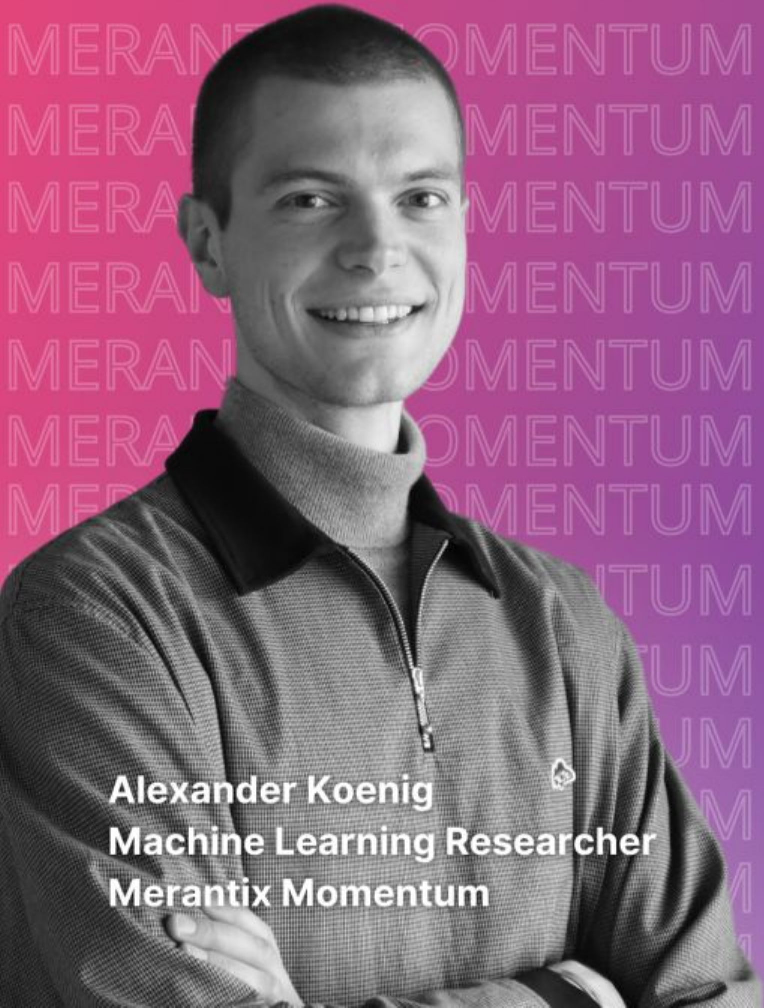


Alexander Koenig
Machine Learning Researcher
Merantix Momentum

PAPER READING GROUP

Wednesday July 5 | 6:30 pm | AI Campus Berlin

CVPR Highlights 2023

MERANTIX MOMENTUM

MERANTIX AI CAMPUS BERLIN

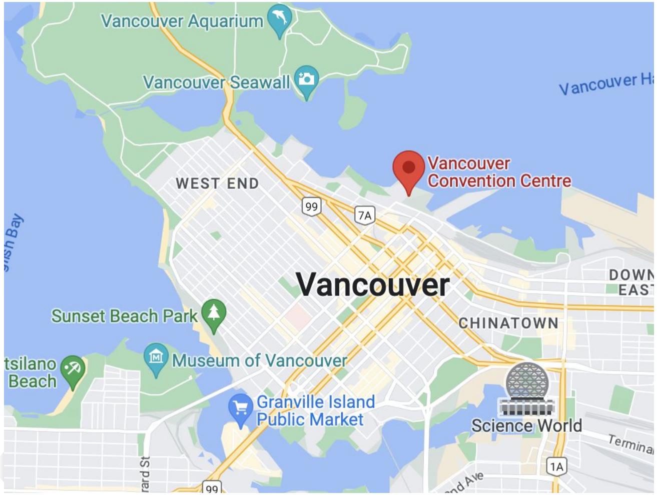
CVPR Impressions

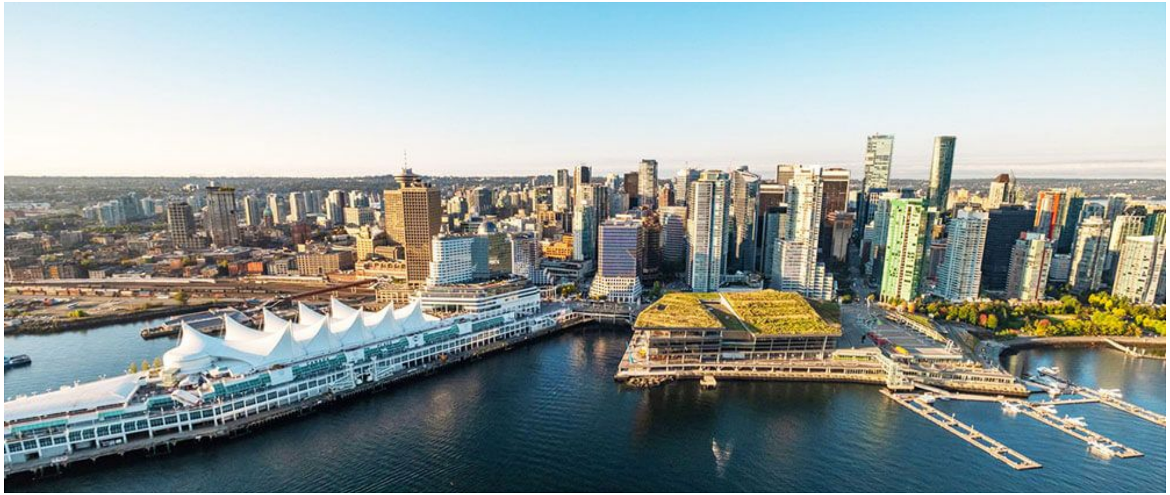
Papers

Food for Thought

Vancouver


ON
QC
WA
MT
ND
MN
NB
PE
ME
NS
SD
WI
MI
OR
ID
WY
NY
NH
MA
NE
IA
PA
NV
UT
United States
IL
IN
OH
MD
DE
CA
MO
KY
WV
VA
AZ
NM
OK
AR
TN
NC
MS
AL
SC
GA
TX
LA
FL
Gulf of Mexico
Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea


Vancouver Aquarium
Vancouver Seawall
Vancouver Convention Centre
WEST END
99
7A
Vancouver
Sunset Beach Park
CHINATOWN
Museum of Vancouver
Granville Island Public Market
Science World
1A



CVPR 2023

- Papers Submitted 9155
- Papers Accepted 2360
- Acceptance Rate 25.78%

- Attendance CVPR 23: 7088 in-person, 3215 virtual
- Attendance CVPR 19: 9375 (pre-COVID)

- Companies at CVPR 23: 116, 21200 square feet expo
- Companies at CVPR 19: 181, 41200 square feet expo

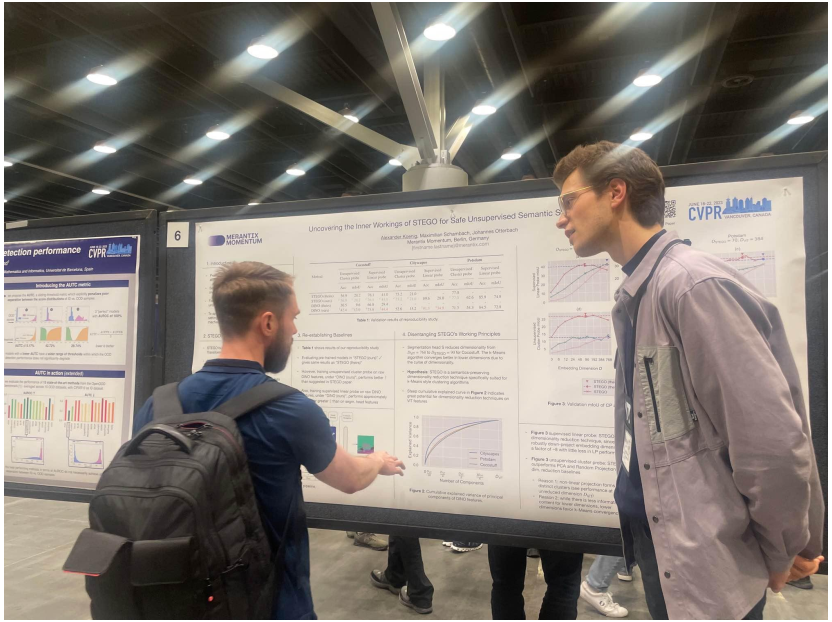
CVPR Impressions

Papers

Food for Thought

Disclaimers

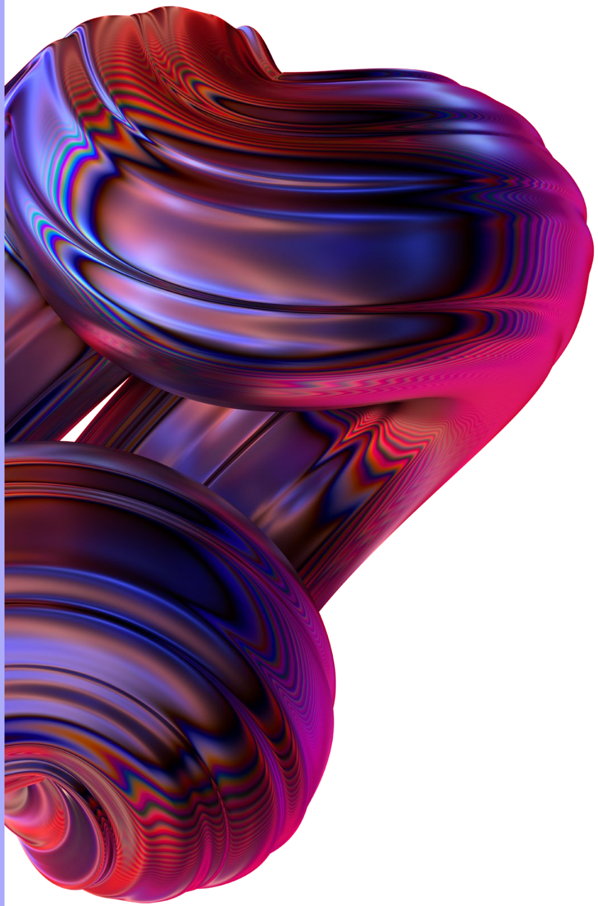
- Not my work! Just want to share some “aha” moments
- Hope to convey the paper’s message
- If you want to deep dive, read the paper



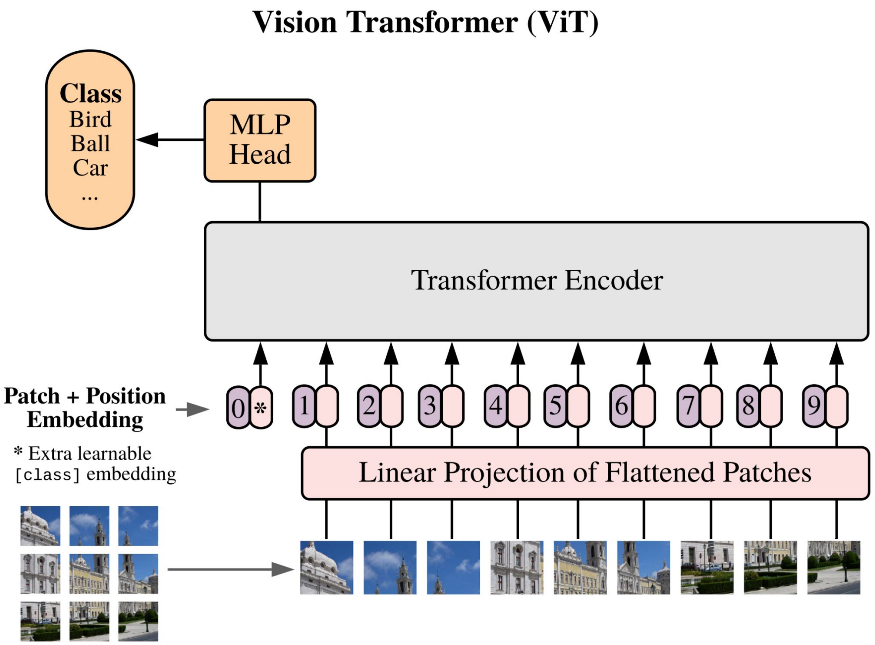
FlexiViT: One Model for All Patch Sizes

Lucas Beyer[⋆,1] Pavel Izmailov[⋆,1,3] Alexander Kolesnikov[⋆,1] Mathilde Caron[⋆,2] Simon Kornblith[⋆,1]
Xiaohua Zhai[⋆,1] Matthias Minderer[⋆,1] Michael Tschannen[⋆,1] Ibrahim Alabdulmohsin[⋆,1] Filip Pavetic[⋆,1]

Google Research



Vision Transformers 101



Source: Dosovitskiy et al. ICLR 2021

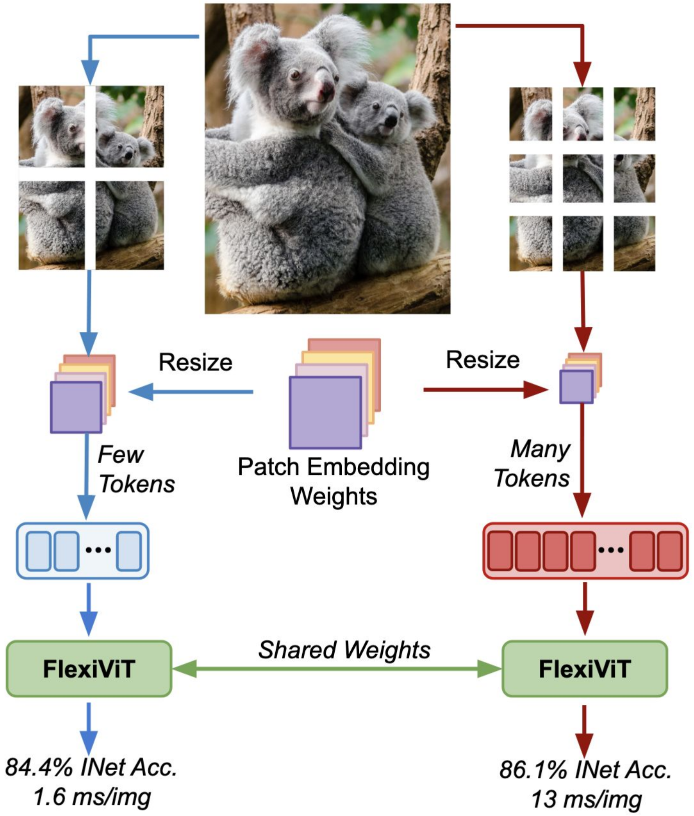
	ViT-B/16	ViT-B/32	ViT-L/16	ViT-L/32
CIFAR-10	98.13	97.77	97.86	97.94
CIFAR-100	87.13	86.31	86.35	87.07
ImageNet	77.91	73.38	76.53	71.16

Family of ViT Models

- **Problem**: need to train one model for each patch size (expensive, inflexible, must scale image s.t. 16 or 32 are a factor of resolution)

- **Trade-off**: small patch size → high performance, but expensive compute, and vice versa for large patch sizes

FlexiViT - Key Idea



Algorithm 1 Minimal FlexiViT pseudo-implementation.

```
model = ViT(...)
for batch in data:
  ps = np.random.choice([8, 10, ..., 40, 48])
  logits = model(batch["images"], (ps, ps))
  # [...] backprop and optimize as usual

class ViT(nn.Module):
  def __call__(self, image, patchhw):
    # Patchify, flexibly:
    w = self.param("w_emb", (32, 32, 3, d))
    b = self.param("b_emb", d)
    w = resize(w, (*patchhw, 3, d))
    x = conv(image, w, strides=patchhw) + b
    # Add flexible position embeddings:
    pe = self.param("posemb", (7, 7, d))
    pe = resize(pe, (*x.shape[1:3], d))
    return TransformerEncoder(...)(x + pe)
```

Notes: Changes to existing code highlighted via violet background.

