Self-supervised Learning from Images and Videos

YUKI M ASANO
AT
SBA WORKSHOP, UNIVERSITY OF BIRMINGHAM

UNIVERSITY OF AMSTERDAM

Hi, I'm Yuki

- Assistant Professor with Video & Image Sense (VIS) Lab
  - Self-supervised Learning
  - Multi-modal Learning
  - Large Model Adaptation

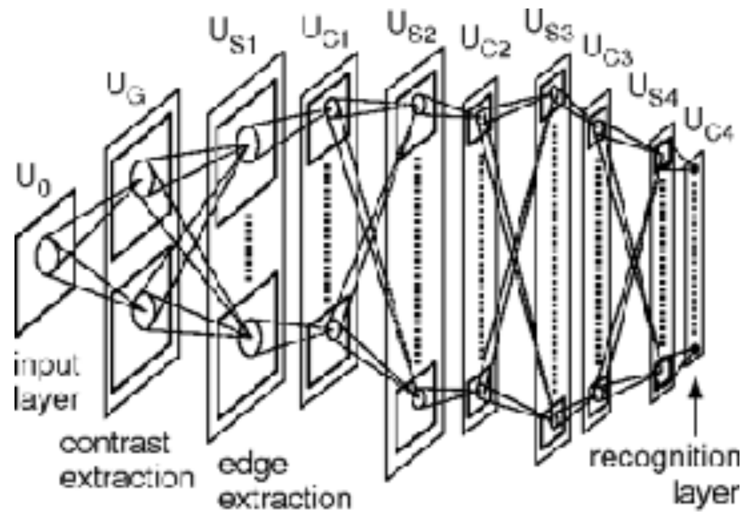
focus of today

- More info: https://yukimasano.github.io/

The field of AI has made rapid progress, the crucial fuel is data

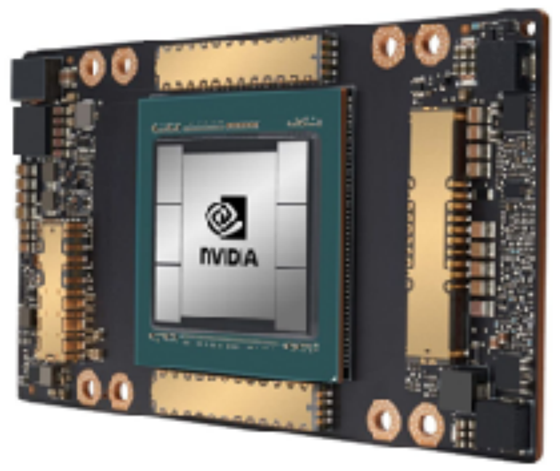
Algorithms

Deep neural networks



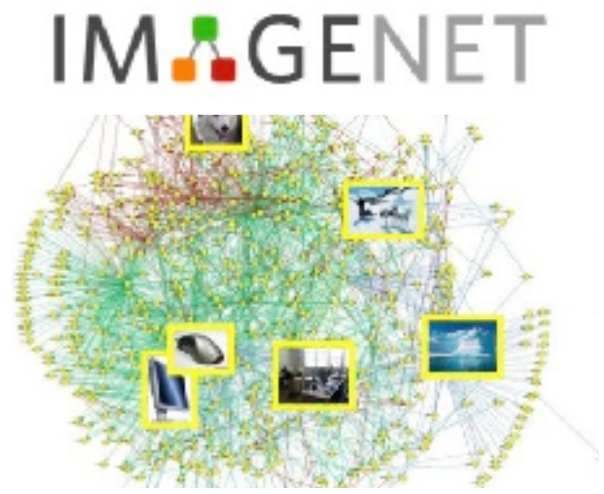
Hardware

GPUs



Data

Large scale datasets



UNIVERSITY OF AMSTERDAM
YUKI ASANO

Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, K., Biol. Cybernetics 1980
Object Recognition with Gradient-Based Learning. LeCun et al. Shape, Contour and Grouping in Computer Vision 1999
ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.
ImageNet Classification with Deep Convolutional Neural Networks., Krizhevsky et al., NeurIPS 2012

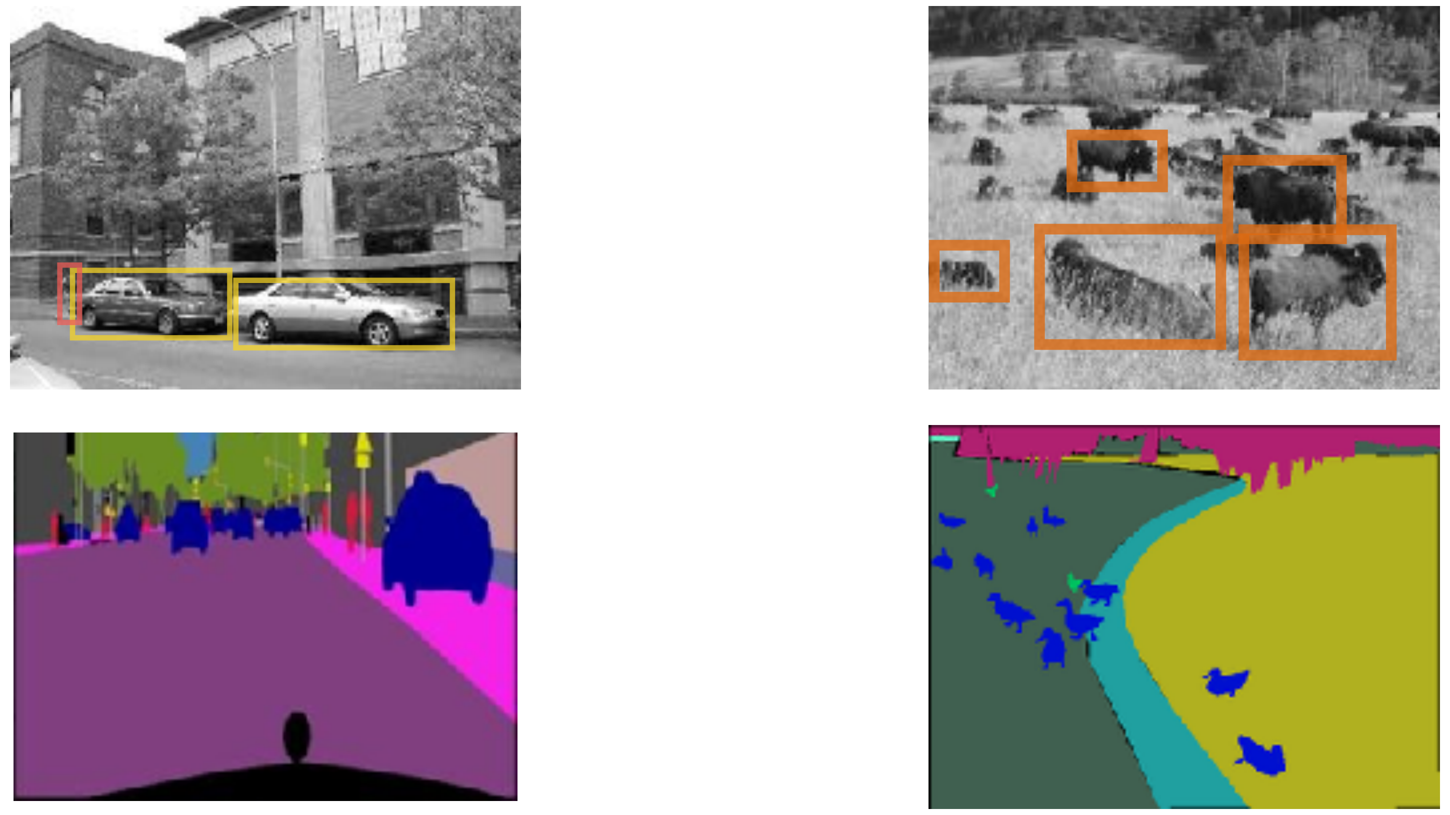
Manual annotations for the data are limiting.

Supervised Learning

Images are often cheap

"Railroad crossing"

"Trogon"



But manual annotations are expensive:
e.g. 30min per image / requiring experts

ImageNet: A Large-Scale Hierarchical Image Database. Dong et al. CVPR 2009
The Cityscapes Dataset for Semantic Urban Scene Understanding. Cordts et al. CVPR 2016
Scene parsing through ADE20K dataset. Zhou et al. CVPR 2017.

Self-supervised Learning replaces the need for labels & annotations.

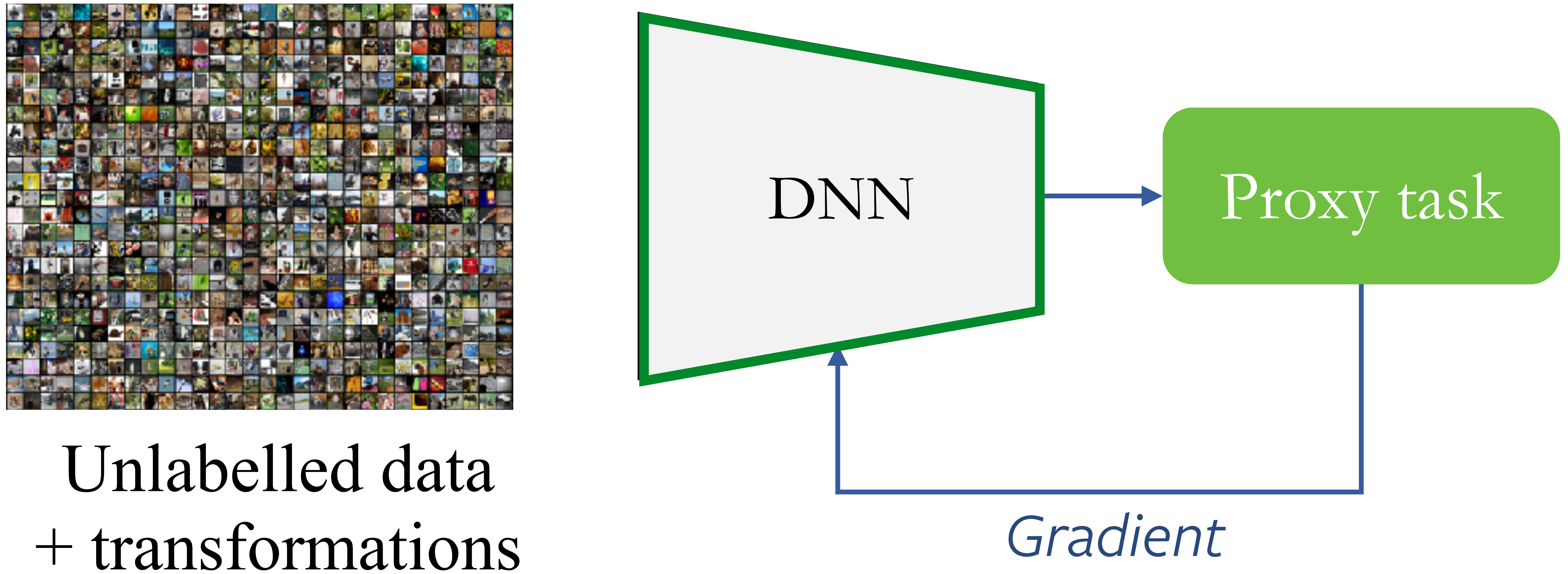
Self-supervised Learning

Extract a free supervisory signal from the raw data

ImageNet: A Large-Scale Hierarchical Image Database. Dong et al. CVPR 2009
The Cityscapes Dataset for Semantic Urban Scene Understanding. Cordts et al. CVPR 2016
Scene parsing through ADE20K dataset. Zhou et al. CVPR 2017.

General procedure of self-supervised learning.

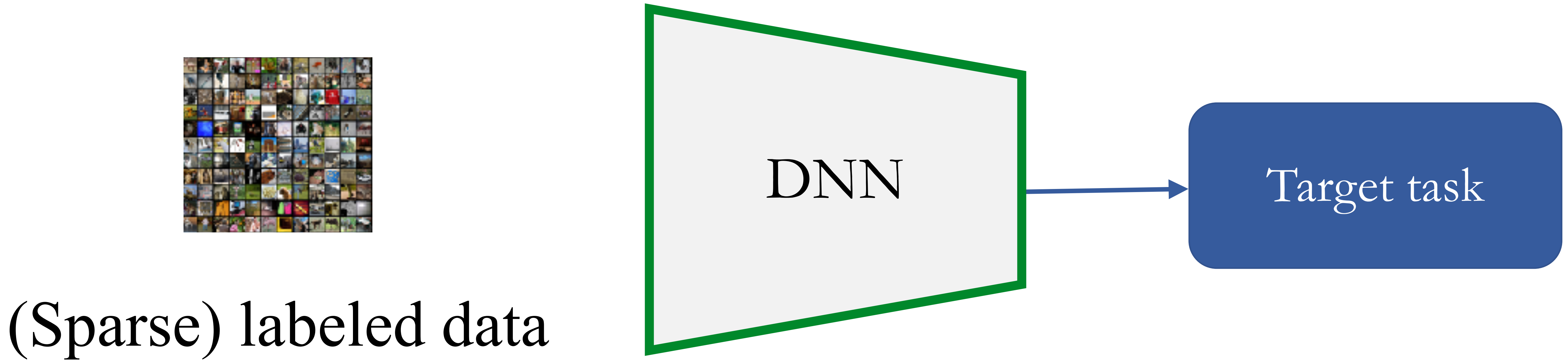
Phase 1: Pretraining



For example:

1) rotate image by multiple of 90 degrees
2) have network predict the rotation

Phase 2: Downstream tasks

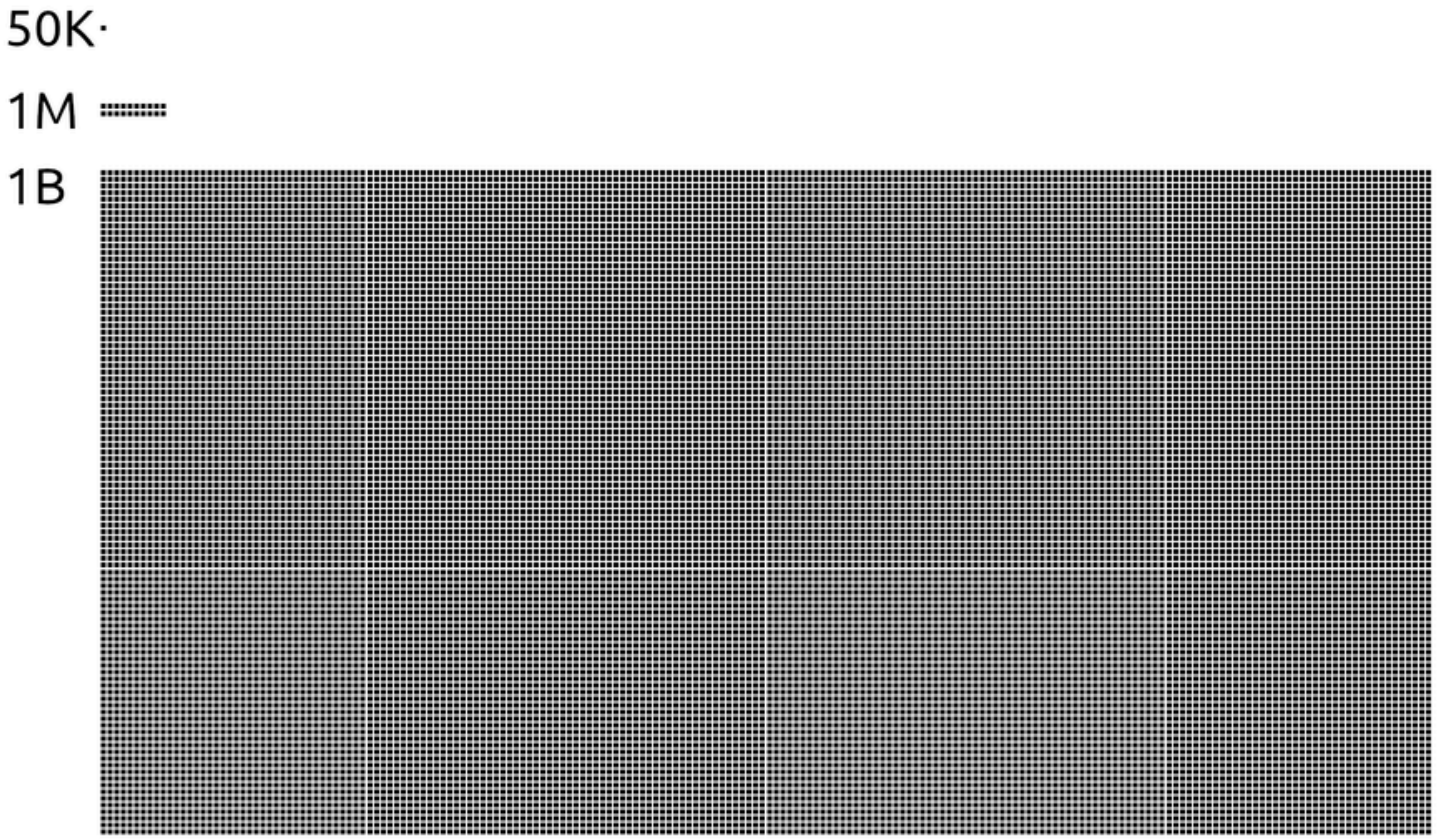


Typically has less data

Why do we want to do self-supervised learning?

