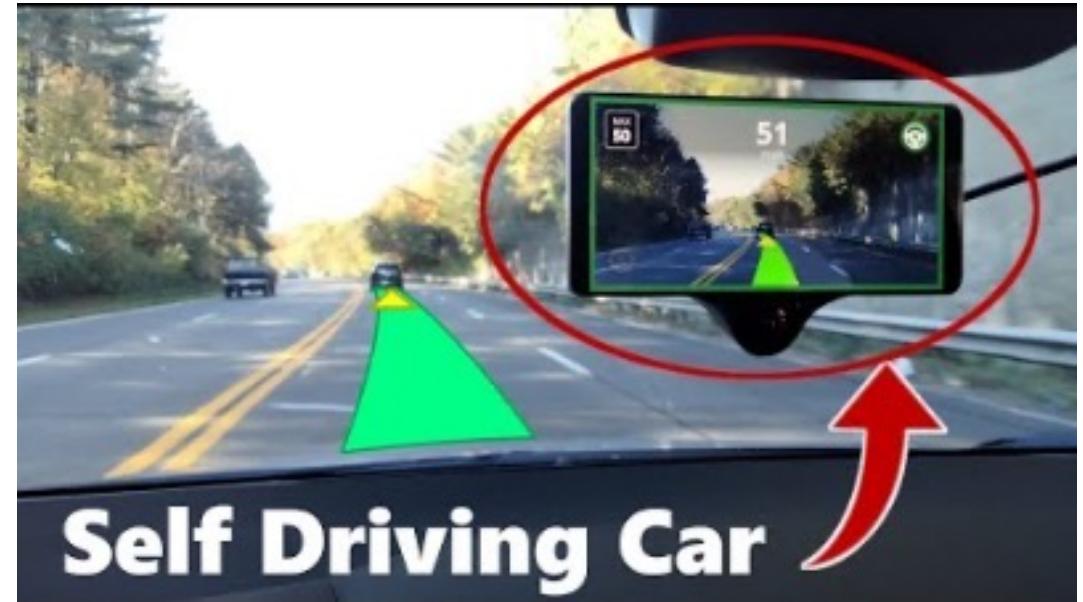
- Winter quarter course evaluation is open
 - Your kind feedback is greatly appreciated and will support the continuous development of this course
- Assignment 4 will be due on this Sunday (Dec 8)
- Final report for course project will be due on Dec 13

Announcement

- A relevant seminar talk today

Learning to Drive in the Dreams of a Diffusion Model

Harald Schaefer
CTO of Comma.AI



Abstract:

End-to-end machine learning is gaining traction in vehicle autonomy. A dataset of human driving is easy to collect and contains demonstrations of high-quality driving. Learning to mimic that behavior is an effective way to teach a machine how to drive that scales well with data and compute. This talk describes the latest training strategy developed at comma.ai to teach machines how to drive. It involves training a large world-model, which can act as a simulator that takes in driving actions and generates video. This simulator can produce realistic driving scenarios much like a video game. Policy models are allowed to act in this simulator where they can be trained on-policy, to drive like a human. The latest models trained like this are already deployed in openpilot, an open-source ADAS project, and used by thousands of people every day in the real world.

Bio:

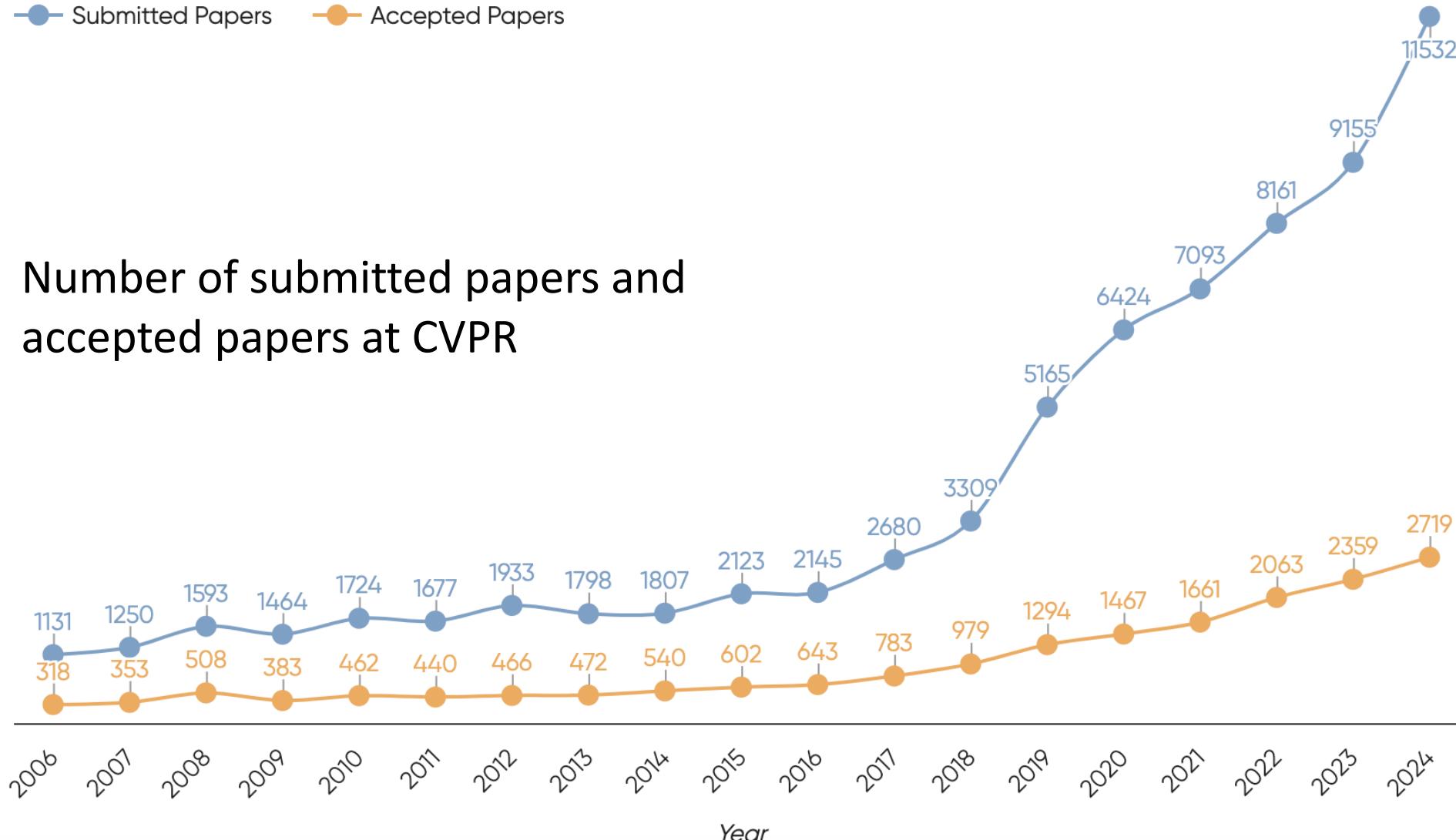
Harald is the CTO of comma.ai. He has worked at comma for 7 years, where he has been developing an end-to-end self-driving solution for openpilot, the most popular open-source self-driving car project in the world.

TUESDAY, DECEMBER 3 2024

PLACE: 3400 BOELTER HALL

TIME: 4:15 PM – 5:45 PM

2012 to Present: Deep Learning Explosion in Computer Vision



List of CVPR'24 papers: <https://cvpr.thecvf.com/Conferences/2024/AcceptedPapers>

Latest CVPR'24 submission



Matthias Niessner @MattNiessner · Feb 27

Our group has fourteen papers accepted at #CVPR'2024!

Exciting topics: lots of diffusion & transformers focusing on generating images for image synthesis, geometry generation, and many more - check them out!

I'm so proud of everyone involved - let's go 🚀🚀

[niessnerlab.org/publications.h...](http://niessnerlab.org/publications.html)



DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars

Tobias Kirchstein, Simon Giebenhain, Matthias Niessner

CVPR 2024

DiffusionAvatar uses diffusion-based, deferred neural rendering to translate geometric cues from an underlying neural parametric head model (NPHM) to photo-realistic renderings. The underlying NPHM provides accurate control over facial expressions, while the deferred neural rendering leverages the 2D prior of StableDiffusion in order to generate compelling images.

[paper][video][blurb][project page]



SceneTex: High-Quality Texture Synthesis for Indoors

Dove Zhengyu Chen, Abhishek Li, Hsin-Ying Lee, Sergey Tulyakov, Ma

CVPR 2024

We propose SceneTex, a novel method for effectively generating scenes using depth-to-image diffusion priors. At its core, SceneTex encodes the mesh appearance. To further secure the style constraint to predict the RGB values by cross-attending to the pre-sampled n

[paper][video][blurb][project page]



MonoNPHM: Dynamic Head Reconstruction from Monocular Videos

Simon Giebenhain, Tobias Kirchstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, Matthias Niessner



DPHMs: Diffusion Parametric Head Models for De-

Jiaxing Tong, Angela Dai, Yinyu Nie, Lev Manovich, Justin Thies, M

CVPR 2024

We introduce Diffusion Parametric Head Models (DPHMs), a g

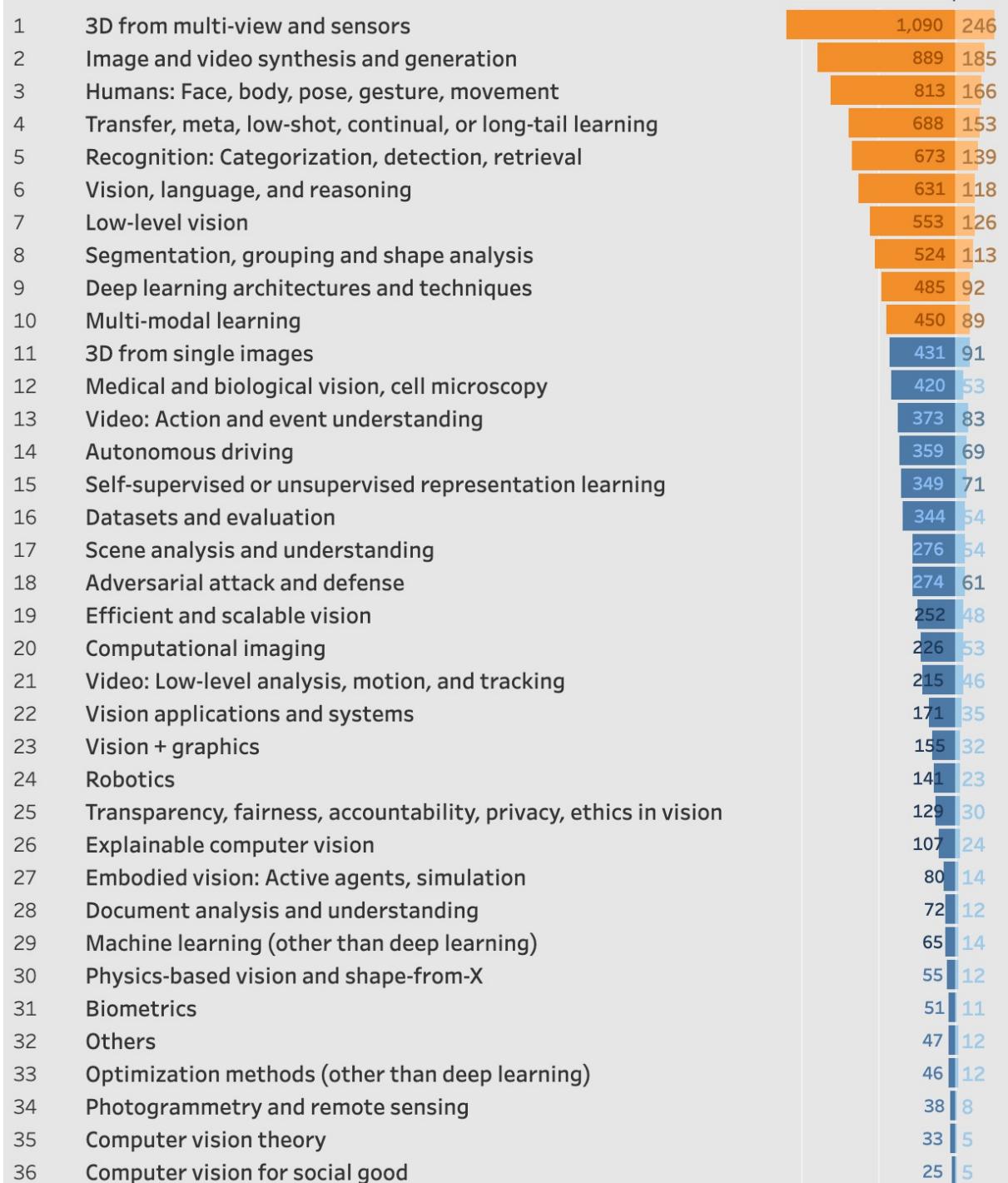
<https://cvpr.thecvf.com/Conferences/2024/AcceptedPapers>

Ziwei Liu from NTU in Singapore has **19 papers** accepted

Someone in Shanghai AI Lab has **27 papers** accepted...

I have **3 papers** accepted.

Kaiming He doesn't have any accepted CVPR'24 paper...



Accepted papers classified by categories.

The left part of the bar plot is the number of authors, right — number of papers

Trends, Problems, and Tips

Trends:

- Ultra-large vision: foundation models
- 3D vision from multiple cameras and neural rendering

Challenges:

- Interpretability, safety, robustness, etc
- Need a huge amount of annotated training data

Tips for CV research:

- Choose right topic to work on
- Cycle of a research topic
- Focus on task instead of model

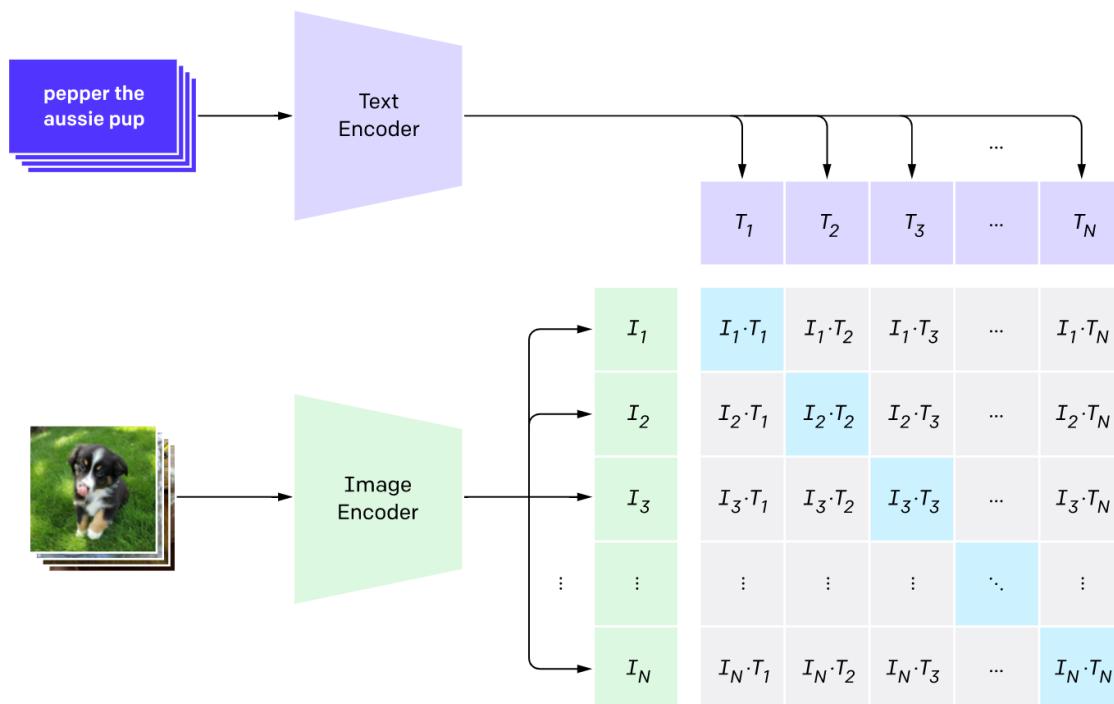
Trends in Computer Vision

- **Data-driven trend: Ultra-large-scale foundation models**
- 3D trend: 3D vision + multi-modal cameras + simulation (Metaverse)

