University of Amsterdam

Vision Foundation Models (with academic compute)

YUKI M ASANO

ECCV 2024

SSLWIN

UNIVERSITY OF AMSTERDAM

Vision Foundation Models (with academic compute)

YUKI M ASANO

ECCV 2024

SSLWIN

Instruction tuning: the difference between GPT-3 and

giant training corpora

small IT datasets

(raw) LLMs
(e.g. GPT-3)

Instruction tuning →

Chat LLMs
(e.g. ChatGPT)

❌ requires industrial compute

❌ model is relatively useless

✅ requires fraction of compute

✅ makes LLM useful, customisable, bette

Computer vision needs more *post-pretraining*

(raw) LLMs

❌ requires industrial compute

❌ model is relatively

Instruction tuning →

Chat LLMs
(e.g. ChatGPT)

✅ requires fraction of compute

✅ makes LLM useful, customisable, better

1st-gen Vision Foundation Models

❌ requires industrial compute

✅ model is already useful

Post-pretraining →

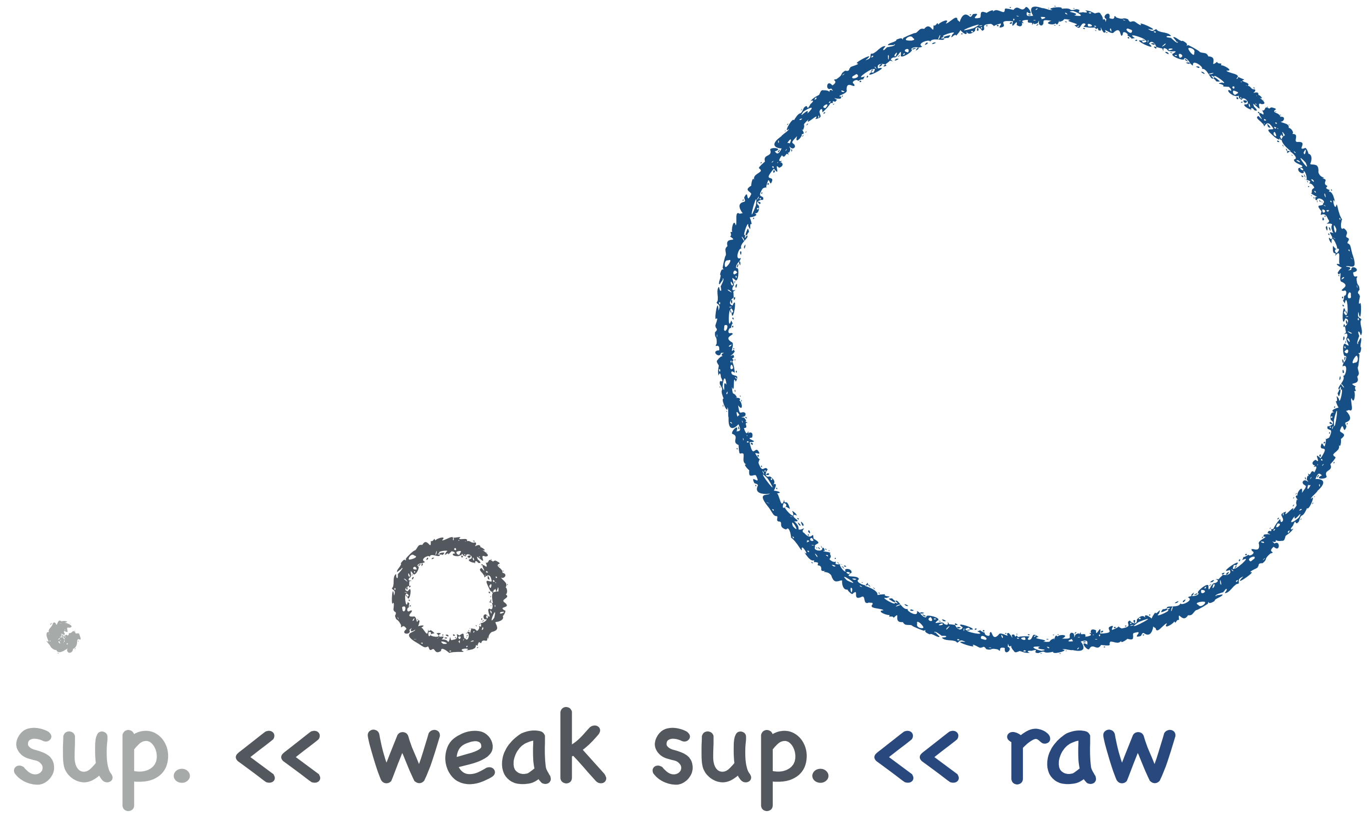
2nd-gen Vision Foundation Models
(e.g. Neco, TimeTuning)

✅ requires **fraction of compute**

✅ makes vision model **even better, across various tasks**

Why Self-supervised Learning still matters, despite CLIP and

Massive scale



Cost of (re)labelling

Problems of labels

Fundamentals

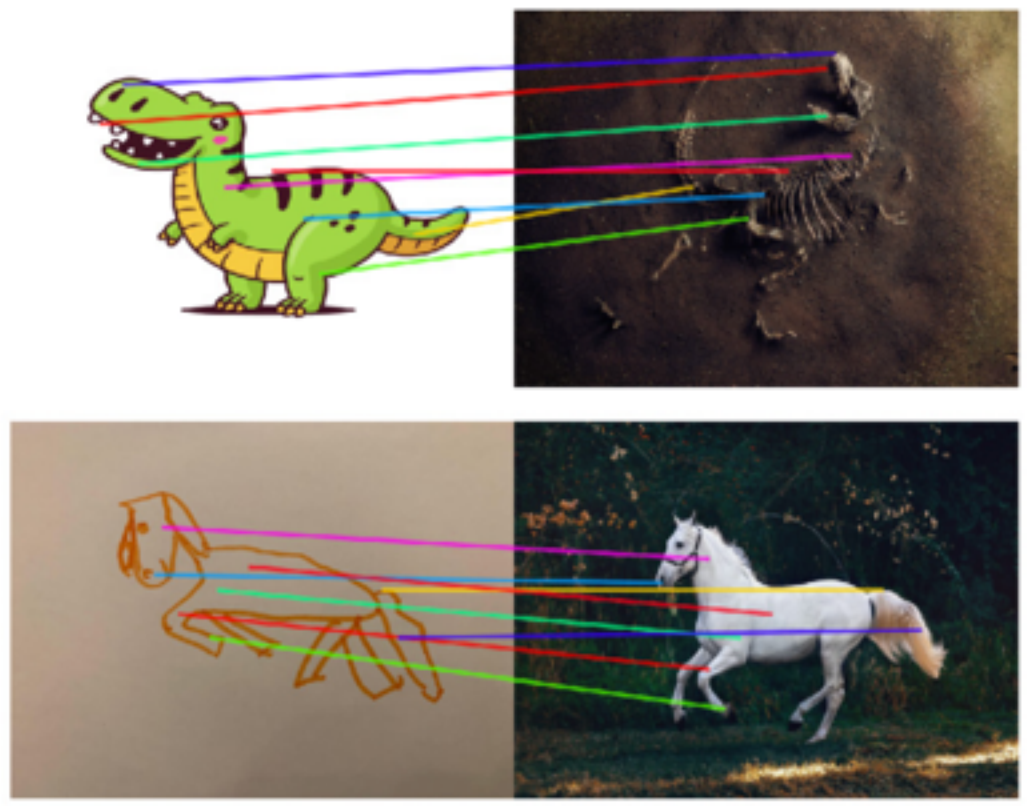
NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency.

Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano.

https://www.taste.com.au/recipes/creamy-bacon-carbonara/12c27b1e-5fb6-48c9-ac42-82a37cc8ab30?nk=3b0399578618abc93c6fc9abab4a14a7-1727548620

How semantic are patch representations?

Qualitative results in DINOv2



(Drawings / Animals)

But often...