→ bilinear interpolation to resize patch embedding weights and positional embeddings

FlexiViT - Results

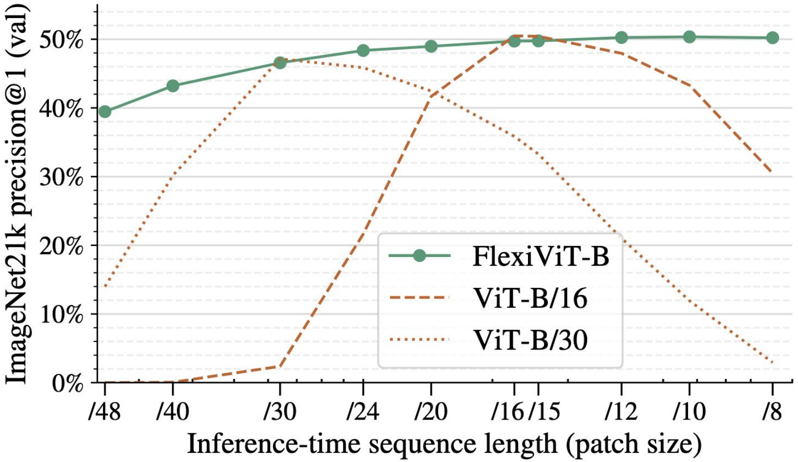


Figure 3. **Standard ViTs are not flexible** in patch size. However, FlexiViT can train them to be flexible without loss of performance.

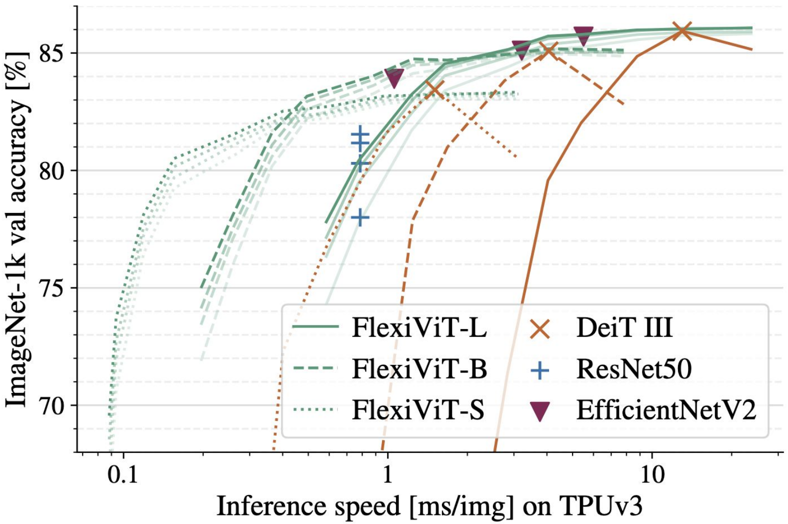


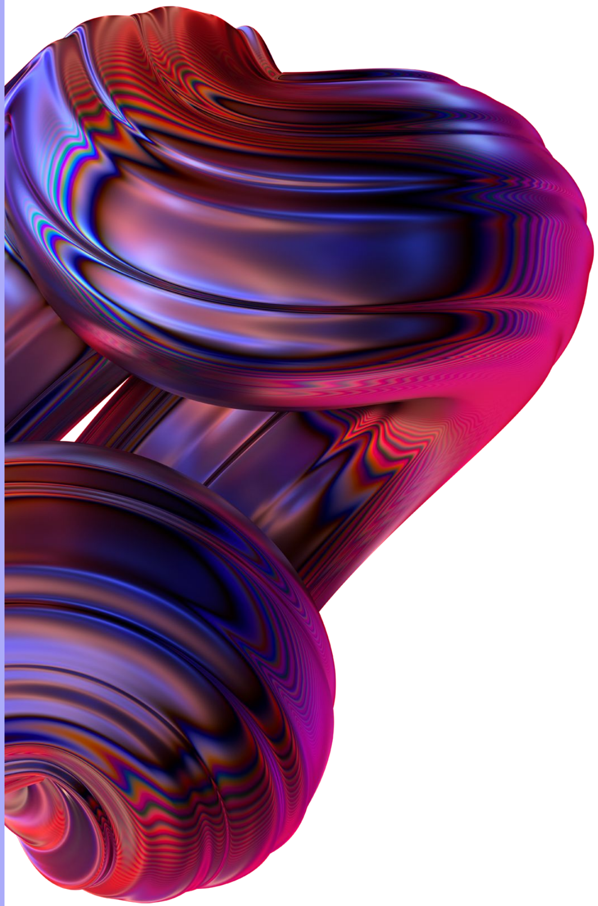
Figure 2. **FlexiViT results on ImageNet-1k.** We train three FlexiViTs based on DeiT III on ImageNet-1k and show their speed-accuracy tradeoff when evaluated at various patch sizes.

Heuristic: choose smallest patch size that still fits your compute budget ;-)

Uncovering the Inner Workings of STEGO for Safe Unsupervised Semantic Segmentation

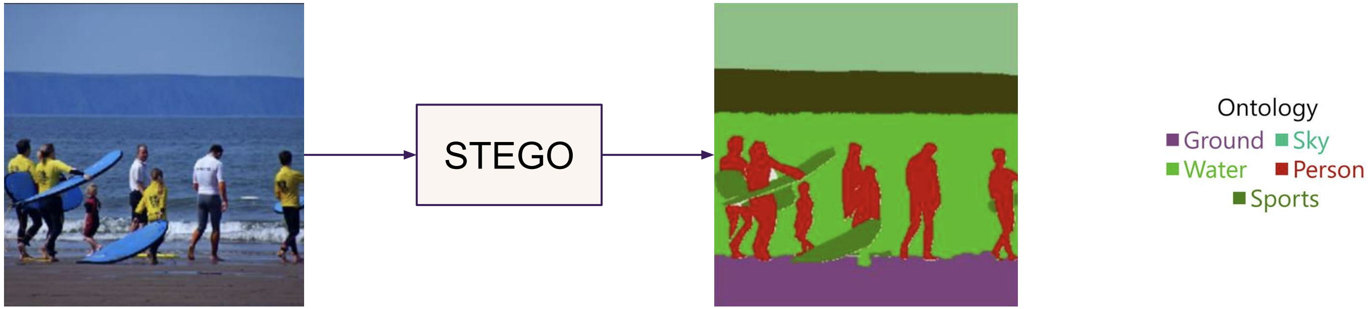
Alexander Koenig Maximilian Schambach Johannes Otterbach

Merantix Momentum
{firstname.lastname}@merantix.com

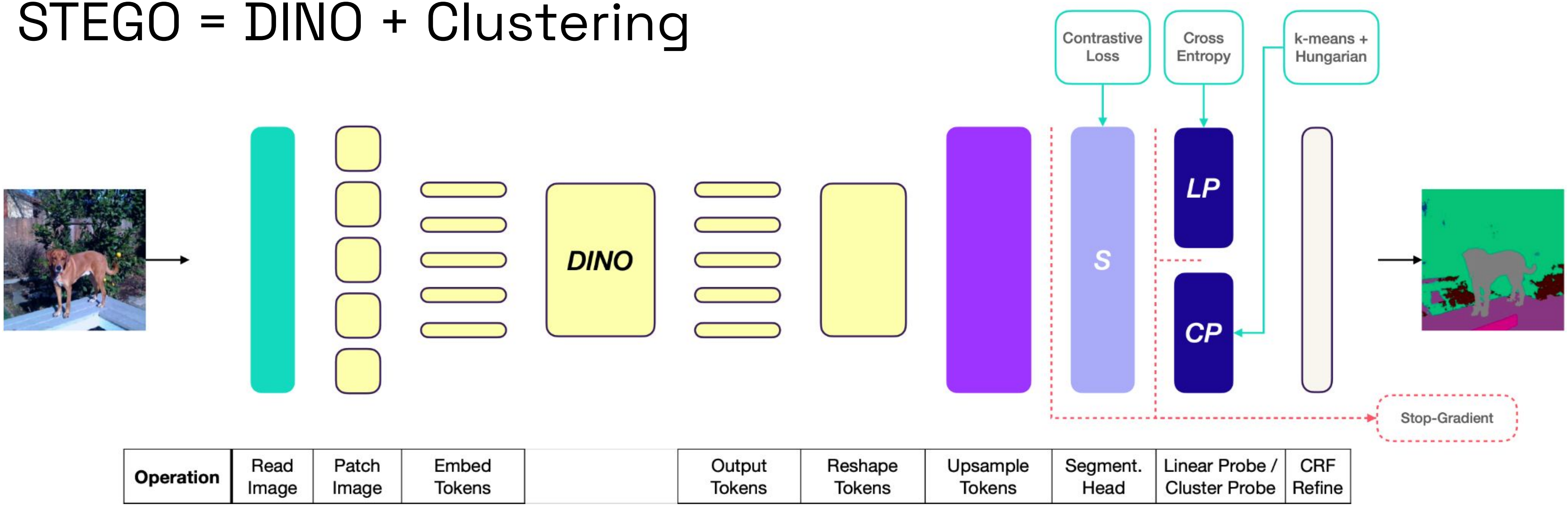


STEGO Follow-Up: Motivation

- Problem: labeled data is scarce, but unlabeled data is abundant
- Self-supervised learning recently demonstrated impressive results on unlabeled datasets
- STEGO (Hamilton et al., ICLR 2022) does unsupervised semantic segmentation
- To apply STEGO safely in real-world, it’s crucial to understand its working mechanisms



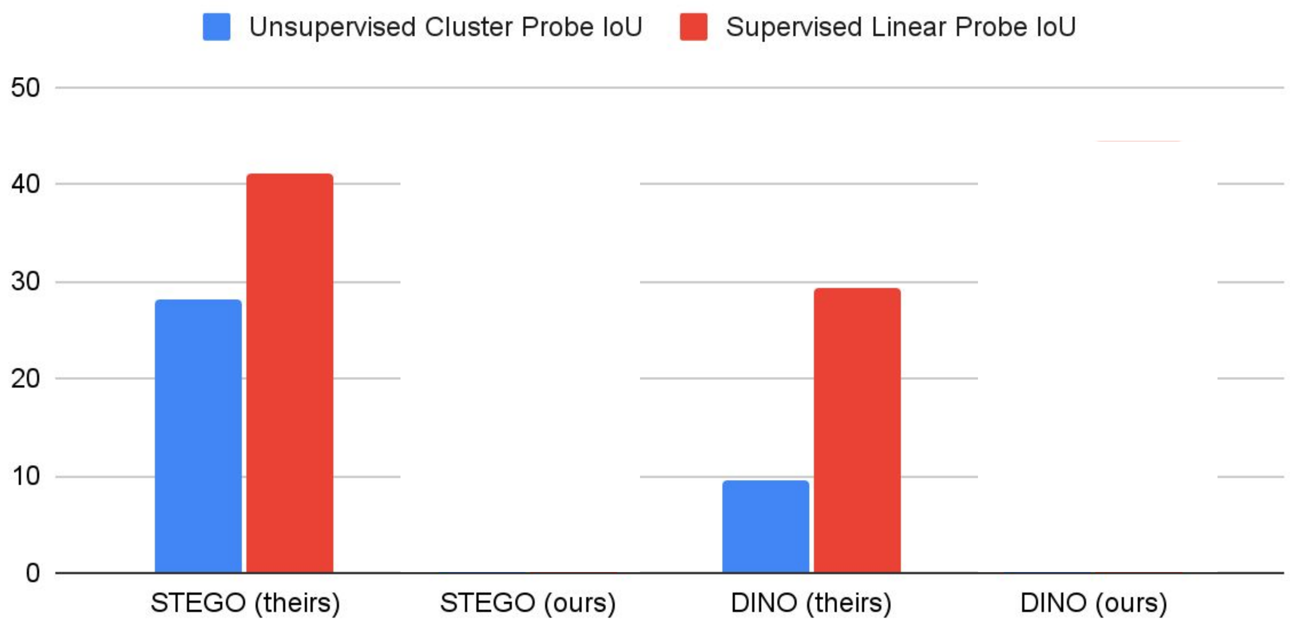
STEGO = DINO + Clustering



- STEGO builds on DINO (Caron et al., ICCV 2021) pre-trained Vision Transformer
- Segmentation head S projects DINO feats into lower-dimensional space, “distilling” DINO feature correspondences
- Cluster Probe maps STEGO features to ontologies using k-Means