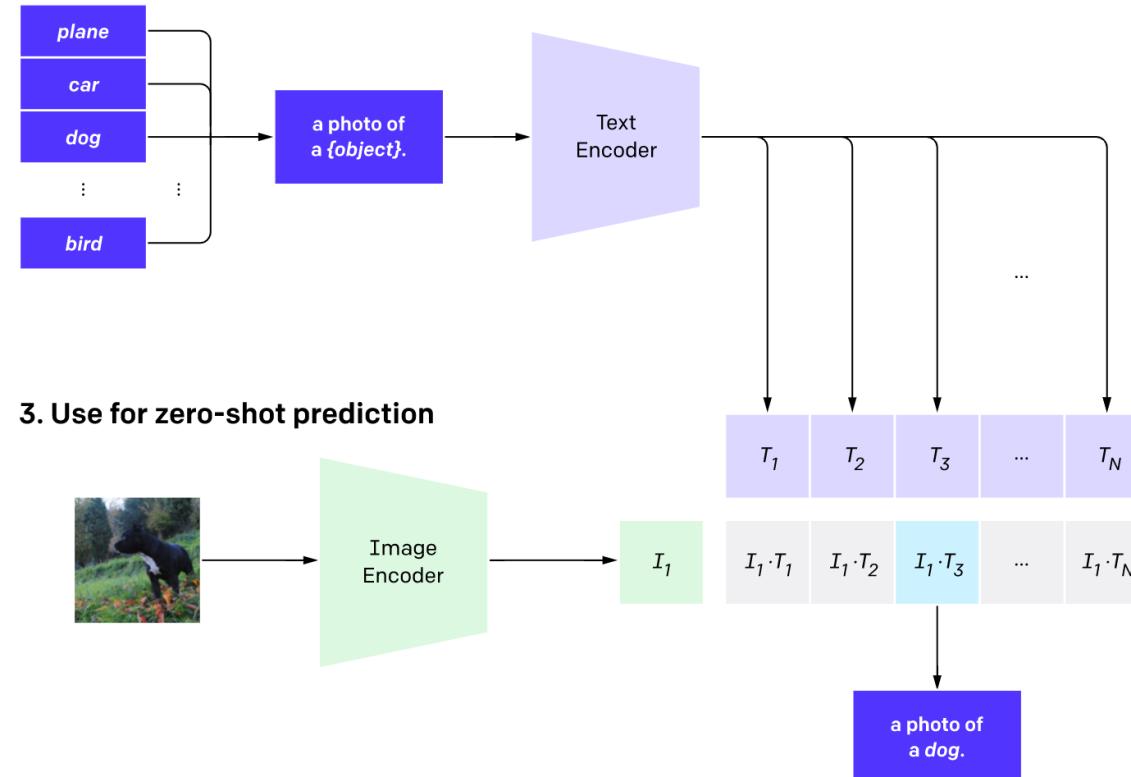
Scaling training data from 1 million images to 100 million images and more

Milestone method: OpenAI's CLIP, with 650 million image-caption pairs as training data

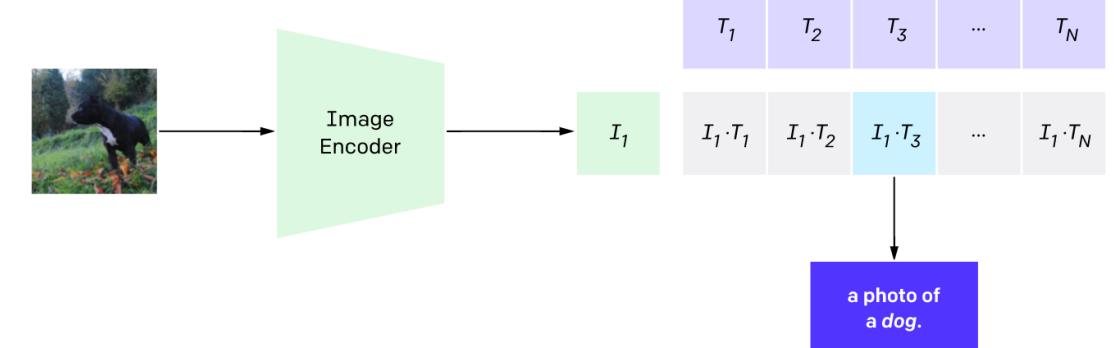
1. Contrastive pre-training



2. Create dataset classifier from label text

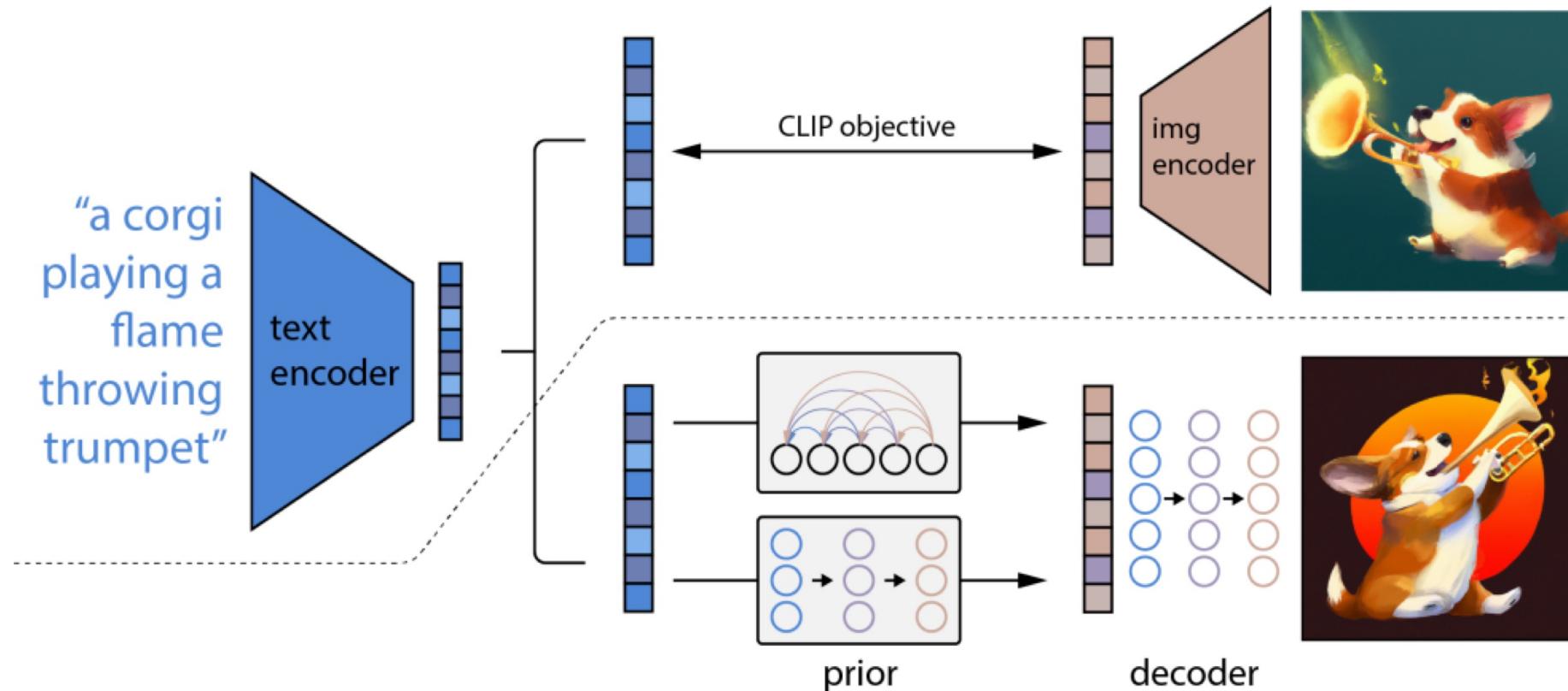


3. Use for zero-shot prediction



Scaling training data from 1 million images to 100 million images and more

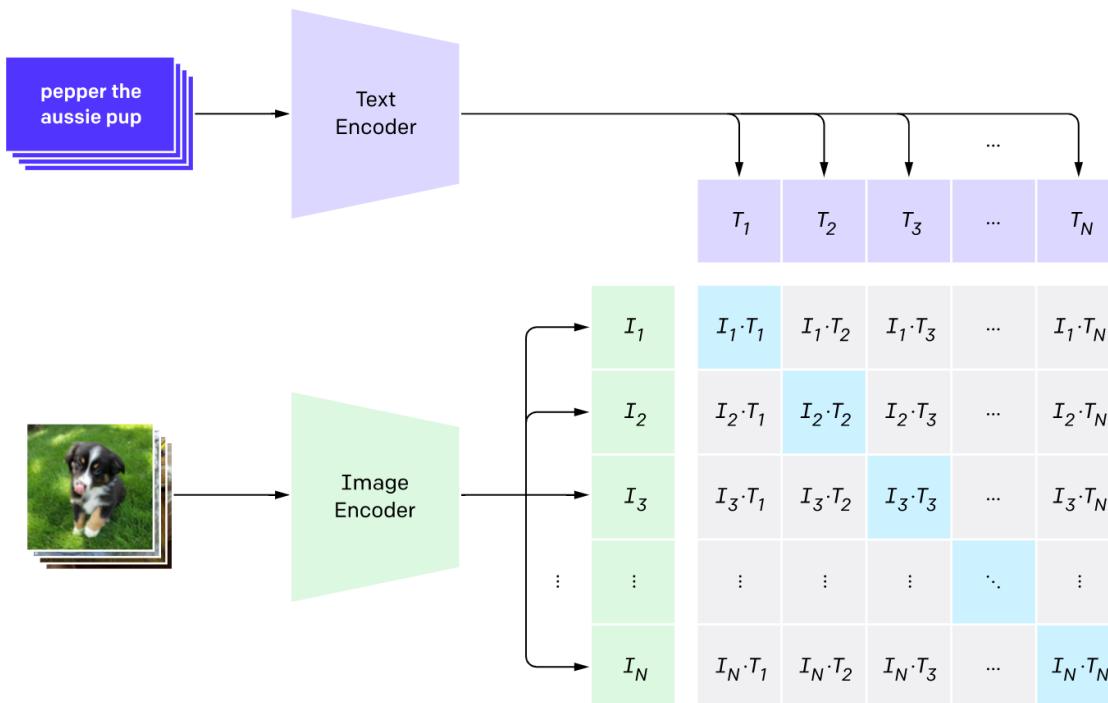
OpenAI's DALLE2: 650 million images for CLIP , 250 million text-image pairs for diffusion model



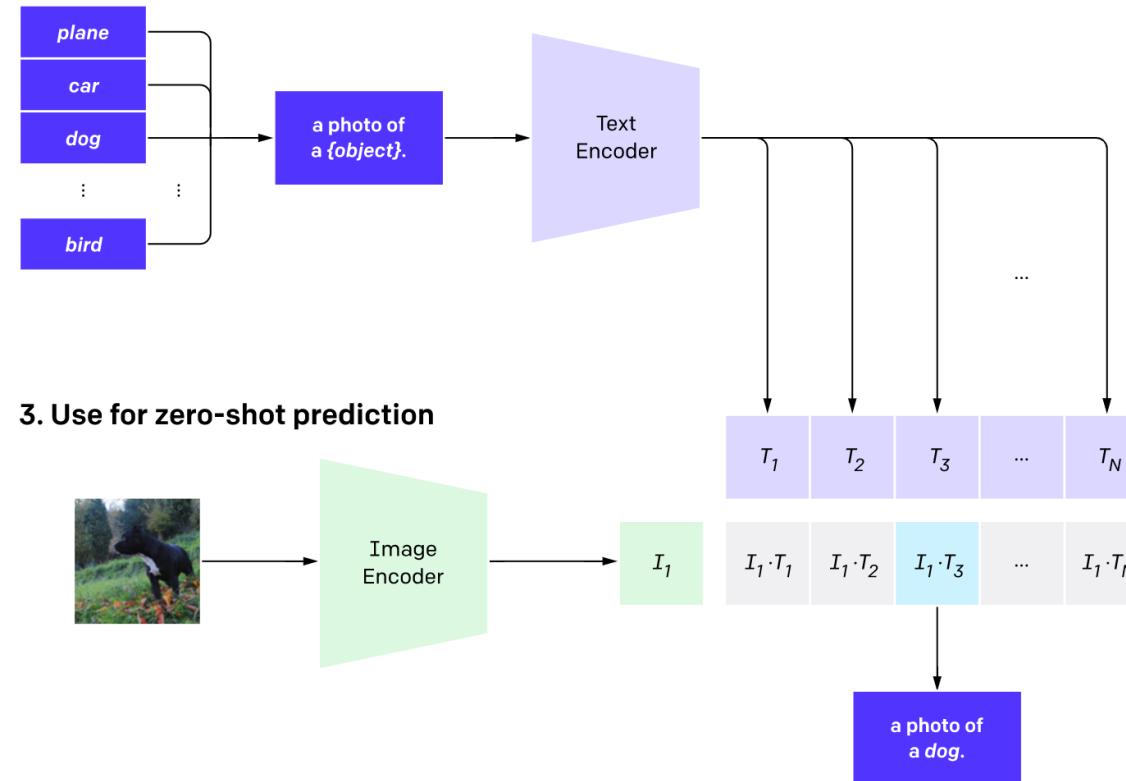
Foundation Models in Computer Vision (Open-Source)

Pretrained image features: CLIP model

1. Contrastive pre-training



2. Create dataset classifier from label text



Foundation Models in Computer Vision (Open-Source)

Multimodal image understanding models (LLaVA)



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User
LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

<https://llava-vl.github.io/>

Foundation Models in Computer Vision (Open-Source)

Image segmentation model: Segmenting Anything Models (SAMs)

11M images, 1B+ masks

After annotating enough masks with SAM's help, we were able to leverage SAM's sophisticated ambiguity-aware design to annotate new images fully automatically. To do this, we present SAM with a grid of points on an image and ask SAM to segment everything at each point. Our final dataset includes more than 1.1 billion segmentation masks collected on ~11 million licensed and privacy preserving images.