How it should be

How it is

Which patch from the whole dataset is the closest?

with SoTA DINOv2-R model

Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. TMLR 2023
Darcet et al. Vision Transformers Need Registers. ICLR 2024

Idea of Patch Nearest Neighbor Consistency: intuitive

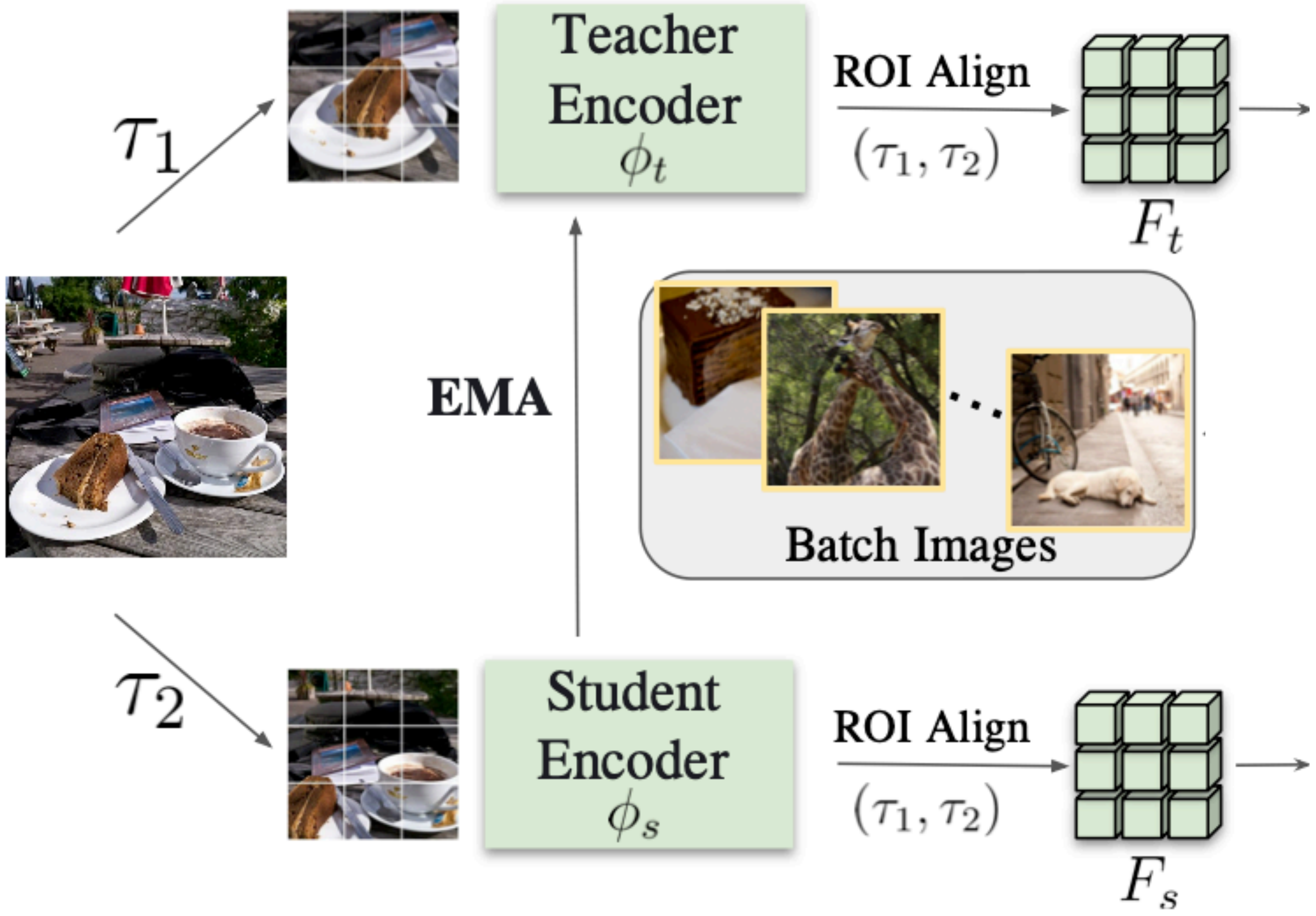
Given a **query patch of a right shoulder**, top neighbors should be in the following order:

(1) All Right Shoulder Patches, (2) All Left Shoulder Patches, (...) (3) Everything Else