Reason 1: Scalability and GPT as proof-of-concept

Instagram: >50B images



Annotation is expensive, yet datasets keep getting bigger.

Reason 2: Constantly changing domains

Unclear when & what to relabel. Again, large costs just to "keep up".

Reason 3: Ambiguity of labels and captions

"A house"?

"A boat"?

Example from Flickr30k

A hot, blond girl getting criticized by her boss.

Problematic captions

Labels are ambiguous at best, discriminating and bias-propagating at worst.
Do we really wish to provide our models with these priors?

https://en.wikipedia.org/wiki/List_of_house_styles
https://www.shutterstock.com/image-illustration/flat-ships-sailing-yachts-marine-sailboats-1903407259
https://excavating.ai/ Crawford & Paglen
Stereotyping and Bias in the Flickr30K Dataset. Miltenburg. MMC. 2016.

Especially videos open exciting new directions

Visual development for AI

"Get" physics

Embodied AI

Bonus: *insane* scale:

Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.

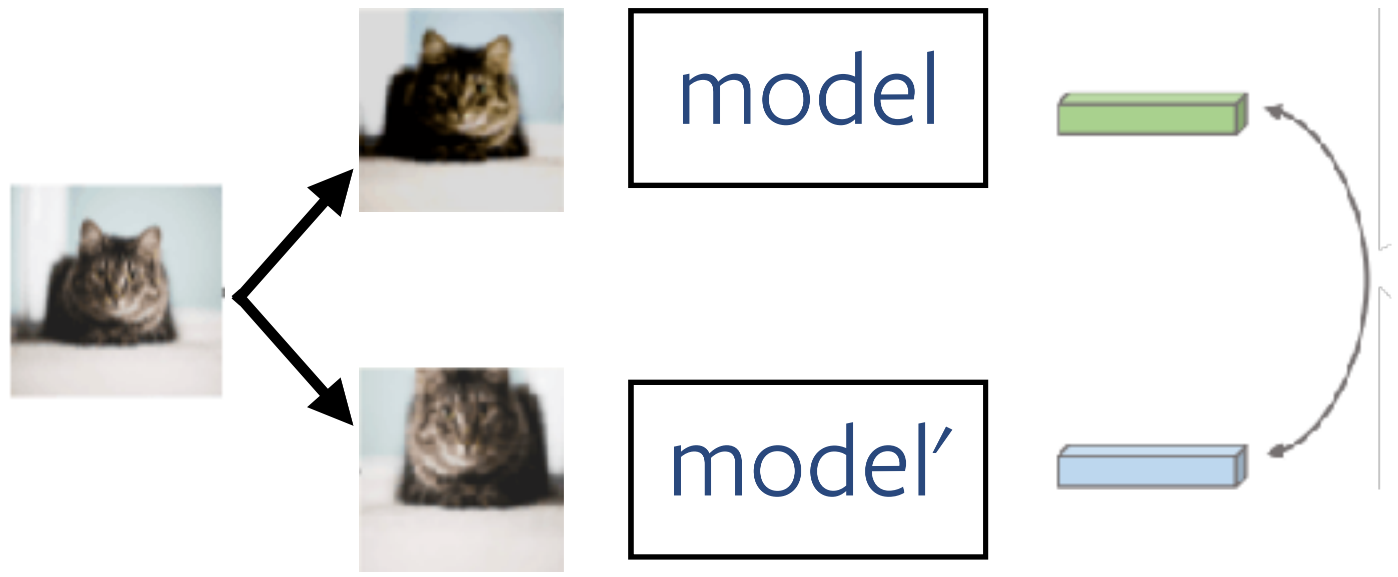
VENKATARAMANAN, RIZVE, CARREIRA, AVRITHIS*, ASANO*.

ICLR 2024

UNIVERSITY OF AMSTERDAM

Augmentations are crucial in classic image-SSL, but forcing frames to be invariant is limiting

Images: SimCLR, MoCo, SwaAV et al.



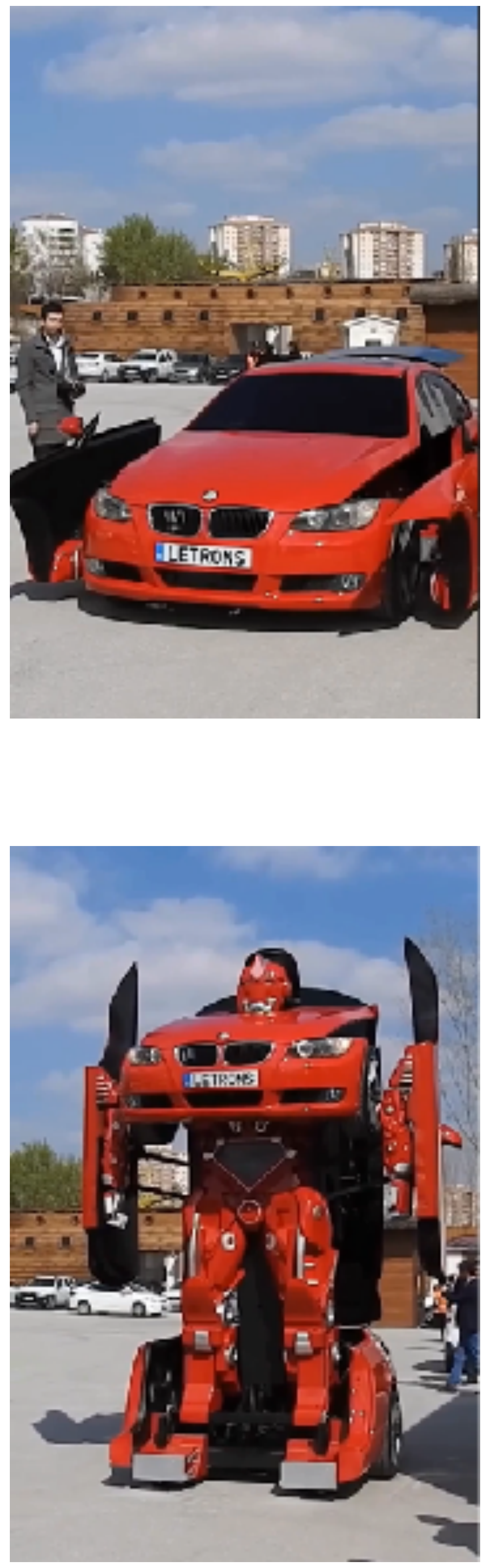
key principle: view-invariance

Might be ok for videos like this:

But does this generally make sense?

Frame 1

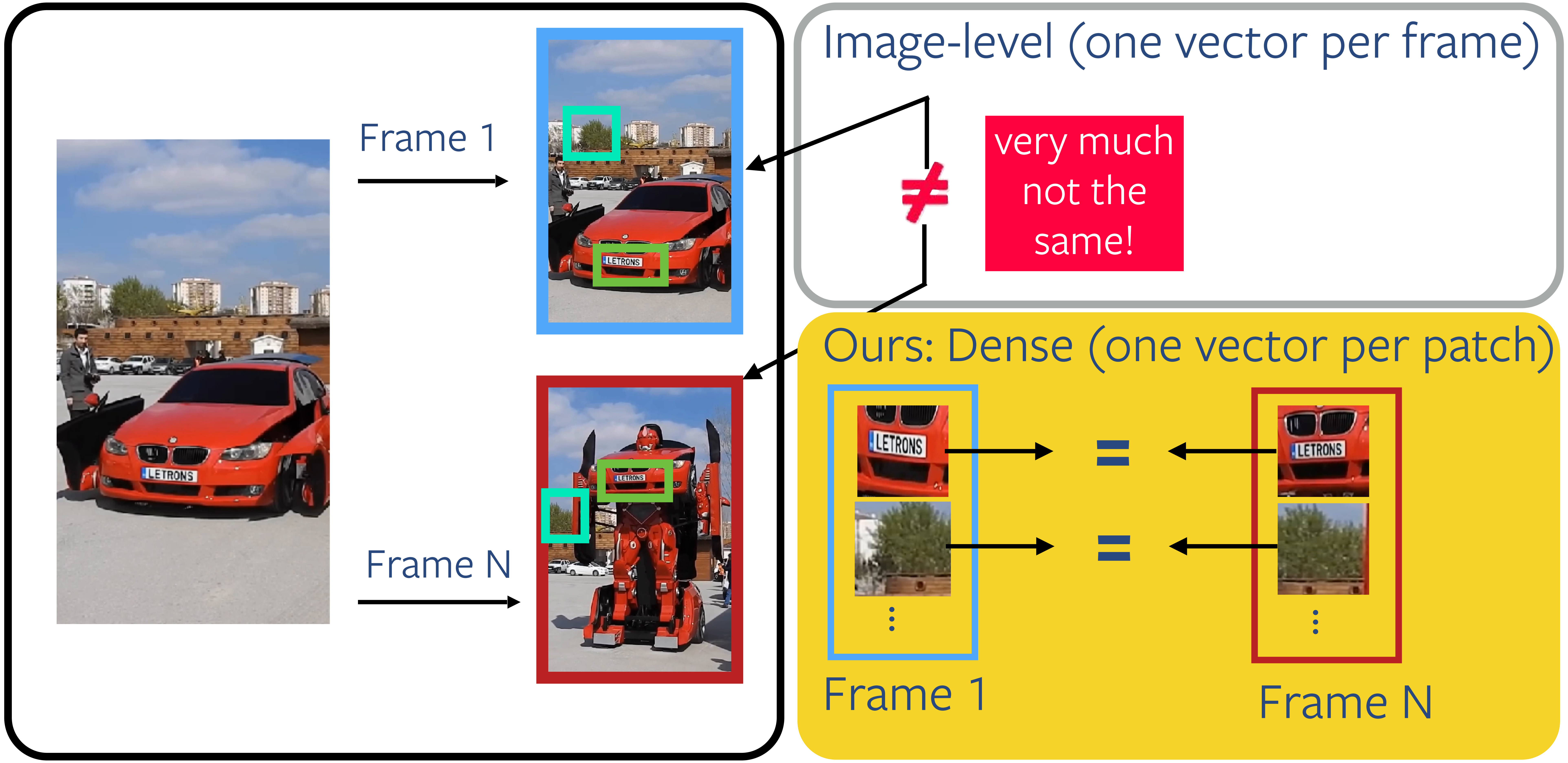
Frame N



≠

very much not the same!

Solution is obvious



Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

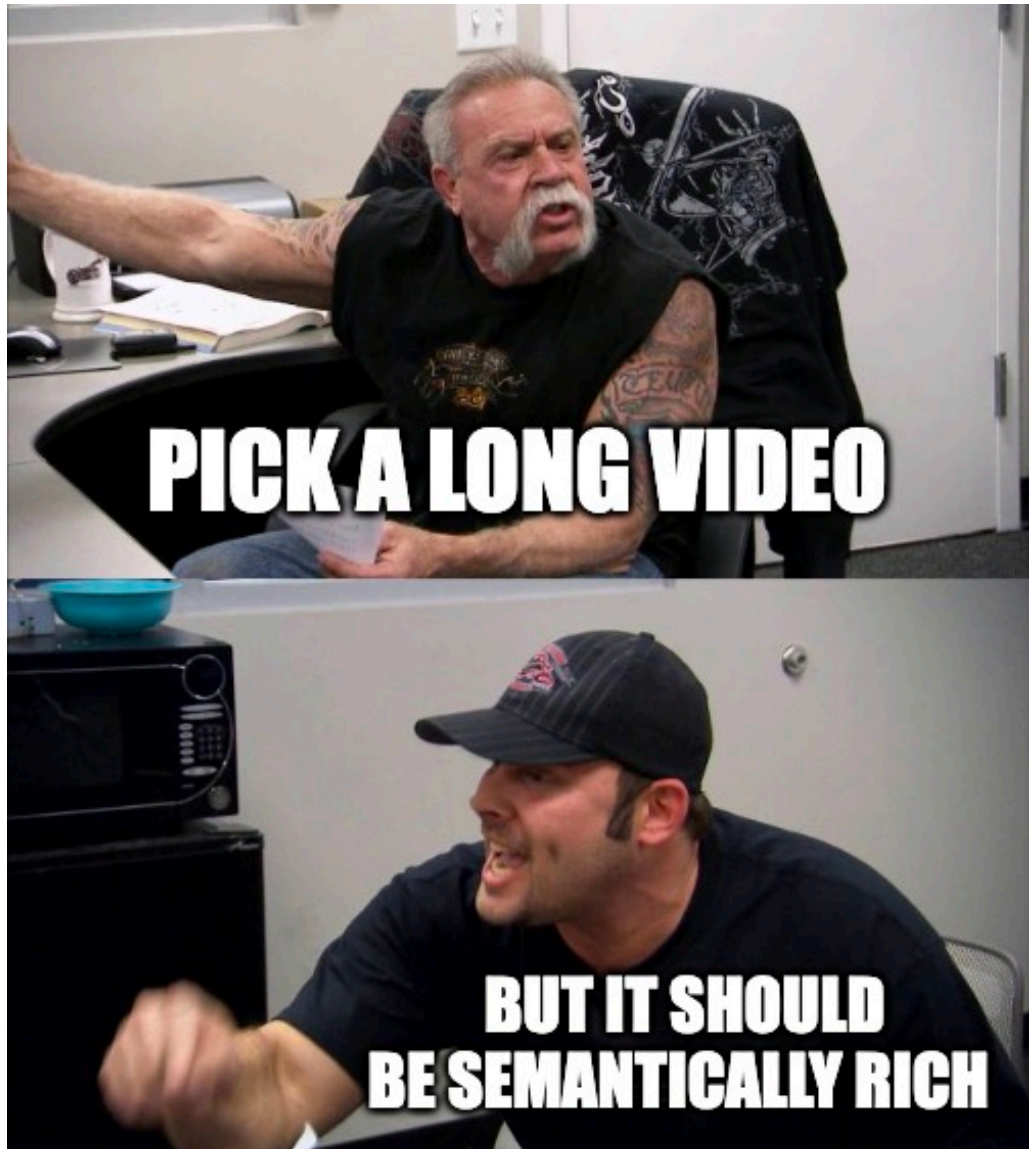
TimeTuning: pretrained model & use temporal info of videos.