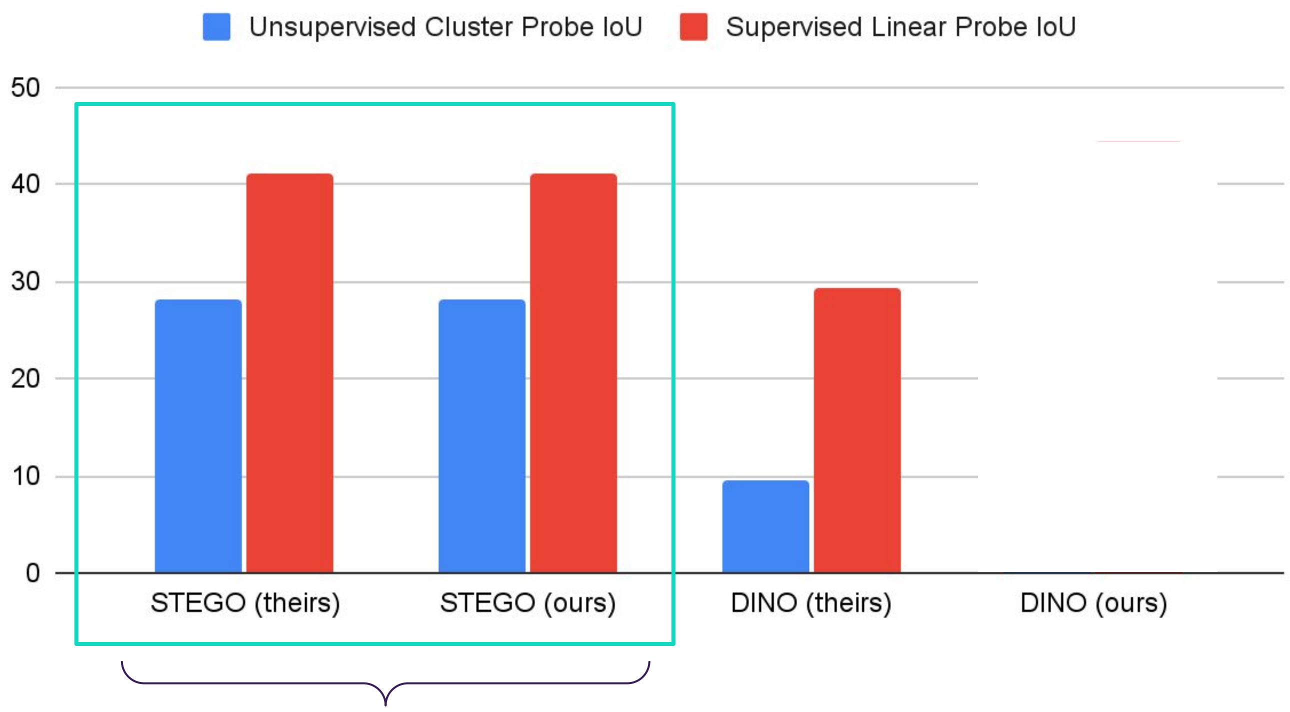
Reproducibility Cocostuff

Cluster Probe = SegHead + K-Means + Hungarian
Linear Probe = SegHead + Lin. Layer + X Entropy



Reproducibility Cocostuff

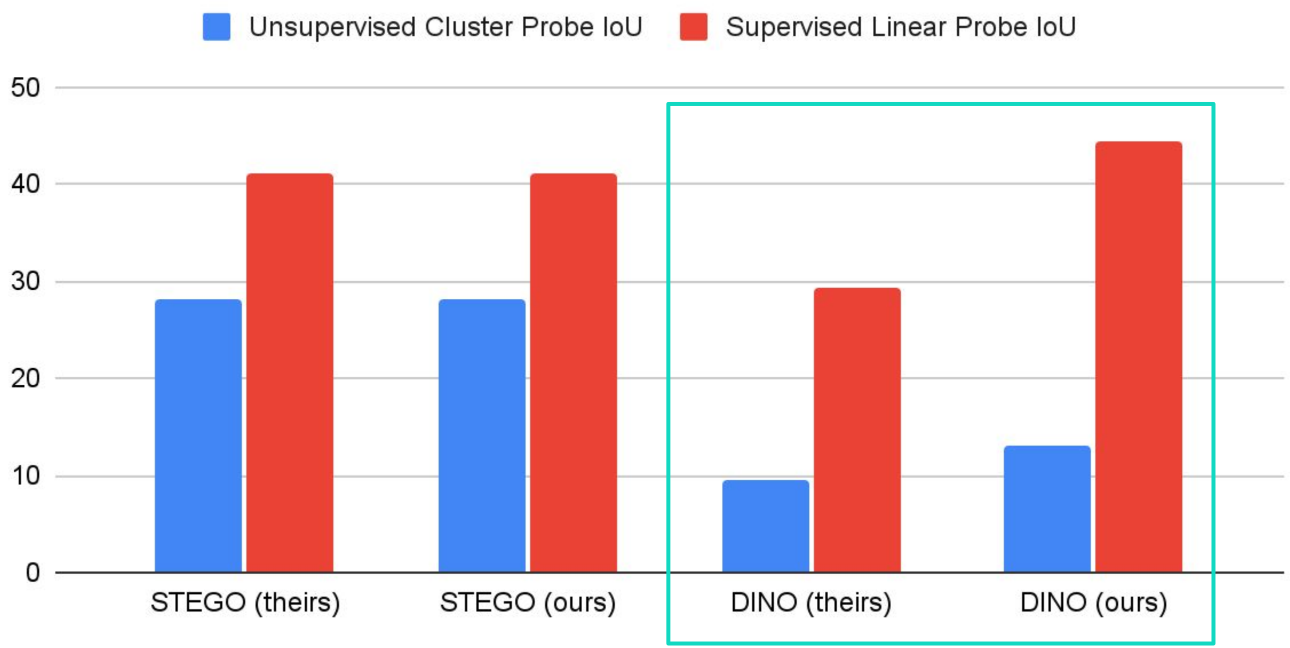
Unsupervised = SegHead + K-Means + Hungarian
Linear Probe = SegHead + Lin. Layer + X Entropy



reproduced!

Reproducibility Cocostuff

Unsupervised = SegHead + K-Means + Hungarian
Linear Probe = SegHead + Lin. Layer + X Entropy

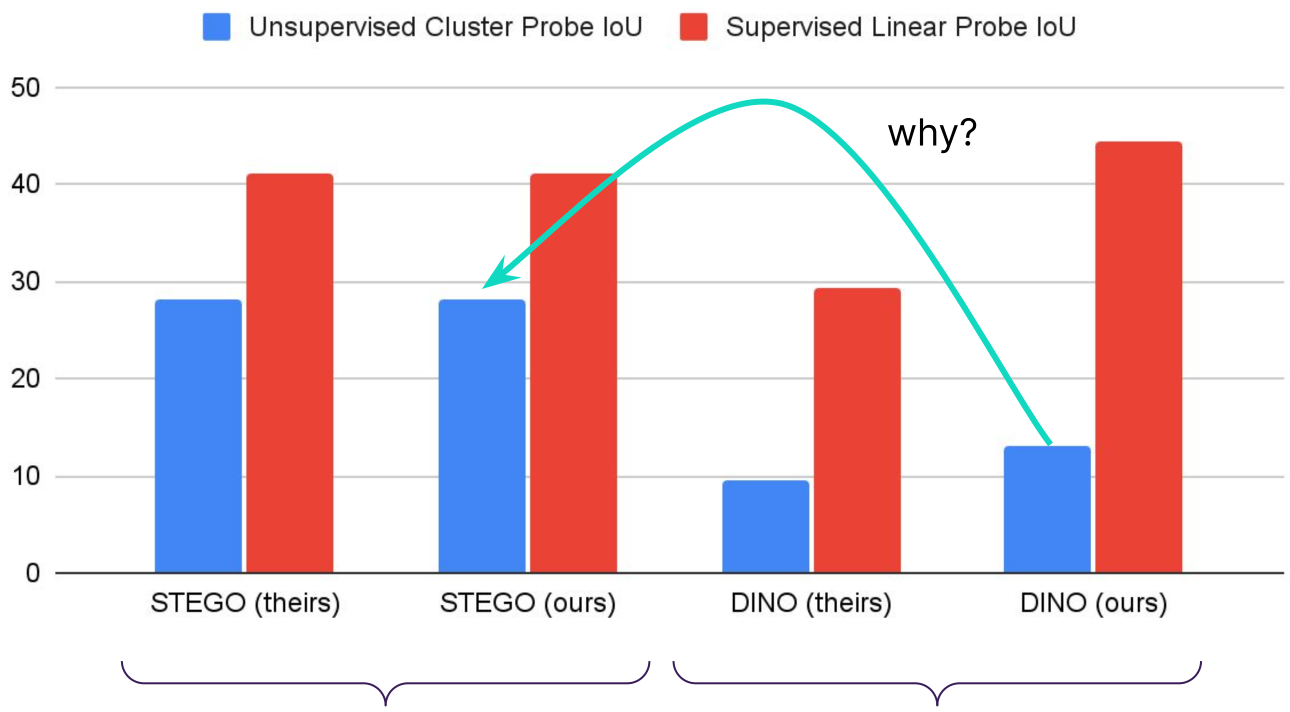


reproduced!

not reproduced!

Reproducibility Cocostuff

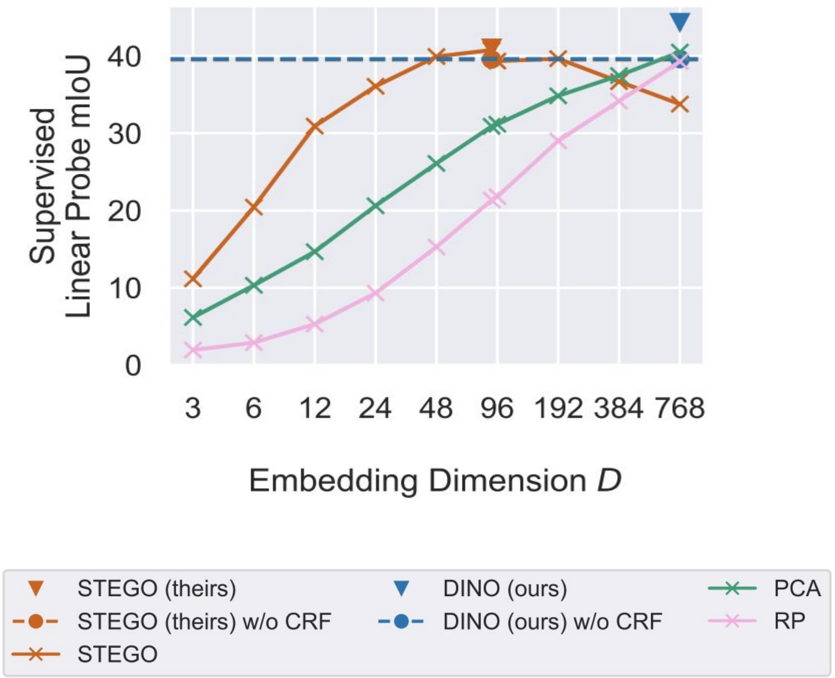
Unsupervised = SegHead + K-Means + Hungarian
Linear Probe = SegHead + Lin. Layer + X Entropy



reproduced!

not reproduced!

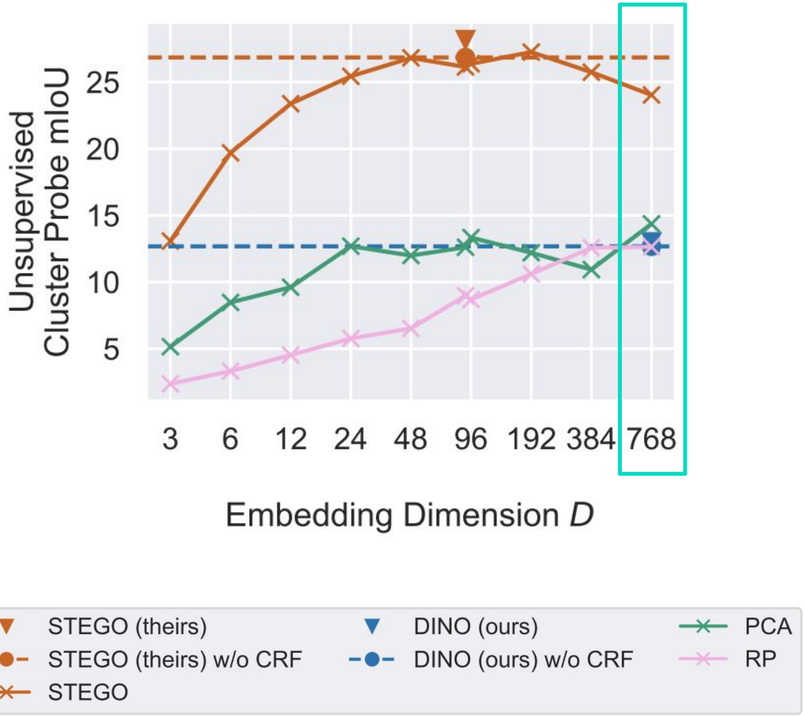
STEGO’s Working Mechanisms



<u>Working Mechanism 1:</u>

- STEGO is a dimensionality reduction technique
- k-Means converges better in fewer dimensions

STEGO’s Working Mechanisms

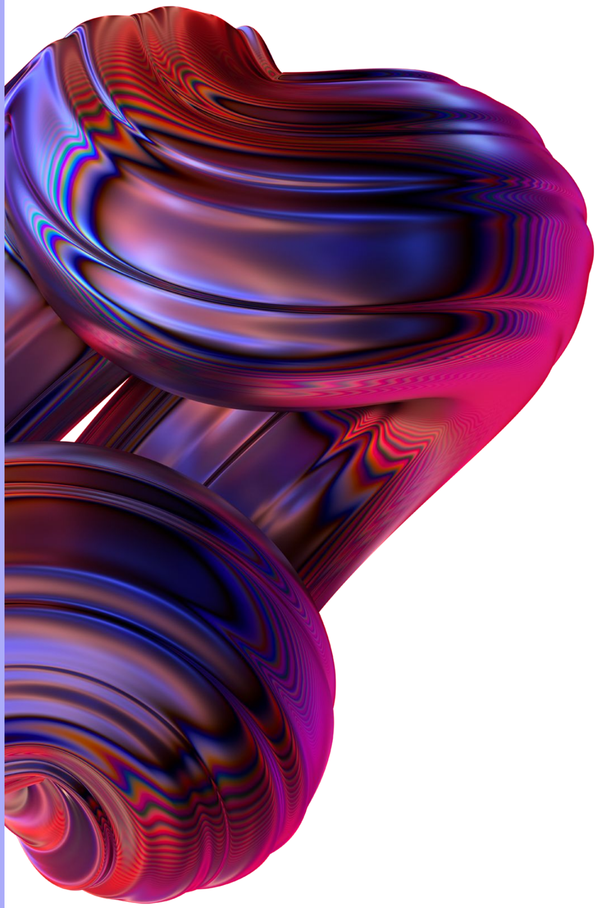


<u>Working Mechanism 2:</u>

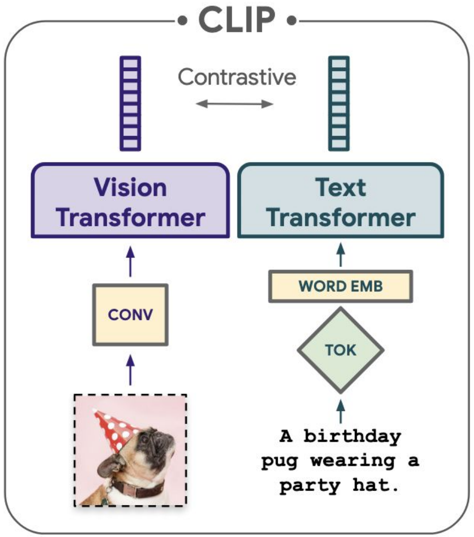
- Segmentation head output forms more distinct clusters