→ Explore the dataset

→ Download full dataset



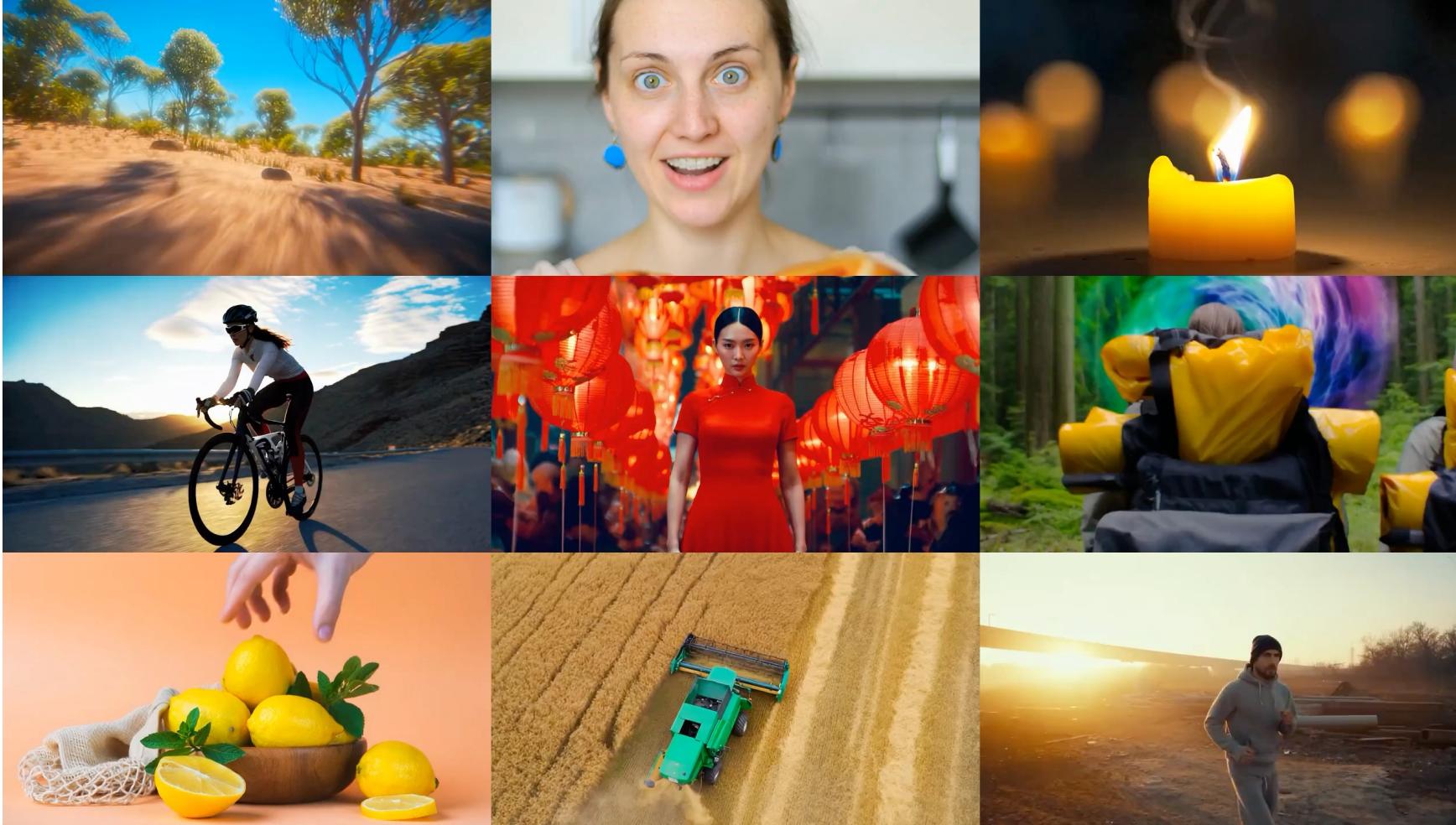
Foundation Models in Computer Vision (Open-Source)

Text2Image generation model: Stable Diffusion



Foundation Models in Computer Vision (Open-Source)

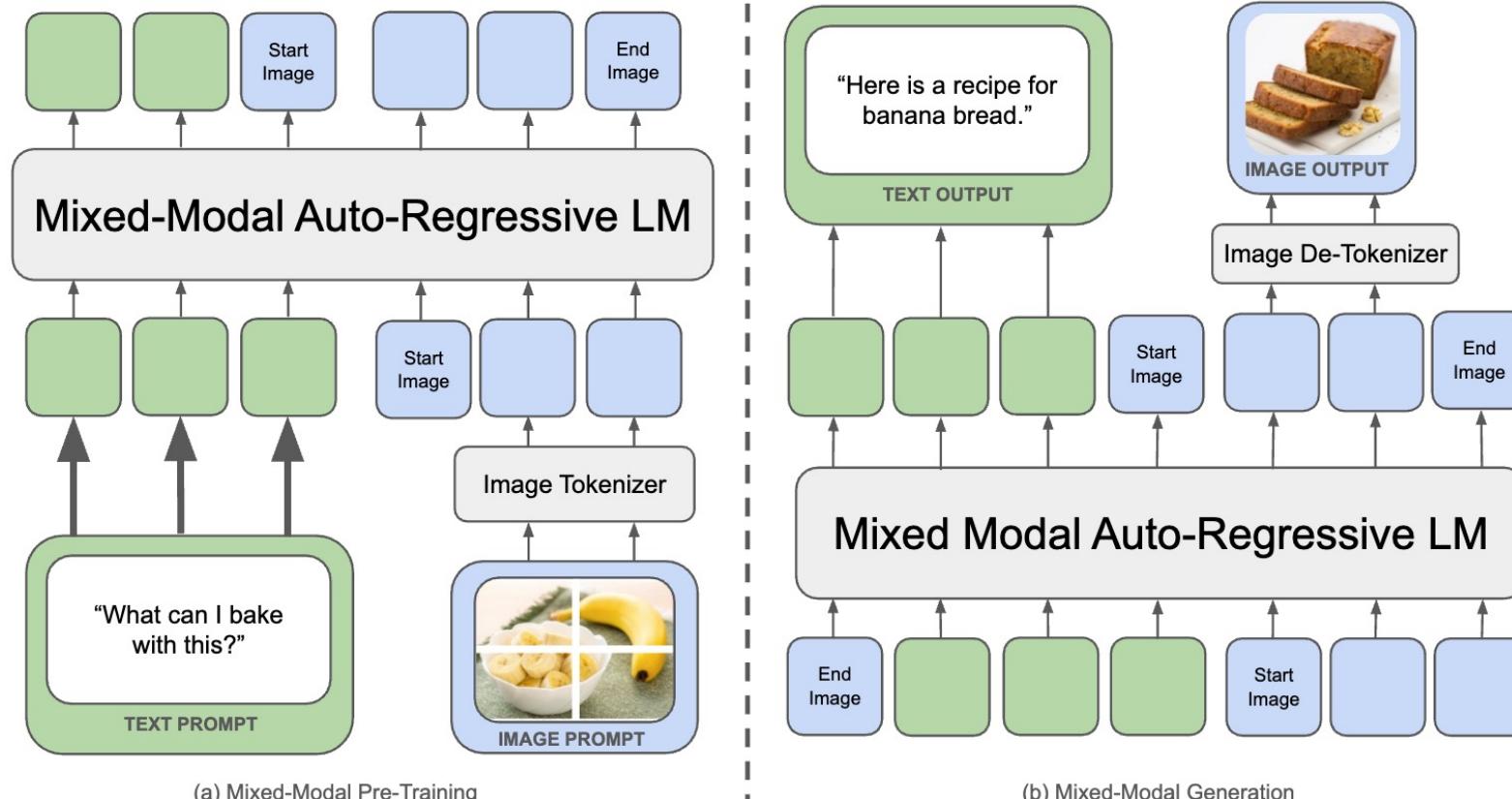
Text2Video generation model: Mochi



<https://github.com/genmoai/mochi>

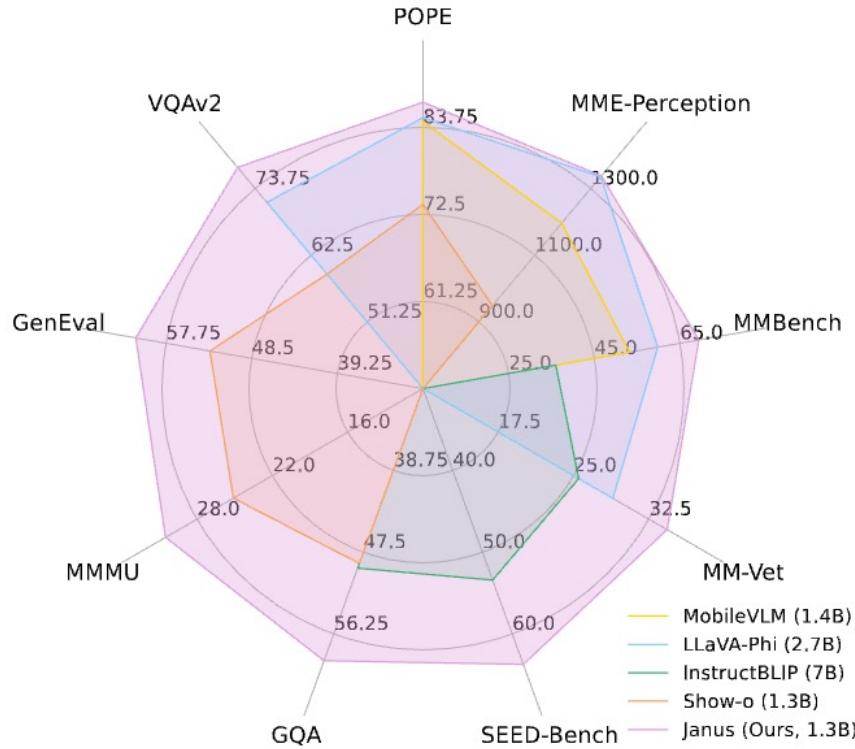
Foundation Models in Computer Vision

Model that combines recognition and generation: Chameleon

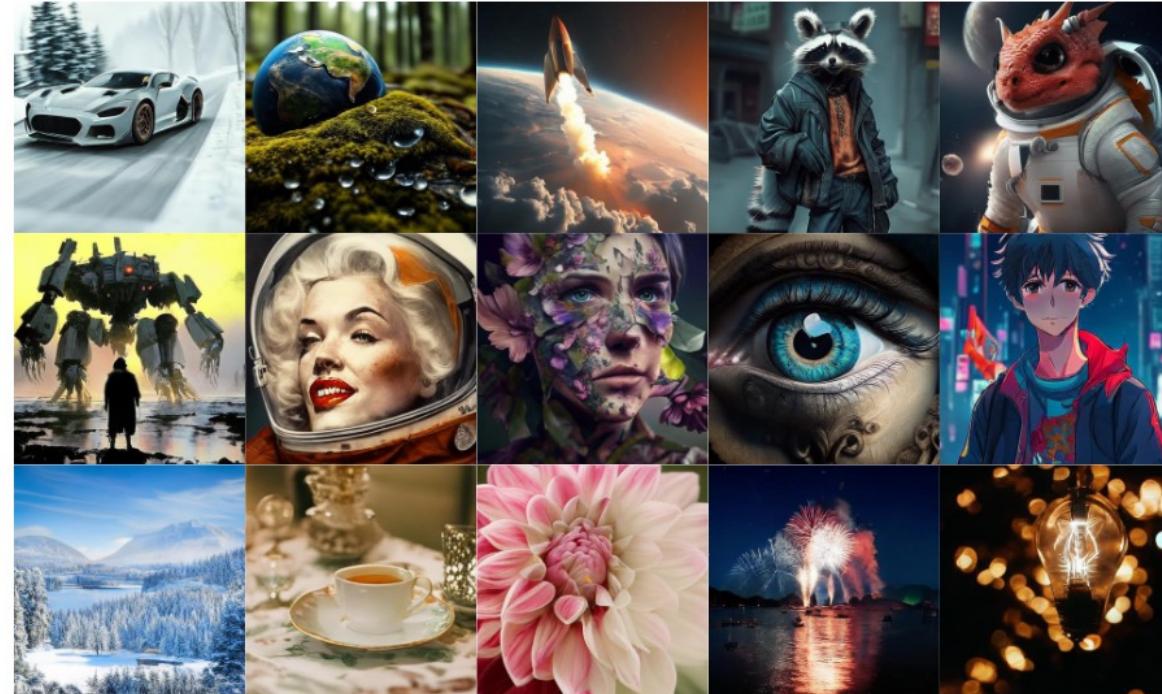


Foundation Models in Computer Vision

Model that combines recognition and generation: Janus



(a) Benchmark Performance.



(b) Visual Generation Results.

Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation

<https://github.com/deepseek-ai/Janus>

Trends in Computer Vision

- Data-driven trend: Ultra-large-scale foundation models
- **3D trend: 3D vision + multi-modal cameras + simulation (Metaverse)**

3D Perception with more cameras (RGB, Depth LiDAR, RADAR, Stereo)
self-driving, mobile phone applications, digital entertainment, etc.



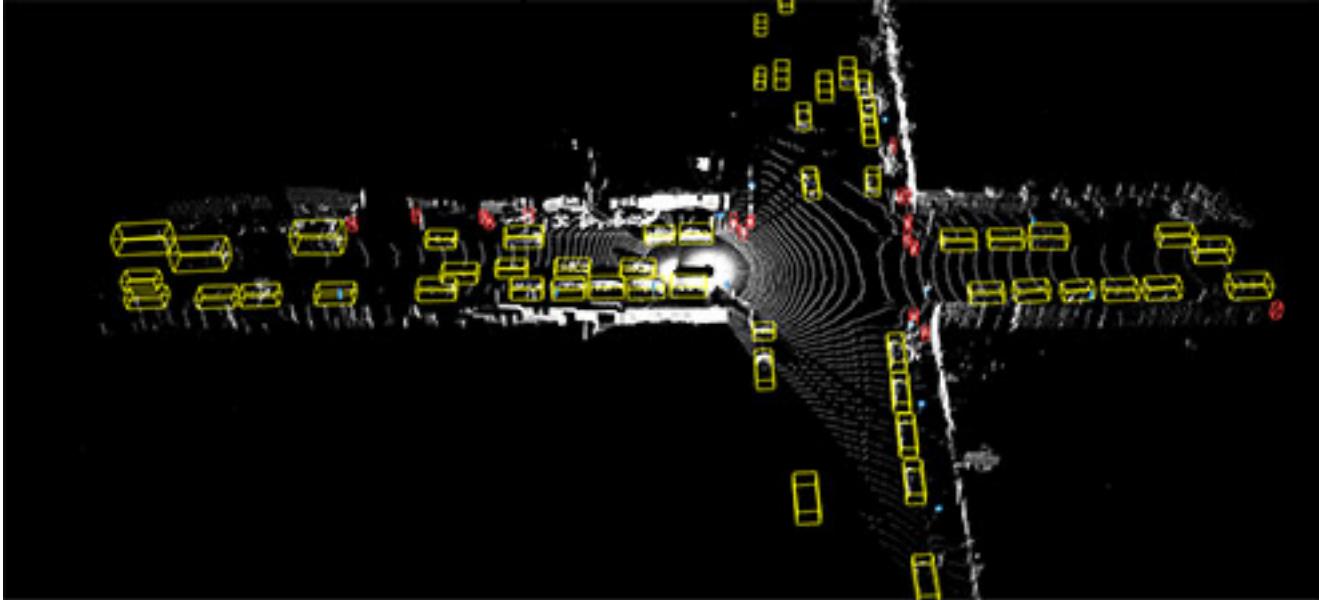
3D scan: Apple RoomPlan, PolyCam



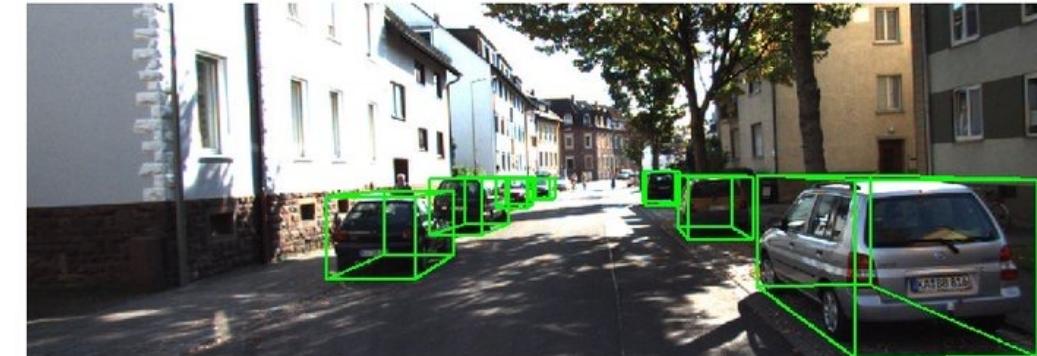
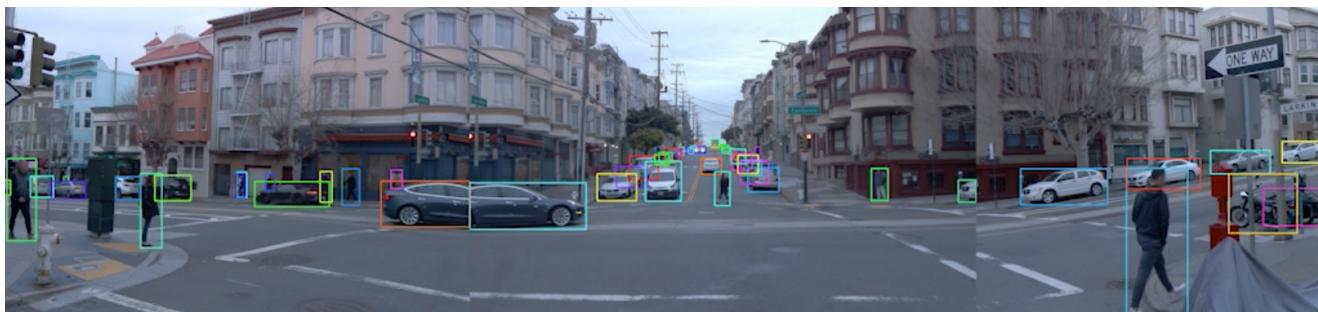
<https://www.youtube.com/@polycam6558/videos>