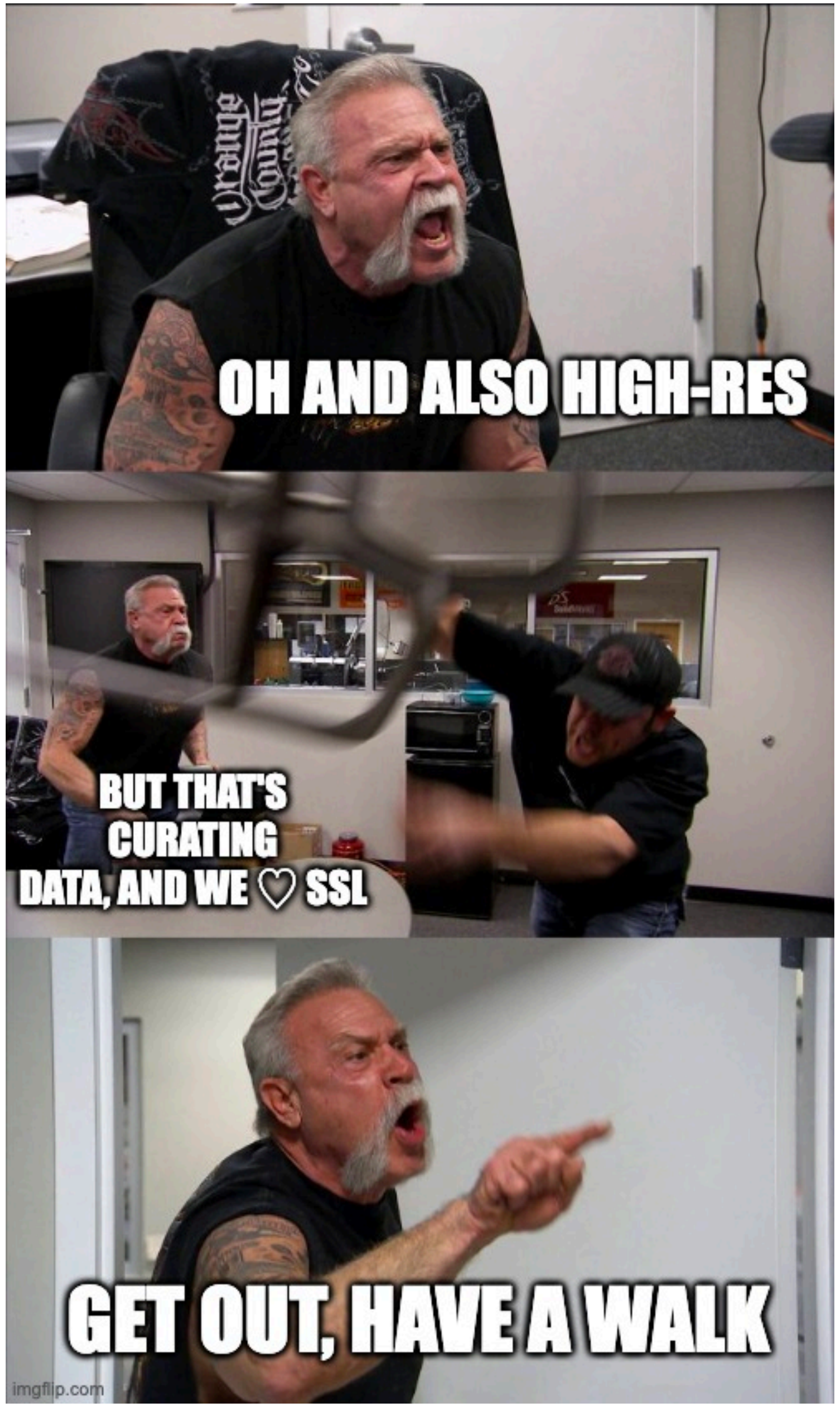
How powerful is time without image-pretraining?

Study the extreme: try to learn from a **single video**, **from scratch**.

Motivated by: Asano Rupprecht, Vedaldi. *A critical analysis of self-supervision, or what we can learn from a single image*. ICLR 2020
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations*. ICCV 2023

Us figuring out which video to use



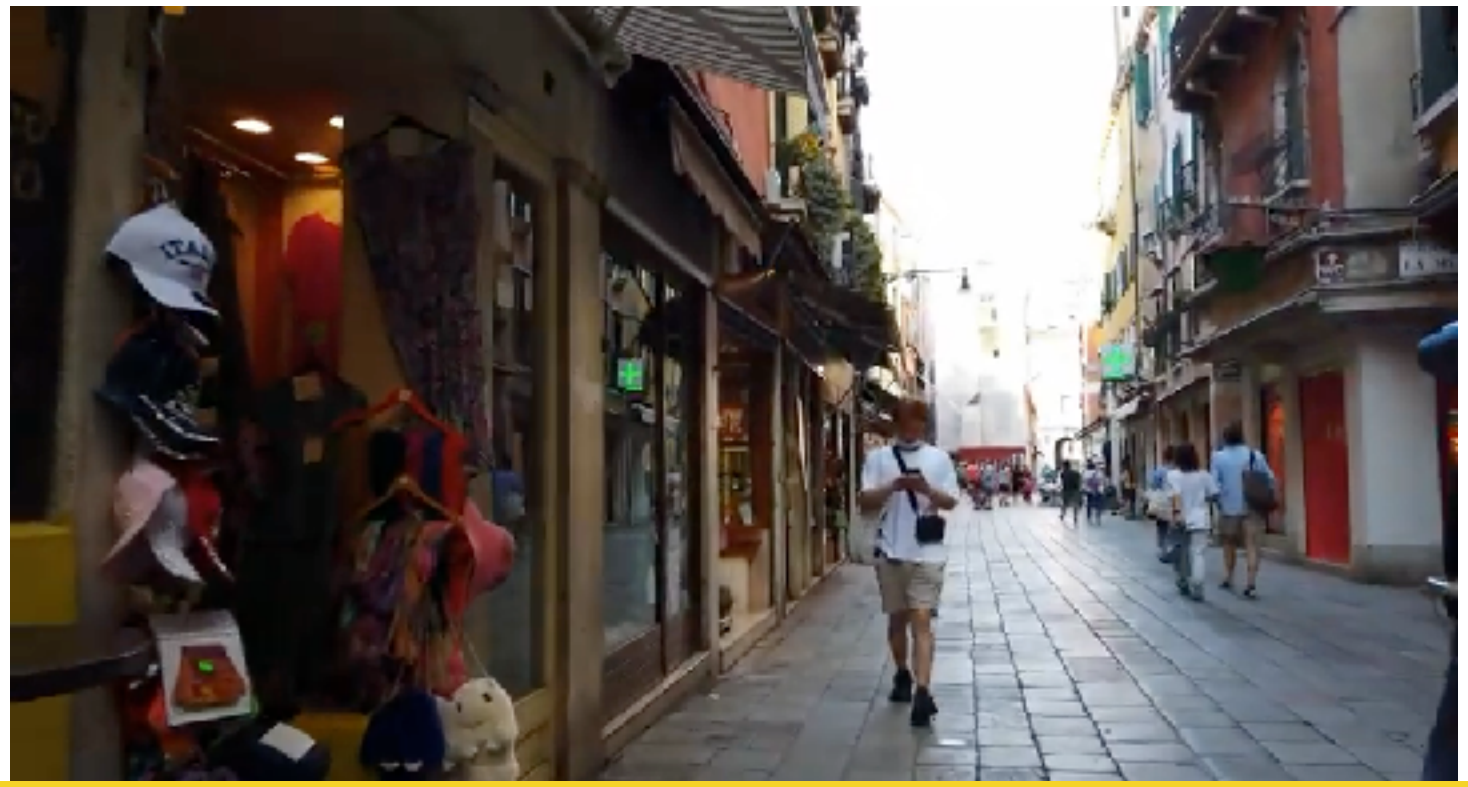


- ✓ Long
- ✓ High-res, smooth
- ✓ Semantically rich
- ✓ Scalable (for SSL)

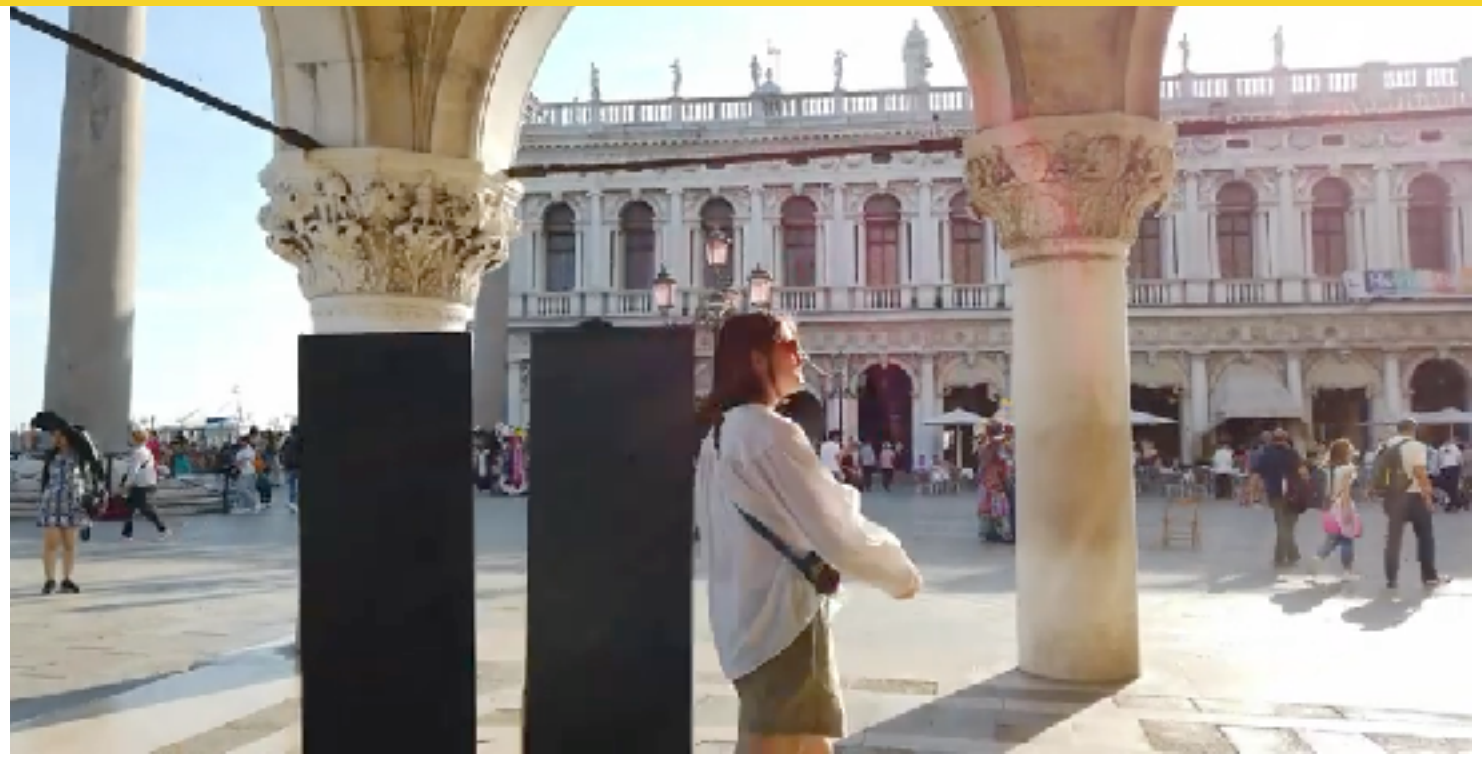
↓

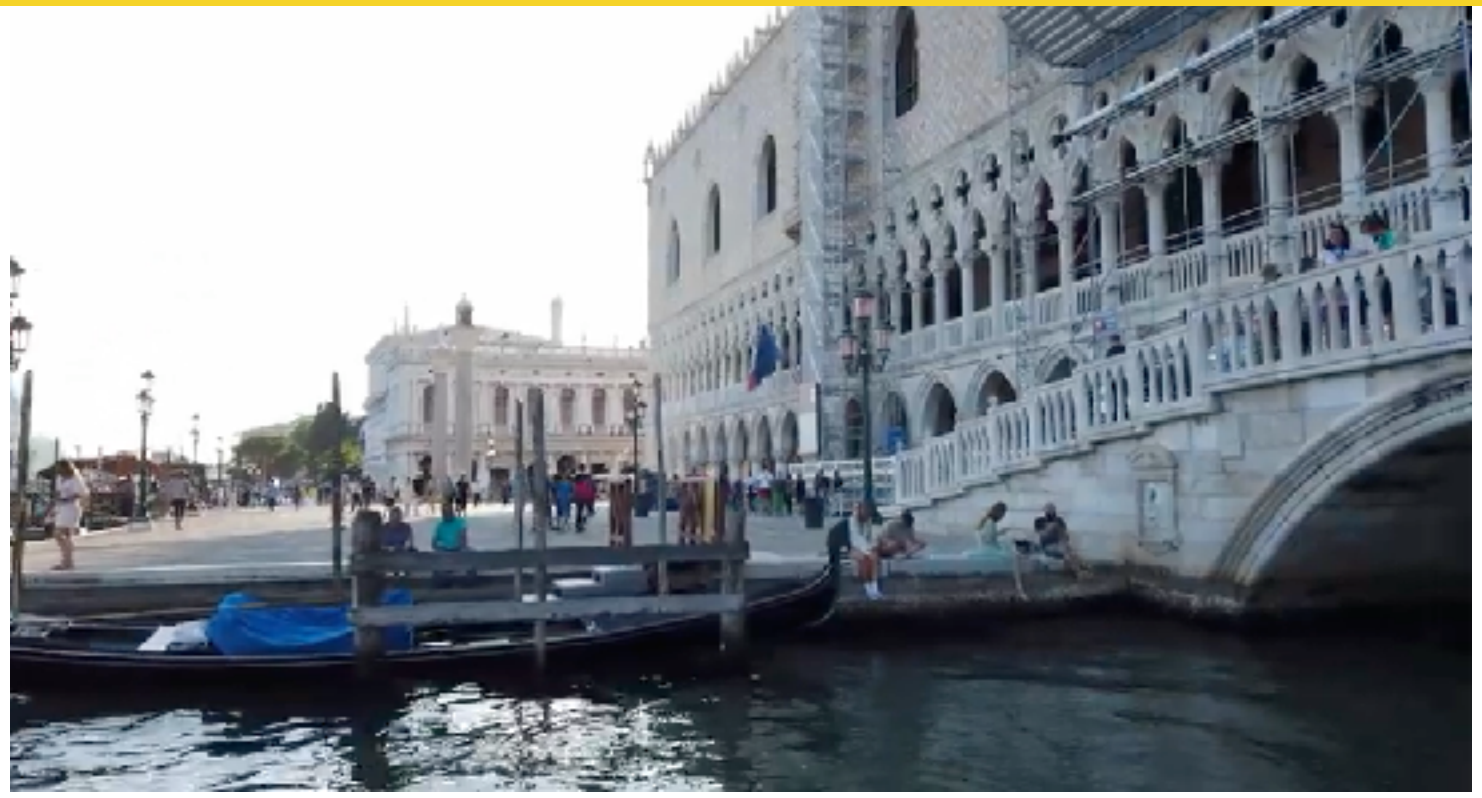
Walking Tours

WTours also used for learning video compression in ACCV 2022: Wiles et al. Compressed Vision for Efficient Video Understanding.