CLIPPO: Image-and-Language Understanding from Pixels Only

Michael Tschannen, Basil Mustafa, Neil Houlsby
Google Research, Brain Team, Zürich



CLIP-Pixels Only (CLIPPO) - Key Idea



- CLIP (Radford et al. 2021) trains separate image and text encoder

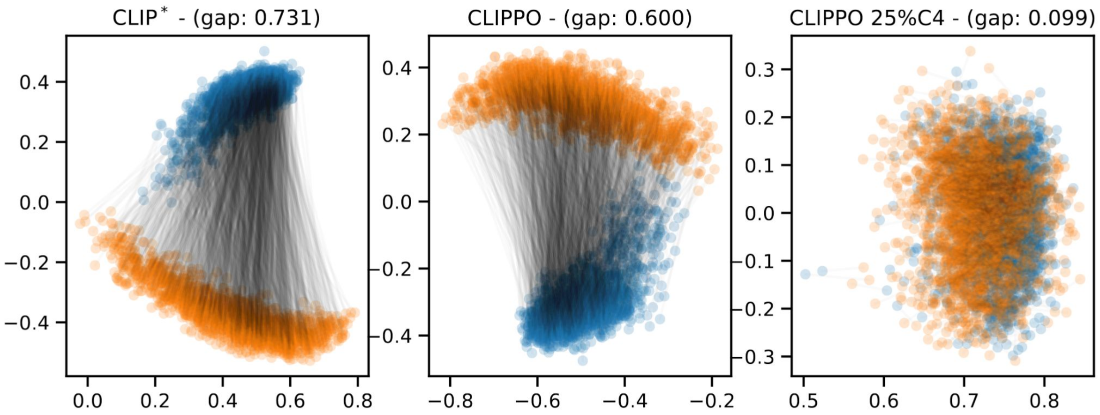
CLIPPO - Results

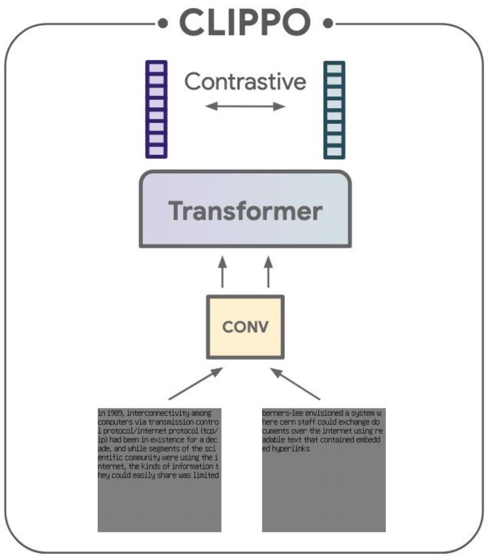
- CLIPPO approaches BERT performance on GLUE benchmark
- “CLIPPO performs similarly to CLIP-style models (within 1-2%) on the main tasks CLIP was designed for - image classification and text/image retrieval”
- Good results on VQA despite never trained on that

VQAv2 dataset: Classifying CLIPPO feats

CLIPPO - Modality Gap

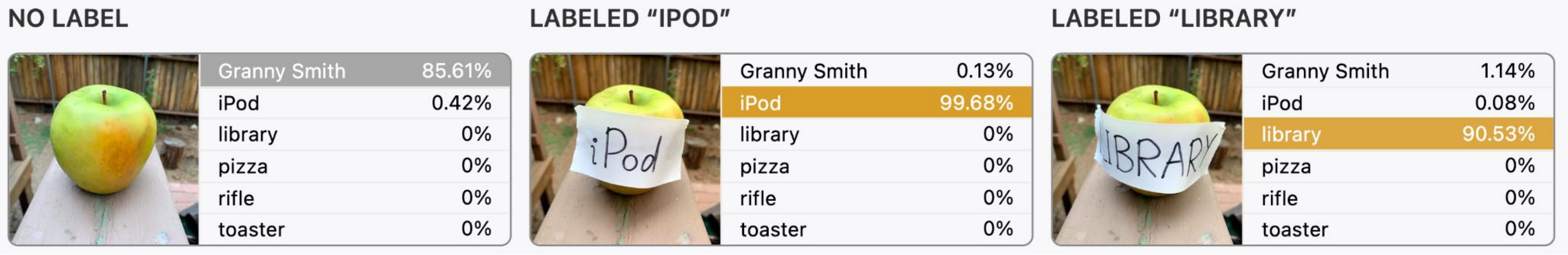
also see: Lian et al. “Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning”, NeurIPS 2022





Pre-training on text-text pairs with C4 (Colossal Clean Crawled Corpus) reduces modality gap

CLIPPO - Typographic Attacks



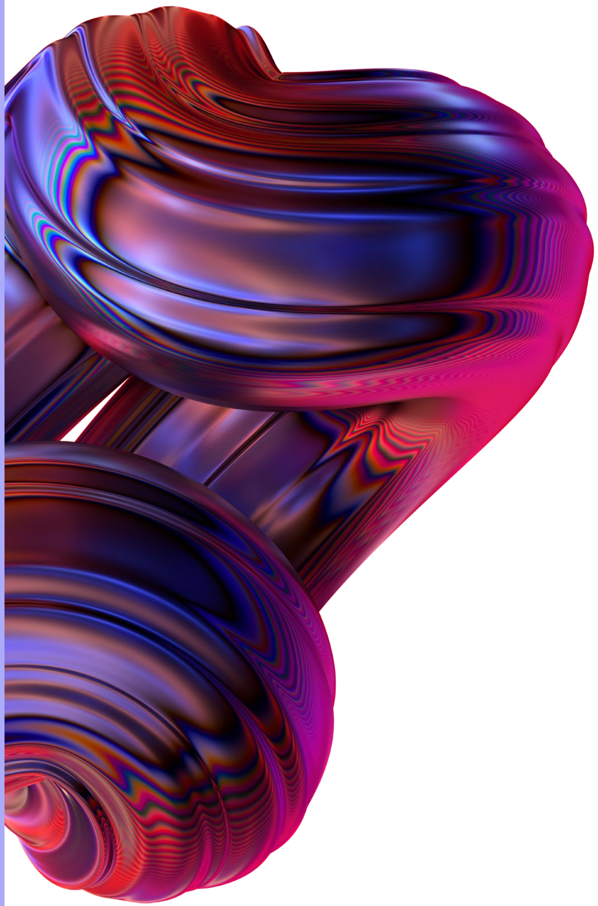
Source: https://distill.pub/2021/multimodal-neurons/

Typographic attack: "the tendency of CLIP-style models to zero-shot classify an image according to adversarially injected scene text unrelated to the scene"

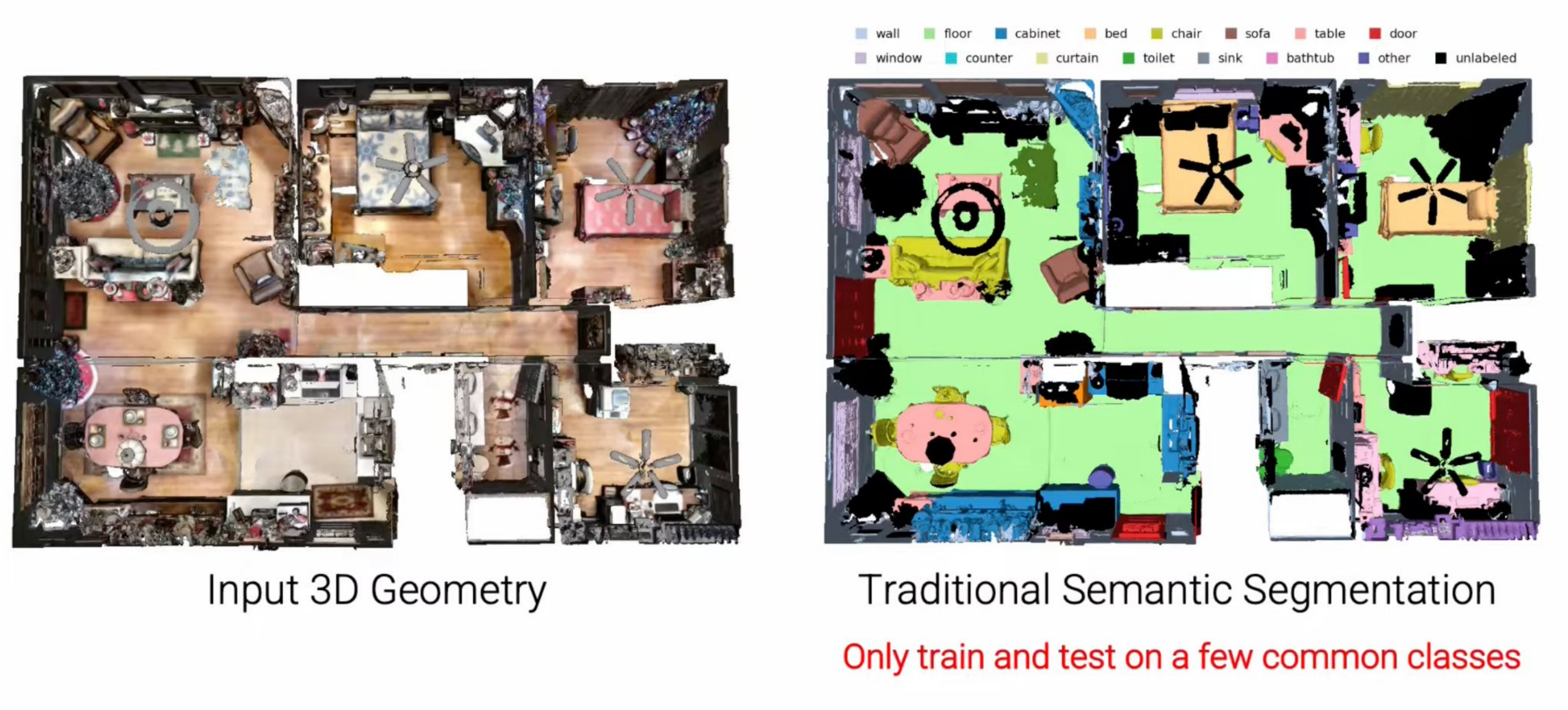
CLIPPO Result: "All models are largely able to ignore the typographic attack, and the CLIPPO models are on par with or better than the counterparts relying on a tokenizer."

OpenScene: 3D Scene Understanding with Open Vocabularies

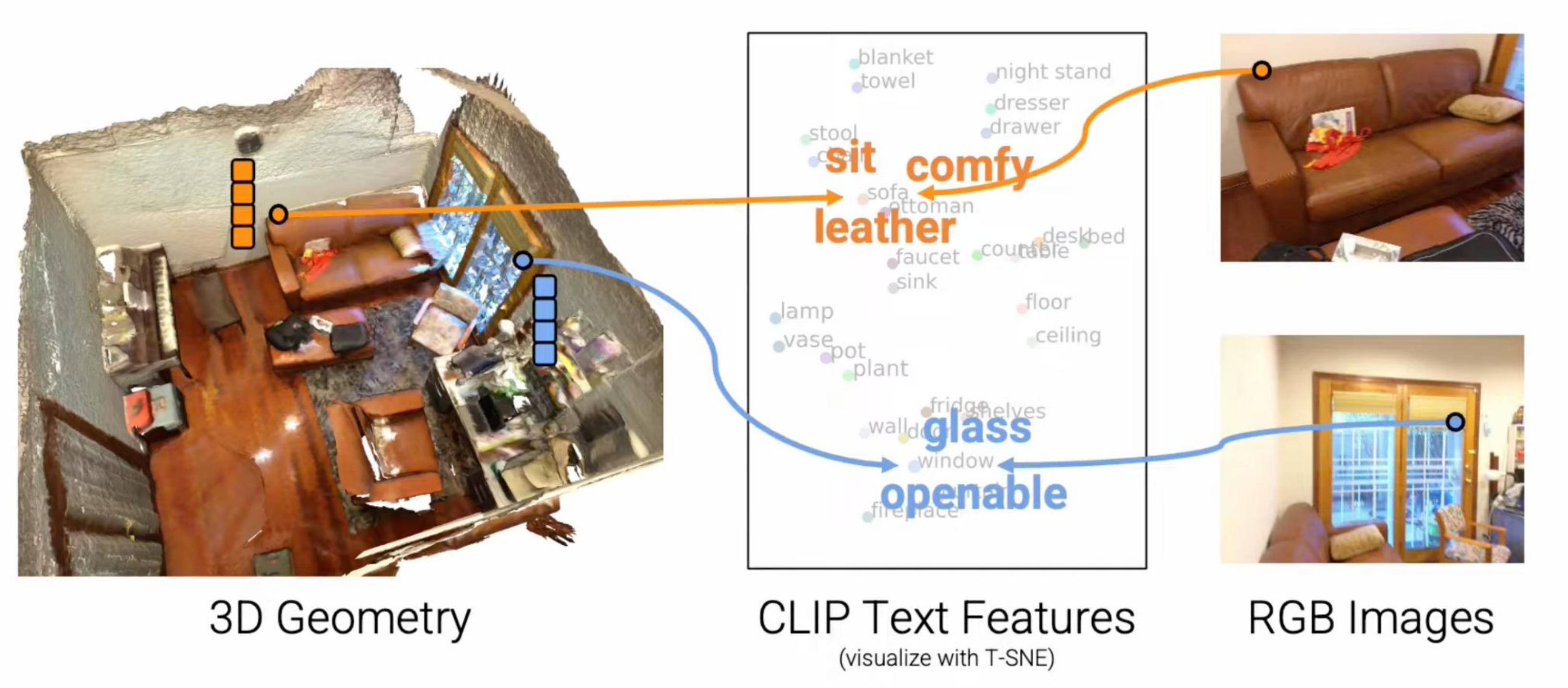
Songyou Peng[1,2,3] Kyle Genova[1] Chiyu “Max” Jiang[4] Andrea Tagliasacchi[1,5]
Marc Pollefeys[2] Thomas Funkhouser[1]
[1] Google Research [2] ETH Zurich [3] MPI for Intelligent Systems, Tübingen [4] Waymo LLC [5] Simon Fraser University
pengsongyou.github.io/openscene