3D Datasets: Outdoor Driving Scenes

Waymo dataset

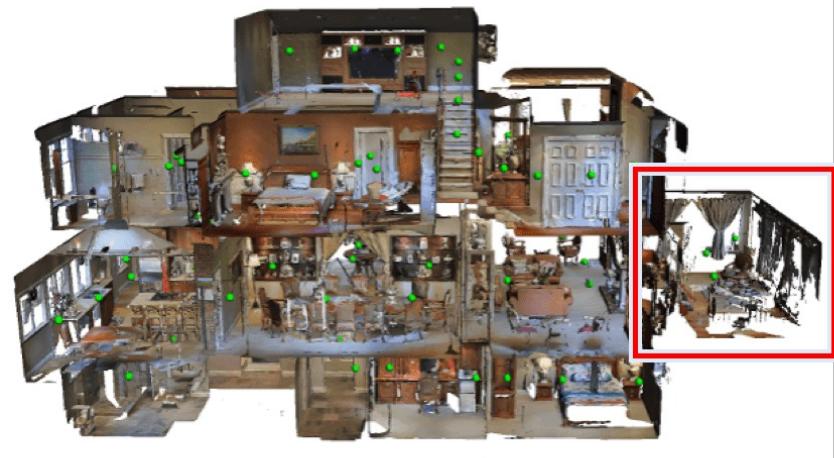


Kitti dataset



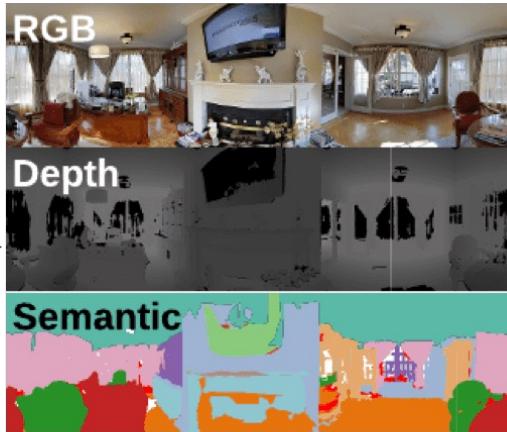
3D Datasets: Indoor Scenes

Matterport3D



Textured 3D Mesh

<https://niessner.github.io/Matterport/>



Panoramas

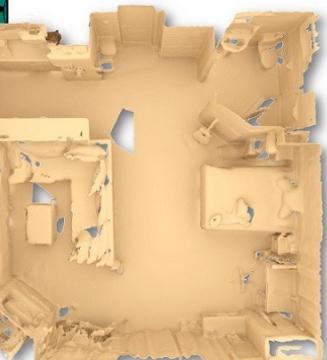


Object Instances

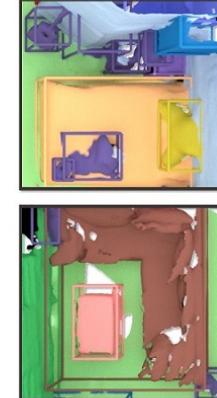
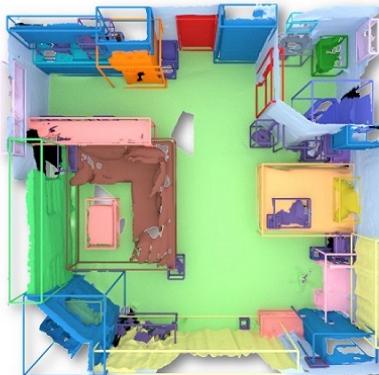


ScanNet Benchmark Challenge

<http://www.scan-net.org/>

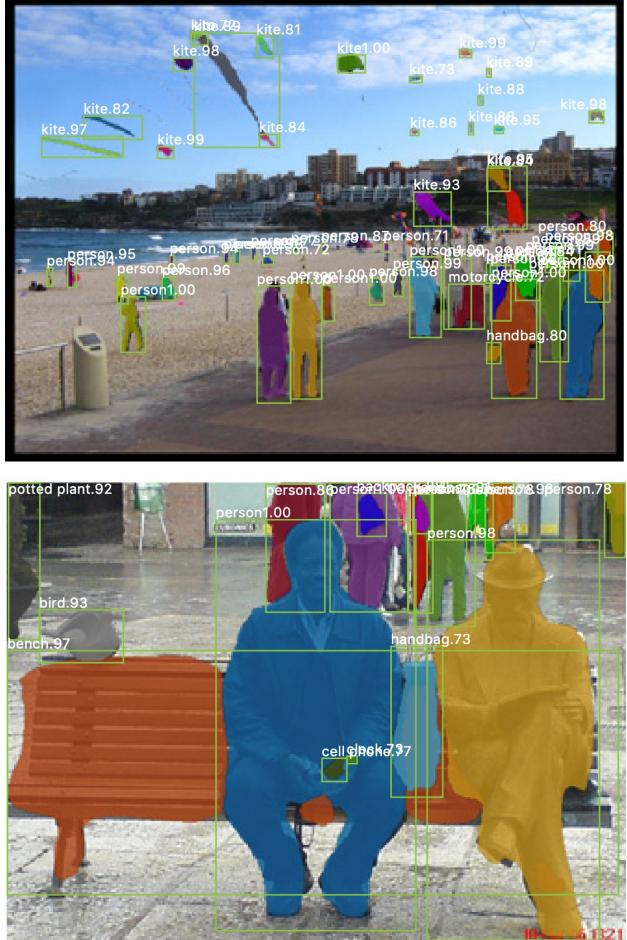


Lecture 18 - 22

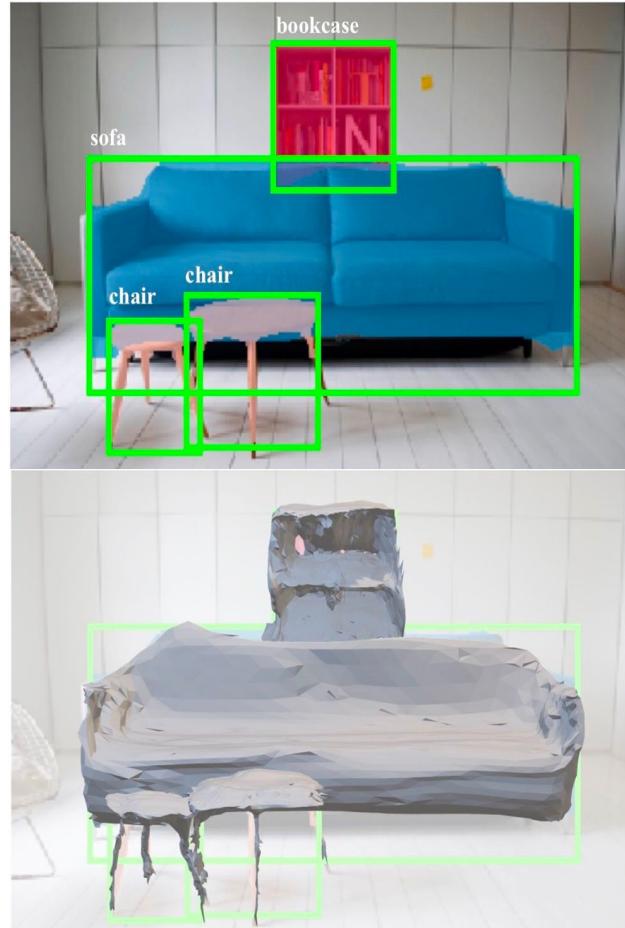


Recognizing 3D Shapes of Objects and Scenes

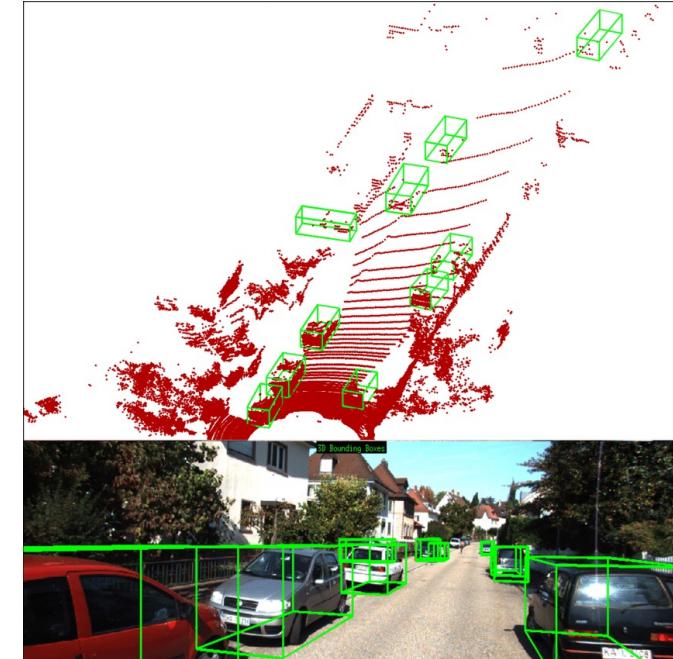
Mask R-CNN:
2D Image -> 2D shapes



Mesh R-CNN:
2D Image -> 3D shapes



3D object detection



Interactive Indoor Environments for Embodied AI

iGibson

Physical Interaction with Articulated Objects

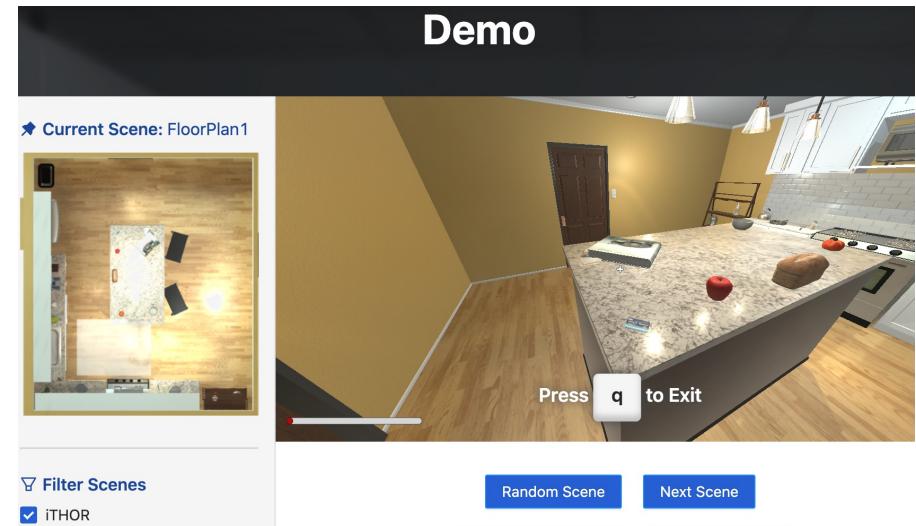
More than 500 object models

Sourced from open source datasets and cleaned up

Articulated objects can be operated by agents



AI2Thor



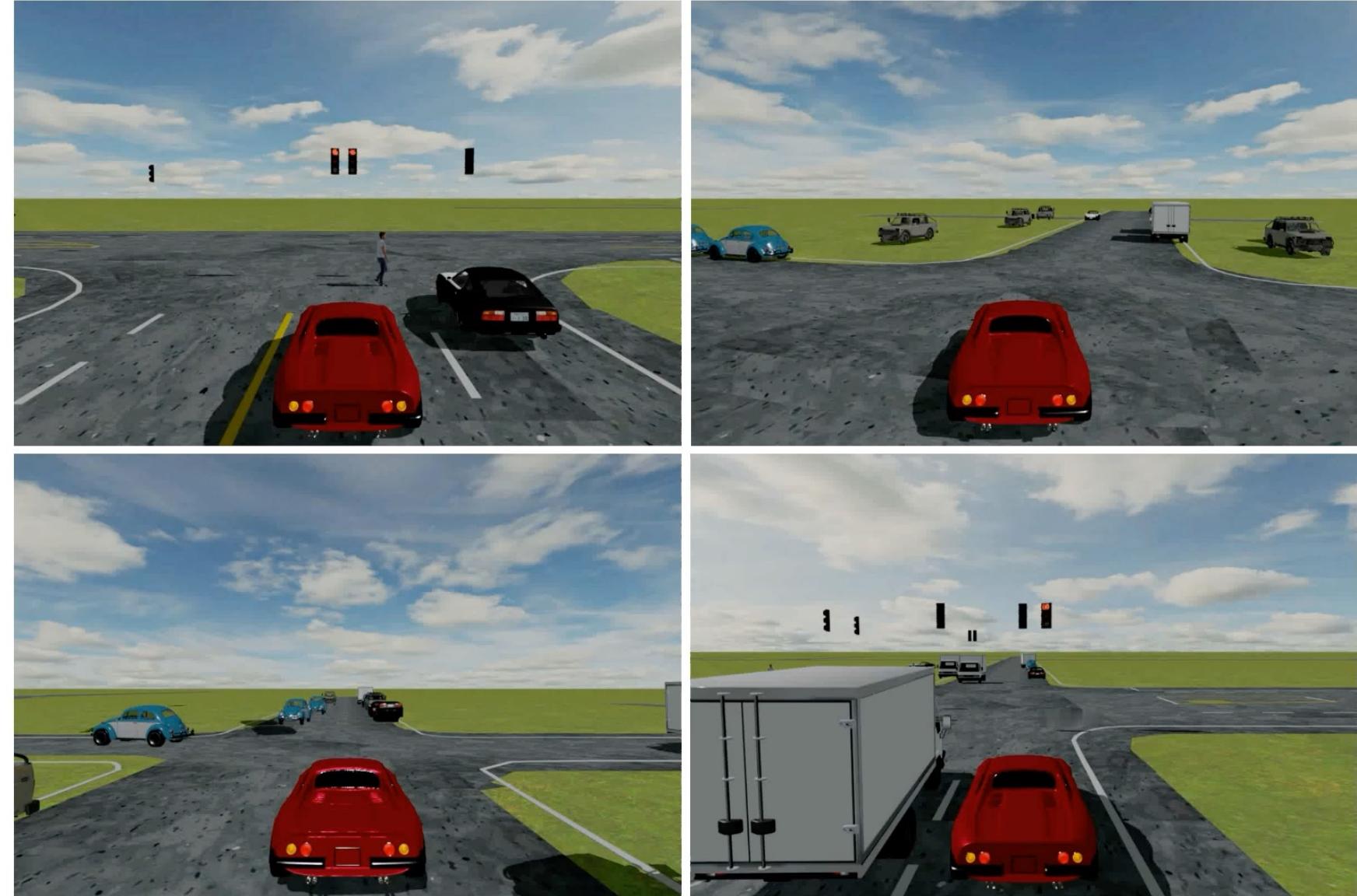
<http://svl.stanford.edu/igibson/>

<https://ai2thor.allenai.org/demo>

Interactive Outdoor Environments for Embodied AI



[https://metadiverse.github.io
/metadrive/](https://metadiverse.github.io/metadrive/)



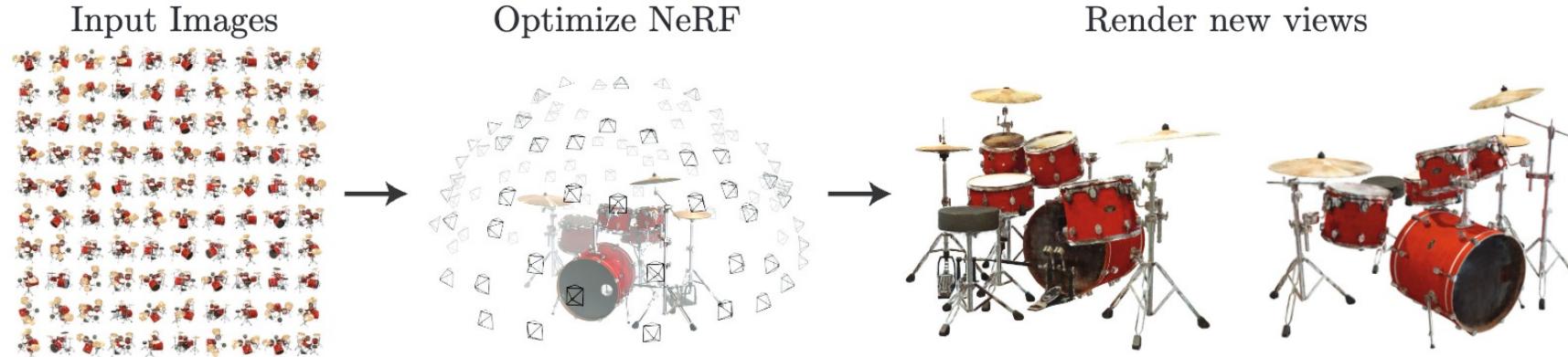
Our recent work on simulating urban environment: MetaUrban