The dataset consists of 10x 4K videos of different cities' Walking Tours.



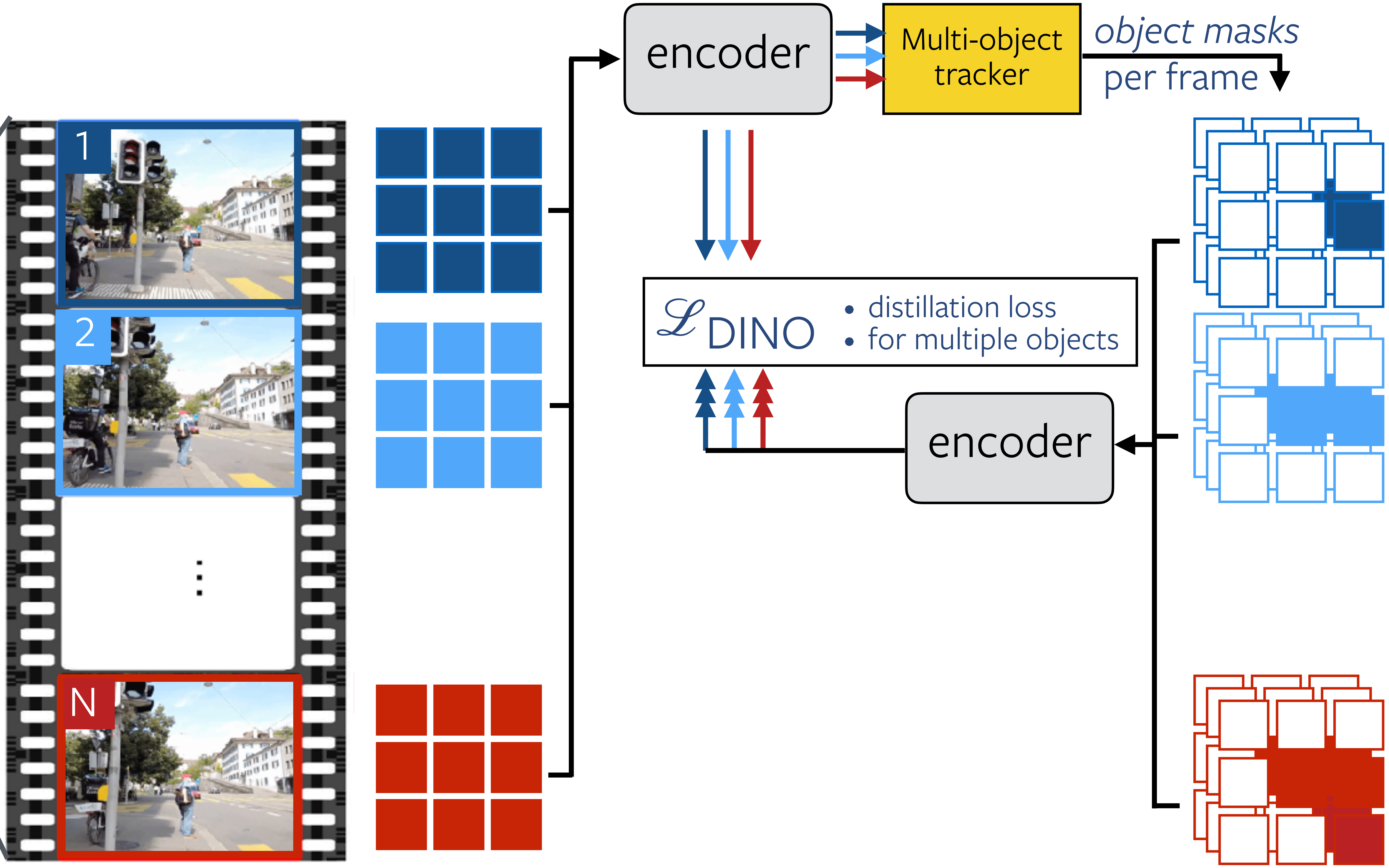


WT Venice: https://www.youtube.com/watch?v=fGX0Te6pFvk. CC-BY Poptravel.

Dora: **D**isc**o**ver and T**ra**ck

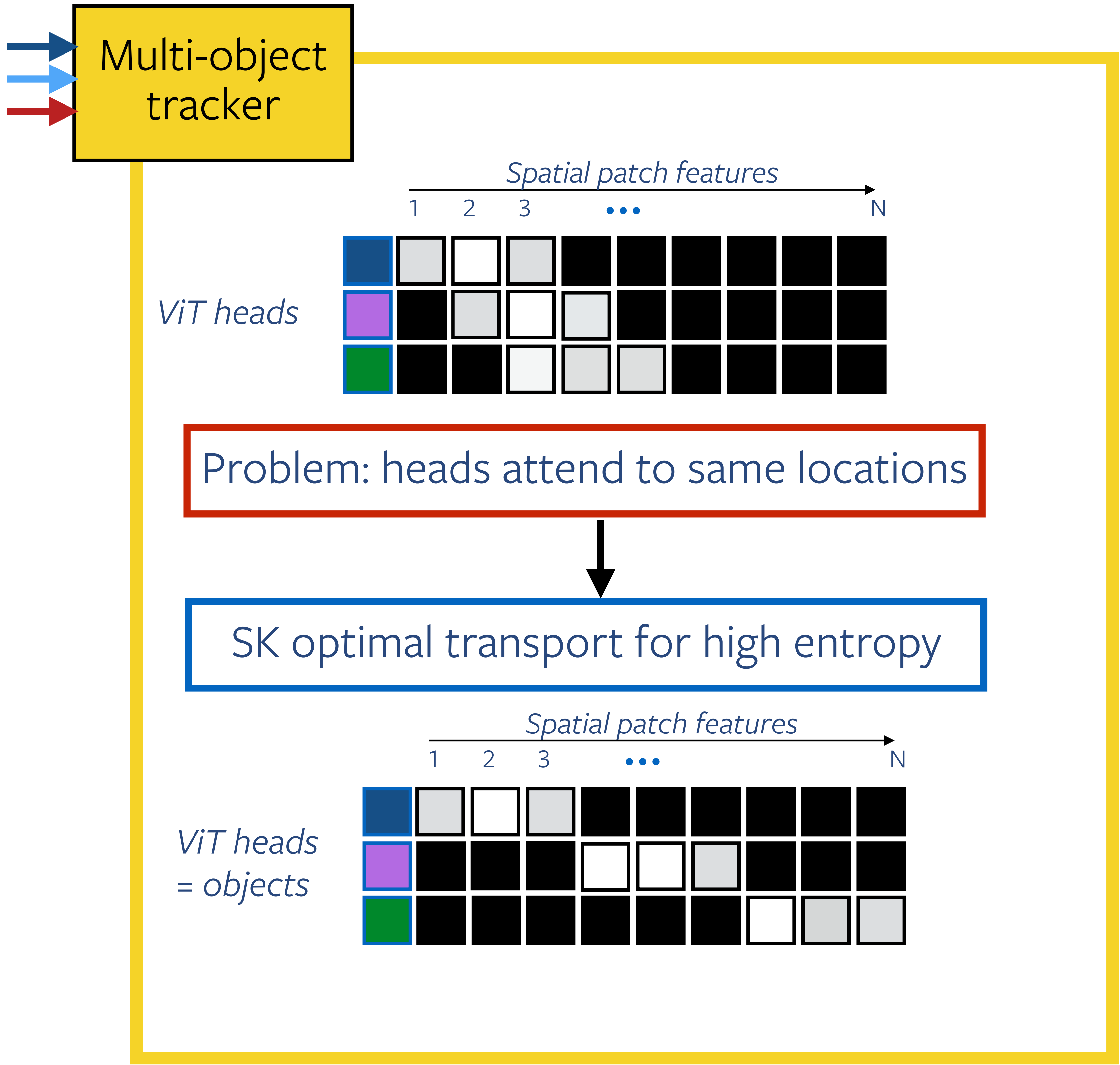
High-level idea:
1) track multiple objects across time
2) enforce invariance of features across time

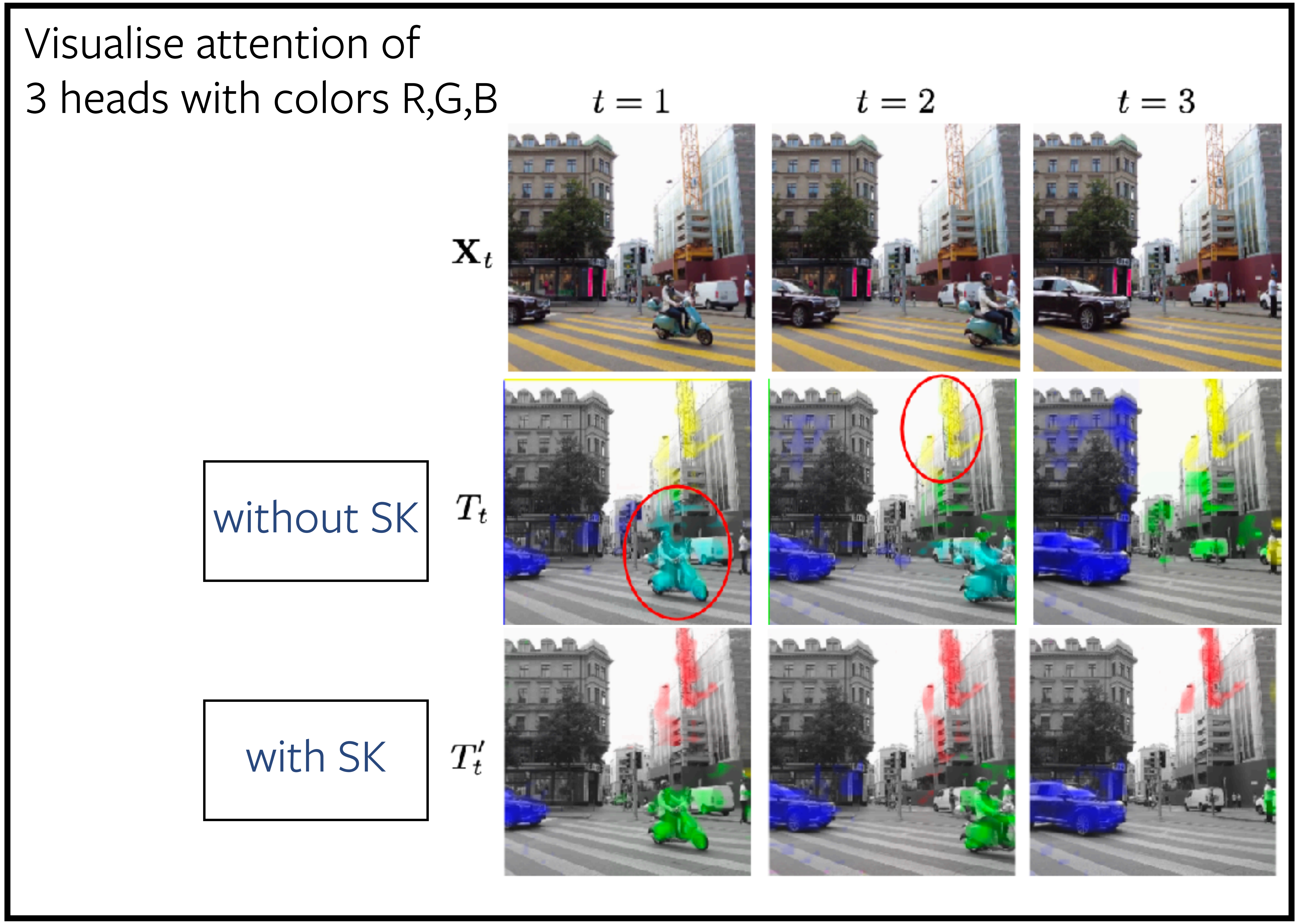
Much like Dora, we walk around and learn from what we see.



Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

Spreading attention with Sinkhorn-Knopp





Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

More examples: multi-object tracking in a ViT *emerges*

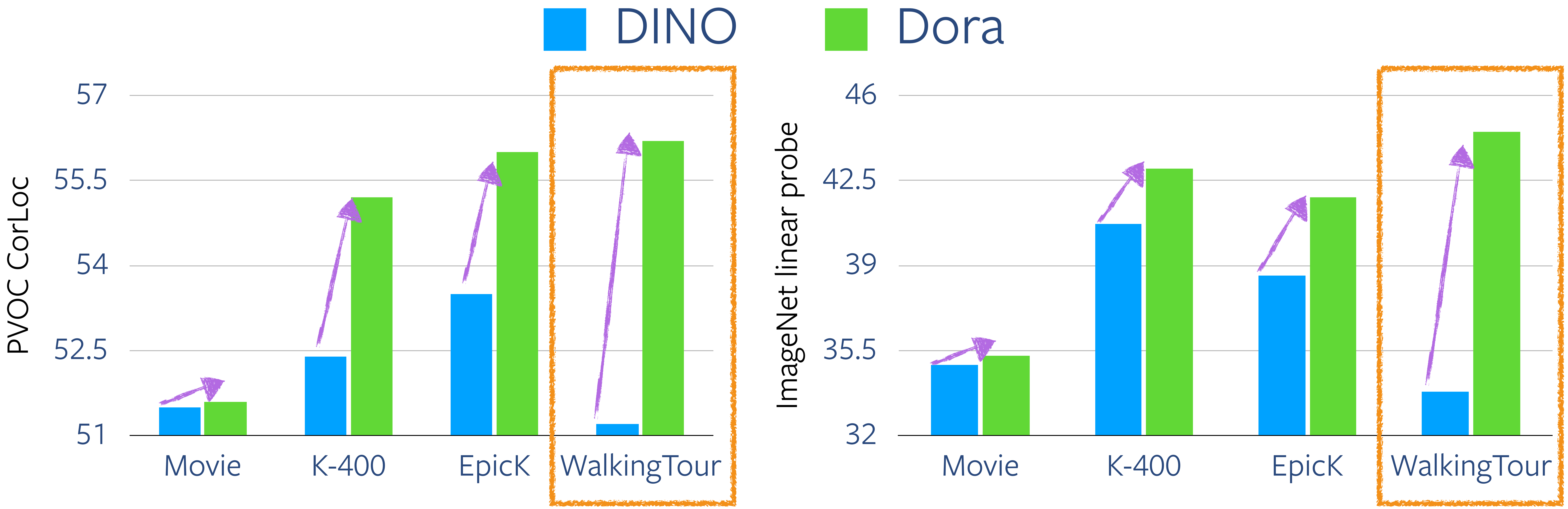
Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

Can we obtain performances better than training on images?

yes!

(I have no idea where this gif is from)

Dora better than DINO
WT+ Dora: great match



Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

The Augmented Image Prior: Distilling 1000 Classes by Extrapolating from a Single Image.

SAEED*, ASANO*.

ICLR 2023

UNIVERSITY OF AMSTERDAM

≈

Wood rabbit

≈

0.4 x n02087394

0.2 x n01443537

0.1 x n07697537

...

!?