Traditional (3D) Semantic Segmentation

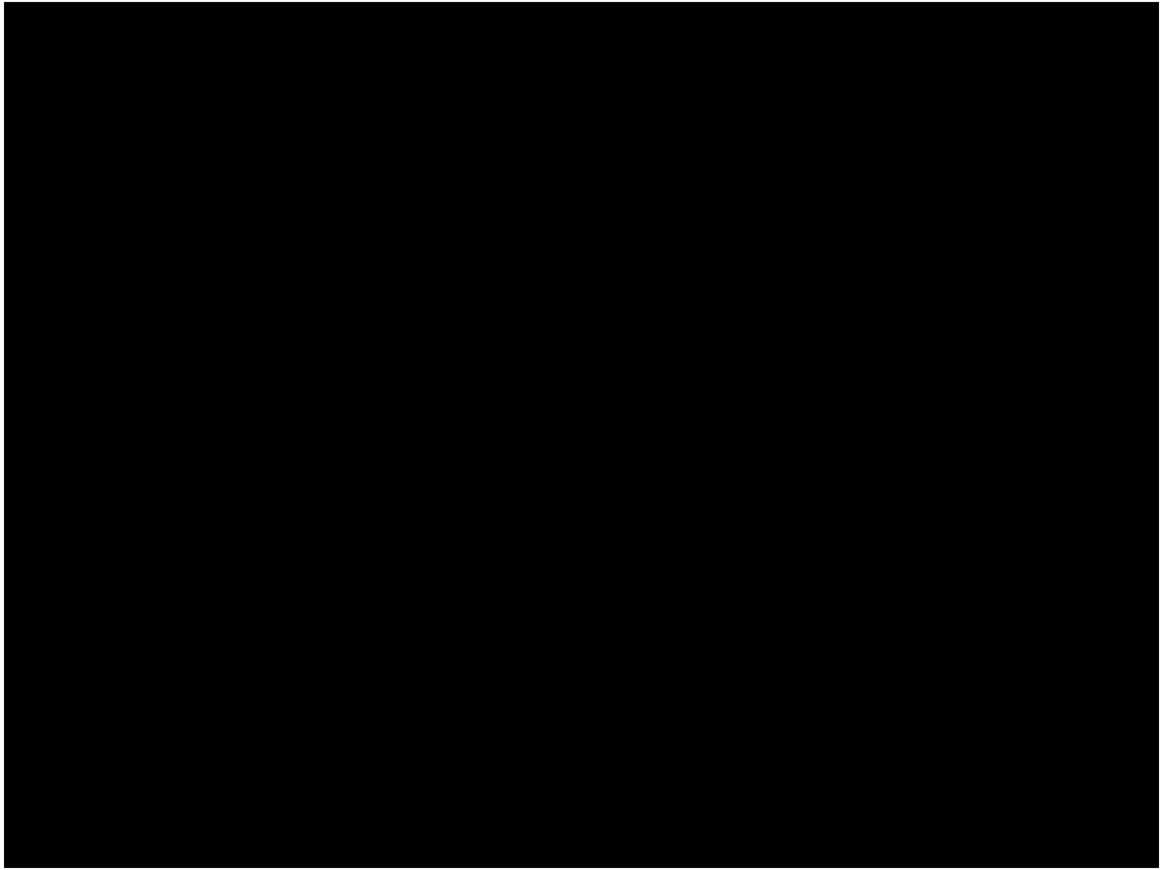


OpenScene - Key Idea



1. Co-embed 3D text-image features
2. Reason about properties of 3D points via cosine-similarity

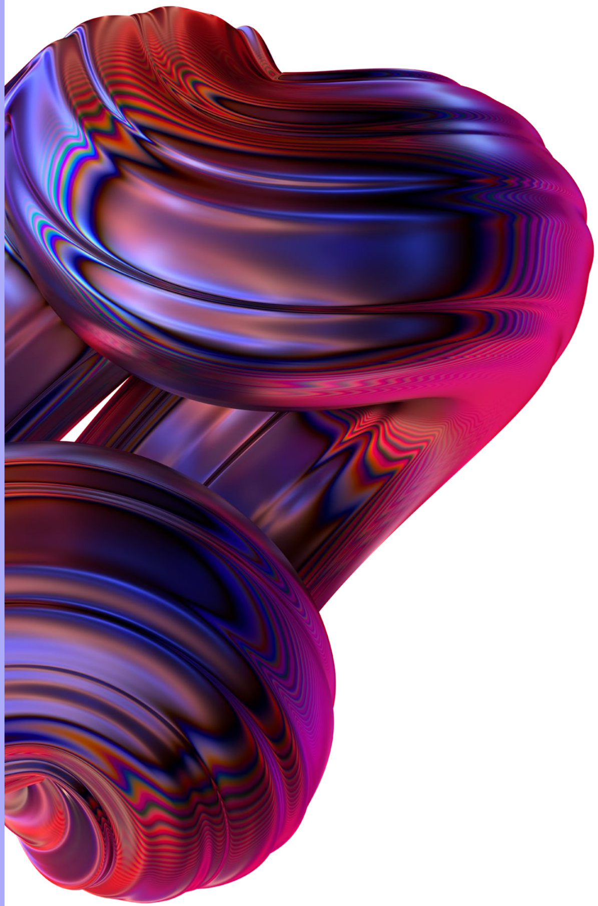
OpenScene - Demo



Visual Programming: Compositional visual reasoning without training

Tanmay Gupta, Aniruddha Kembhavi
PRIOR @ Allen Institute for AI
https://prior.allenai.org/projects/visprog

CVPR Award Candidate



VisProg

- VisProg - a framework that builds CV pipelines from natural language
- "uses the in-context learning ability of GPT3 to generate python programs"
- Each line invokes functions s.a. CV models, openCV or PIL routines, …

HandsOff: Labeled Dataset Generation With No Additional Human Annotations

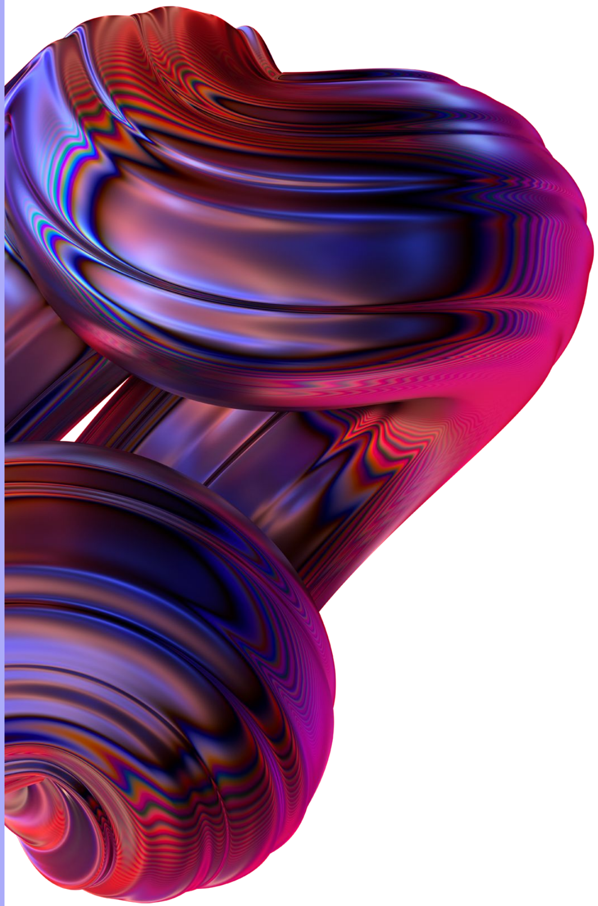
Austin Xu*
Georgia Institute of Technology

Mariya I. Vasileva
Amazon AWS

Achal Dave†
Toyota Research Institute

Arjun Seshadri
Amazon Style

CVPR Highlight



HandsOff - Key Idea

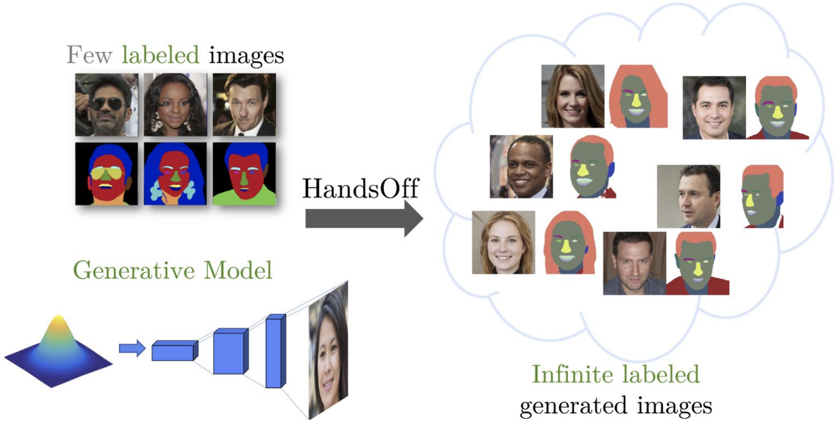


Figure 1. The HandsOff framework uses a small number of existing labeled images and a generative model to produce **infinitely** many labeled images.

- Trained on less than 50 labeled images
- GAN inversion for dataset generation

HandsOff - GAN Inversion 101

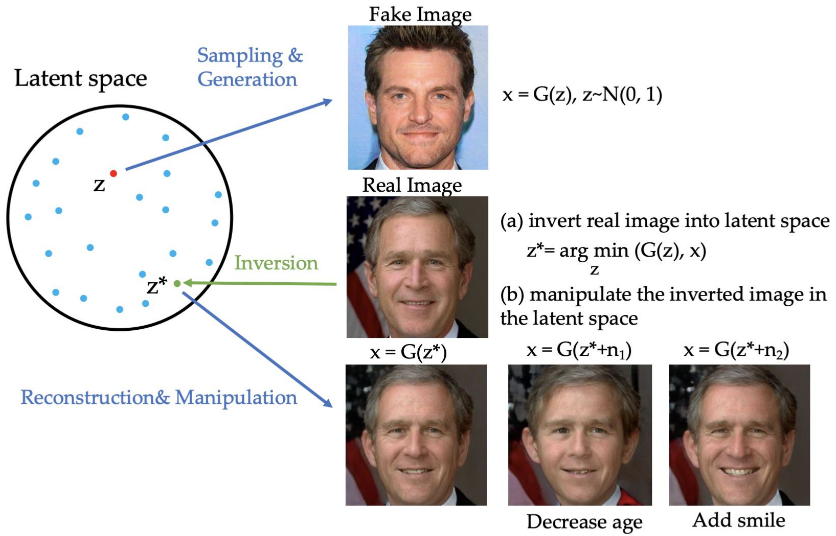


Figure 1: GAN inversion overview

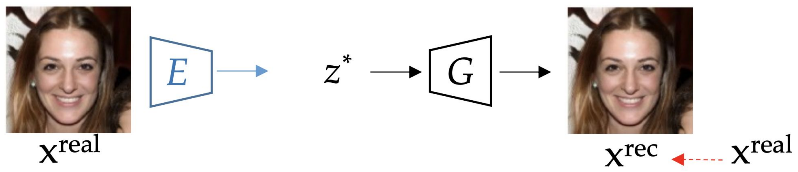


Figure 2: Invert GAN with encoder E, trained by min. rec. los

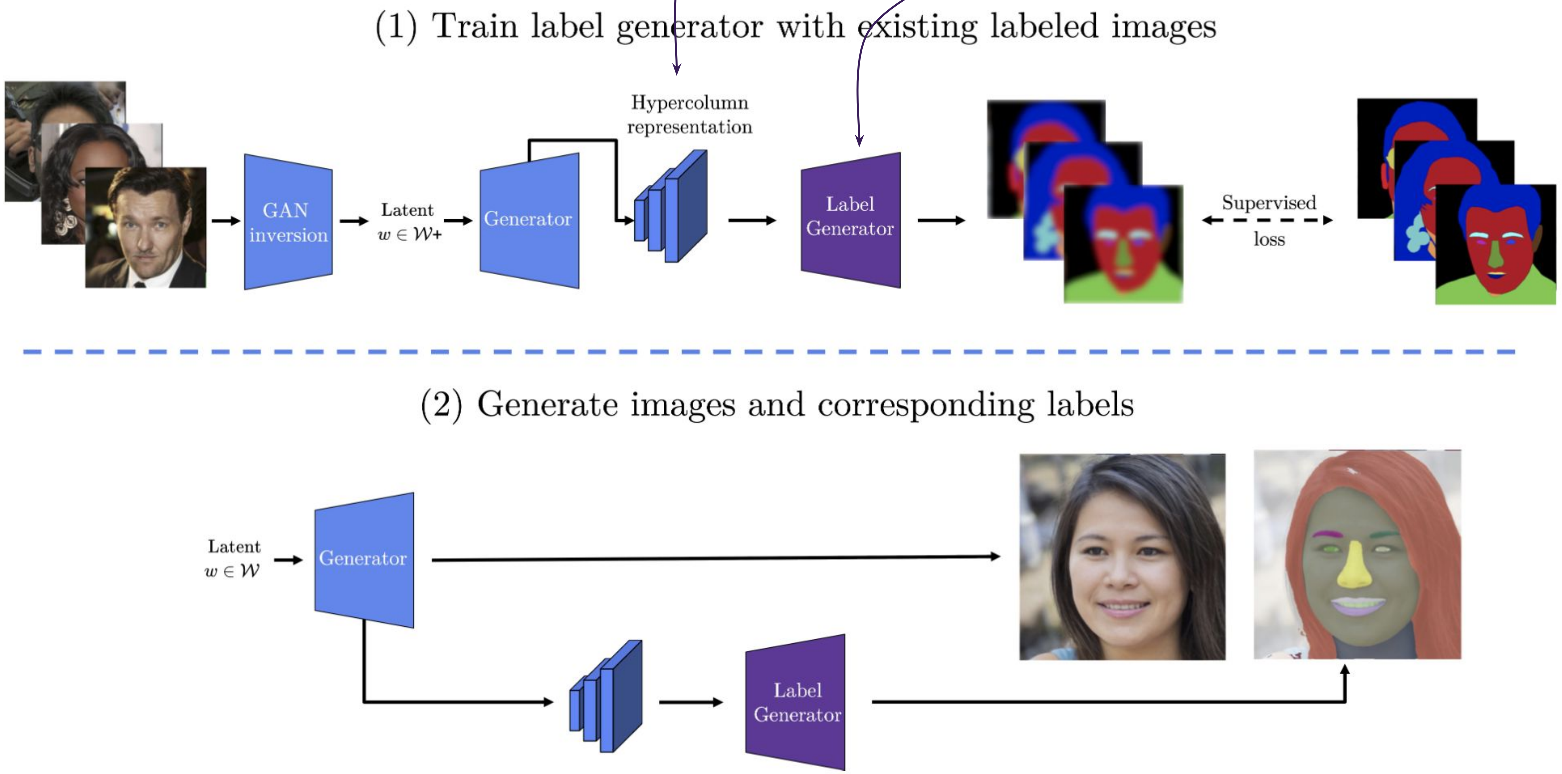
- “**GAN inversion** aims to invert a given image back into the latent space of a pretrained GAN model so that the image can be faithfully reconstructed from the inverted code by the generator”