Query Patch

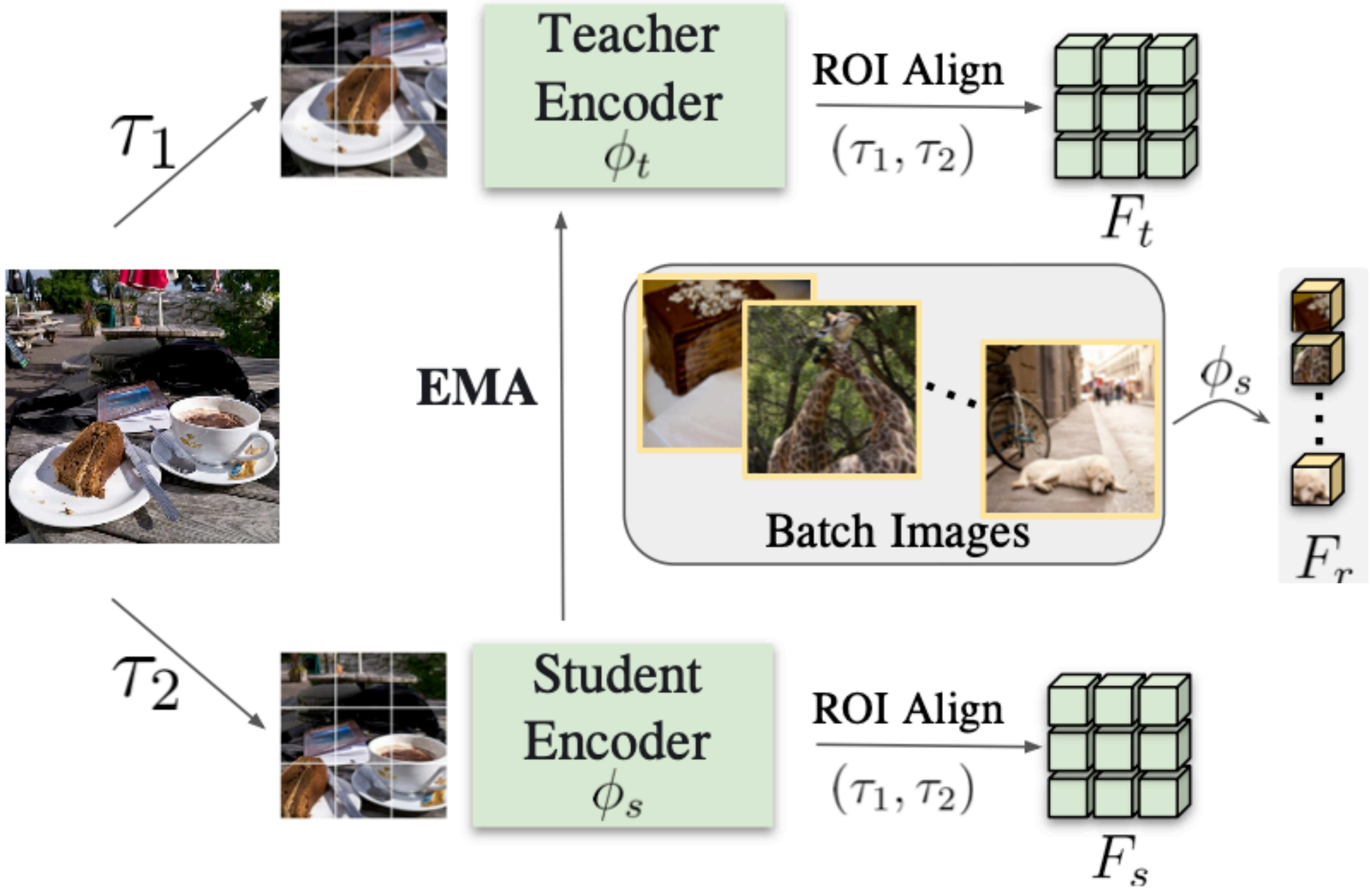
Example Patches

PaNeCo: Patch Nearest neighbor Consistency

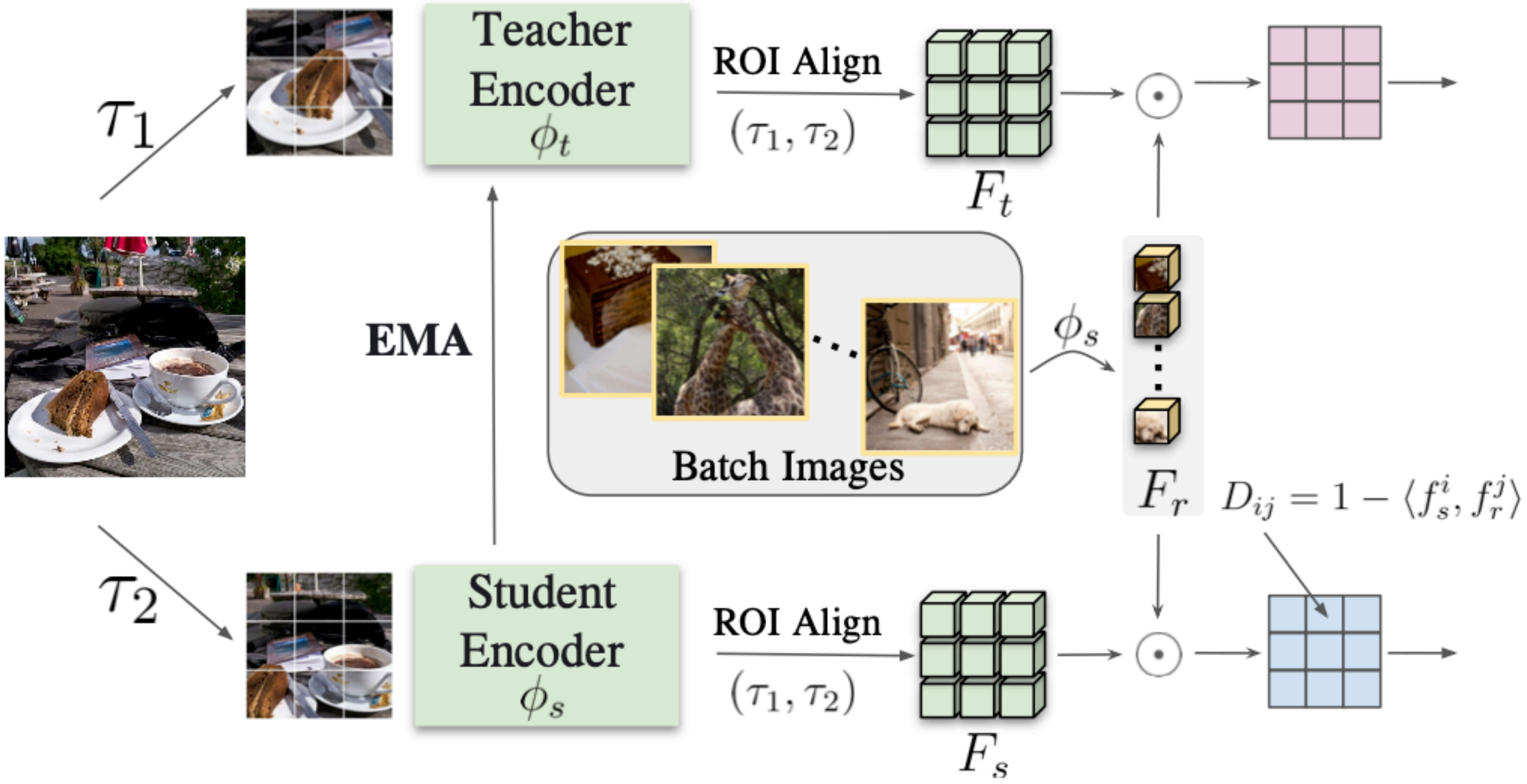


NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo: Patch Nearest neighbor Consistency