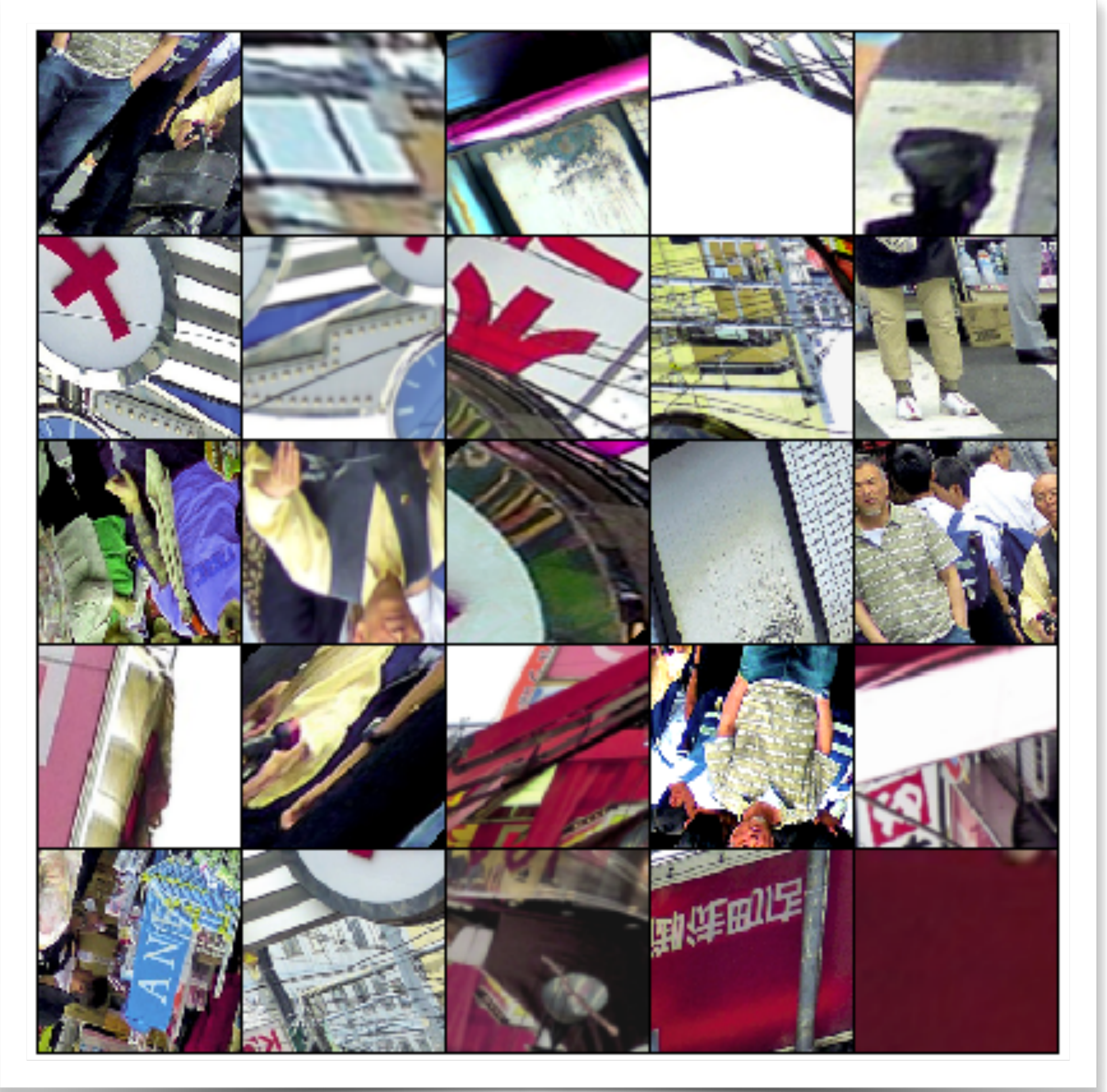
How can we test this fairytale?

A single image I

Pretrained neural network (teacher)

Randomly initialized network (student)

Augmentations $\mathscr{A}(I)$

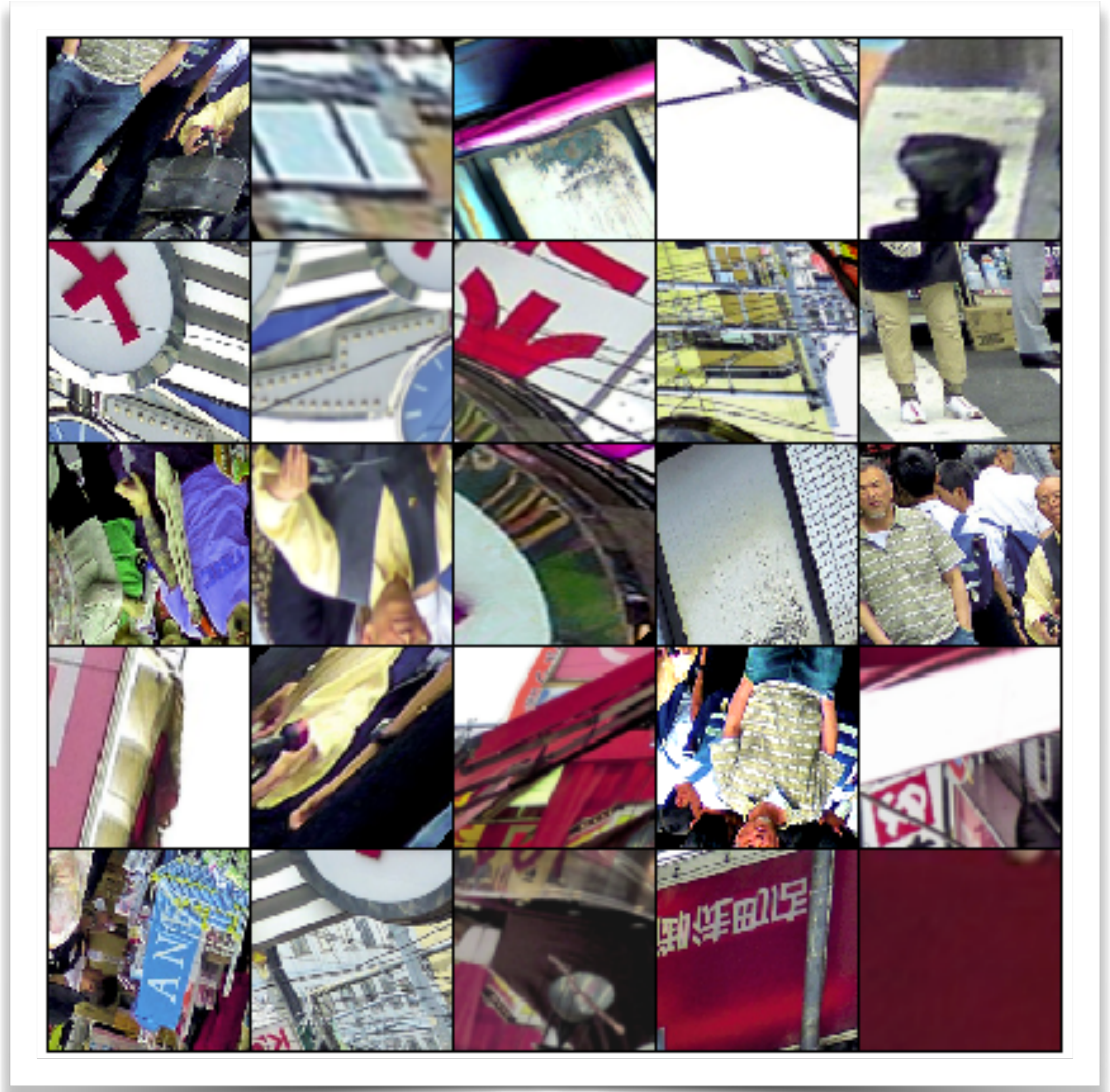


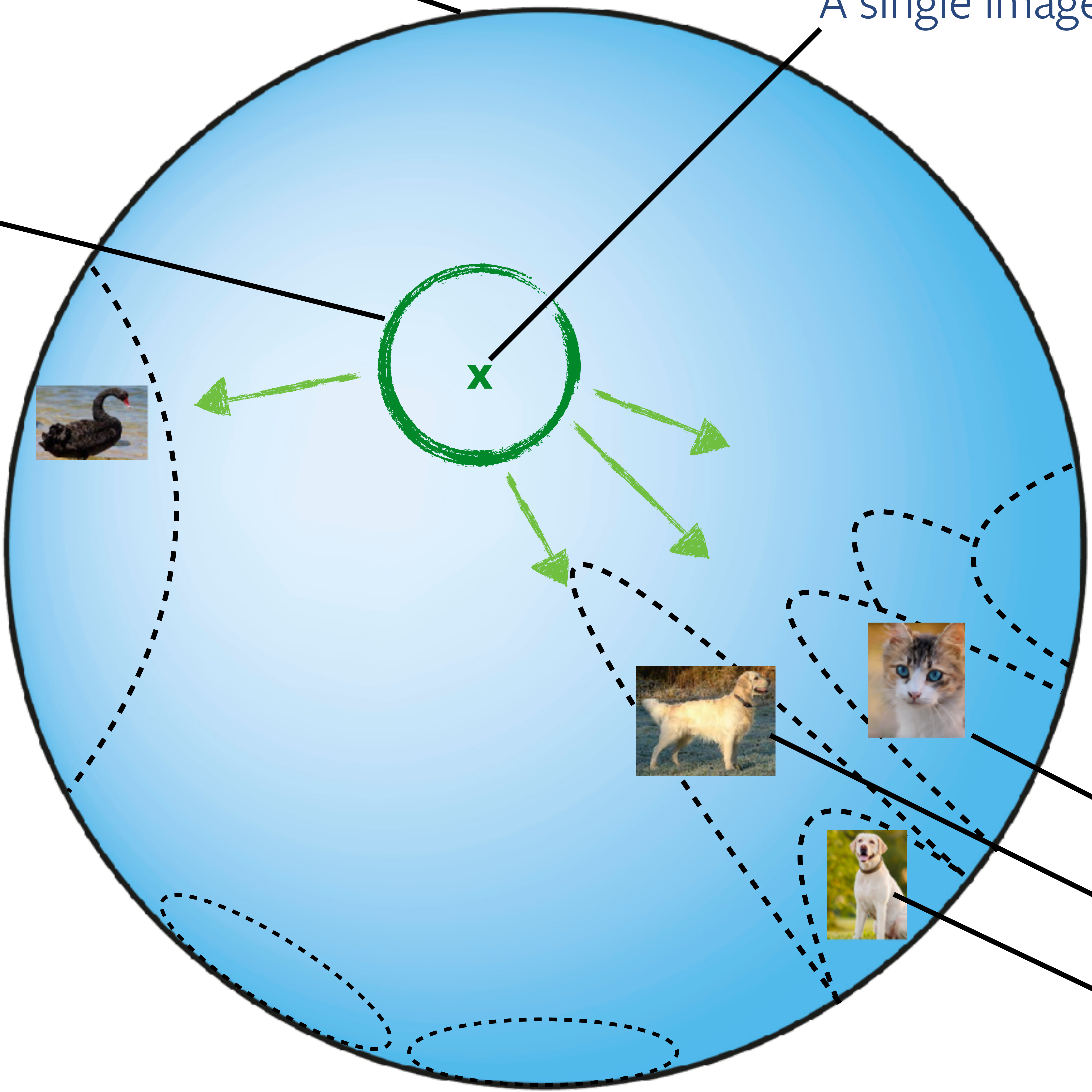
≈ *ImageNet,*
≈ *Kinetics*
≈ *UCF*
.≈ ..

The *real* motivation

Can we extrapolate from a single image to semantic categories?

Augmentations $\mathscr{A}(I)$





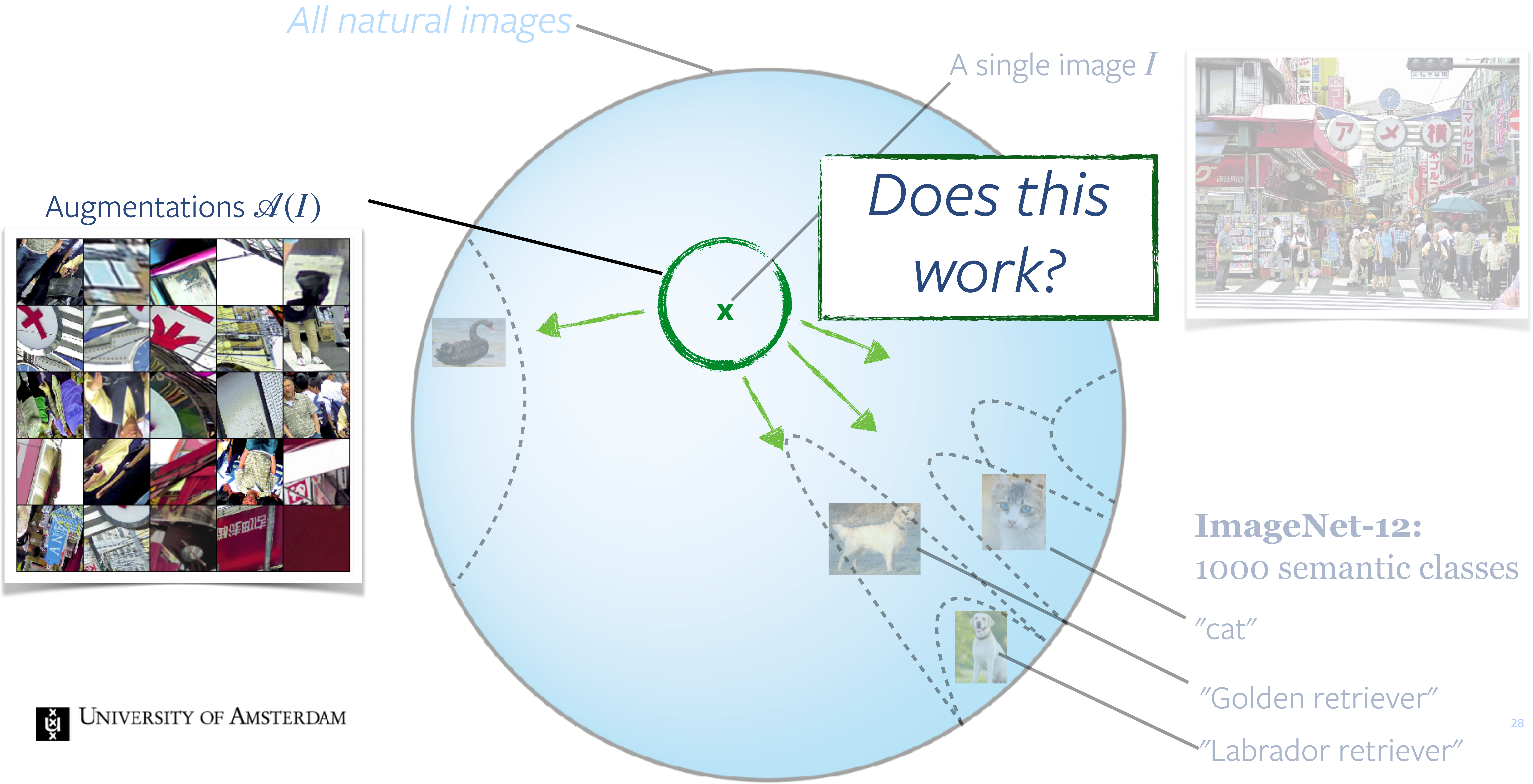
ImageNet-12:
1000 semantic classes

"cat"

"Golden retriever"