Source: Xia et al. “GAN Inversion: A Survey”, 2022

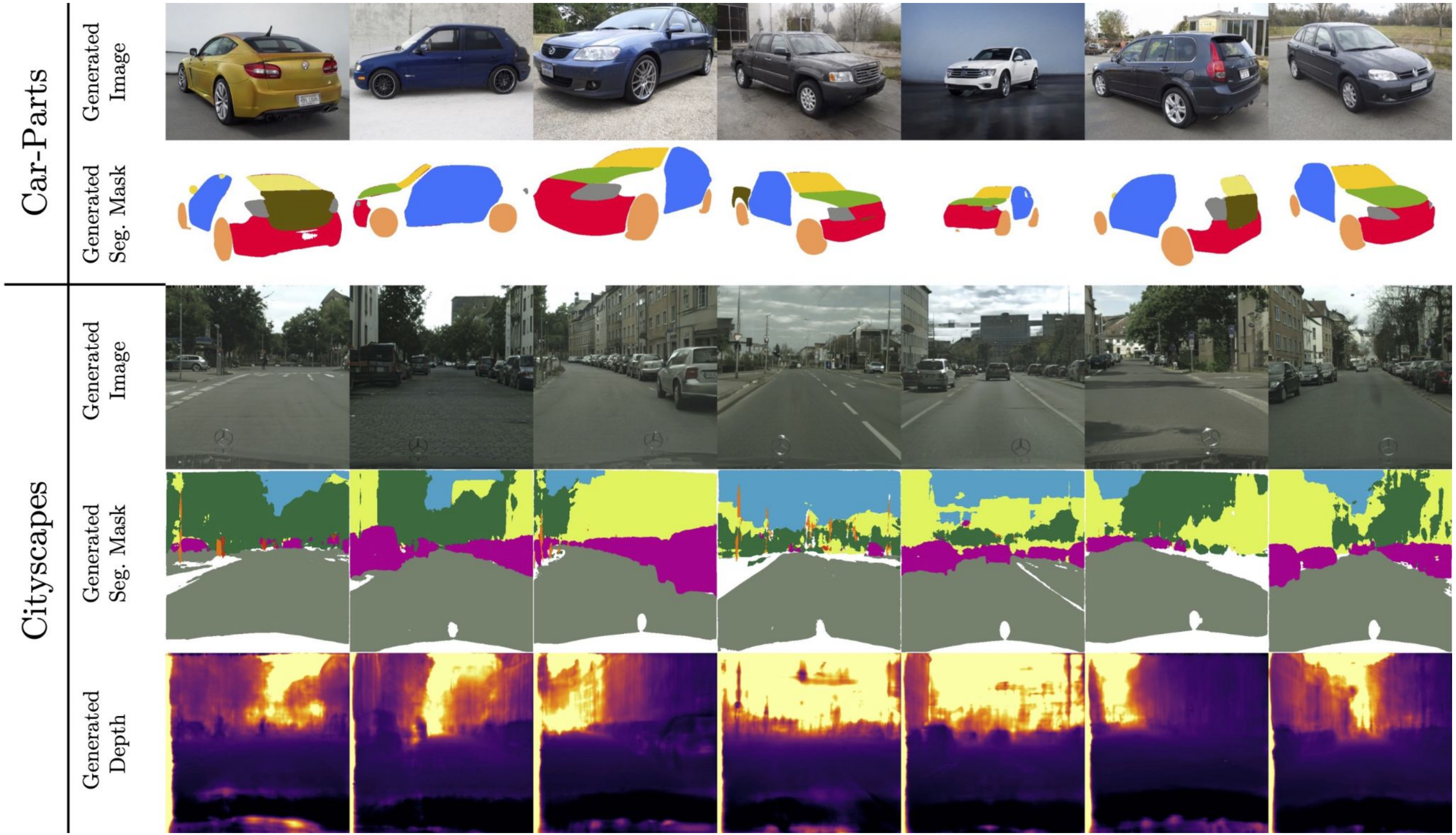
HandsOff - Details

grabs intermediate layers of StyleGAN2 generator,
up-samples them, forms pixel-wise features

MLP head maps pixel-wise feature to label



HandsOff - Results



HandsOff - Long Tail Improvement

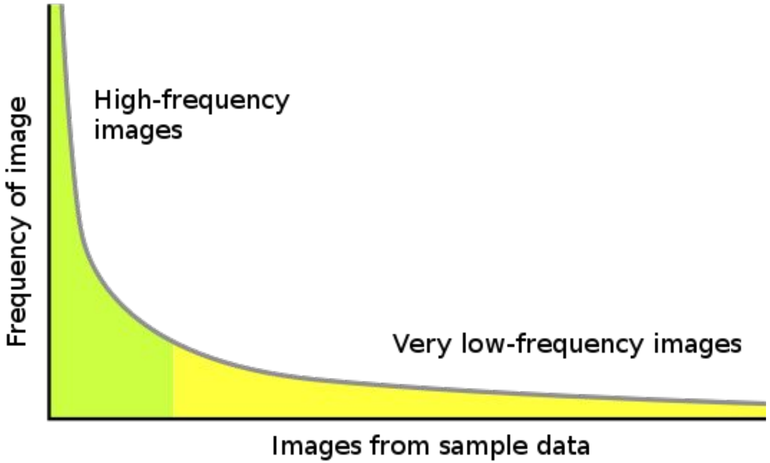


Figure 1: Long-tailed data distrubtions.

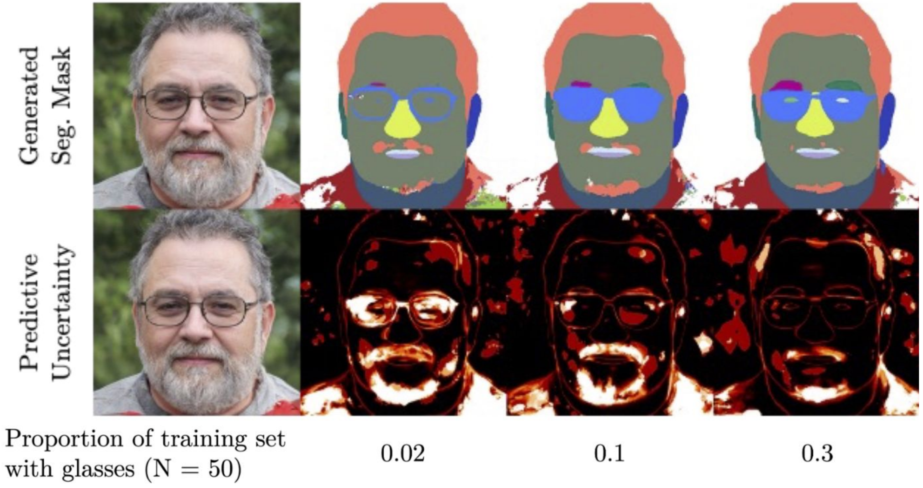


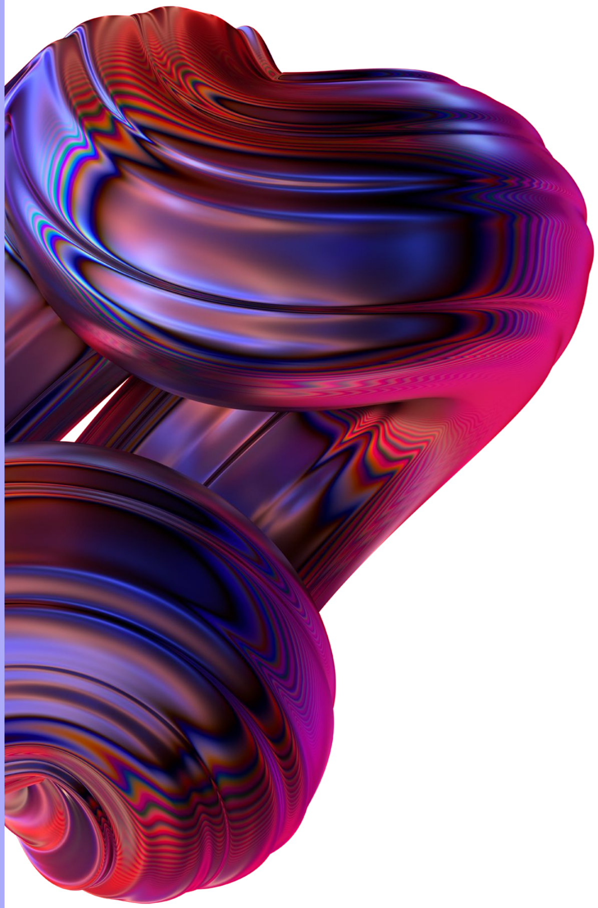
Figure 2: Improved Jensen-Shannon divergence and mask quality with more synthetic training data.

Source Graph: https://www.marksayson.com/blog/advances-in-computer-vision-and-chasing-long-tail/

IMAGEBIND: One Embedding Space To Bind Them All

Rohit Girdhar* Alaaeldin El-Nouby* Zhuang Liu Mannat Singh
Kalyan Vasudev Alwala Armand Joulin Ishan Misra*
FAIR, Meta AI
https://facebookresearch.github.io/ImageBind

CVPR Highlight



ImageBind - Key Idea

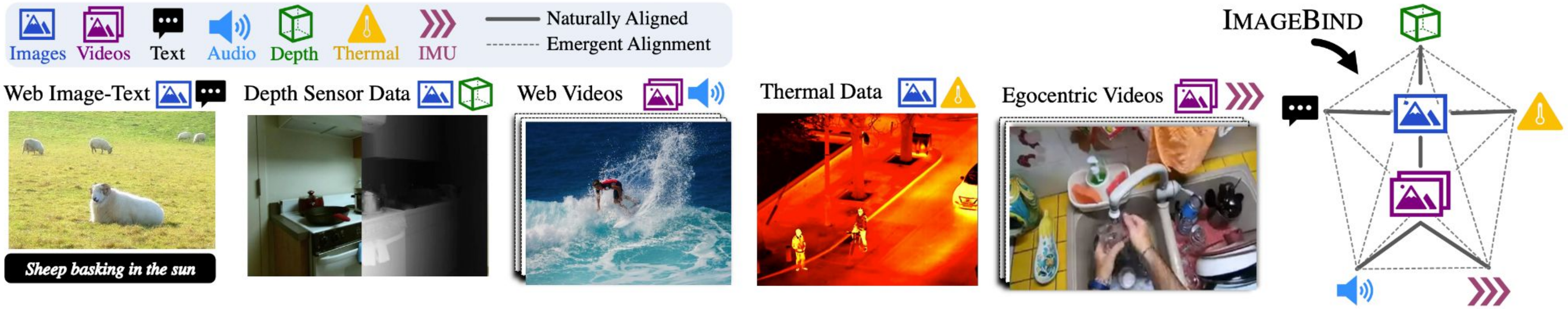
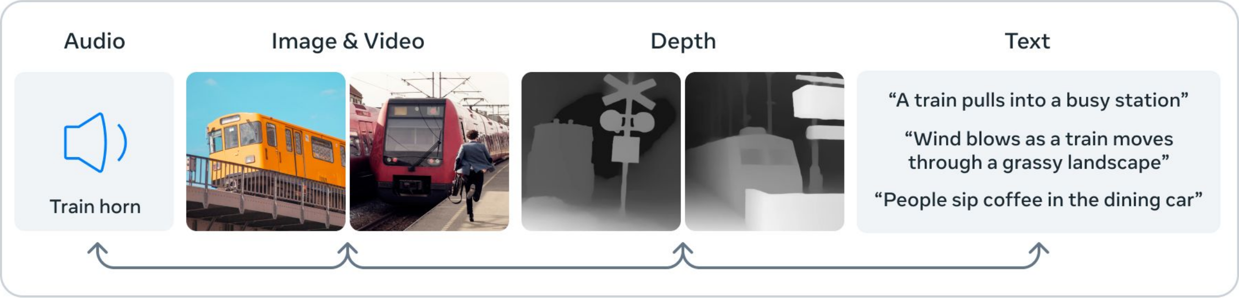


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc.* IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

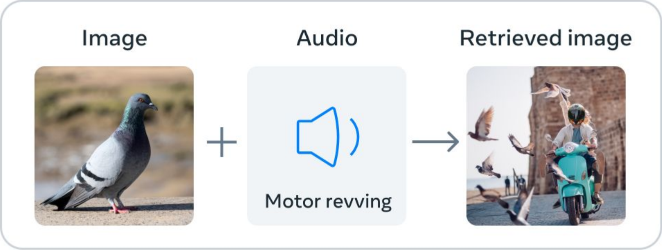
- Goal: multimodal representation learning (i.e. have single aligned feature space)
- But: no dataset couples modalities s.a. Vision, Audio, IMU, Depth, Thermal, … → self-supervision
- Idea: contrastive learning on (I, M) pairs, where I=image and M=some other modality

ImageBind - Emergent Properties

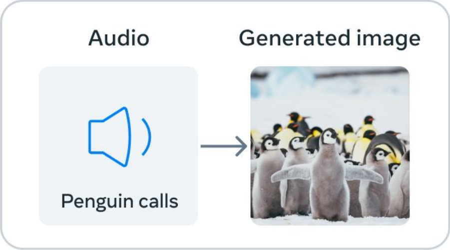
Cross-modal retrieval



Embedding-space arithmetic



Audio to image generation



now you can use diffusion model (DALLE-2) as image generator from audio!