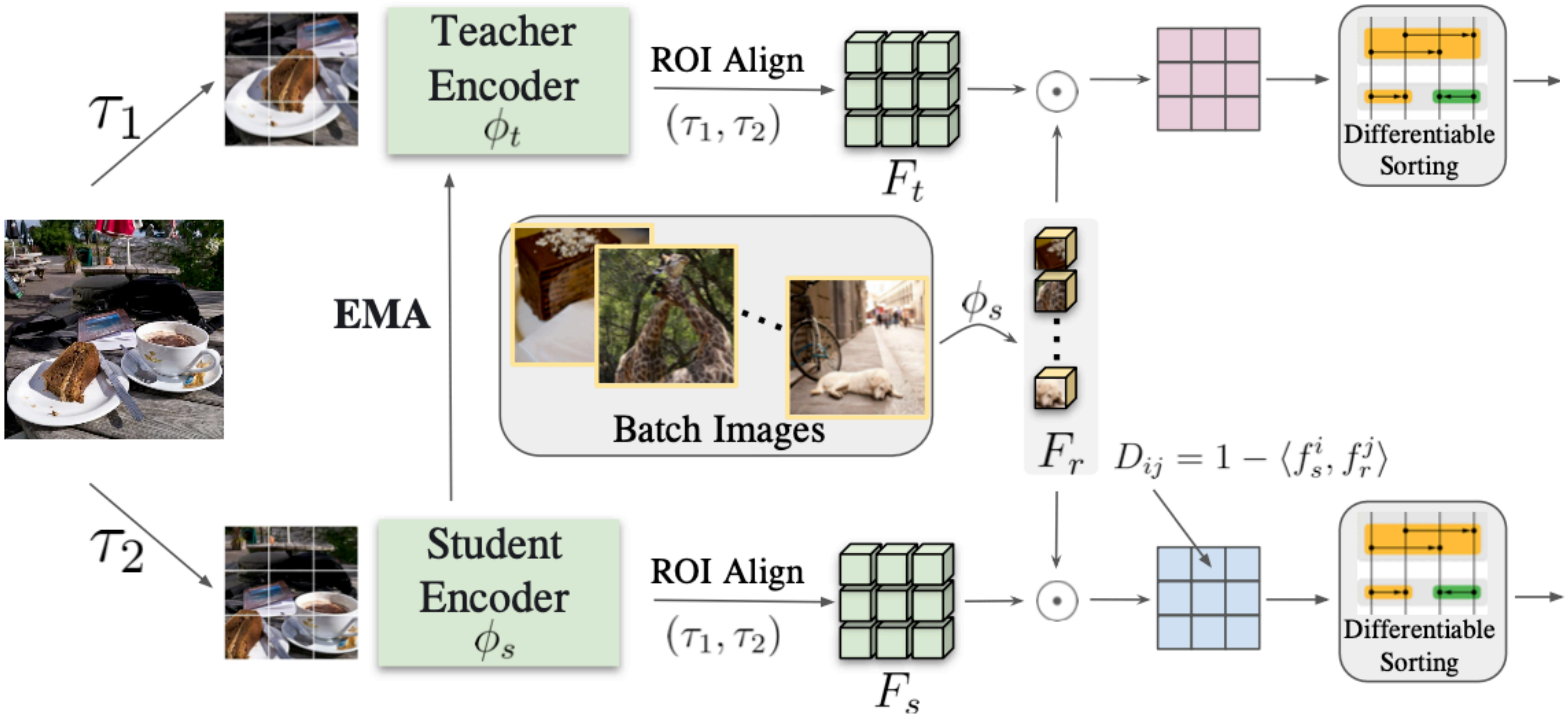


PaNeCo: Patch Nearest neighbor Consistency



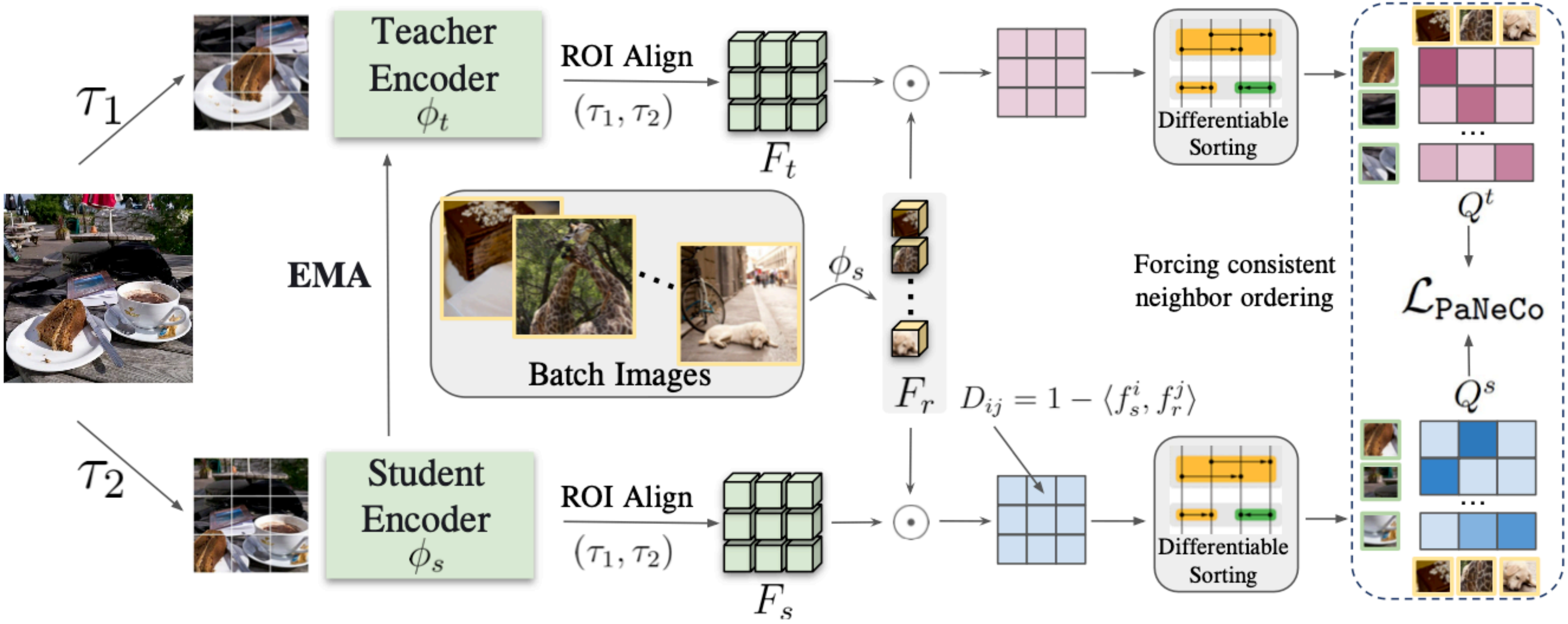
NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo: Patch Nearest neighbor Consistency



NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

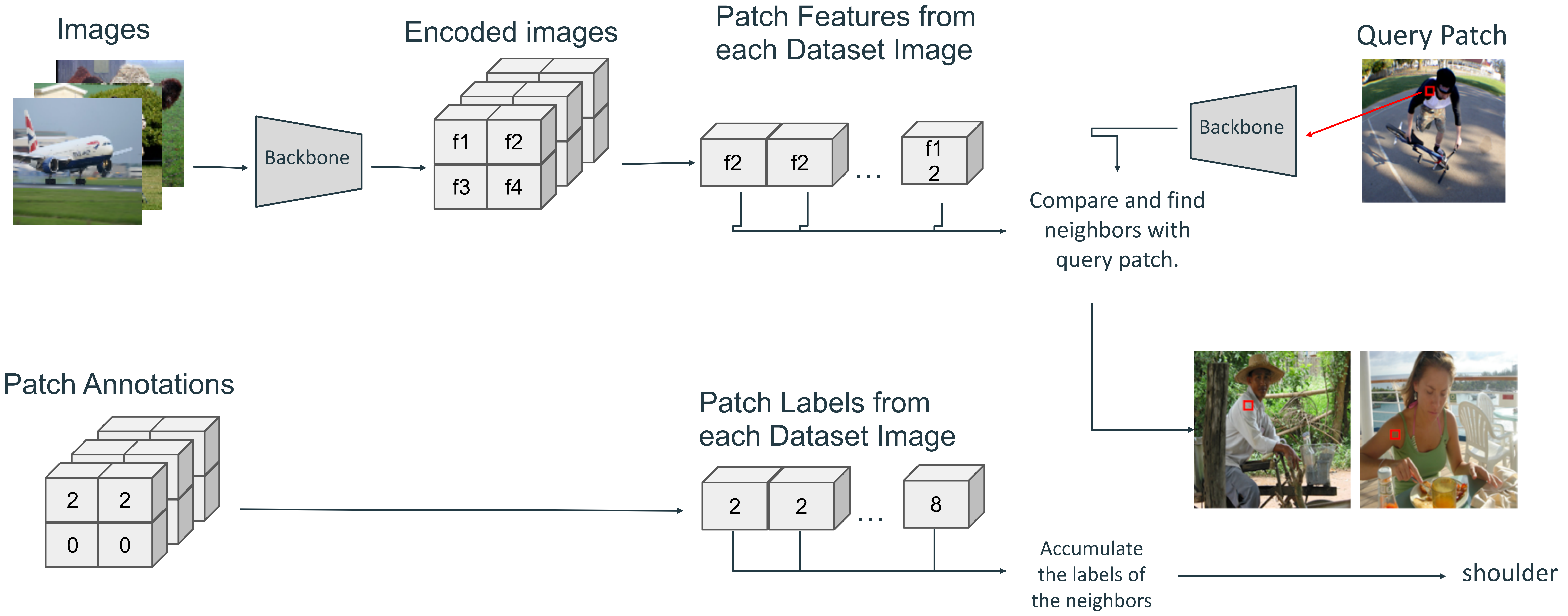
PaNeCo: Patch Nearest neighbor Consistency



NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

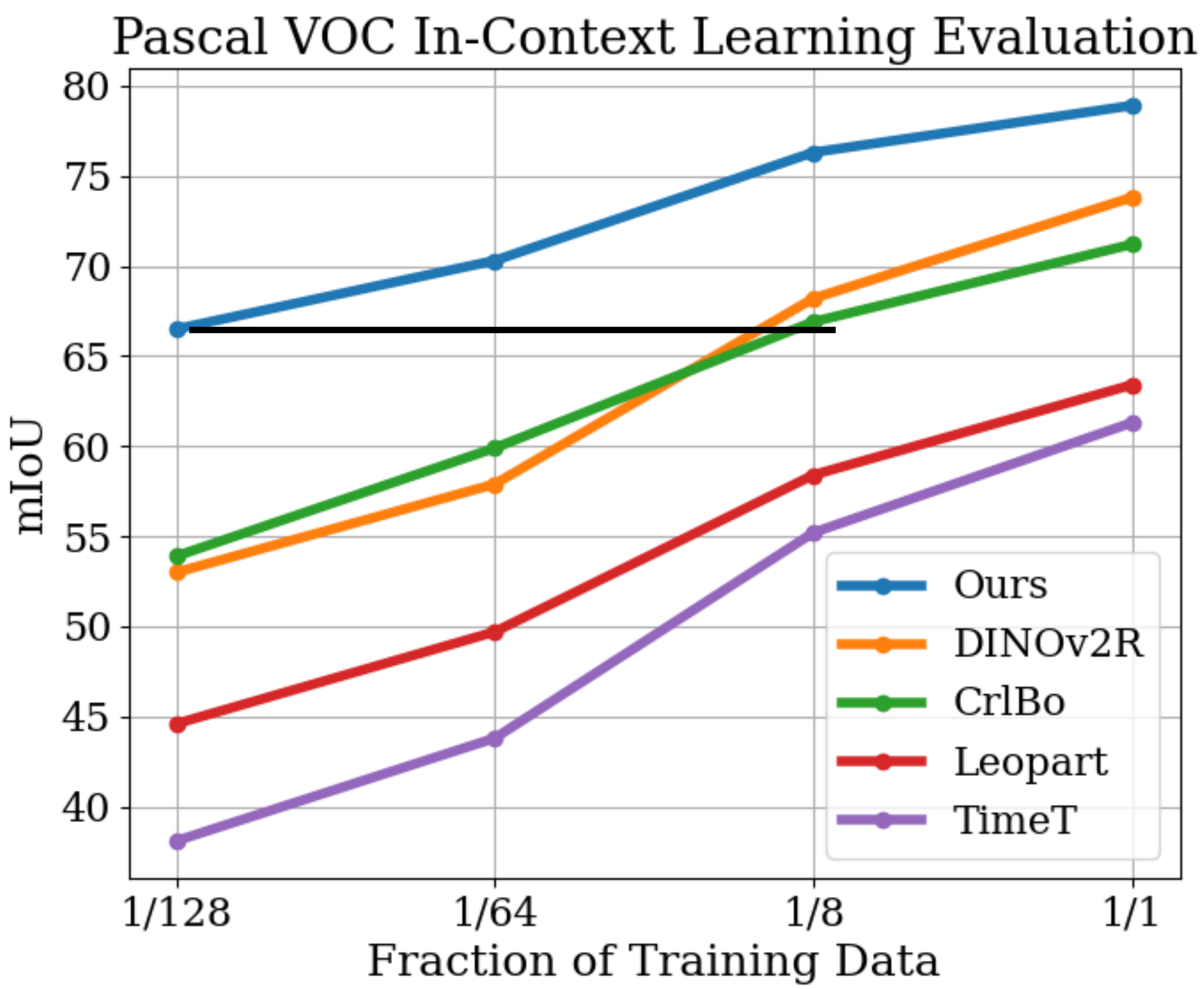
Results

Evaluation 1: Visual in-context segmentation via dense NN retrieval



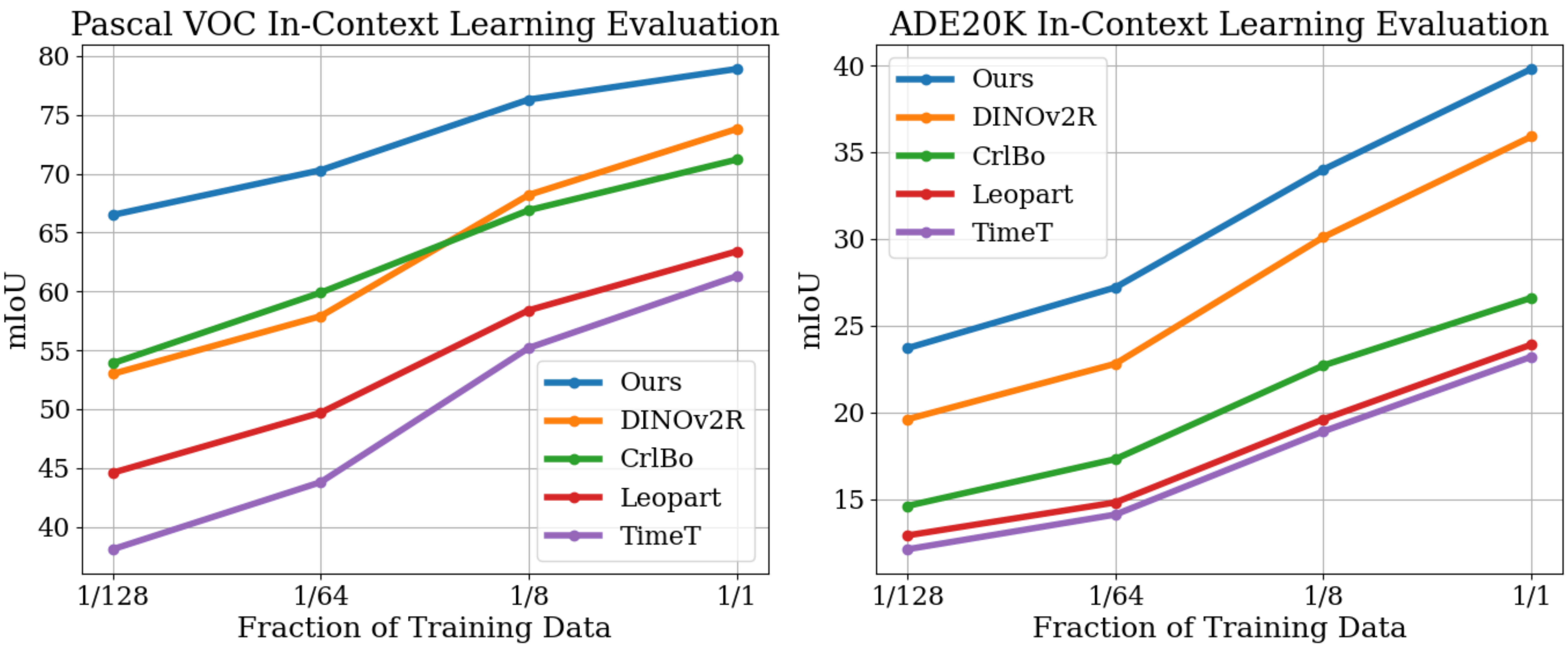
Towards In-context Scene Understanding. Ivana Balažević, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, Olivier J. Hénaff. NeurIPS 2023

In-context scene understanding benchmark



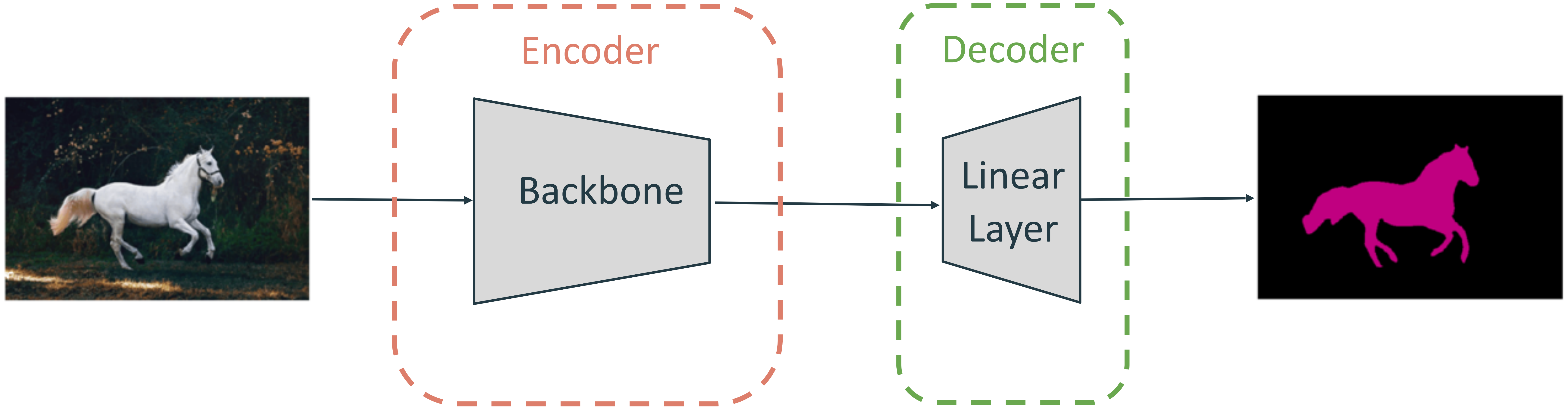
matches performances of DINOv2-R with ~15x less data

In-context scene understanding benchmark



NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

Linear Segmentation Evaluation



- Encode Image to patch-level features,
- Decode with a linear layer the per pixel semantic labels of the image,
- Supervised training of the linear layer of the decoder for this task.