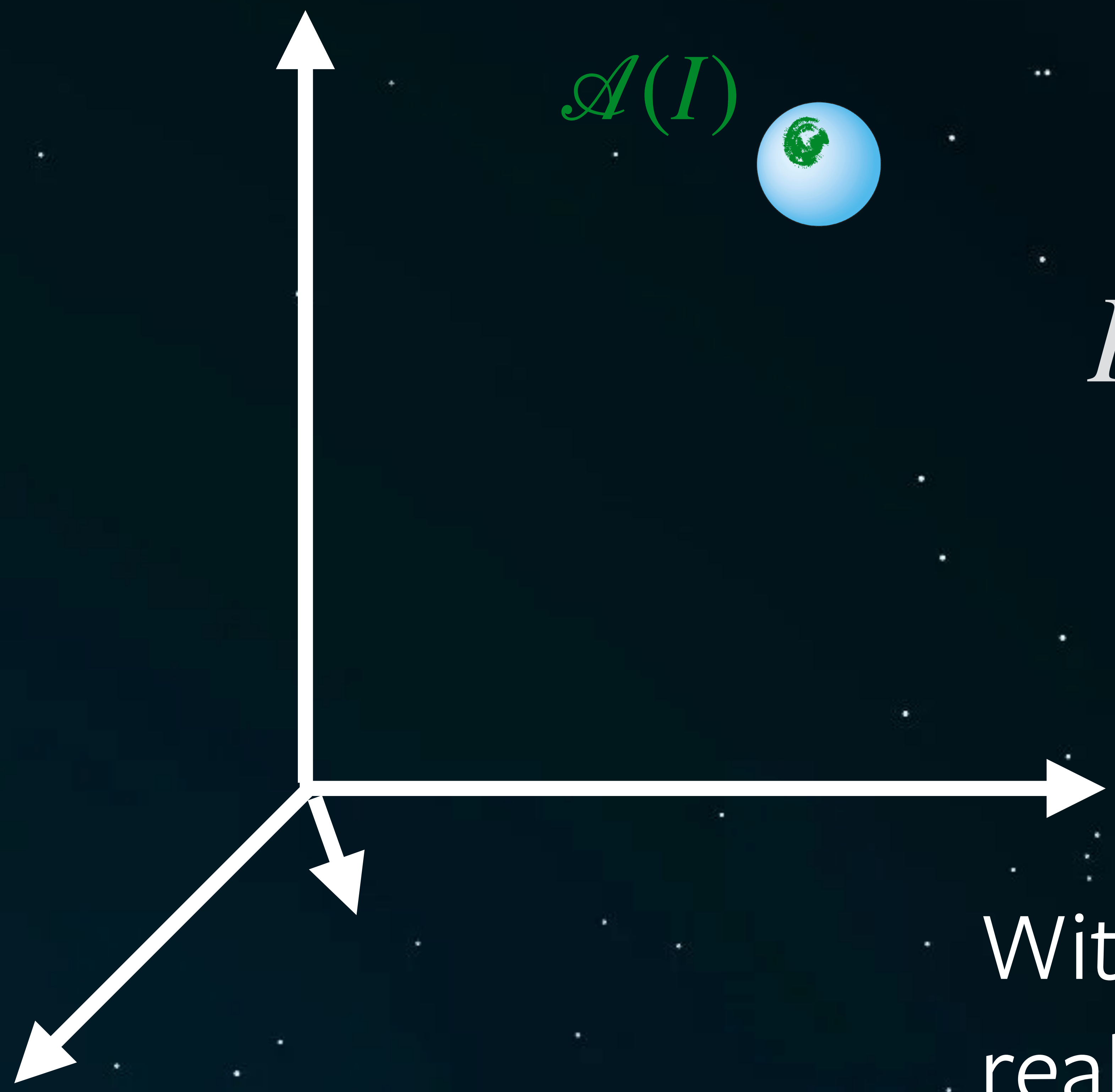
"Labrador retriever"

Can we extrapolate from a single image to semantic categories?



UNIVERSITY OF AMSTERDAM

Why it might work: *The augmented image prior*



$$I \in \mathscr{I}, \quad \mathscr{I} = \{0,..,255\}^{3\times224\times224}$$

Within the space of all possible images $\mathscr{I}$, a single real image I and its augmentations $\mathscr{A}(I)$ provide a very informative prior about all real images for extrapolation

Method

Our pipeline is kept as simple as possible.

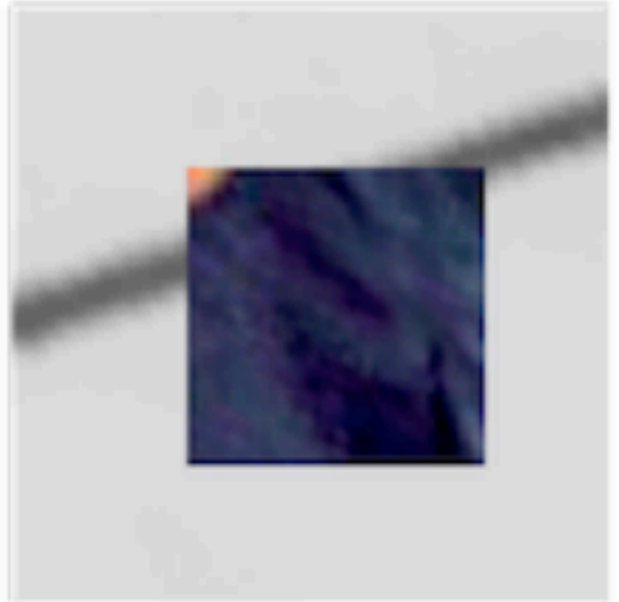
Training data: very varied, as are the teacher's predictions.

0.2803%: packet (692)
0.1994%: rubber eraser (767)
0.1947%: envelope (549)
0.1910%: Band Aid (419)
0.1865%: lighter (626)

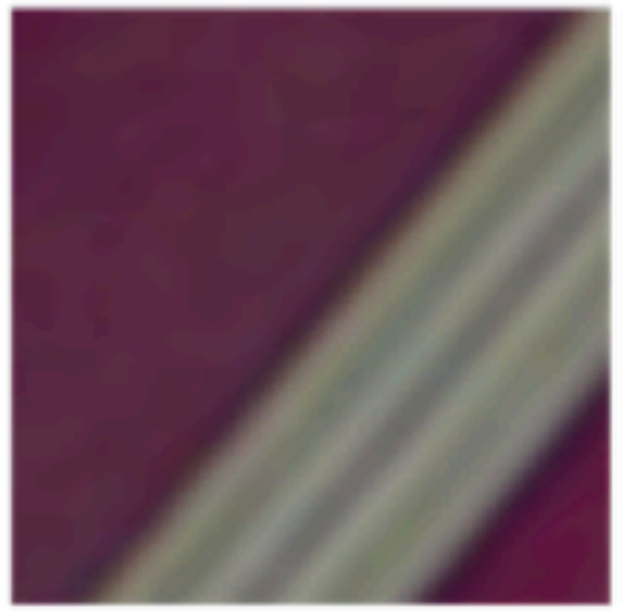
0.1563%: cleaver (499)
0.1549%: can opener (473)
0.1521%: whistle (902)
0.1511%: screw (783)
0.1507%: spatula (813)

0.2653%: comic book (917)
0.2453%: sarong (775)
0.2436%: sombrero (808)
0.2344%: shopping basket (790)
0.2336%: toyshop (865)

0.3454%: grocery store (582)
0.3175%: jinrikisha (612)
0.2830%: restaurant (762)
0.2367%: toyshop (865)
0.2322%: barbershop (424)

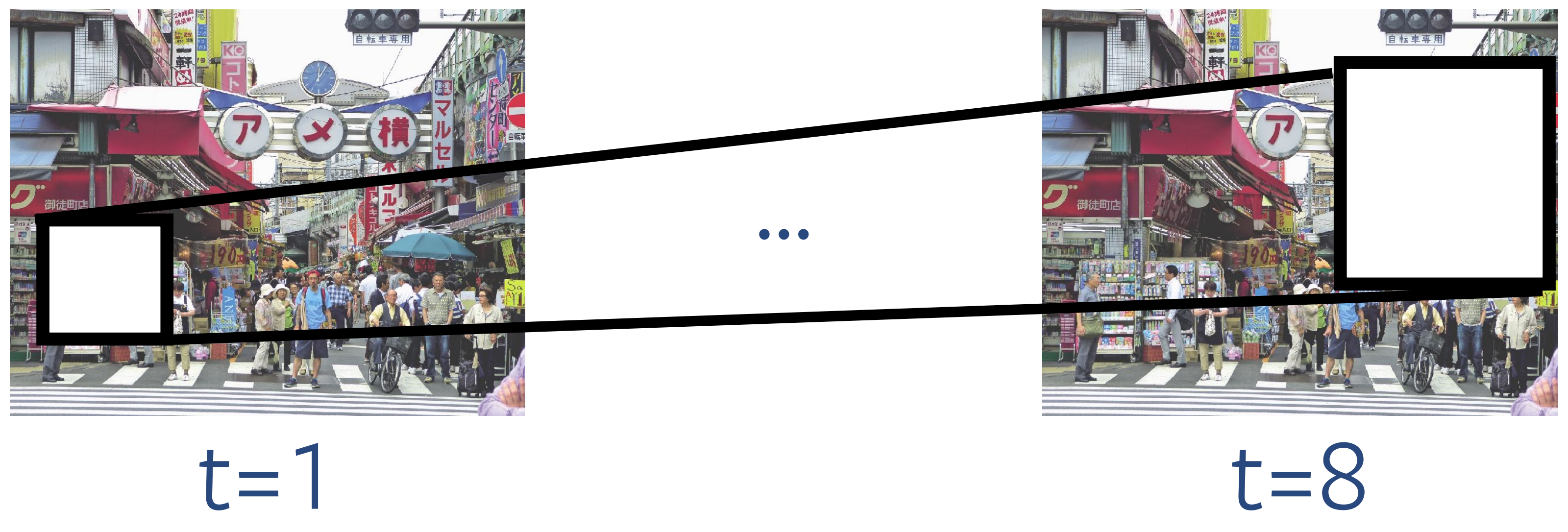


0.2736%: nail (677)
0.2657%: screw (783)
0.2121%: hook (600)
0.2035%: knot (616)
0.1901%: letter opener (623)



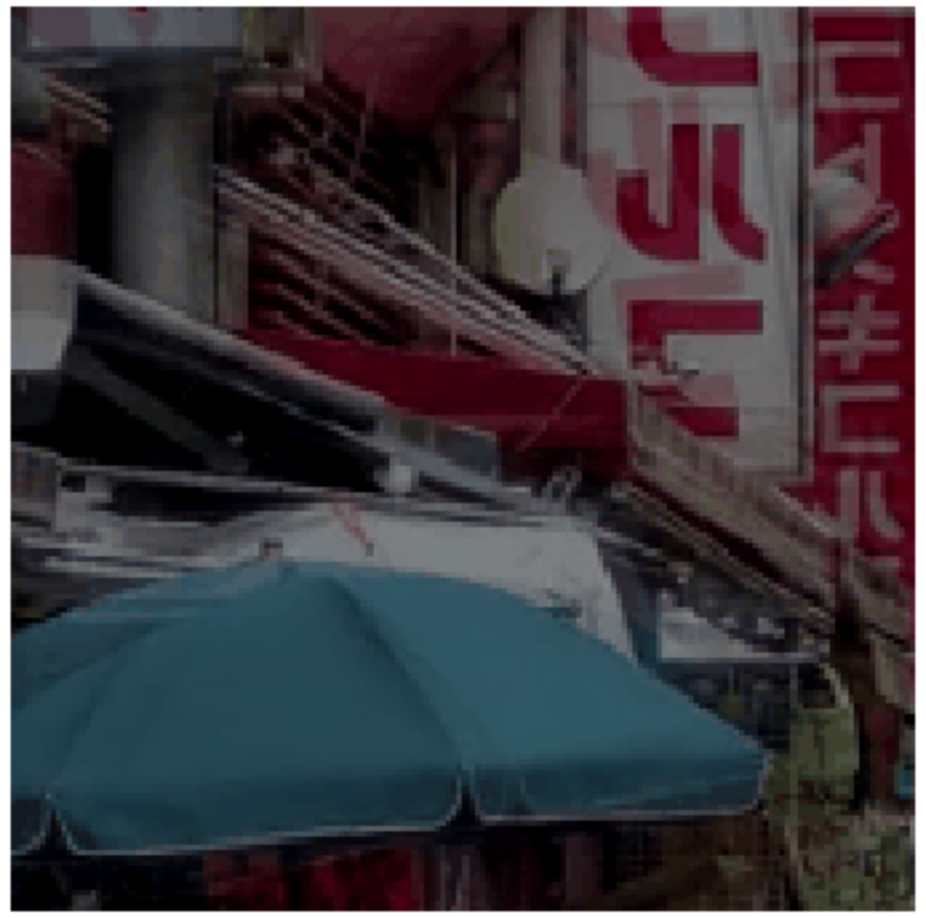
0.2557%: slide rule (798)
0.2362%: rule (769)
0.2135%: letter opener (623)
0.2116%: Windsor tie (906)
0.2061%: matchstick (644)

We can also make fake-videos out of images.



pick two crops, smoothly transition between them.

Guessing game! [the actual training data]



(In training we use the temperature-scaled logits)

Results

Comparison of datasets (*number of pixels*)