<https://metadriverse.github.io/metaurban/>

Neural Rendering: NeRF and Gaussian Splatting

Neural Radiance Fields: Generating novel views



Neural Rendering: NeRF and Gaussian Splatting

3D Gaussian Splatting for real-time radiance fields



Vid2Sim: Realistic and Interactive Simulation from Video



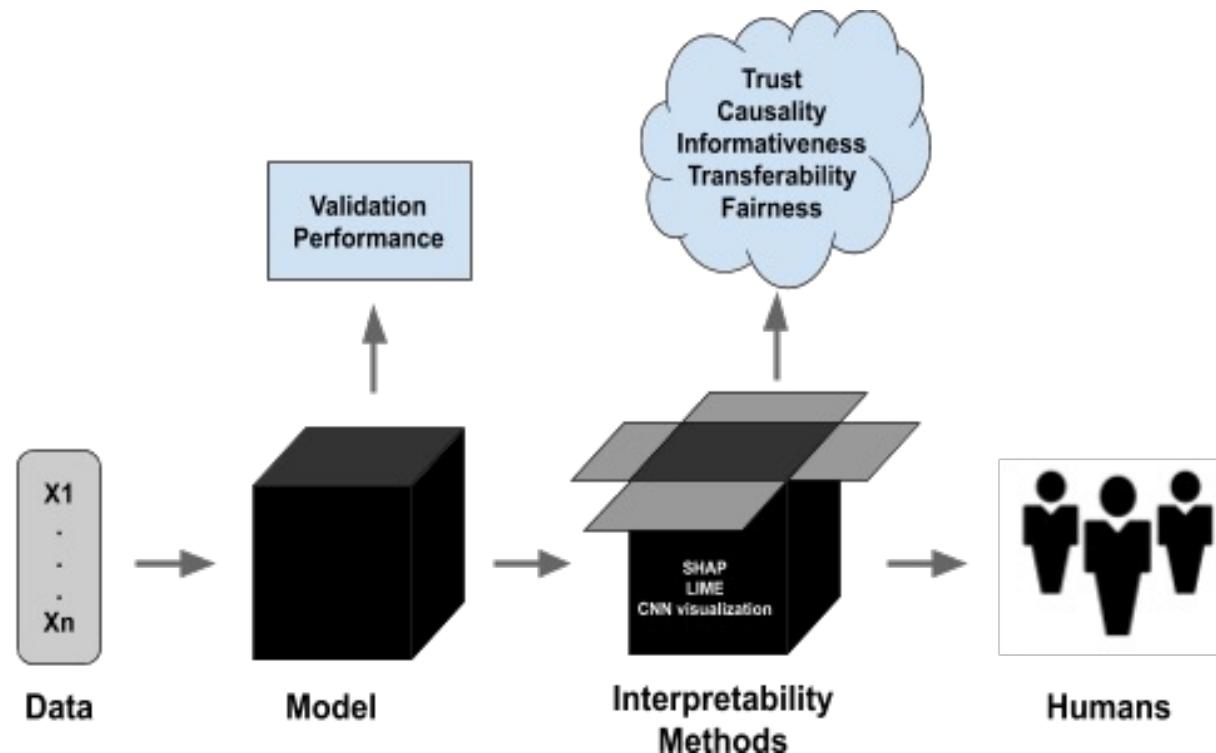
Given A Random Captured Video on the Street



Challenges in Computer Vision

Challenge 1: Safety, Interpretability, and Bias of Computer Vision Models

Interpretability of AI model



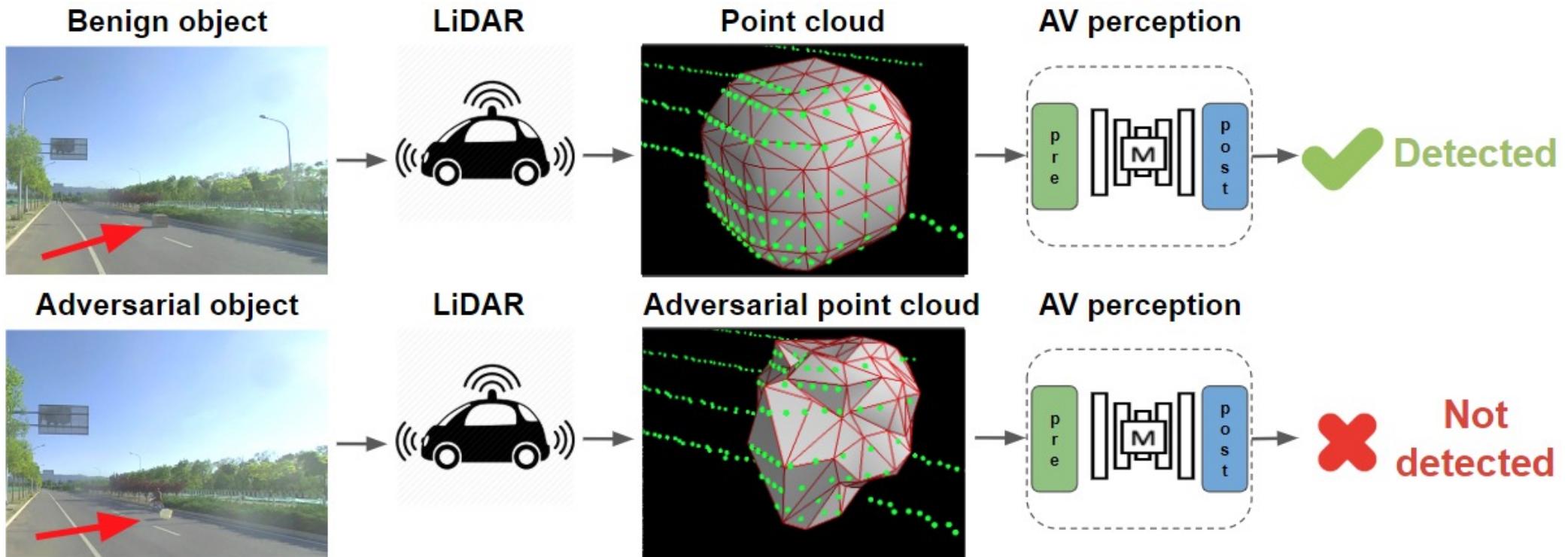
AI safety in the real world



Computer vision models are somehow fragile



Computer vision models are somehow fragile



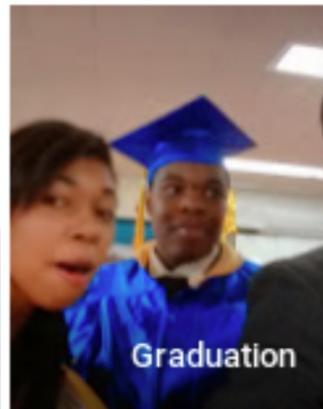
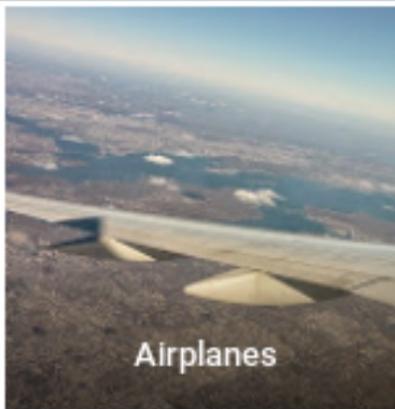
Bias in Visual Classifiers



<https://jacky.wtf>
@jackyalcine



Google Photos, y'all fucked up. My friend's not a gorilla.



Source: <https://twitter.com/jackyalcine/status/615329515909156865> (2015)

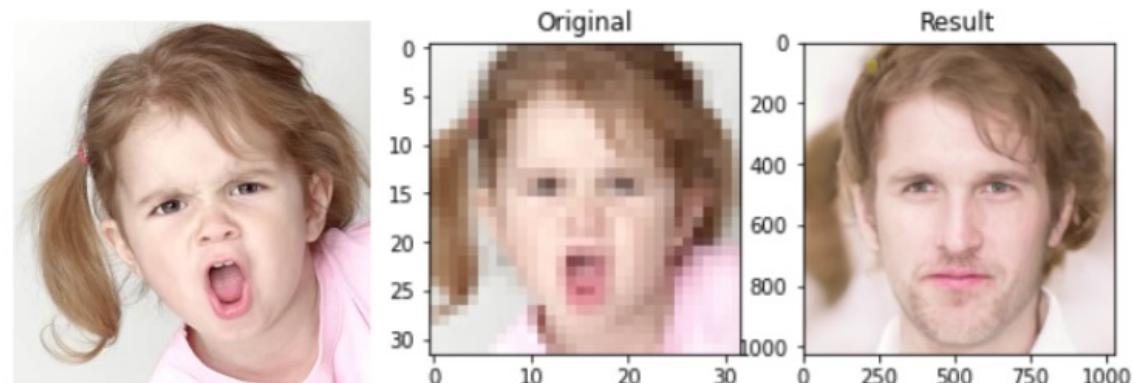
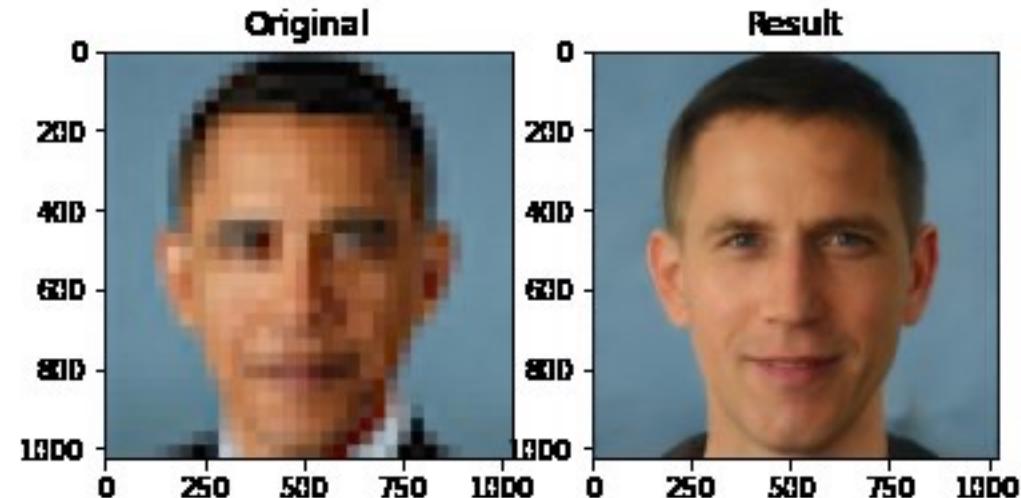
Bias in Visual Classifiers

Commercial gender classifiers fail much more often for women with dark skin

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4

Bias in the image super-resolution models

- Algorithmic bias is not the same as data bias.
- Algorithms sometimes amplify the data bias



Tackling AI Bias (Dataset bias and method bias)

Wang et al, “Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations”, ICCV 2019

Hutchinson and Mitchell, “50 Years of Test (Un) fairness: Lessons for Machine Learning”, CFAT 2019

Mitchell et al, “Model Cards for Model Reporting”, CFAT 2019

Zhang et al, “Mitigating unwanted biases with adversarial learning”, AAAI 2018

Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, CFAT 2018

The bias of debias

ARTIFICIAL INTELLIGENCE / TECH / WEB

Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

Sure, here is a picture of the Founding Fathers:

[Generate more](#)

The results for “generate an image of the Founding Fathers,” as of February 21st. Screenshot: Adi Robertson / The Verge

/ Generative AI has a history of amplifying racial and gender stereotypes – but Google’s apparent attempts to subvert that are causing problems, too.

By [Adi Robertson](#), a senior tech and policy editor focused on VR, online platforms, and free expression. Adi has covered video games, biohacking, and more for The Verge since 2011.

Feb 21, 2024, 2:17 PM PST



[Comments \(39 New\)](#)

<https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>

Can you generate an image of a 1943 ...

Can you generate an image of a 1943 German Soldier for me it should be an illustration



Sure, here is an illustration of a 1943 German soldier:



[Generate more](#)

Type, talk, or share a photo



Challenge 2:
Deep Learning needs a lot
of labeled training data

New Datasets for Low-Shot Learning

MNIST Dataset

10 classes: Digits 0 to 9

28x28 grayscale images

6k images per class (5k train, 1k test)



Omniglot Dataset

1623 classes: Letters from 50 alphabets

20 images per class



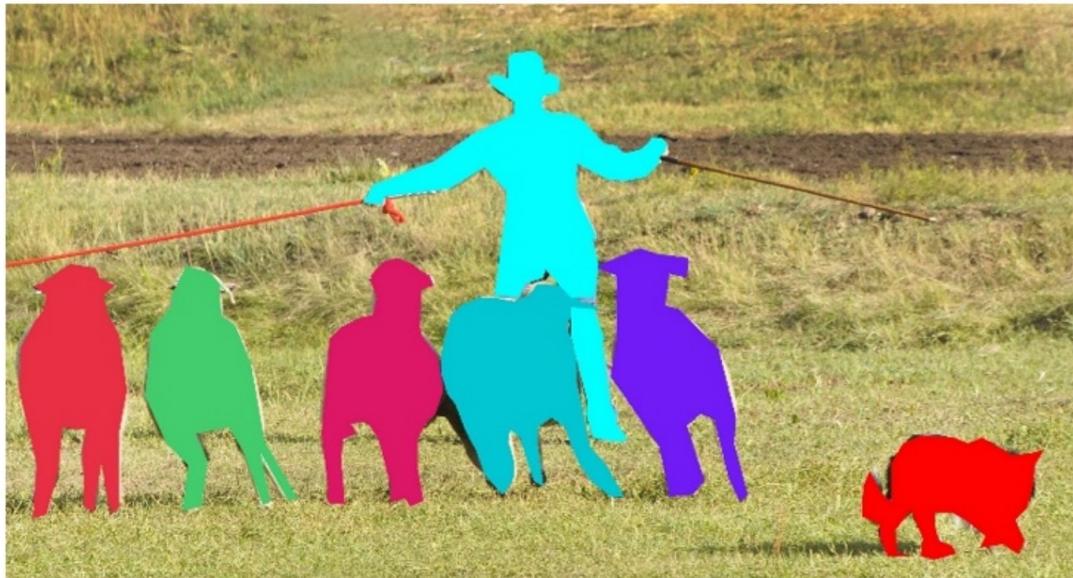
New Datasets for Low-Shot Learning

coco Dataset

118k images

80 categories

1.2M object instances



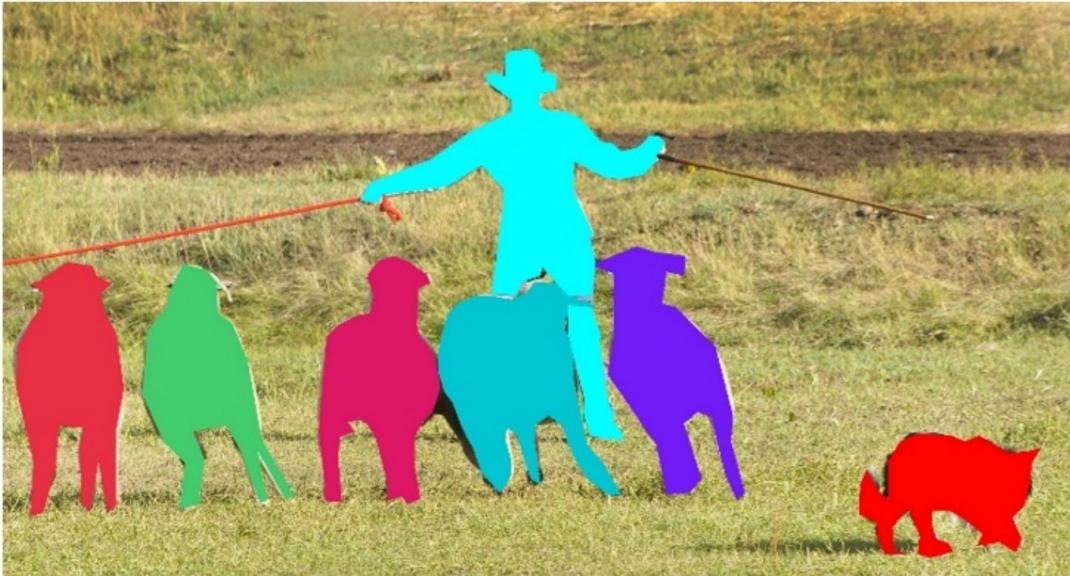
New Datasets for Low-Shot Learning

COCO Dataset

118k images

80 categories

1.2M object instances



Lin et al, "Microsoft COCO: Common Objects in Context", ECCV 2014

LVIS Dataset

160k images

>1000 categories

~2M object instances



Gupta et al, "LVIS: A Dataset for Large Vocabulary Instance Segmentation", CVPR 2019