Linear segmentation performance

Method	Backbone	Params	**COCO-Things**	**COCO-Stuff**	**Pascal VOC**	**ADE20K**
DINO	ViT-S/16	21M	43.9	45.9	50.2	17.5
TimeT	ViT-S/16	21M	58.2	48.7	66.3	20.7
iBOT	ViT-S/16	21M	58.9	51.5	66.1	21.8
CrOC	ViT-S/16	21M	64.3	51.2	67.4	23.1
CrlBo	ViT-S/16	21M	64.3	49.1	71.6	22.7
DINOv2R	ViT-S/14	21M	75.3	56.0	74.2	35.0
PaNeCo	ViT-S/14	21M	**82.3**	**62.0**	**81.3**	**40.1**
DINO	ViT-B/16	85M	55.8	51.2	62.7	23.6
MAE	ViT-B/16	85M	38.0	38.6	32.9	5.8
iBOT	ViT-B/16	85M	69.4	55.9	73.1	30.1
CrIBo	ViT-B/16	85M	69.6	53.0	73.9	25.7
DINOv2R	ViT-B/14	85M	78.3	57.6	79.8	40.3
PaNeCo	ViT-B/14	85M	**85.5**	**63.3**	**83.3**	**44.9**

A linear segmentation head is trained on top of the frozen spatial features obtained from different feature extractors. We report the mIoU scores achieved on the validation sets of 4 different datasets.

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo starting with different pretrained weights.

	Pascal VOC						COCO-Things					
	At Init			+PaNeCo			*At Init*			+PaNeCo		
Pretrain	K=GT	K=500	Lin.	K=GT	K=500	Lin.	K=21	K=500	Lin.	K=21	K=500	Lin.
iBOT [92]	4.4	31.1	66.1	$15.4^{\uparrow 11.0}$	$51.2^{\uparrow 20.1}$	$68.6^{\uparrow 2.5}$	7.6	28.0	58.9	$20.4^{\uparrow 12.8}$	$52.8^{\uparrow 24.8}$	$67.7^{\uparrow 8.8}$

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo starting with different pretrained weights.

	Pascal VOC						COCO-Things					
	At Init			+PaNeCo			*At Init*			+PaNeCo		
Pretrain	K=GT	K=500	Lin.	K=GT	K=500	Lin.	K=21	K=500	Lin.	K=21	K=500	Lin.
iBOT [92]	4.4	31.1	66.1	15.4$^{\uparrow 11.0}$	51.2$^{\uparrow 20.1}$	68.6$^{\uparrow 2.5}$	7.6	28.0	58.9	20.4$^{\uparrow 12.8}$	52.8$^{\uparrow 24.8}$	67.7$^{\uparrow 8.8}$
DINO [15]	4.3	17.3	50.2	14.5$^{\uparrow 10.2}$	47.9$^{\uparrow 30.6}$	61.3$^{\uparrow 11.1}$	5.4	19.2	43.9	16.9$^{\uparrow 11.5}$	50.0$^{\uparrow 30.8}$	62.4$^{\uparrow 18.5}$

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo starting with different pretrained weights.

	Pascal VOC						COCO-Things					
	At Init			+PaNeCo			*At Init*			+PaNeCo		
Pretrain	K=GT	K=500	Lin.	K=GT	K=500	Lin.	K=21	K=500	Lin.	K=21	K=500	Lin.
iBOT [92]	4.4	31.1	66.1	15.4 $^{\uparrow 11.0}$	51.2 $^{\uparrow 20.1}$	68.6 $^{\uparrow 2.5}$	7.6	28.0	58.9	20.4 $^{\uparrow 12.8}$	52.8 $^{\uparrow 24.8}$	67.7 $^{\uparrow 8.8}$
DINO [15]	4.3	17.3	50.2	14.5 $^{\uparrow 10.2}$	47.9 $^{\uparrow 30.6}$	61.3 $^{\uparrow 11.1}$	5.4	19.2	43.9	16.9 $^{\uparrow 11.5}$	50.0 $^{\uparrow 30.8}$	62.4 $^{\uparrow 18.5}$
TimeT [66]	12.2	46.2	66.3	17.9 $^{\uparrow 5.7}$	52.1 $^{\uparrow 5.9}$	68.5 $^{\uparrow 2.2}$	18.4	44.6	58.2	20.6 $^{\uparrow 2.2}$	54.3 $^{\uparrow 9.7}$	64.8 $^{\uparrow 6.6}$

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo starting with different pretrained weights.

	Pascal VOC						COCO-Things					
	At Init			+PaNeCo			*At Init*			+PaNeCo		
Pretrain	K=GT	K=500	Lin.	K=GT	K=500	Lin.	K=21	K=500	Lin.	K=21	K=500	Lin.
iBOT [92]	4.4	31.1	66.1	15.4 ↑11.0	51.2 ↑20.1	68.6 ↑2.5	7.6	28.0	58.9	20.4 ↑12.8	52.8 ↑24.8	67.7 ↑8.8
DINO [15]	4.3	17.3	50.2	14.5 ↑10.2	47.9 ↑30.6	61.3 ↑11.1	5.4	19.2	43.9	16.9 ↑11.5	50.0 ↑30.8	62.4 ↑18.5
TimeT [66]	12.2	46.2	66.3	17.9 ↑5.7	52.1 ↑5.9	68.5 ↑2.2	18.4	44.6	58.2	20.6 ↑2.2	54.3 ↑9.7	64.8 ↑6.6
Leopart [93]	15.4	51.2	66.5	21.0 ↑5.6	55.3 ↑4.1	68.3 ↑1.8	14.8	53.2	63.0	18.8 ↑4.0	53.9 ↑0.7	65.4 ↑2.4

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo starting with different pretrained weights.

	Pascal VOC						COCO-Things					
	At Init			+PaNeCo			*At Init*			+PaNeCo		
Pretrain	K=GT	K=500	Lin.	K=GT	K=500	Lin.	K=21	K=500	Lin.	K=21	K=500	Lin.
iBOT [92]	4.4	31.1	66.1	$15.4^{\uparrow 11.0}$	$51.2^{\uparrow 20.1}$	$68.6^{\uparrow 2.5}$	7.6	28.0	58.9	$20.4^{\uparrow 12.8}$	$52.8^{\uparrow 24.8}$	$67.7^{\uparrow 8.8}$
DINO [15]	4.3	17.3	50.2	$14.5^{\uparrow 10.2}$	$47.9^{\uparrow 30.6}$	$61.3^{\uparrow 11.1}$	5.4	19.2	43.9	$16.9^{\uparrow 11.5}$	$50.0^{\uparrow 30.8}$	$62.4^{\uparrow 18.5}$
TimeT [66]	12.2	46.2	66.3	$17.9^{\uparrow 5.7}$	$52.1^{\uparrow 5.9}$	$68.5^{\uparrow 2.2}$	18.4	44.6	58.2	$20.6^{\uparrow 2.2}$	$54.3^{\uparrow 9.7}$	$64.8^{\uparrow 6.6}$
Leopart [93]	15.4	51.2	66.5	$21.0^{\uparrow 5.6}$	$55.3^{\uparrow 4.1}$	$68.3^{\uparrow 1.8}$	14.8	53.2	63.0	$18.8^{\uparrow 4.0}$	$53.9^{\uparrow 0.7}$	$65.4^{\uparrow 2.4}$
CrIBo [49]	18.3	54.5	71.6	$21.7^{\uparrow 3.4}$	$59.6^{\uparrow 5.1}$	$72.1^{\uparrow 0.5}$	14.5	48.3	64.3	$21.1^{\uparrow 6.6}$	$54.0^{\uparrow 5.7}$	$68.0^{\uparrow 3.7}$

frozen clustering and linear segmentation results on Pascal VOC and COCO-Things.