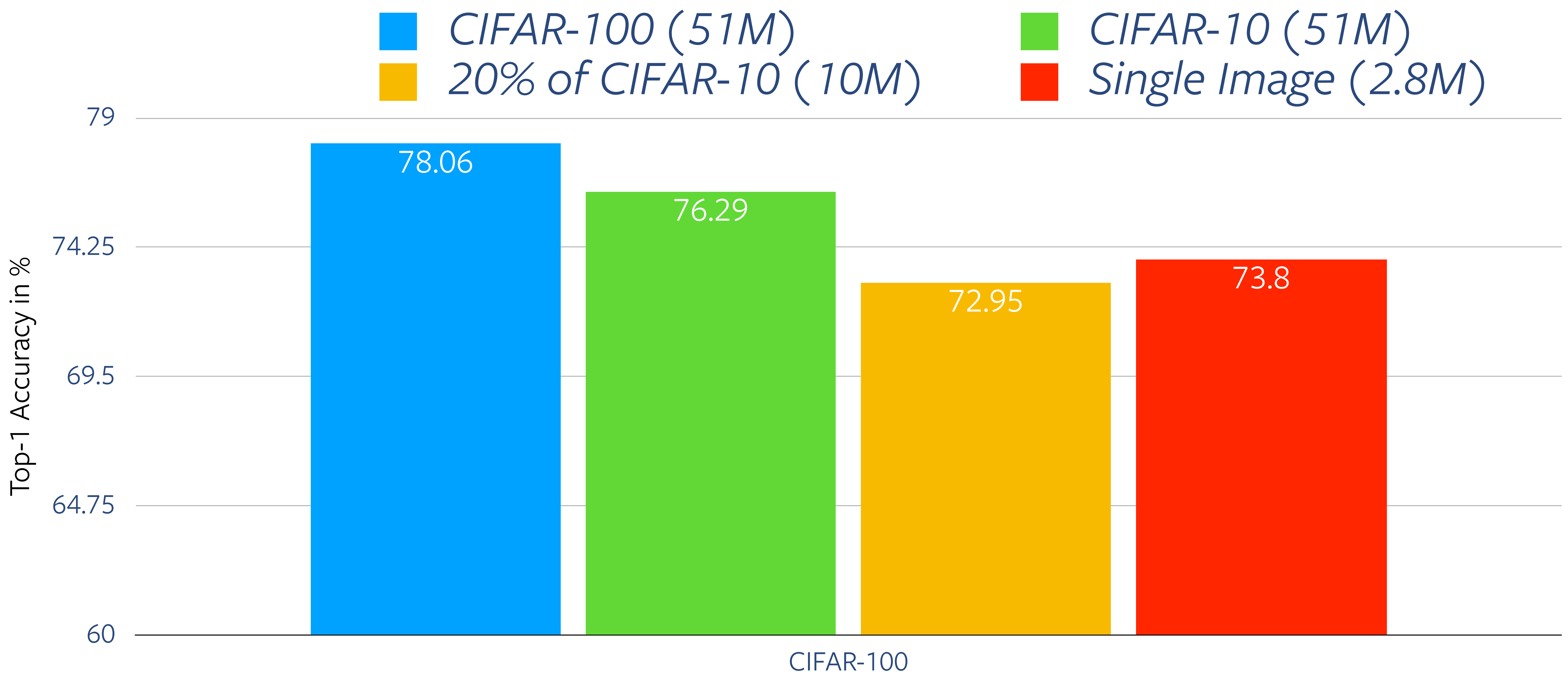


Image complexity matters

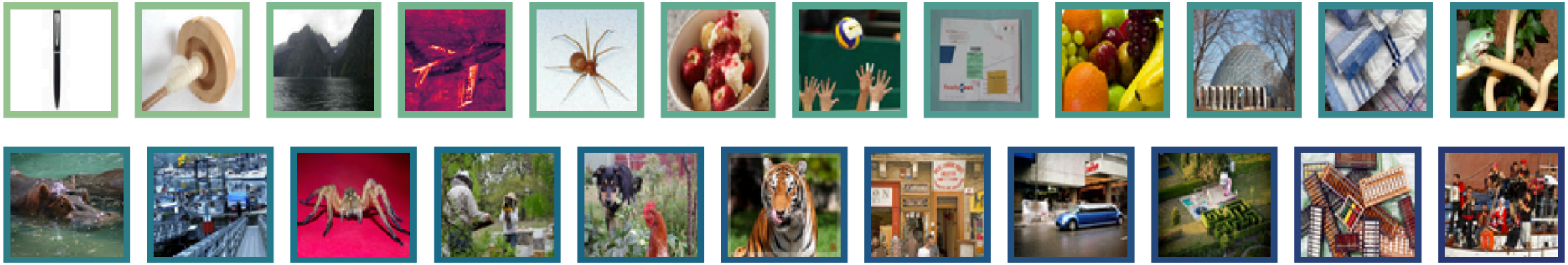
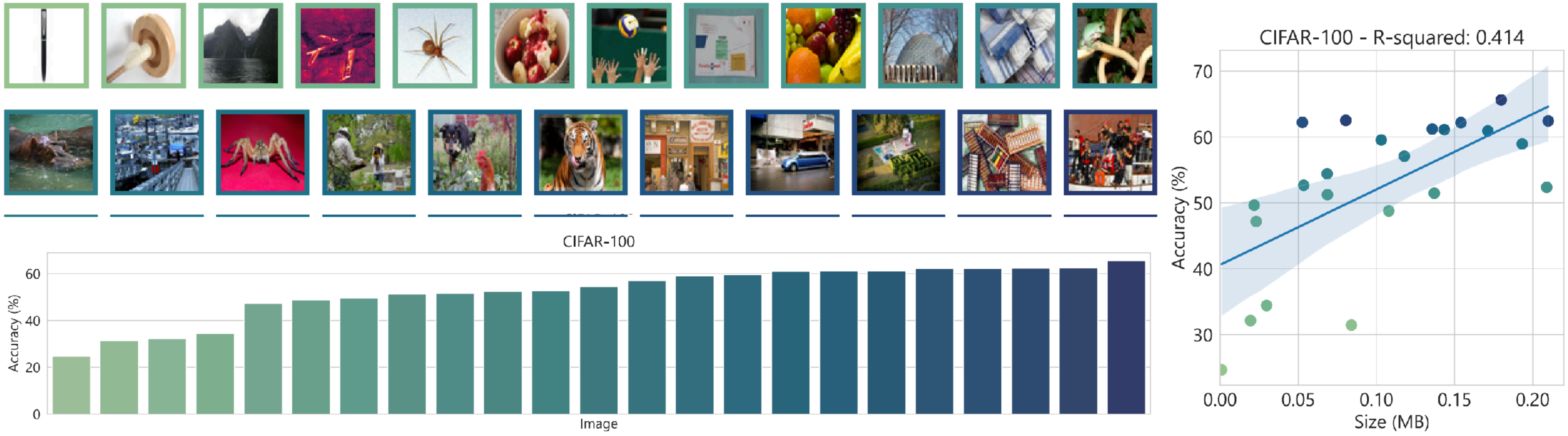
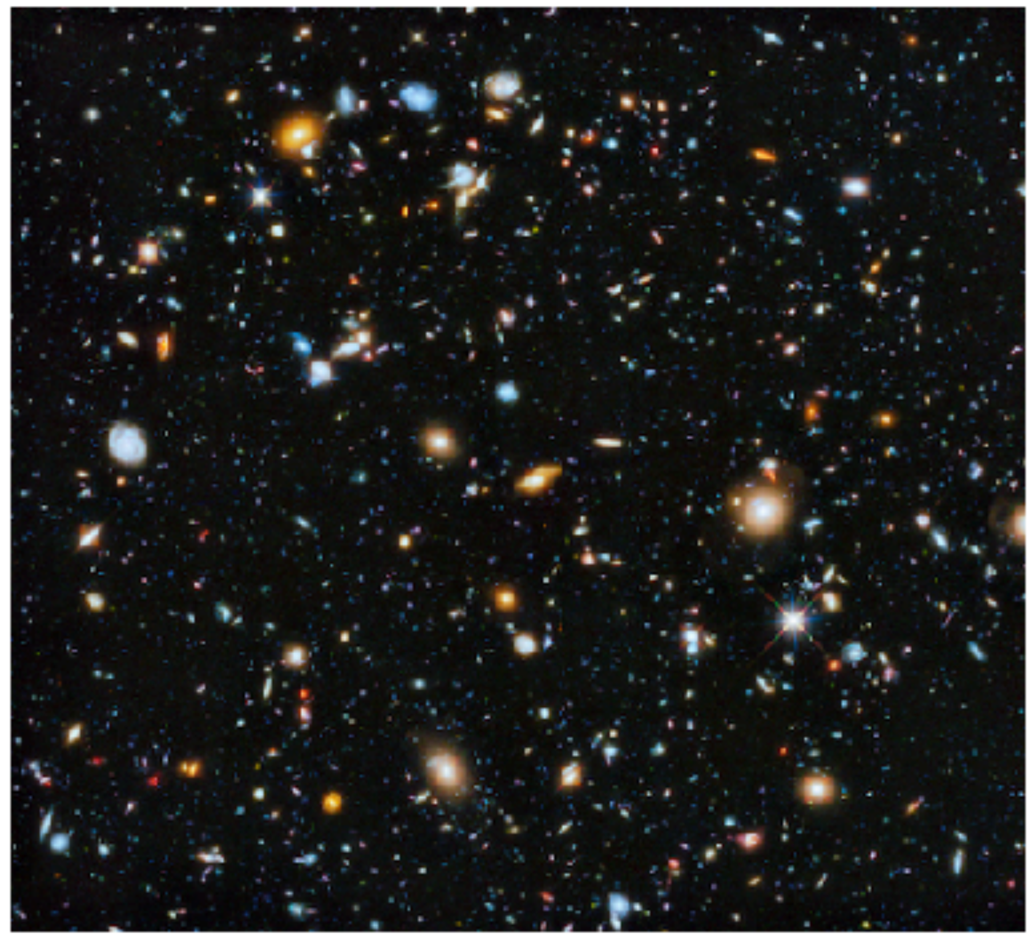


Image complexity matters



Other images

(a) The "Noise" Image. From uniform noise [0,255]. Size: 2,560x1,920, PNG: 16.3MB.



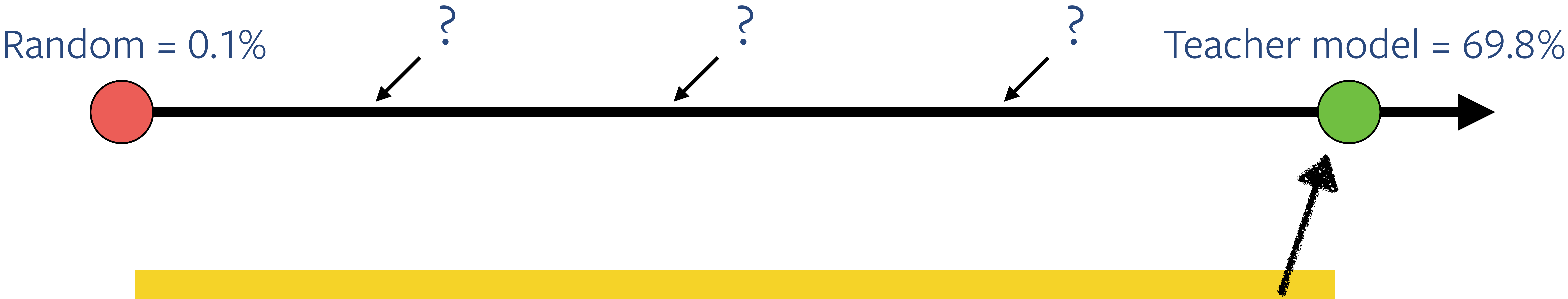
(b) The "Universe" Image. Size: 2,300x2,100, JPEG: 7.2MB.

(c) The "Bridge" Image. Size: 1,280x853, JPEG: 288KB.

(d) The "City" Image. Size: 2,560x1,920, JPEG: 1.9MB.

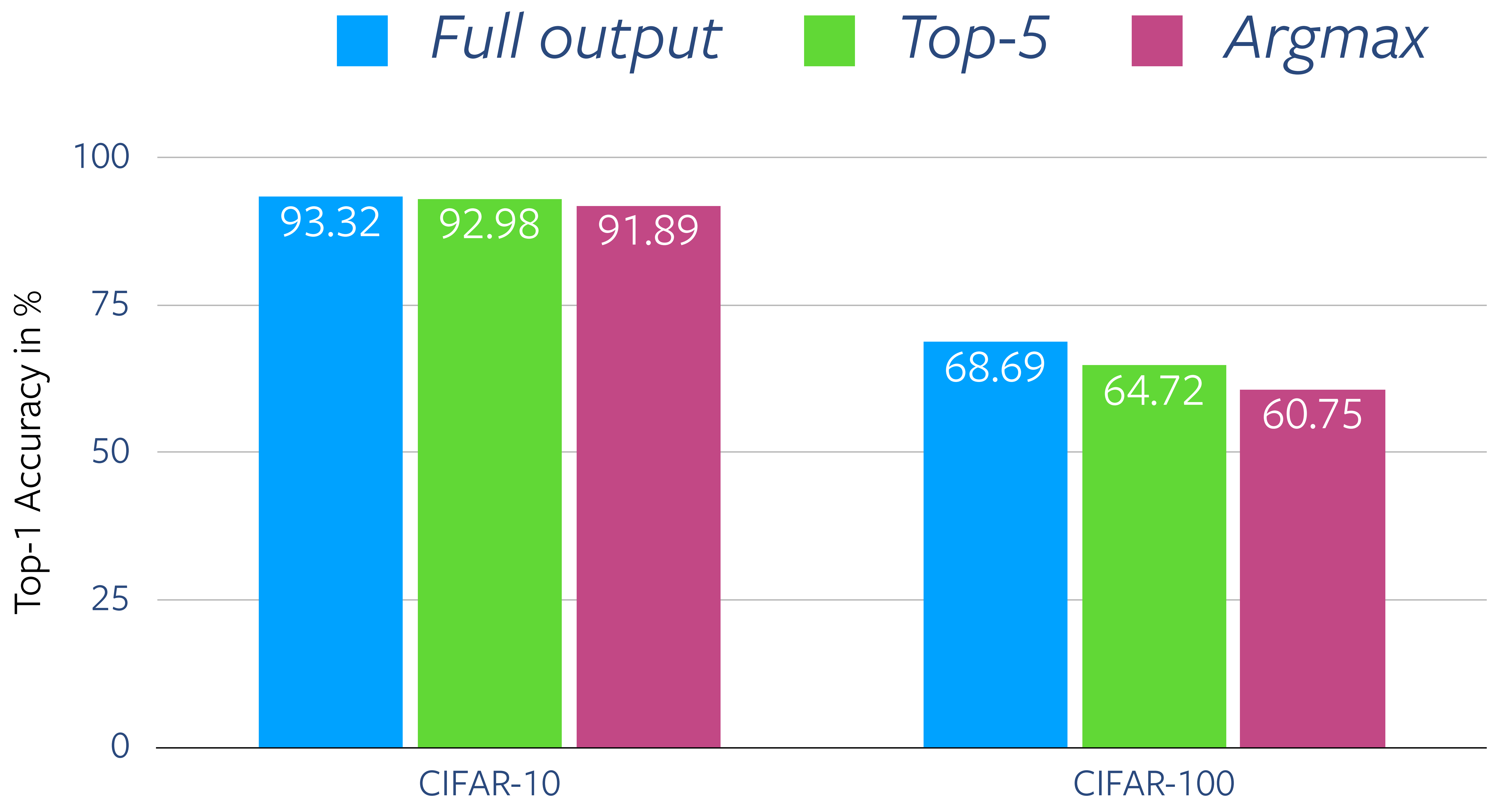
(e) The "Animals" Image. Size: 1,300x600, JPEG: 267KB.

So what accuracy do we get on ImageNet?



Our model achieves 69.0% top-1 accuracy on ImageNet-1k val.
Without ever having seen more than that one image.

Learning signal: even top-5 or argmax works well.