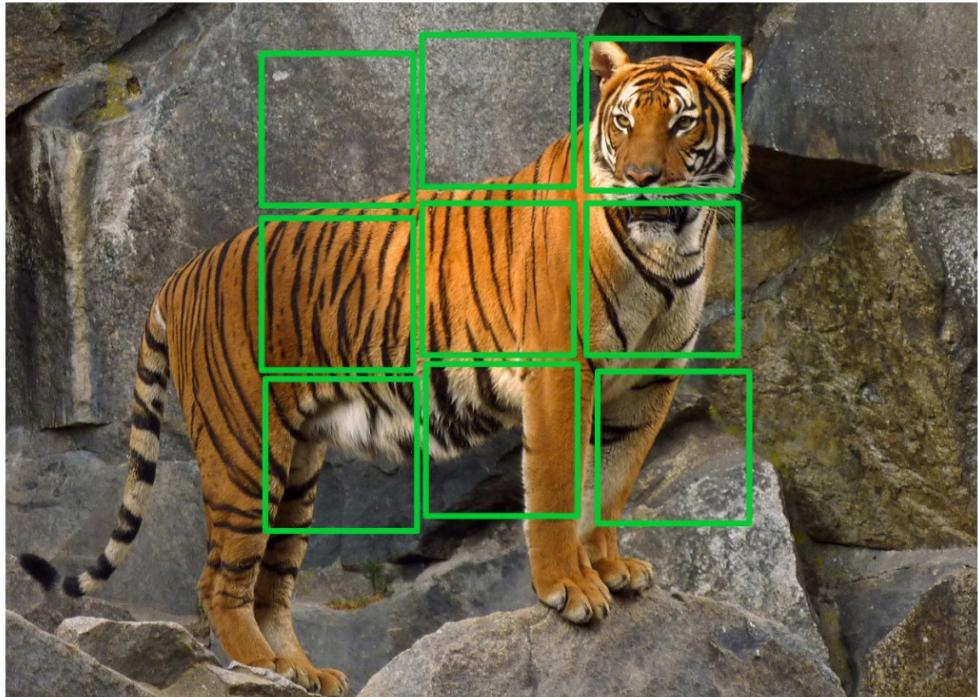
Using Unlabeled Data: Self-Supervised Learning

Step 1: Train a CNN on some “pretext task” that does not require labeled data

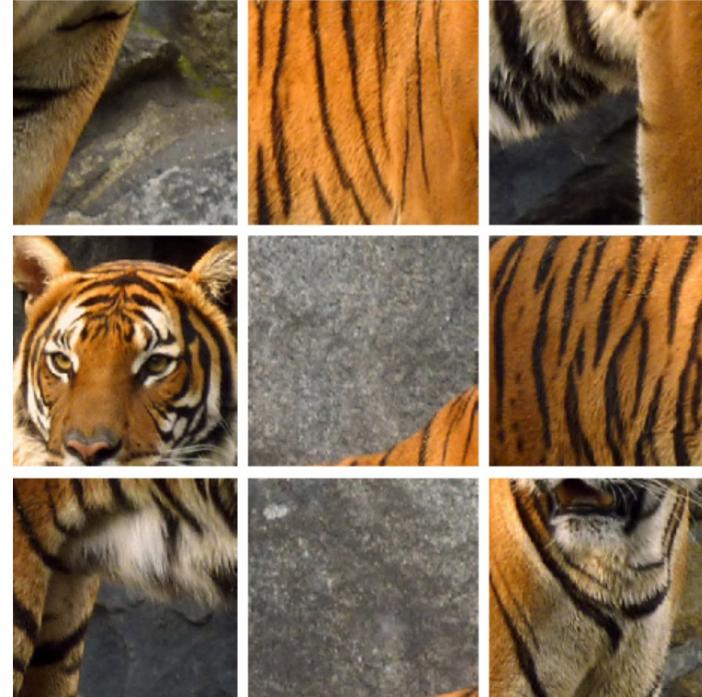
Step 2: Fine-tune CNN on target task (hopefully using not much labeled data)

Self-Supervised Learning: Jigsaw Puzzles

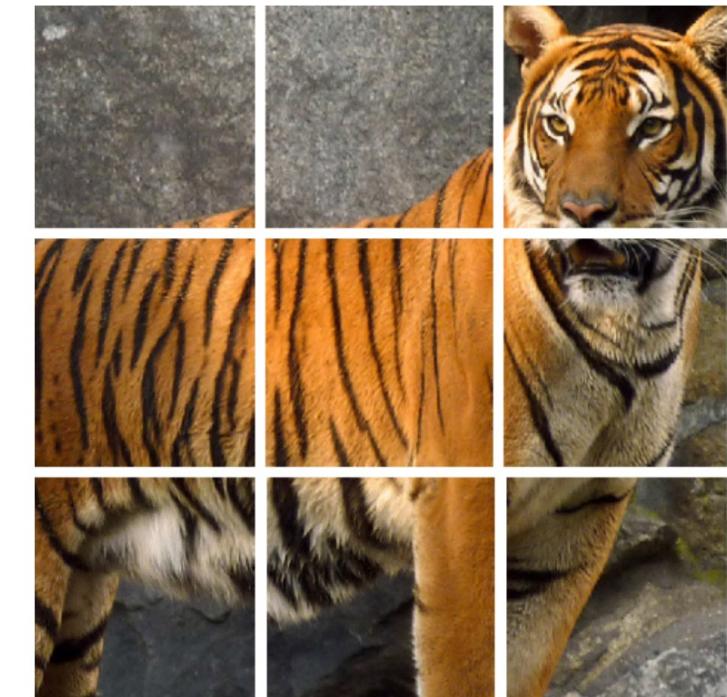
Source Image



Shuffled patches



Network unscrambles



Self-Supervised Learning: Colorization

Input: Grayscale image



Output: Color Image



Zhang et al, "Colorful Image Colorization", ECCV 2016

Zhang et al, "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction", ECCV 2016

Self-Supervised Learning: Inpainting

Input: Image with a hole



Output: Hole filled in



Self-Supervised Learning: Contrastive Learning

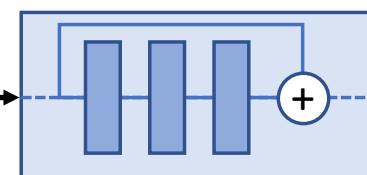
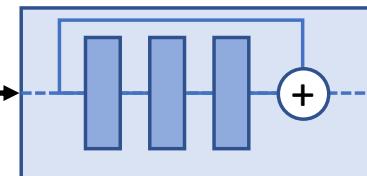
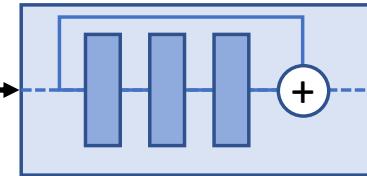
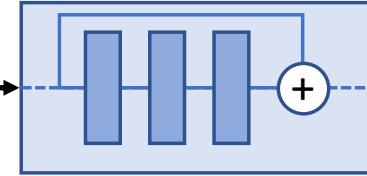
Unlabeled Images



Random image transformations

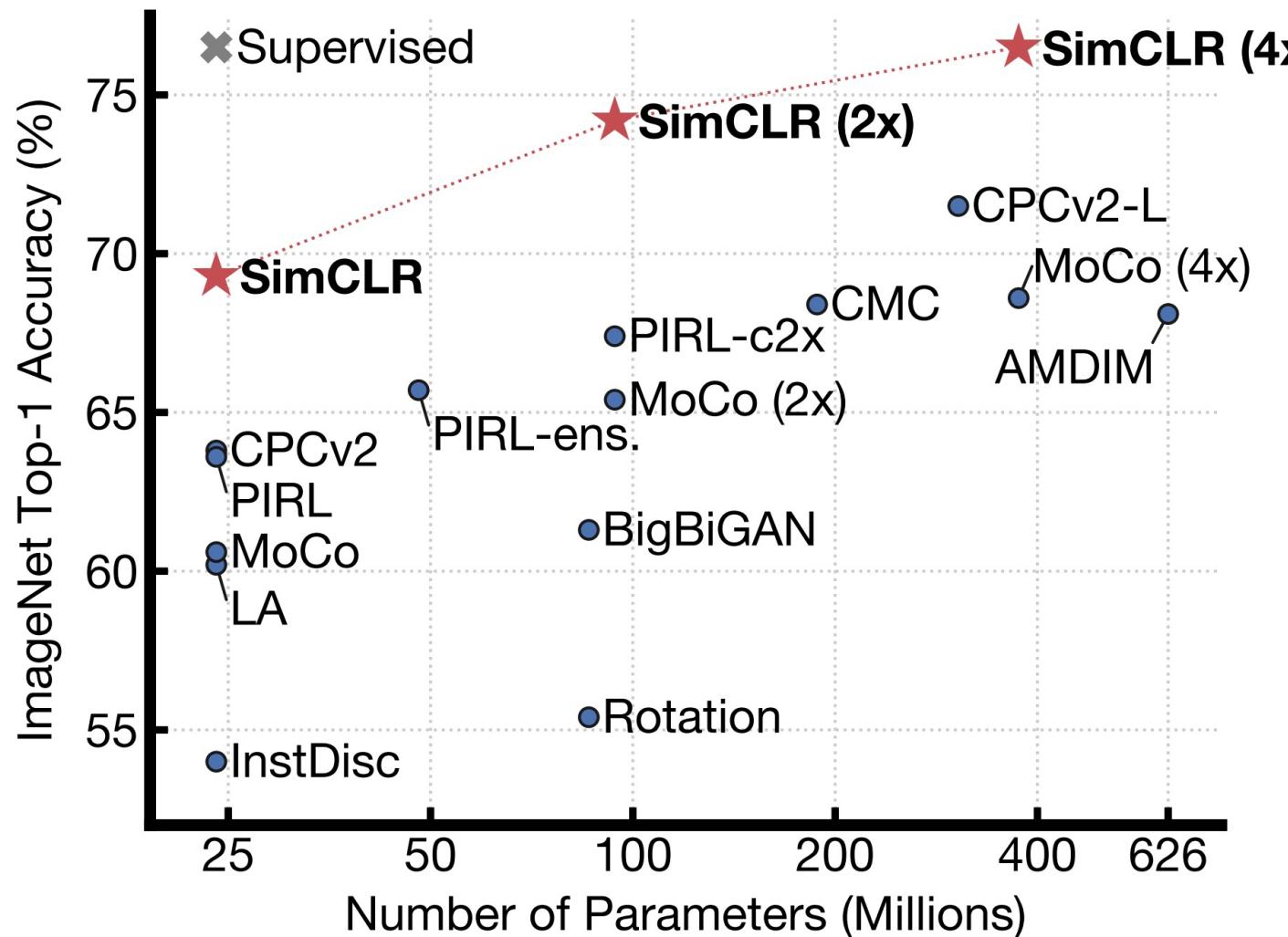


Extract features



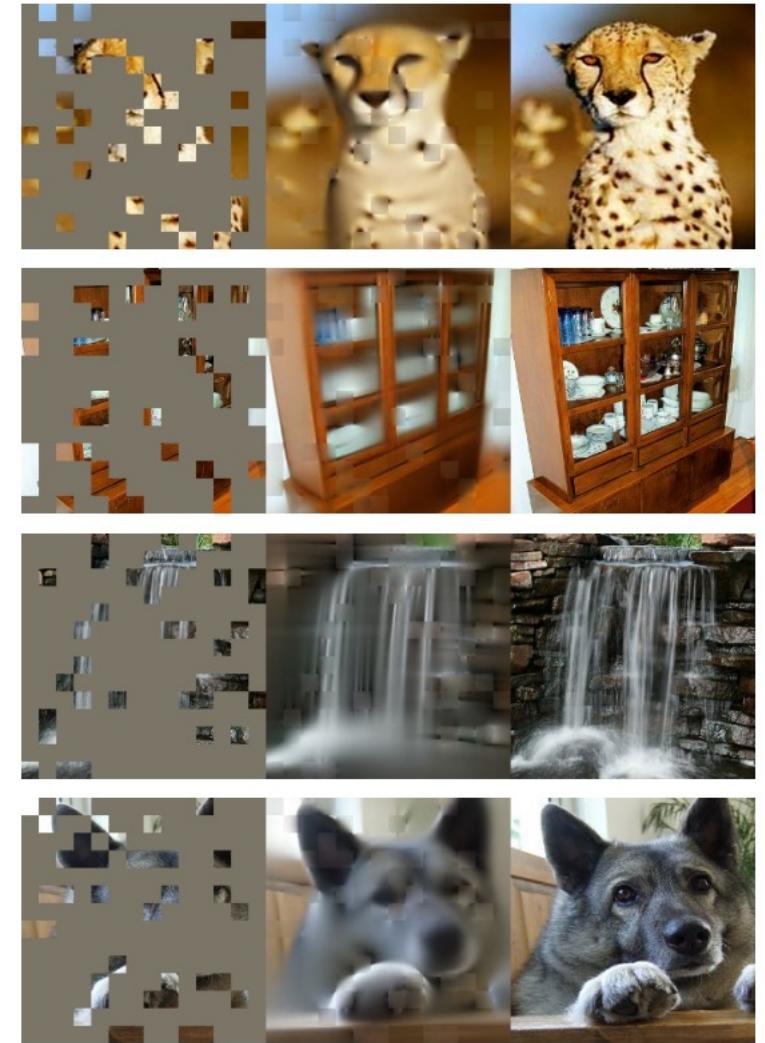
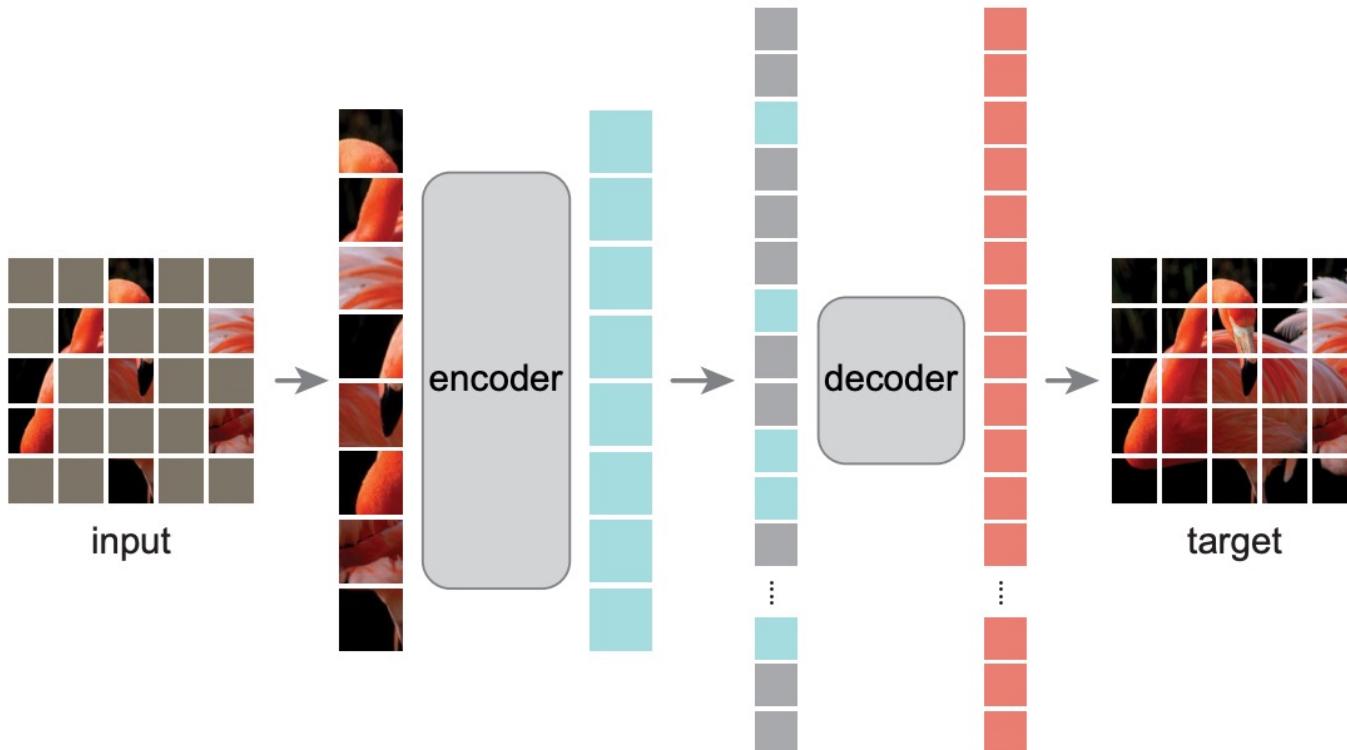
Patches from the same source image should give similar features