→ PaNeCo considerably boosts (↑) the performance of **different backbones**

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

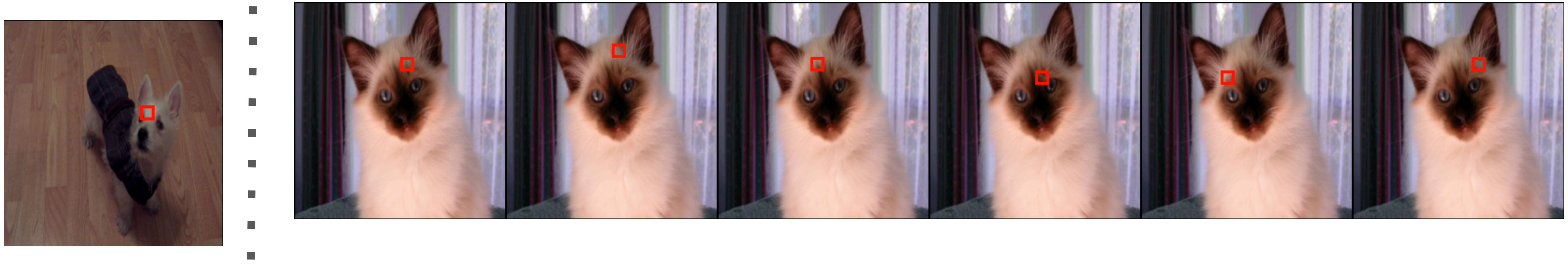
Qualitative Results

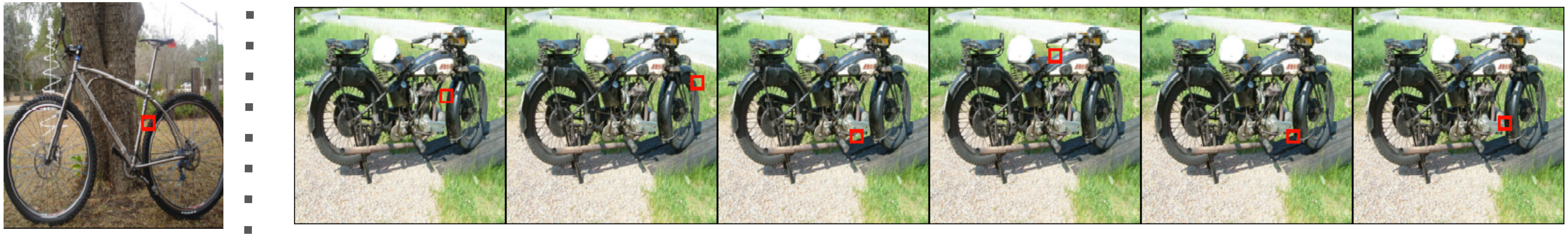
Nearest Neighbors of Patches from representations

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

PaNeCo rarely confuses semantically close patches

Query | **Retrieved Nearest Neighbors**

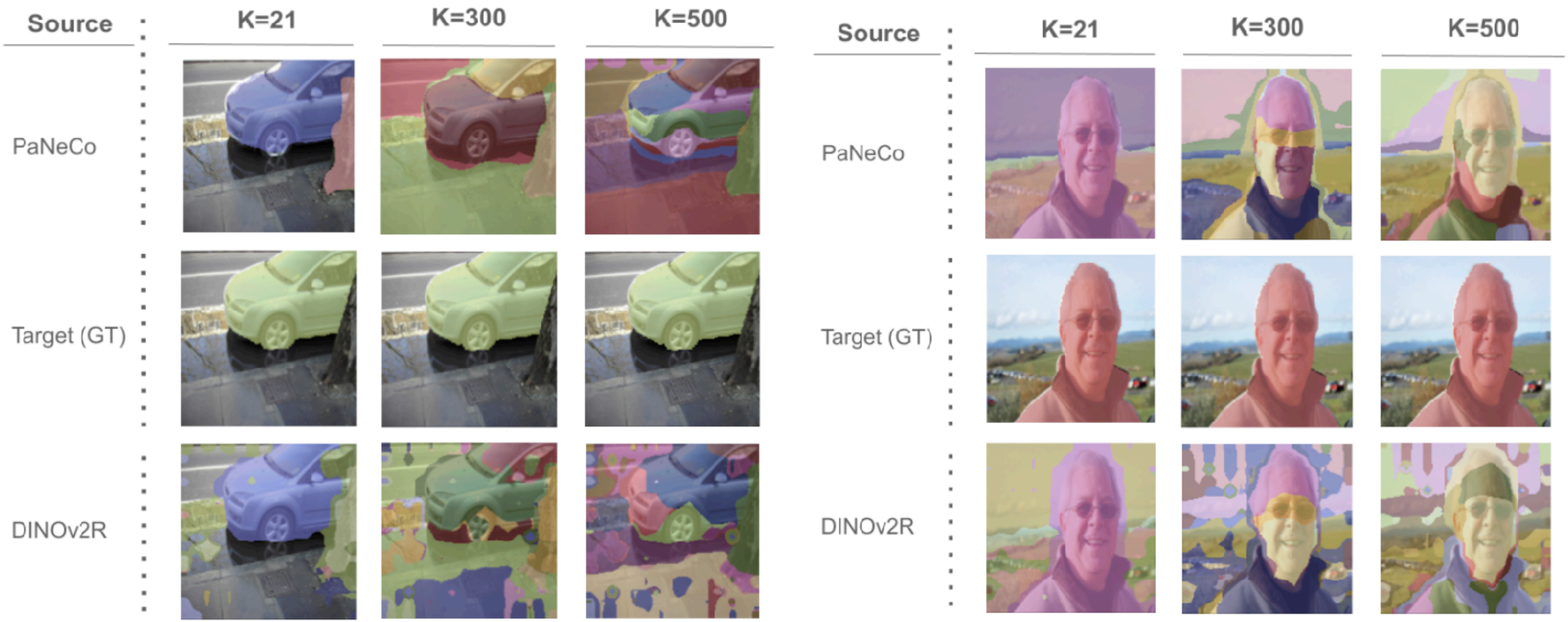




On average such cases appear around 6% of the times from Pascal VOC retrieval cases.

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

k-Means Semantic Segmentation



NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

What's the sauce?

- **Dense Patch-ordering** is loss well suited for post-pretraining
- We can **improve upon (very strong) DINO/ DINOv2R** models
- Strongest improvements in in-context semantic segmentation and even full-finetuning
- also: code/models now available!

NeCo: Improving DINOv2's spatial representations in 19 GPU hours with Patch Neighbor Consistency. Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, Yuki M. Asano. arxiv 2024

https://tomandanitamorgan.blogspot.com/2014/04/black-and-white-pasta-salad.ht

Time does tell: self-supervised time-tuning of dense image representations.

Mohammadreza Salehi, Efstratios Gavves, Cees G.M. Snoek, Yuki

Current Vision Foundation Models are trained with images. Videos can enable new directions