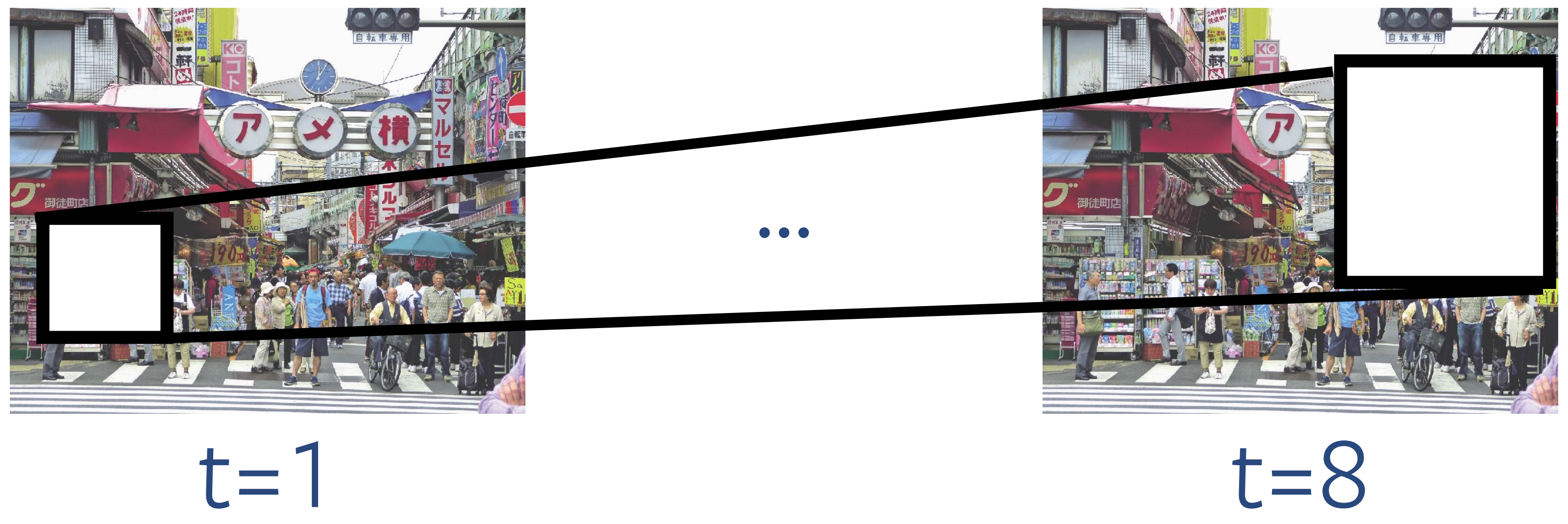
On ImageNet-1k: using argmax:
still 44% top-1 Acc. (compared to 69.%)

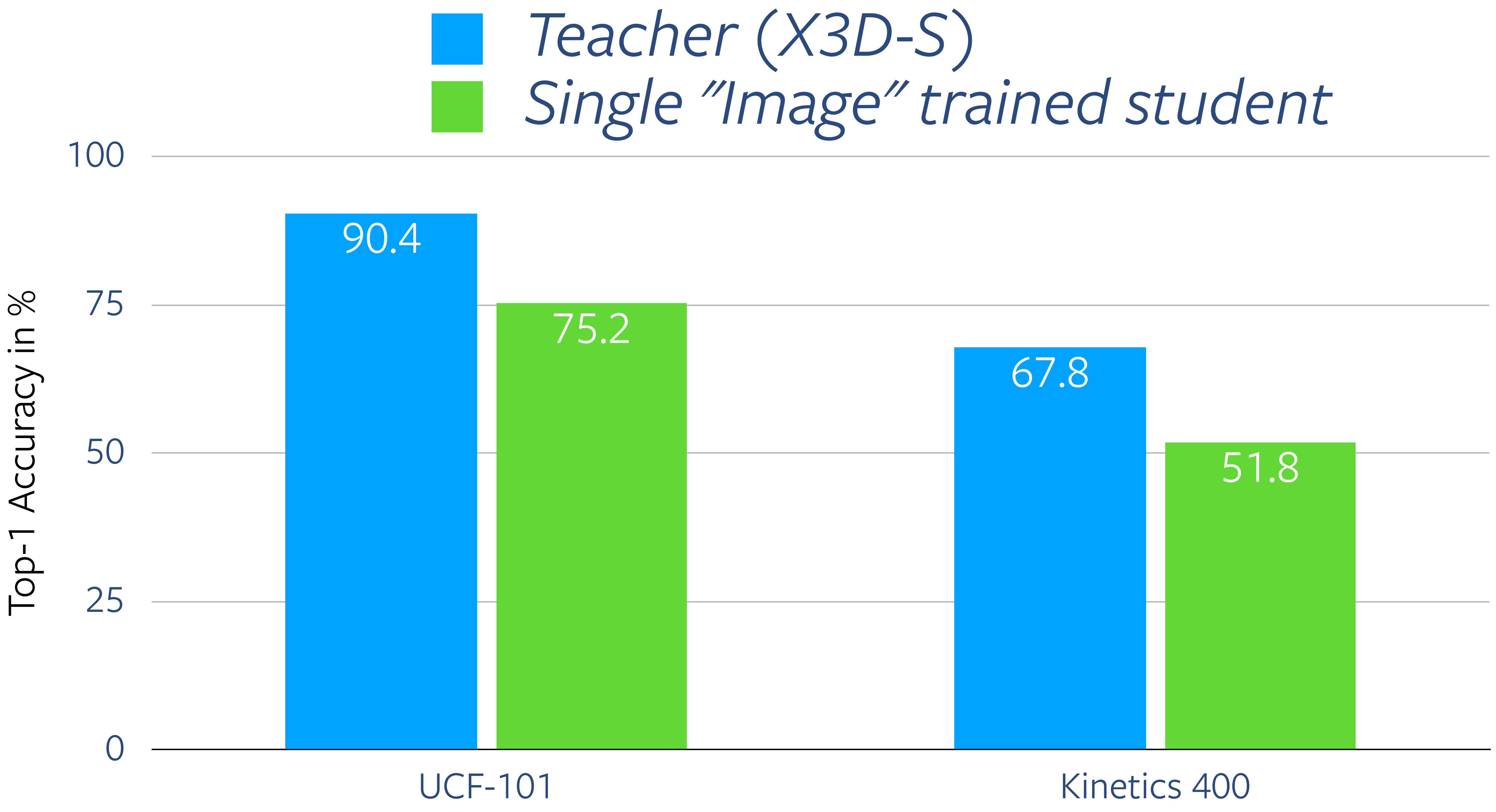
Even with only top-5 predictions (and confidence) or hard distillation, performance only slightly degrades.

⚠️ API providers! (c.f. Orekondy et al.)

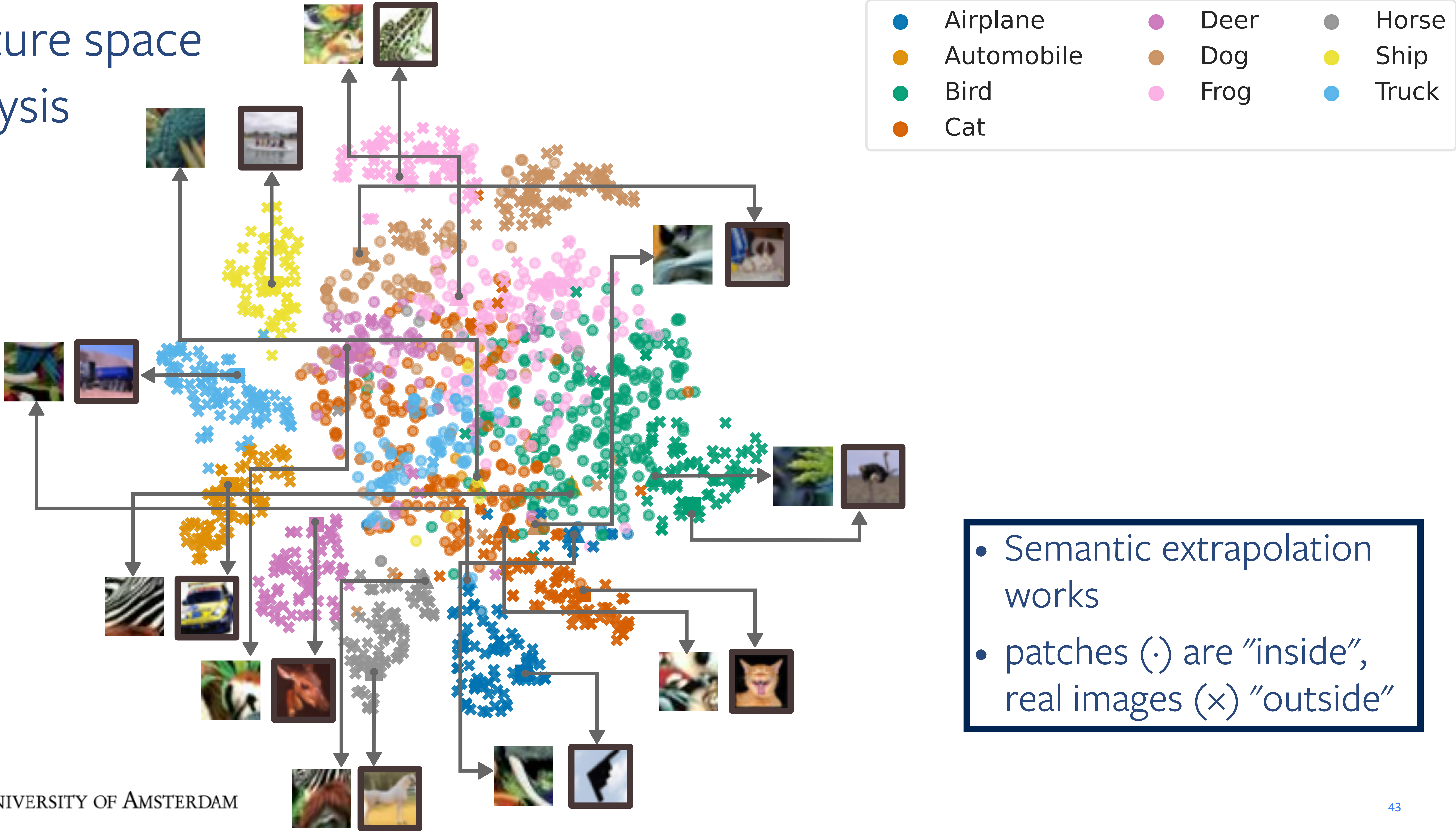
Orekondy et al. Knockoff nets: Stealing functionality of black-box models. CVPR 2019

Performance on video action classification benchmarks.



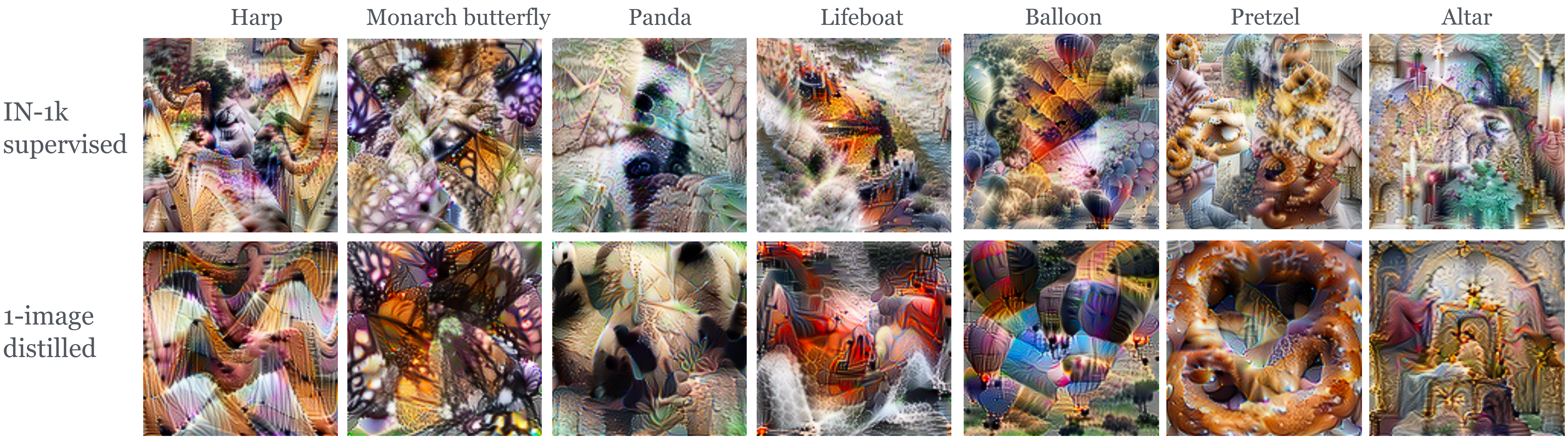


Feature space analysis



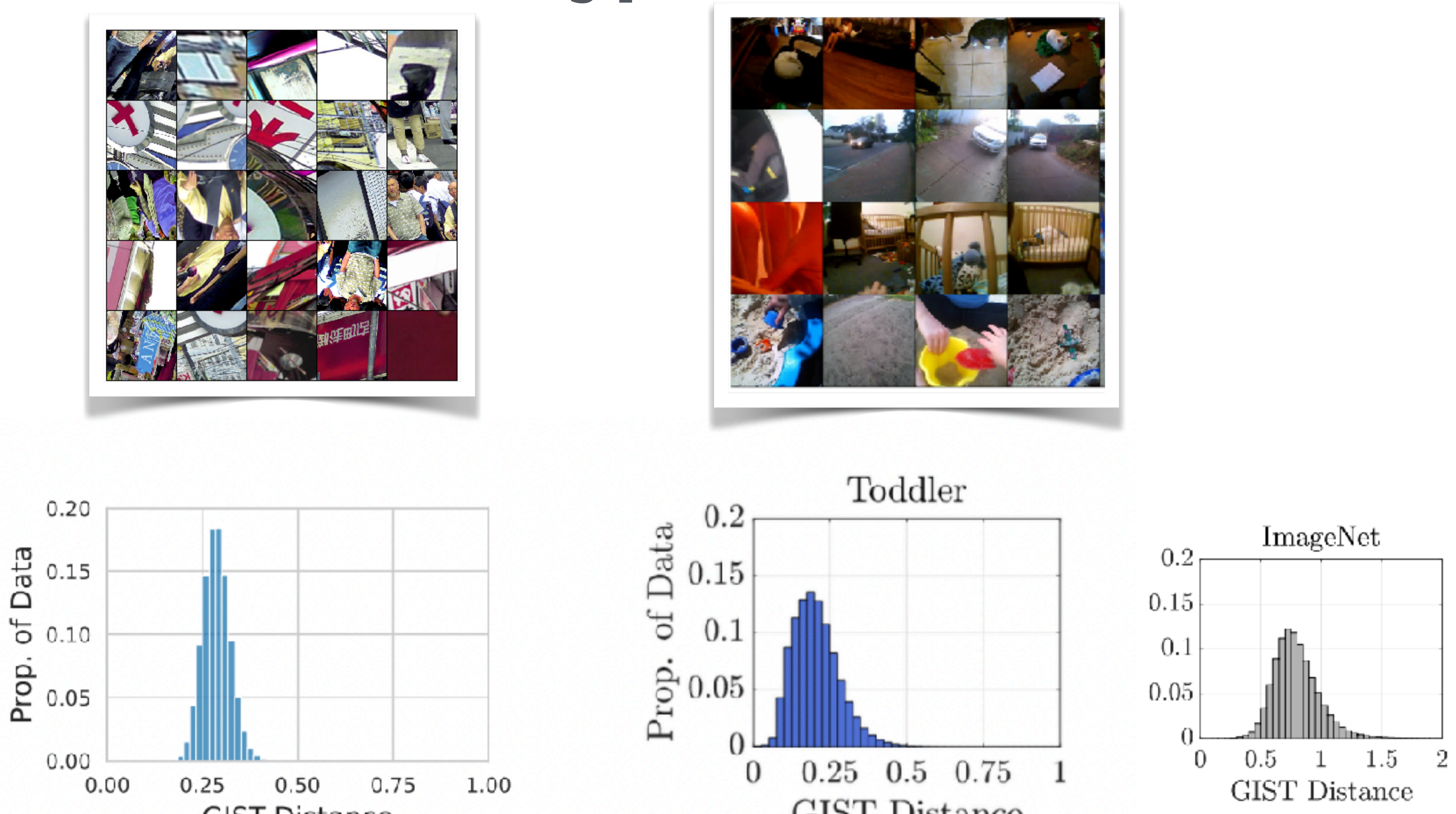
- Semantic extrapolation works
- patches (·) are "inside", real images (×) "outside"

Neuron activation maximisation: learning pandas from the city-jungle.



For intermediate layers, see paper.

One hypothesis



(a) **Our 1-image dataset.** (b) **Toddler data.** Figure from [3].

Toddlers data looks much more like 1-augmented image, compared to ImageNet. Maybe this high correlation helps with learning?

Bambach et al. *Toddler-Inspired Visual Object Learning.* NeurIPS 2018

Team for the works presented

TimeTuning

Mohammadreza Salehi

Efstratios Gavves

Cees G. M. Snoek

Yuki M. Asano

WTour Dora

Shashanka Venkataramanan

Mamshad N Rizve

Joao Carreira

Yuki M. Asano*

Yannis Avrithis*

1-image distill

Yuki M. Asano[1]

Aaqib Saeed[1]

Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations*. ICCV 2023
Venkataramanan, Rizve, Carreira, Avrithis*, Asano*. Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video. ICLR 2024 [oral]
Asano & Saeed. *The Augmented Image Prior: Distilling 1000 Classes by Extrapolating from a Single Image*. ICLR 2023.
[1]: co-first authors; *: co-last authors