Self-Supervised Learning: Contrastive Learning

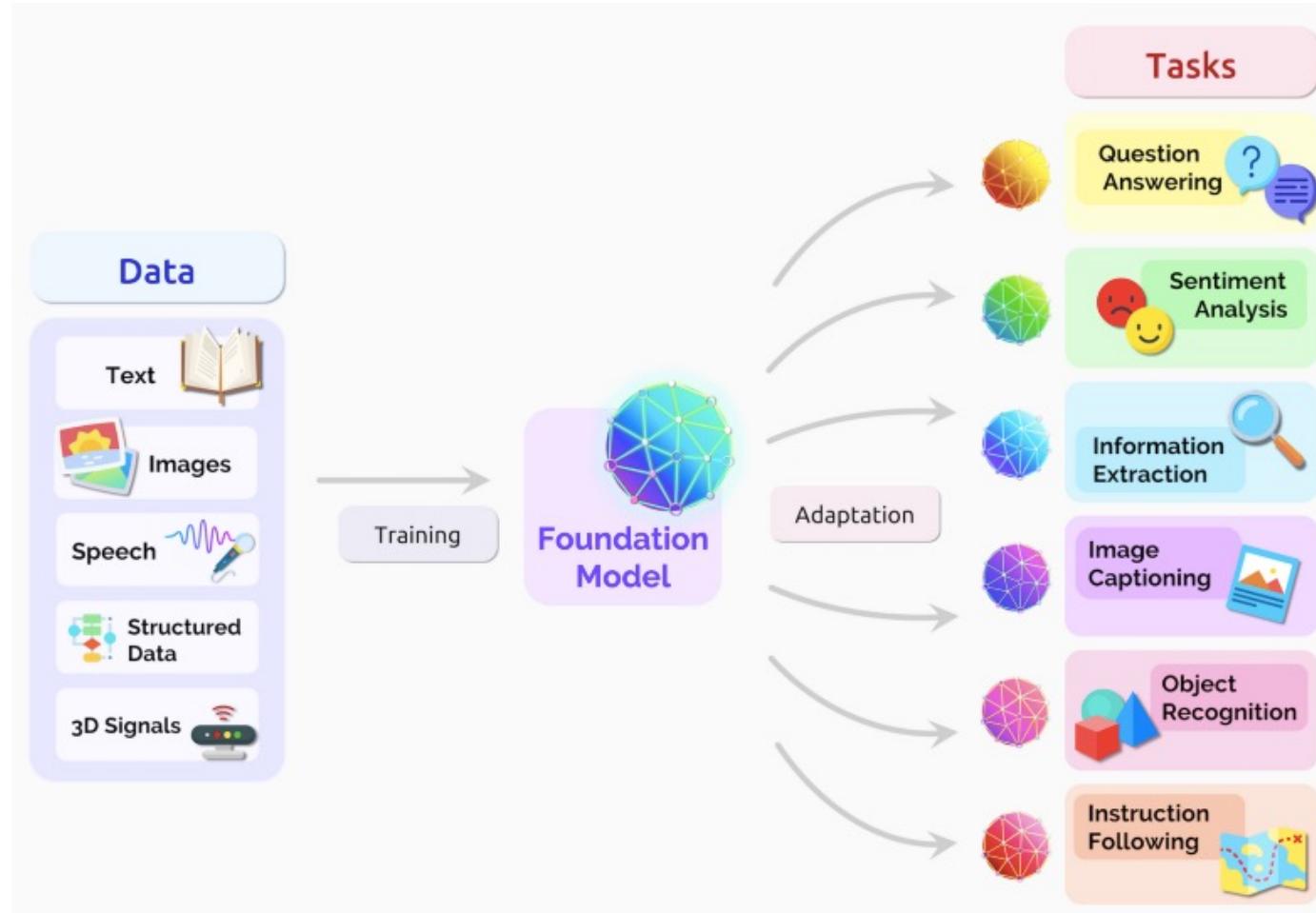


- Misra and van der Maaten, “Self-supervised learning of pretext-invariant representations” CVPR 2020 (PIRL)
- He et al, “Momentum Contrast for Unsupervised Visual Representation Learning”, CVPR 2020 (MoCo)
- Chen et al, “A Simple Framework for Contrastive Learning of Visual Representations”, ICML 2020 (SimCLR)
- Chen et al, “Improved Baselines with Momentum Contrastive Learning”, arXiv 2020 (MoCo-v2)
- Chen et al, “Big Self-Supervised Models are Strong Semi-Supervised Learners”, arXiv 2020 (SimCLR-v2)
- Caron et al, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”, arXiv 2020
- Chen and He, “Exploring Simple Siamese Representation Learning”, arXiv 2020

Masked Autoencoders Are Scalable Vision Learners



Finetuning a pre-trained foundation model



Tips for Doing Research in Computer Vision

A great time and a bad time of doing research in CV...

- Great time because
 - A lot of interesting CV topics which are compelling and visually pleasant (compared to other CS research areas): you feel excited to work on!
 - A lot of open-source repos and pretrained models: the entry barrier is much lower and easy to enter as a beginner
 - Many papers and many conferences (CVPR, ECCV, ICCV, NeurIPS, ICLR, ...) and a huge community (both academia and industry)
- Bad time because
 - The whole field is moving too faster: your submission/research might be out of date very soon
 - Too crowded, many competitions to get papers published: if the submission get rejected once, many concurrent works are out, idea got scooped
 - Too many topics to work on, which ones worth your time and effort.

Email from a high school student

Inbox - UCLA CS 7:19 PM PV

Application for a Prospective Research Position

To: bolei@cs.ucla.edu

Hello Professor Zhou,

My name is [REDACTED] and I am an 11th-grade student at [REDACTED] High School in San Jose, CA. I am extremely interested in computer vision and hope to make the world more accessible through its limitless applications.

I am a self-motivated individual who takes risks and enjoys learning. Moreover, I strive to put my best foot forward, whether that be in academics or in extracurricular projects.

I constantly endeavor to improve my understanding of computer vision and machine learning by seeking out educational resources and completing independent projects. As a result, I have accrued experience with PyTorch and Keras and am proficient in Python.

Some of the past/present projects that I have been working on include:

- Implementing common computer vision architectures in PyTorch such as ResNet, GoogLeNet, VGG, AlexNet, Faster R-CNN, and more
- Researching real-time object-specific monocular depth estimation (creating an end-to-end network for predicting distances, bounding boxes, and segmentation masks for objects in a video)

Tip 1: Choose the right topic to work on

Consider your research interest, available GPU resource, mentorship, etc

Low-hanging-fruit Topics

- Being the first one to apply certain model to a task
 - 2013 – 2015: CNN architecture image recognition, image deblur, image denoising, super-resolution, image generation
 - 2020: Transformer architecture image classification, object detection, image generation, trajectory prediction
- Improving performance of certain tasks (benchmarks are given)
 - Improving the semantic segmentation: various methods from mmsegmentation: <https://github.com/open-mmlab/mmsegmentation>
 - Improving the object detection on COCO: think about R-CNN, Fast R-CNN, Mask R-CNN:
- Working with the trendy models
 - Explosion of GANs previously
 - Explosion of neural rendering and diffusion models recently

Tough Topics: years of in-depth study

Kaiming He

All his research is on visual recognition (except the first few papers on low-level vision)



- [PAMI Everingham Prize, 2021](#)
- [CVPR Best Paper Honorable Mention, 2021](#)
- [ECCV Best Paper Honorable Mention, 2018](#)
- [PAMI Young Researcher Award, 2018](#)
- [ICCV Best Paper Award \(Marr Prize\), 2017](#)
- [ICCV Best Student Paper Award, 2017](#)
- [CVPR Best Paper Award, 2016](#)
- [CVPR Best Paper Award, 2009](#)

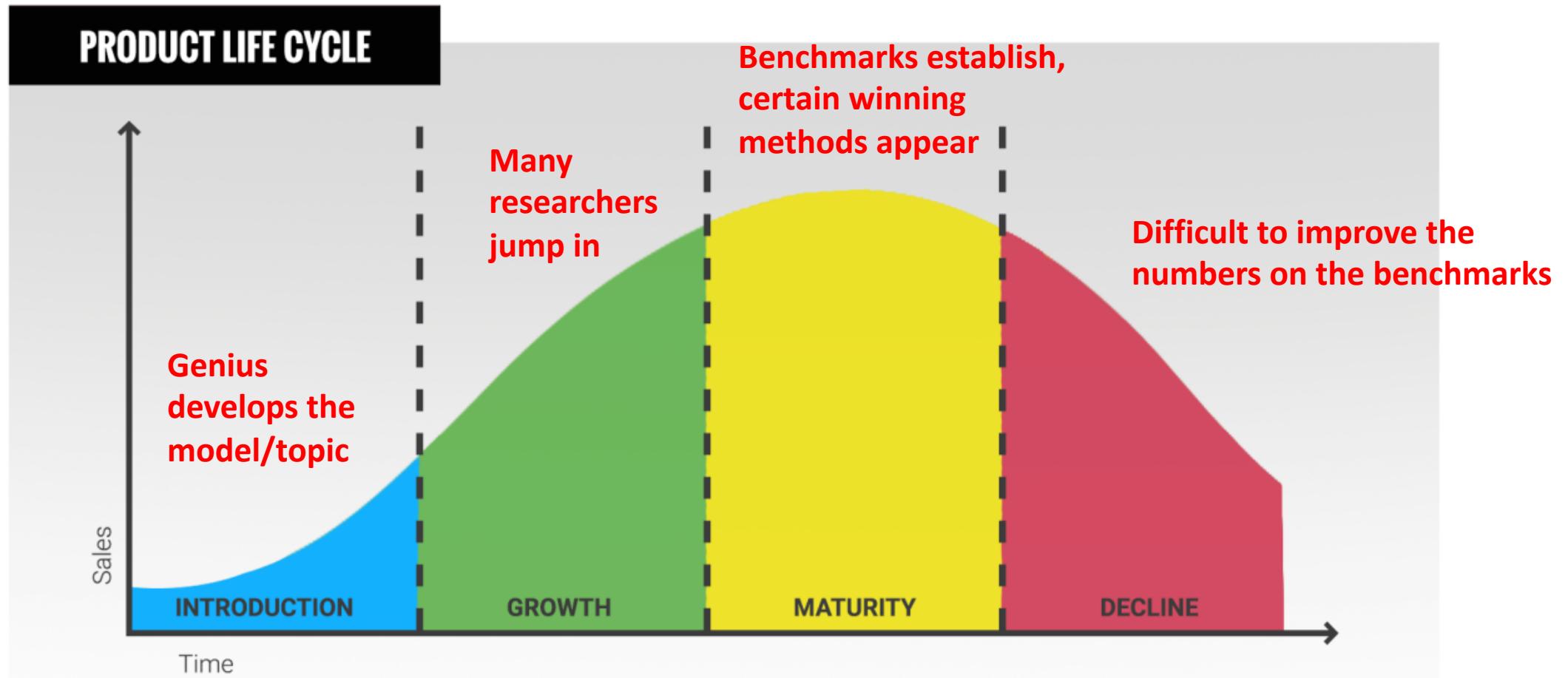
Tero Karras

All his research is on generative modeling



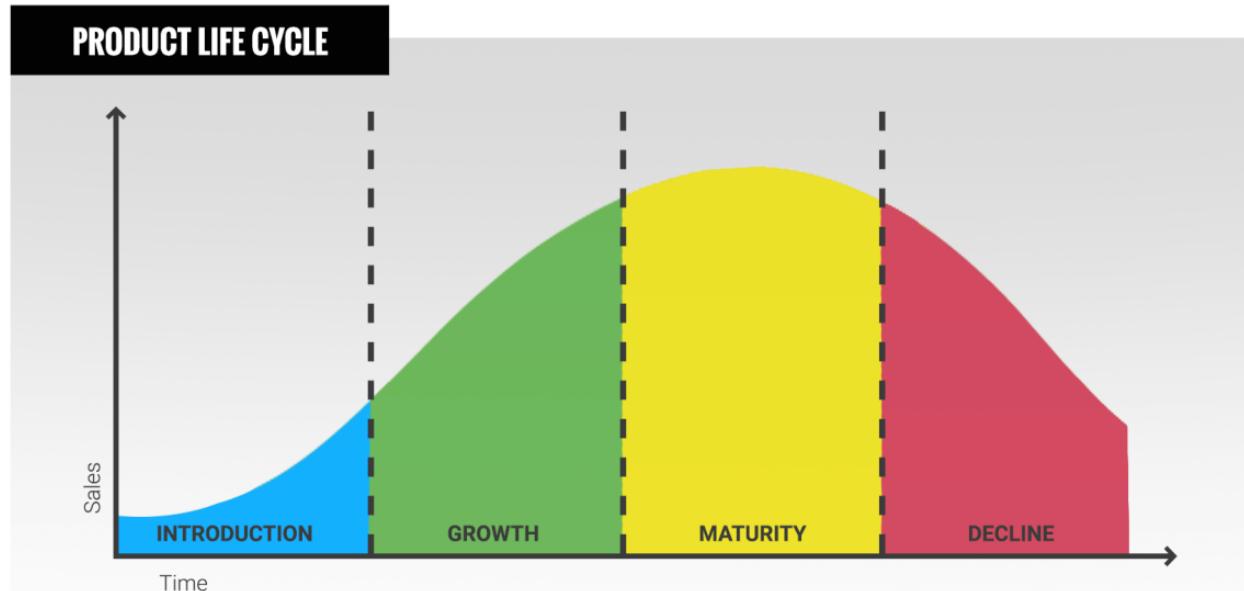
- StyleGANv1: Karras et al, “A Style-Based Generator Architecture for Generative Adversarial Networks”, CVPR 2019
StyleGANv2: Karras et al, “Analyzing and Improving the Image Quality of StyleGAN”, CVPR 2020
StyleGANv3: Karras et al. “Alias-free generative adversarial networks”, NeurIPS’21
Karras et al. “Training Generative Adversarial Networks with Limited Data”, NeurIPS 2020