See demo at: https://imagebind.metademolab.com/

ImageBind - Emergent Properties

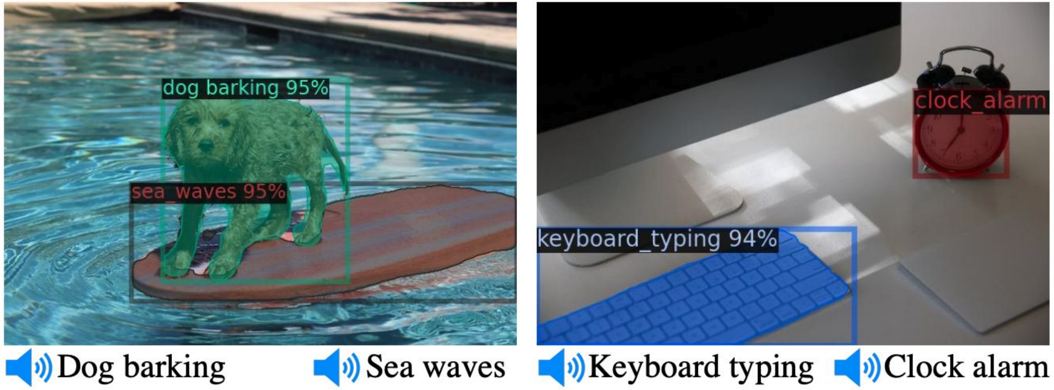


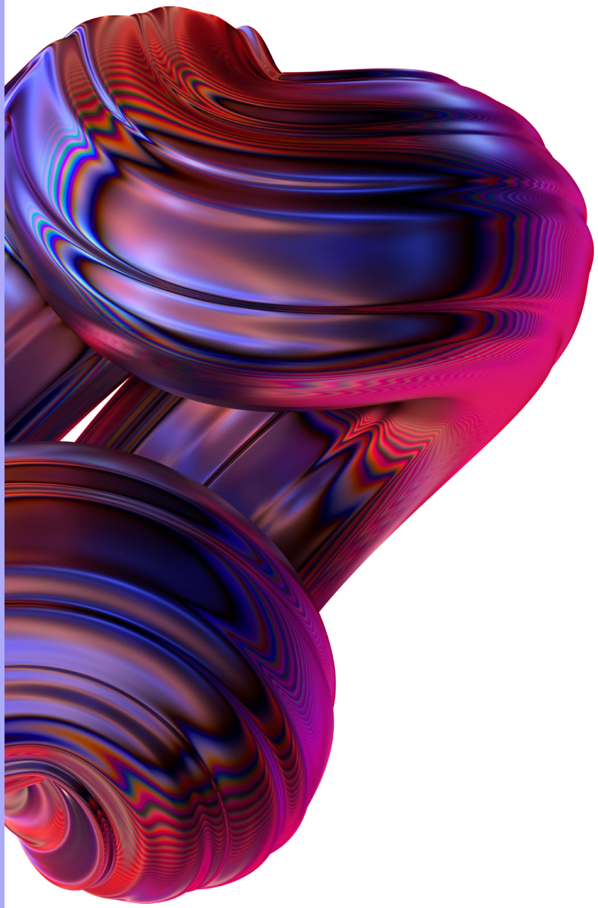
Figure 5. Object detection with audio queries. Simply replacing Detic [88]'s CLIP-based 'class' embeddings with our audio embeddings leads to an object detector promptable with audio. This requires no re-training of any model.

- ImageBind is initialized with CLIP
- Detic = pre-trained text-based detection module uses CLIP embeddings
- Idea: replace Detic's text embeddings with audio embeddings

InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks* Aleksander Holynski* Alexei A. Efros
University of California, Berkeley

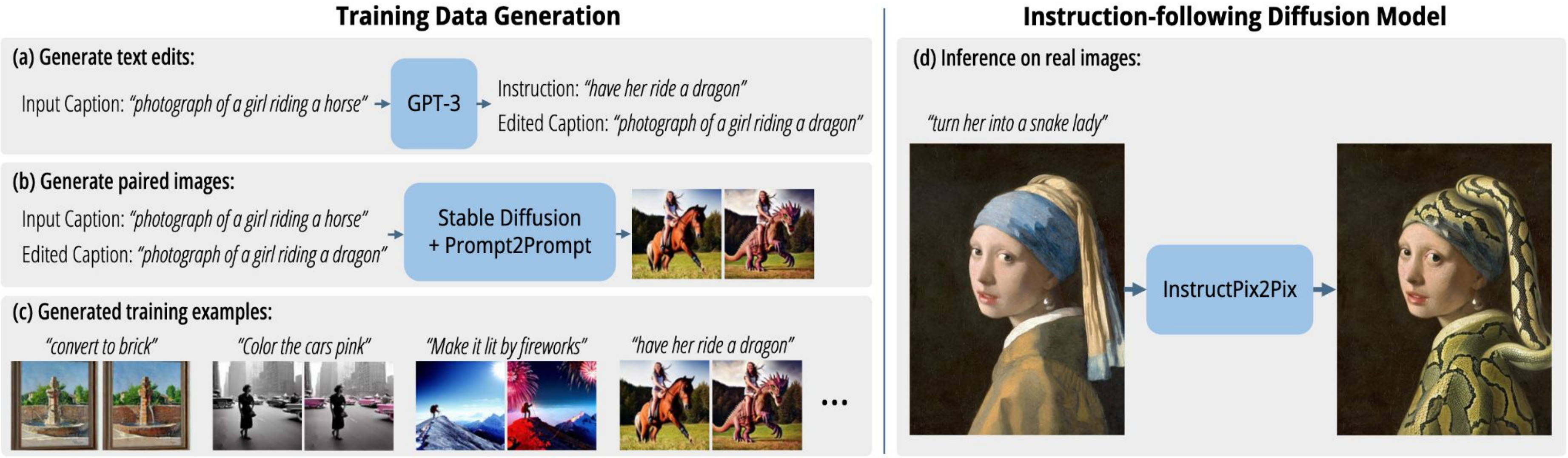
CVPR Highlight



InstructPix2Pix - Goal

Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

InstructPix2Pix - Key Idea



- First generate 450k synthetic training samples
- Then supervised fine-tuning of pre-trained diffusion model conditioned by image
- Zero-shot generalization to real images
- But: performance is bottlenecked by models generating dataset

InstructPix2Pix - More Results

Input | *"Add boats on the water"* | *"Replace the mountains with a city skyline"*

Figure 17. A landscape photograph shown with different contextual edits. Note that isolated changes also bring along accompanying contextual effects: the addition of boats also adds wind ripples in the water, and the added city skyline is reflected on the lake.

InstructPix2Pix - Inherited Biases

Input

"Make them look like flight attendants"

"Make them look like doctors"

Figure 14. Our method reflects biases from the data and models it is based upon, such as correlations between profession and gender.

InstructPix2Pix - More Reading

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

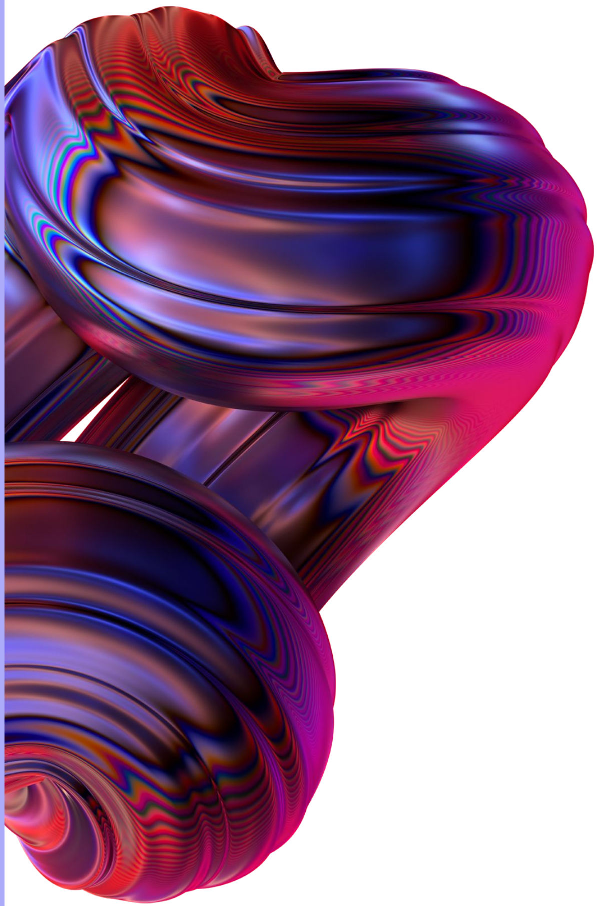
Nataniel Ruiz[*,1,2] Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University

Figure 1. With just a few images (typically 3-5) of a subject (left), *DreamBooth*—our AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and variation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject.

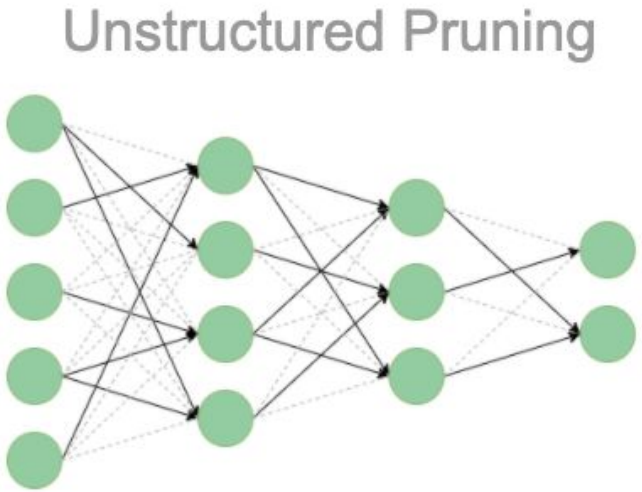
Integral Neural Networks

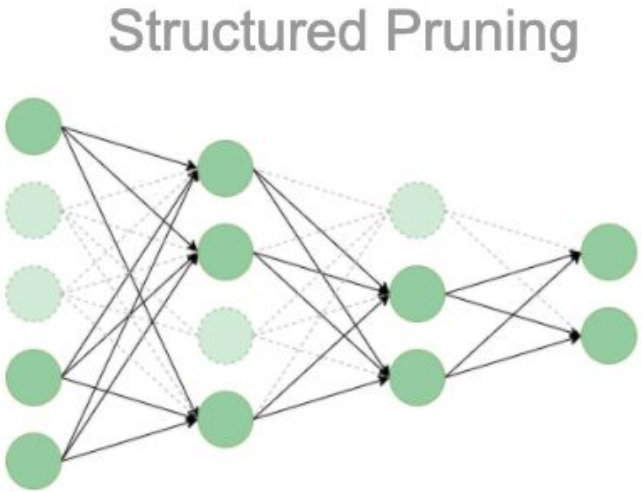
Kirill Solodskikh[*†] Azim Kurbanov[*†] Ruslan Aydarkhanov[†]
Irina Zhelavskaya Yury Parfenov Dehua Song Stamatios Lefkimmiatis
Huawei Noah's Ark Lab

CVPR Award Candidate



Traditional Pruning



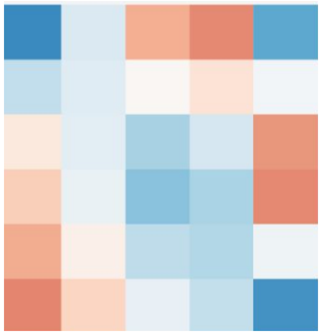


aka “weight pruning”

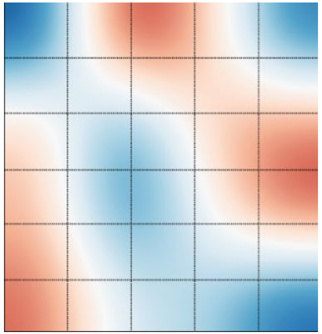
aka “node pruning”

Source: https://neuralmagic.com/blog/pruning-overview/

Integral Neural Networks - Key Idea

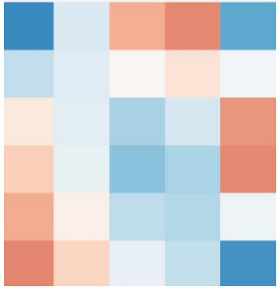
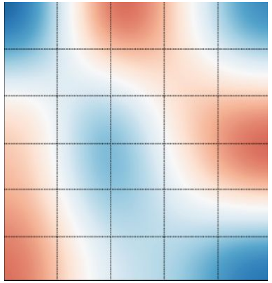


Traditional weight tensor
Discrete weights



Smooth weight representation
Sample weights at grid
Grid res. influences compression

Integral Neural Networks

	Regular NN	Integral NN
		
Weights	Discrete multi-dim. tensors	Smooth multi-dim. functions
Computation	Discrete transformations of inputs	Continuous integration operations Can be discretized at inference
Fine-tuning	Usually necessary after pruning	Not necessary
Deployment	Fixed model size after pruning	Resize model on-the-fly (e.g. on edge device)

Integral Neural Networks

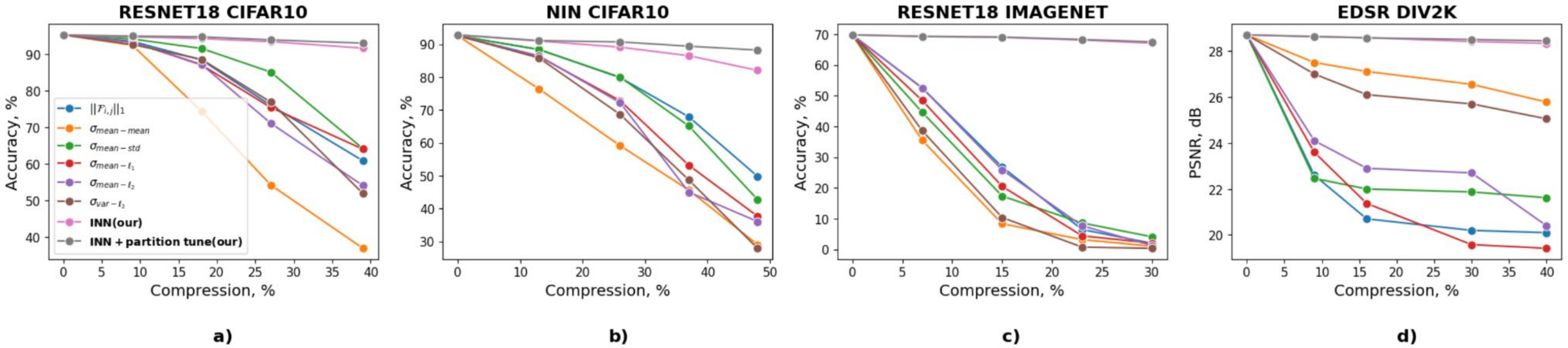


Figure 1. Visualization of different channels selection methods without fine-tuning compared with our proposed integral neural networks. a) ResNet-18 on Cifar10. b) NIN architecture on Cifar10. c) ResNet-18 on ImageNet. d) 4x EDSR on Div2k validation set. By compression we denote the percentage of deleted parameters.