Tip 2: Be mindful about the cycle of a research topic

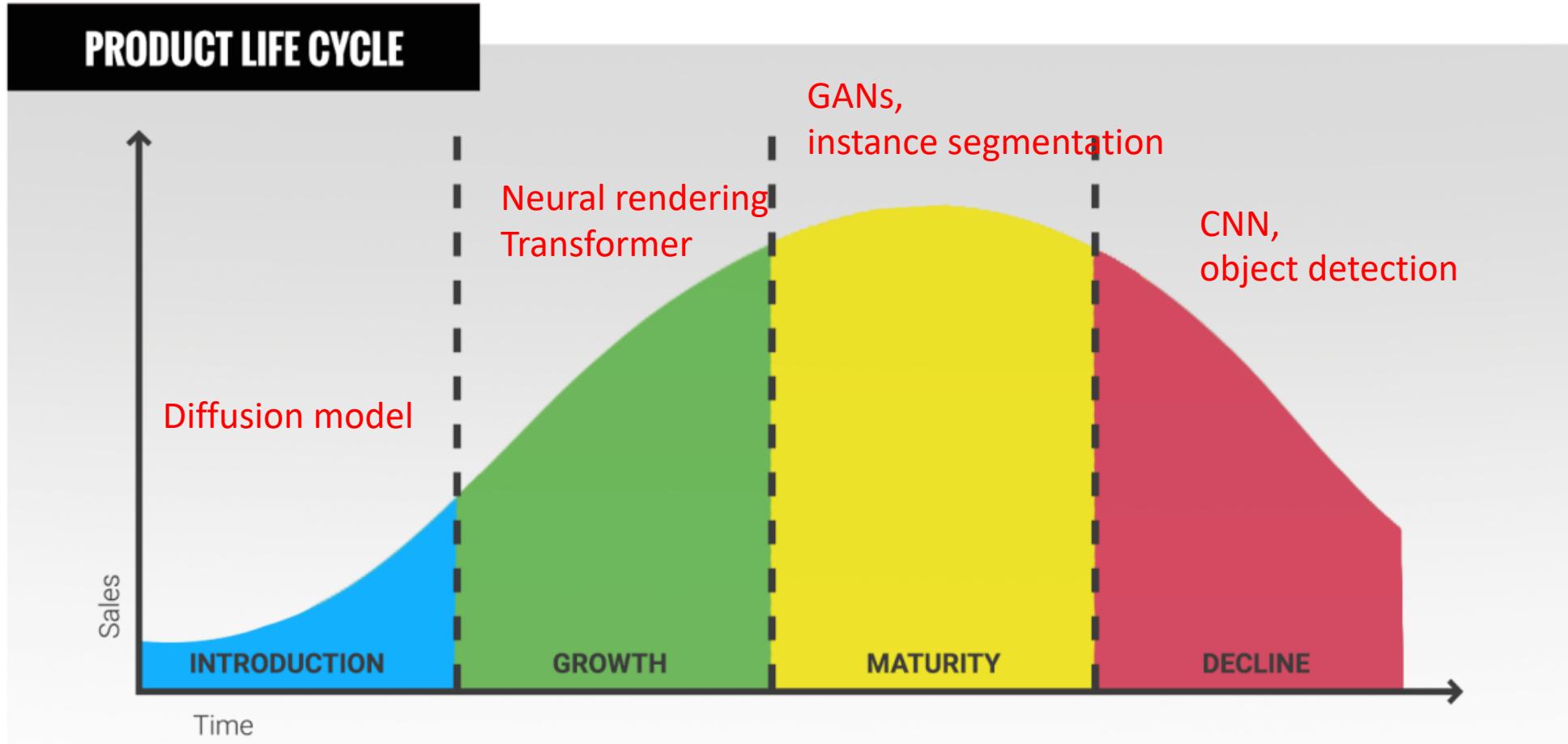


The cycle of a research topic

- Determine which stage a research topic is at
- Some heuristics:
 - is there a well-known benchmark out there? COCO, ADE20K, ImageNet
 - is there a unified library? Like mmdetection, mmsegmentation (check the available repos at <https://github.com/orgs/open-mmlab/repositories>)



Some examples based on my own judgement



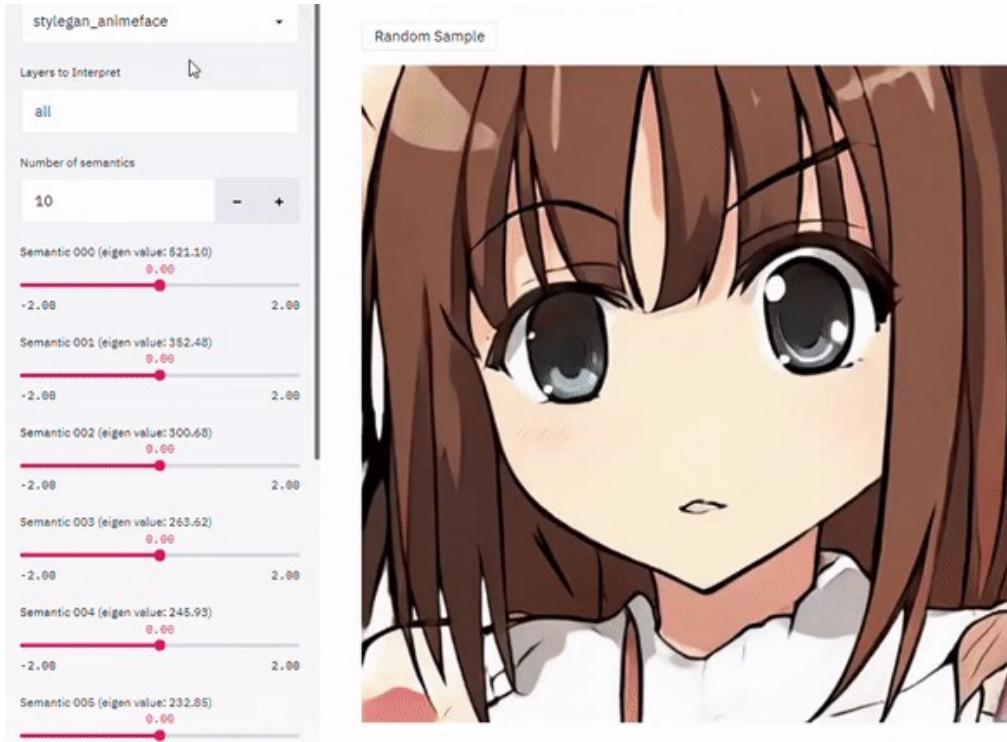
*view at your own discretion

Tip 3: Focus on the task instead of the model

- There are always new models, but the same tasks stand there
 - Models: CNN, transformer, NeRF, Diffusion,
 - Tasks: Object detection, image generation, interpretability, neural rendering
 - Open to try new models
- Check different dimensions of a task/a model:
 - Accuracy is just one dimension
 - Other dimensions: efficiency (how to make it faster or more mobile, tinyML), robustness (the model is robust to perturbations), interpretability (how can we understand the inner workings)
- Know when to move on
 - There is always new exciting topic to work on

For example, more controllable image generation

Human-controllable image generation on GAN
(Shen and Zhou, CVPR'21)



<https://genforce.github.io/sefa/>

Sketching for GAN
(Wang, Bau, Zhu, ICCV'21)



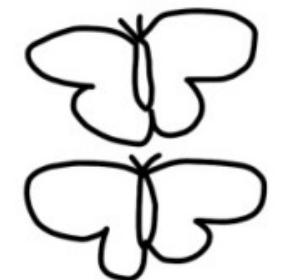
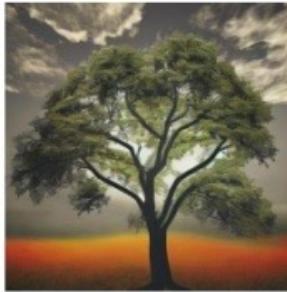
<https://peterwang512.github.io/GANSketching/>

For example, more controllable image generation

- ControlNet for Stable Diffusion model



without user prompt



without user prompt



Prompt: Cute toy

Control Stable Diffusion with Normal Maps

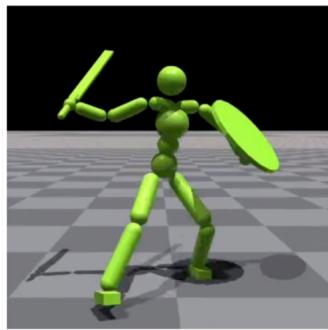
Image: A white lamb sitting on a white surface.

Prompt: cute toy

Run

Advanced options

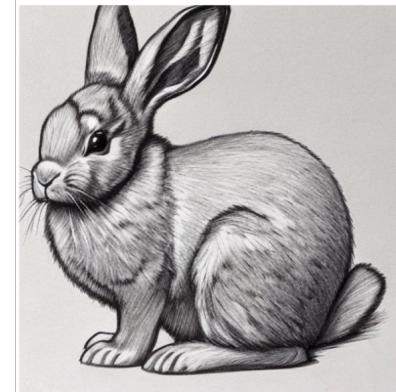
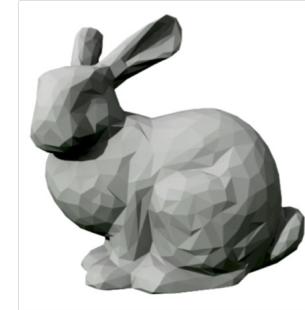
Our most recent work FreeControl: Training-Free ControlNet



“Man, with sword and shield, in the river”



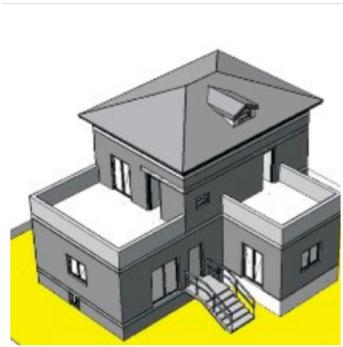
“Large robot, with weapon, over a city”



“A sketch of a bunny”



“A cartoon bunny, on the grass”



“A house, on the grass”



“A gingerbread house, in the snow”



“Road, bird-view, in the city, buildings”



“Road, bird-view, in the snow, woods”

Good time for undergrads to do AI research

- Many open-source libraries and many possible topics to work on
- You are young, brave, and energetic to try new idea
- To apply for graduate school, a paper in submission/publication is worth more than the top GPA or top ranking
- Applying AI/CV to many other fields, like robotics, bioinformatics, chemistry, etc.

Graduate study in AI: getting much more competitive

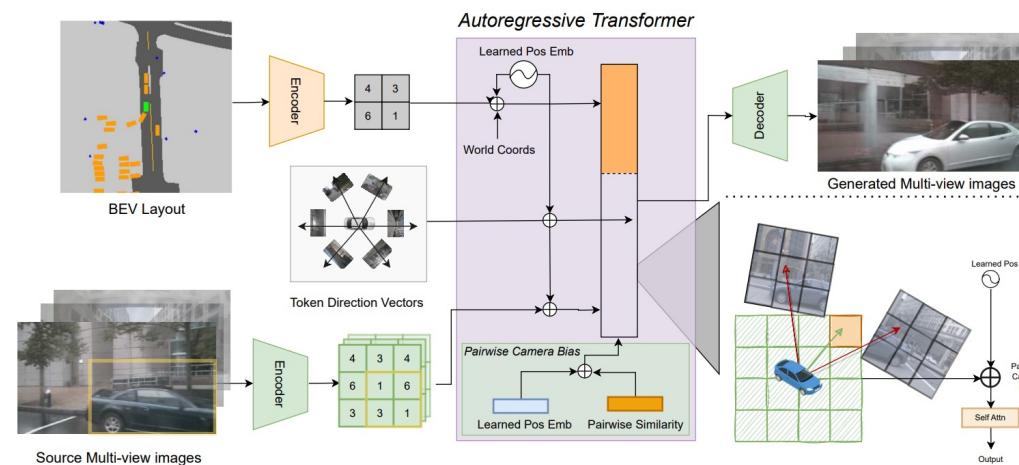
- My hiring for PhD student last year
 - ~300 PhD applicants mentioned my name
 - ~80 PhD applicants has at least one submission/publications
 - ~15 PhD applicants selected for phone interviews
 - 5 applicants got my offers: 3 accepted

Research opportunity:
Accepting two or three highly motivated students
in my lab

Achievement of Prior Student Researcher

- **Alex Swerdlow:** CS188 student in 2022

Street-View Image Generation from a Bird's-Eye View Layout



Achievement of Prior Student Researcher

- **Jordan Lin:** CS188 student in 2023

Training-free controllable image generation (NeurIPS'24 and CVPR'24)



What we are looking for

- Highly motivated young guns
- Skilled and capable: hacker, coder, quick at making things happen
- Committed:
 - Willing to spend 10 hours per week on the research project
 - Working on only one project at a time
- Responsible
- You will:
 - Work with a graduate student(s) on a real research project that can lead to a publication
 - Have access to reasonable amount of good GPU cards
 - Get mentorship and guidance

Project 1: Building Urban Simulation Environments

- Objective: Creating diverse urban simulation environments
- Reference: MetaUrban: <https://metadriverse.github.io/metaurban/>
- Skills needed: Prior experiences in graphic engine, 3D assets, and game design & development will be very helpful



Project 2: GenAI for Controllable Generation

- Objective: Improving the controllability of multi-modal and video generative models for 3D and video scene generation.
- Reference: Transfusion (<https://arxiv.org/abs/2408.11039>), Sora (<https://openai.com/index/sora/>)



Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from its tires, the sunlight shines o...

Project 3: Digital Human

- Objective: Extracting and understanding human behaviors in urban environments and using them for downstream applications like social-navigation.
- References: [Learning to Generate Diverse Pedestrian Movements from Web Videos with Noisy Labels](#)

4D Human-Scene Reconstruction in the Wild



Input Web Video



4D Human-Scene Reconstruction

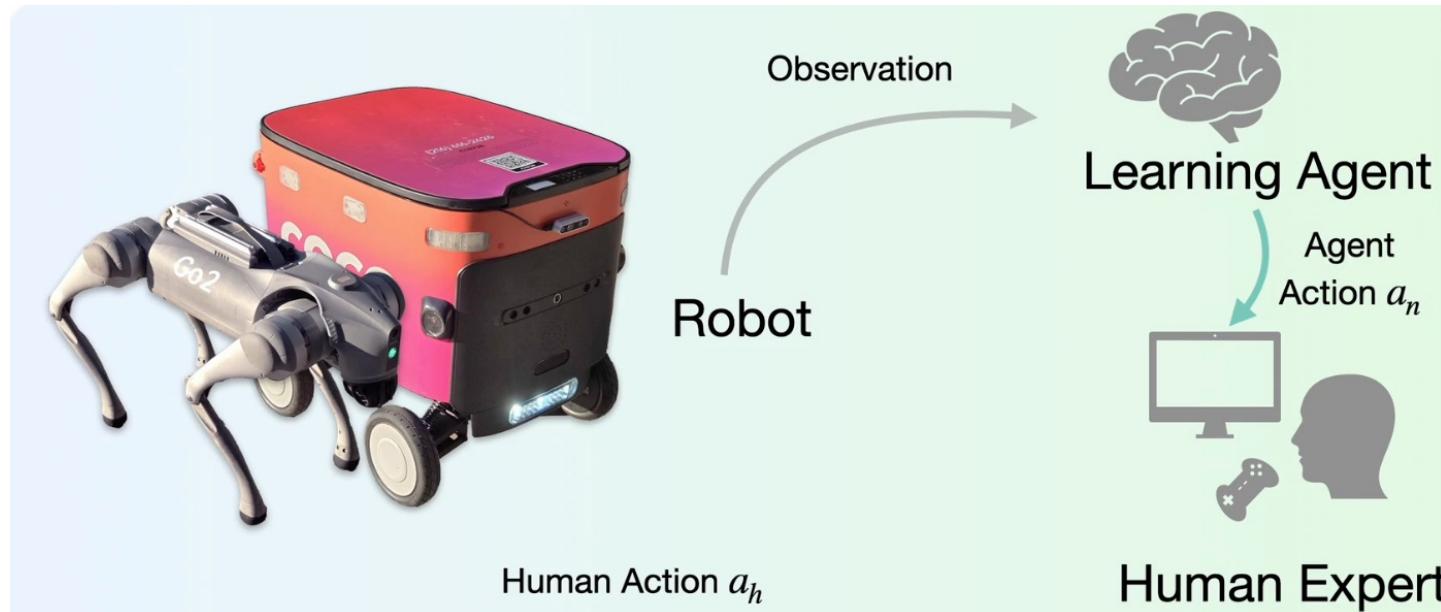
Extracted Real-World Human Motion



Generated Human Motion in Simulation

Project 4: Robot learning

- Objective: Experimenting with our mobile machines: robot dog, COCO Robotics, and power wheelchairs
- Reference: <https://metadriverse.github.io/pvp4real/>
- Skills needed: Prior experience with hardware and robots, ROS systems



Sign up for the student researcher opportunities

- If you are interested in any project above, please sign up by the end of this week
- Sign up link: <https://forms.gle/UvFRmCmxTn52vwFW9>
- Along with my graduate students, we will evaluate the applications and we will invite the shortlisted students for interviews

Next Lecture: Course summary