Compressing pre-trained nets

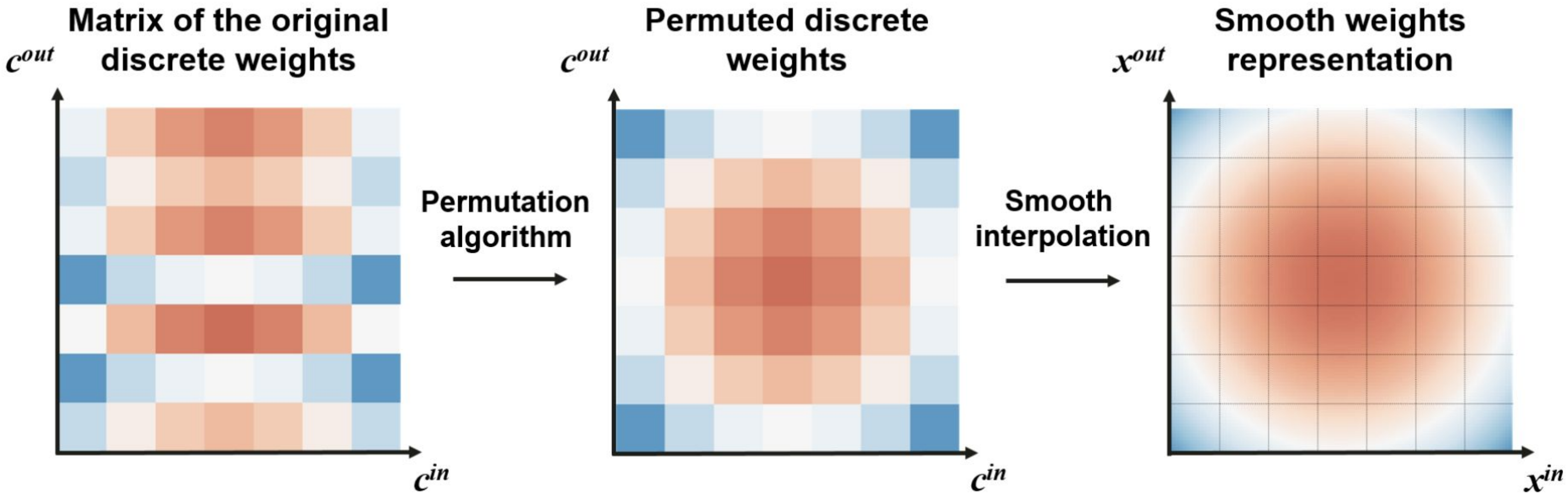


Figure 5. Toy example illustrating the permutation of filters in a discrete weight tensor in order to obtain a smoother structure.

CVPR Impressions

Papers

Food for Thought

Embodied Foundation Models

Robotics and CV

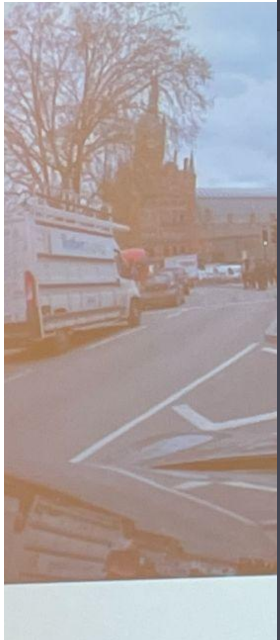
AI and “The Hype”

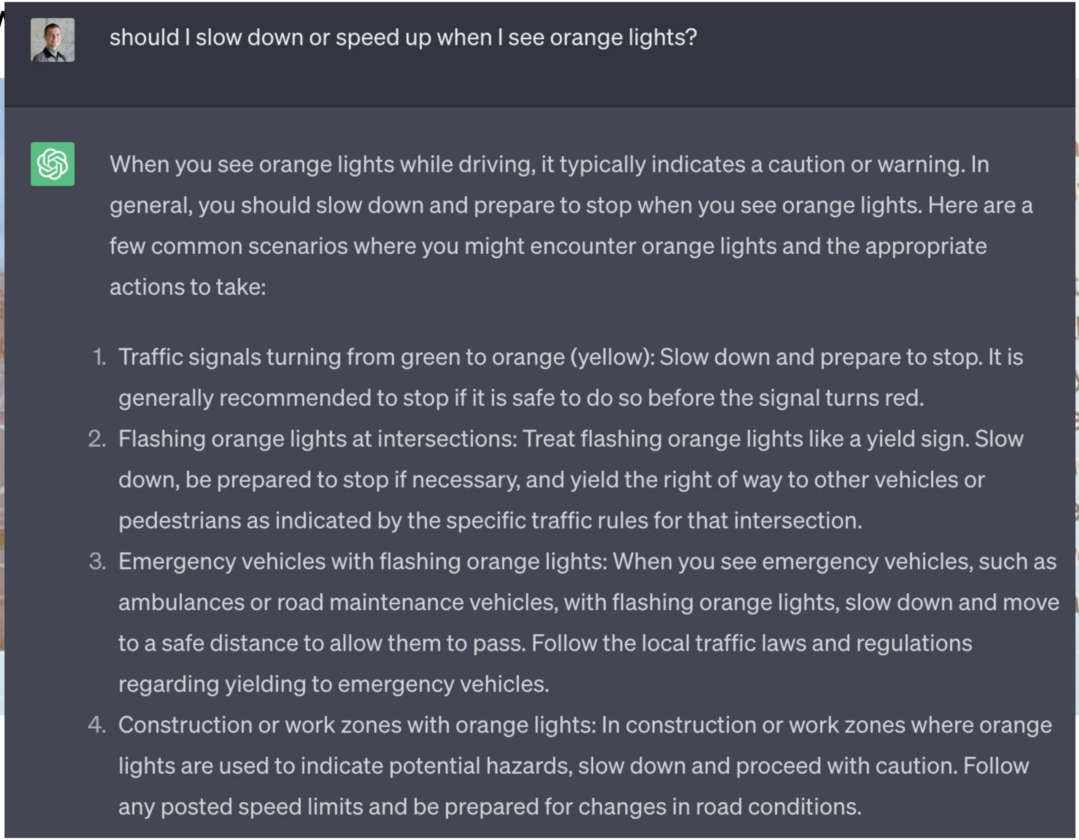
Foundation Models as Motion Planners (1/2)

Driess et al. 2023 “PaLM-E: An Embodied Multimodal Language Model”

Foundation Models as Motion Planners (2/2)

- VQA Model by W





should I slow down or speed up when I see orange lights?

When you see orange lights while driving, it typically indicates a caution or warning. In general, you should slow down and prepare to stop when you see orange lights. Here are a few common scenarios where you might encounter orange lights and the appropriate actions to take:

1. Traffic signals turning from green to orange (yellow): Slow down and prepare to stop. It is generally recommended to stop if it is safe to do so before the signal turns red.
2. Flashing orange lights at intersections: Treat flashing orange lights like a yield sign. Slow down, be prepared to stop if necessary, and yield the right of way to other vehicles or pedestrians as indicated by the specific traffic rules for that intersection.
3. Emergency vehicles with flashing orange lights: When you see emergency vehicles, such as ambulances or road maintenance vehicles, with flashing orange lights, slow down and move to a safe distance to allow them to pass. Follow the local traffic laws and regulations regarding yielding to emergency vehicles.
4. Construction or work zones with orange lights: In construction or work zones where orange lights are used to indicate potential hazards, slow down and proceed with caution. Follow any posted speed limits and be prepared for changes in road conditions.

“World Models” - GAIA-1 by Wayve

- World model = a generative model that predicts what happens next conditioned on an action
- Autoregressive model trained on Wayve’s large unlabeled dataset

“World Models” - GAIA-1 by Wayve

- Autonomous driving may be the first example of where we see embodied AI working

CVPR Impressions

Papers

Food for Thought

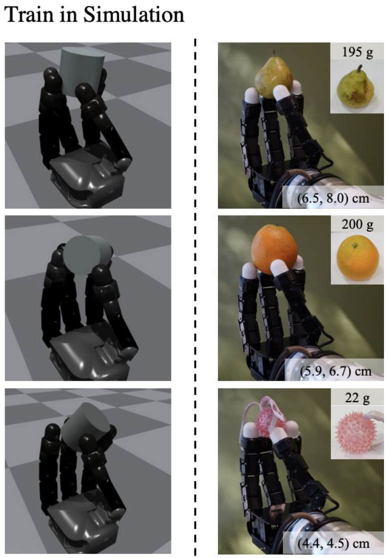
Embodied Foundation Models

Robotics and CV

AI and “The Hype”

Robotics and Computer Vision

- "Robotics is the next big thing"
- Jitendra Malik
  - ~ "Vision has no use by its own. It needs to guide action."
  - ~ "Robotics is 20 years behind computer vision"
  - ~ "Navigation and locomotion are close to being solved."
  - ~ "Manipulation is far from being solved" → Why?
    - Control struggles with making and breaking of contact
    - RL struggles with inaccurate simulations and sim to real gap
    - Lack of dexterous multi-fingered hands
  - Urged the CV community to venture into manipulation
- Differences CV and robotics
  - no standardized benchmarks
  - no large datasets
  - sim to real gap
  - hardware experiments are essential but take long
- Particular hot topics: visual pre-training for robotics, object representations



Qi et al. CoRL 2022

CVPR Impressions

Papers

Food for Thought

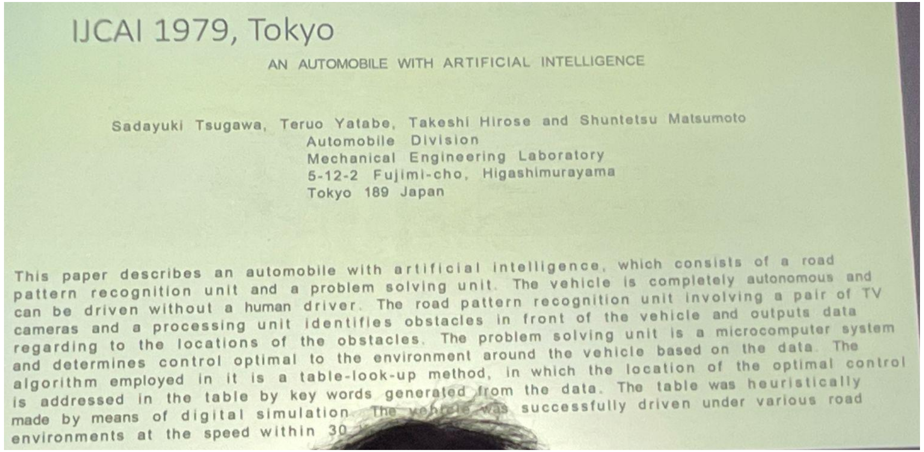
Embodied Foundation Models

Robotics and CV

AI and “The Hype”

AI and "The Hype"

- Rodney Brooks, also see "The Seven Deadly Sins of AI Predictions" blog
  - Roy Amara (1925 – 2007): "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."
  - Example: Fear that computers will replace librarians, librarians kept on going for 40 years until eventually largely impacted by the internet and mobile devices.

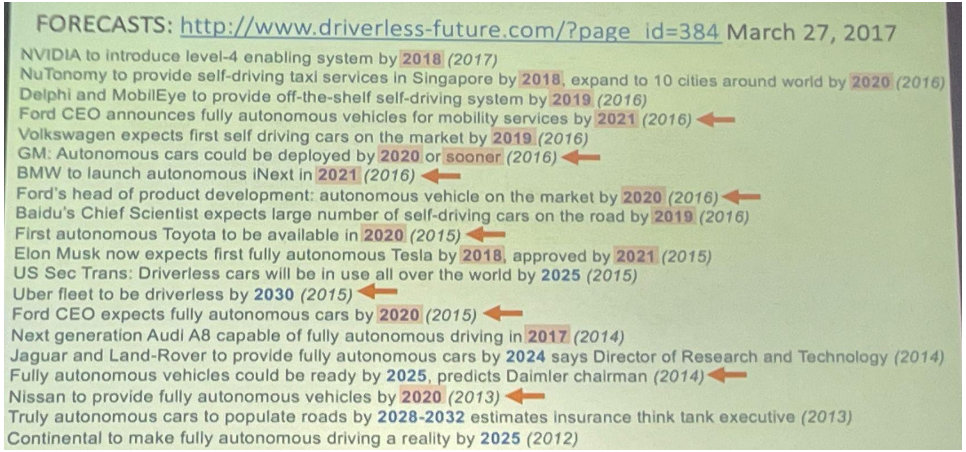


IJCAI 1979, Tokyo

AN AUTOMOBILE WITH ARTIFICIAL INTELLIGENCE

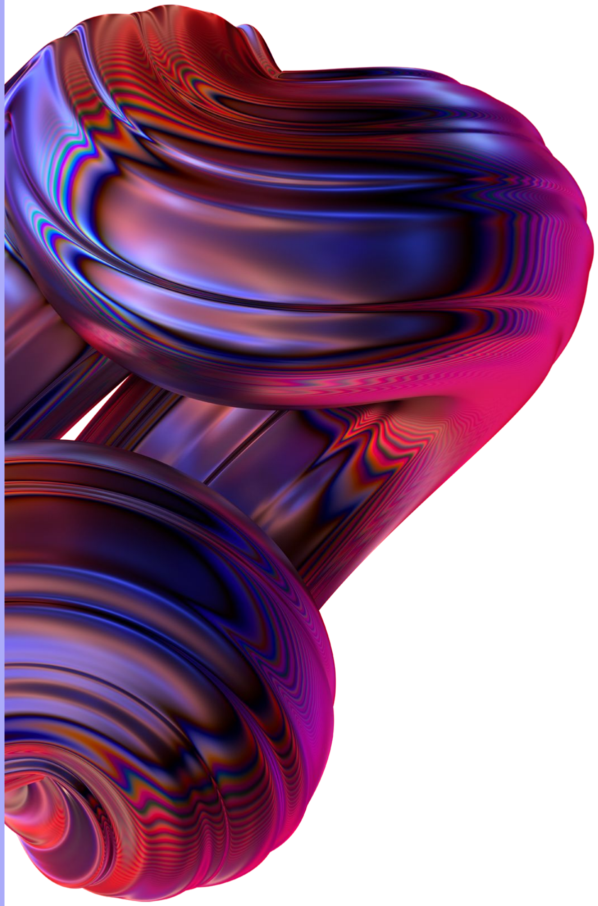
Sadayuki Tsugawa, Teruo Yatabe, Takeshi Hirose and Shuntetsu Matsumoto
Automobile Division
Mechanical Engineering Laboratory
5-12-2 Fujimi-cho, Higashimurayama
Tokyo 189 Japan

This paper describes an automobile with artificial intelligence, which consists of a road pattern recognition unit and a problem solving unit. The vehicle is completely autonomous and can be driven without a human driver. The road pattern recognition unit involving a pair of TV cameras and a processing unit identifies obstacles in front of the vehicle and outputs data regarding to the locations of the obstacles. The problem solving unit is a microcomputer system and determines control optimal to the environment around the vehicle based on the data. The algorithm employed in it is a table-look-up method, in which the location of the optimal control is addressed in the table by key words generated from the data. The table was heuristically made by means of digital simulation. The vehicle was successfully driven under various road environments at the speed within 30



"Don't be the best, be the only!"

Thanks!



Alexander Koenig
alexander.koenig@merantix.com