Visual development for AI

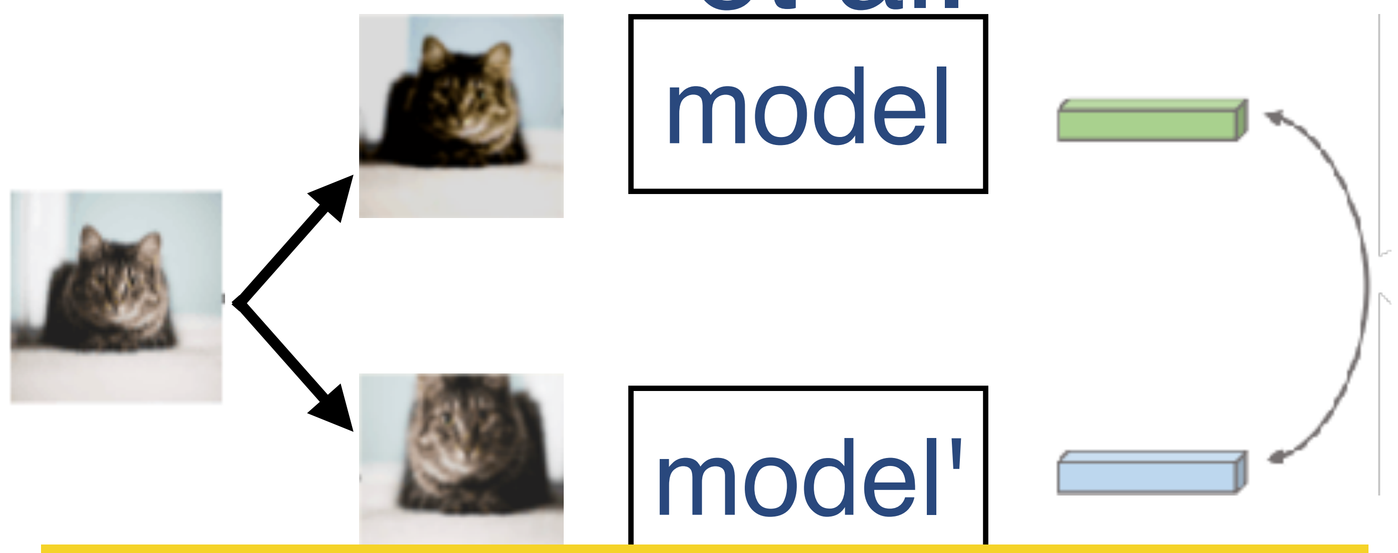
"Get" physics

Embodied AI

+ their *insane* scal

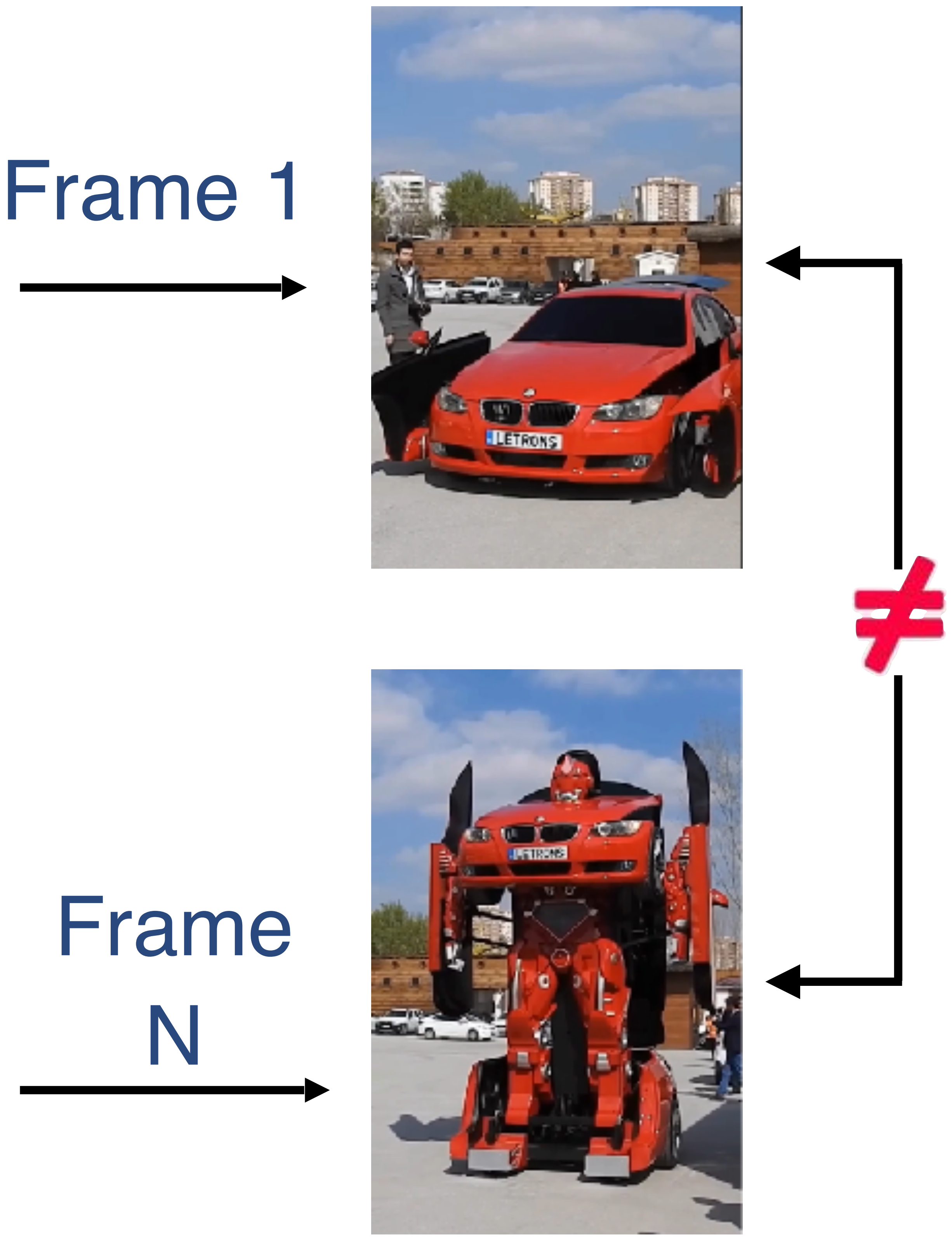
Augmentations are crucial in classic image-SSL,
but forcing frames to be invariant is limiting

Images: SimCLR, MoCo, SwaAV et al.



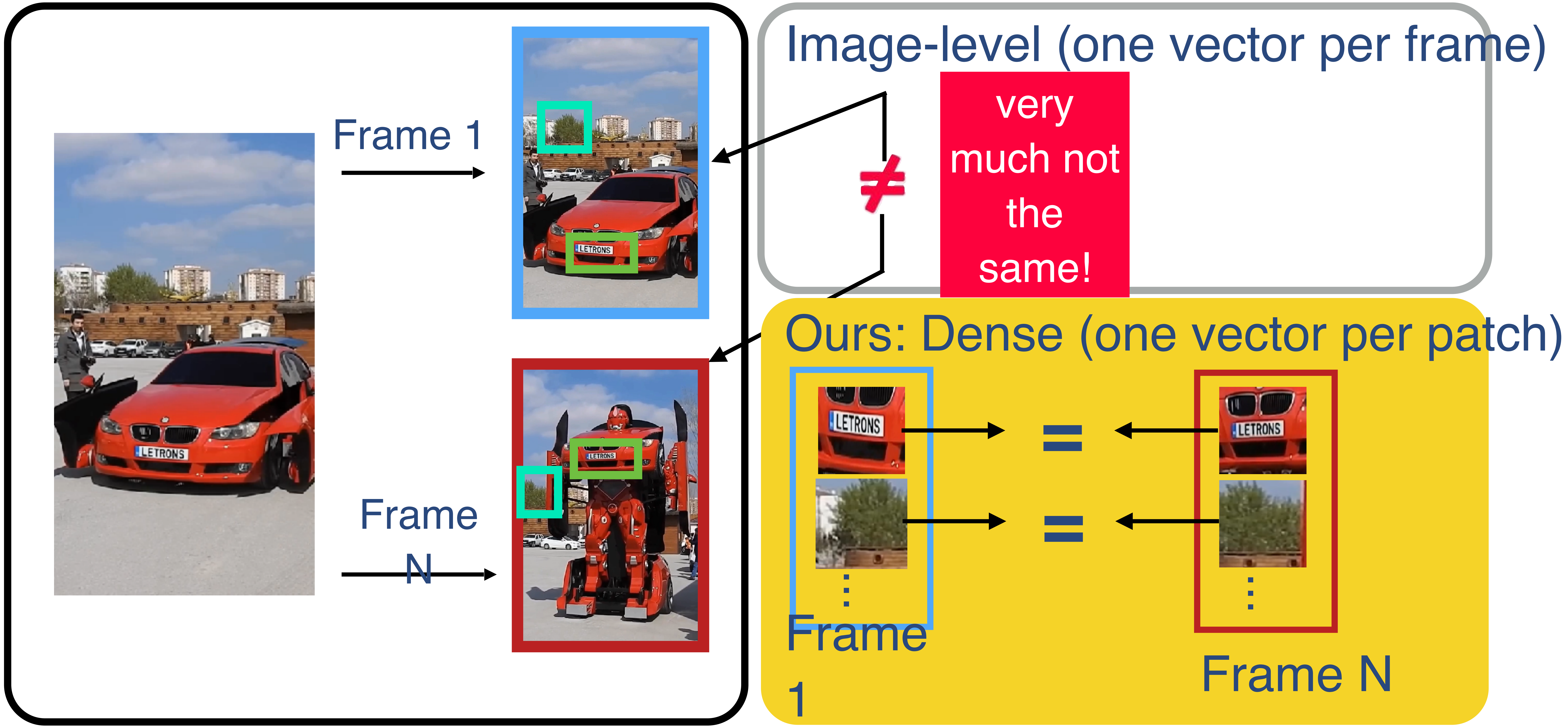
key principle: view-invariance

But does this generally make sense?



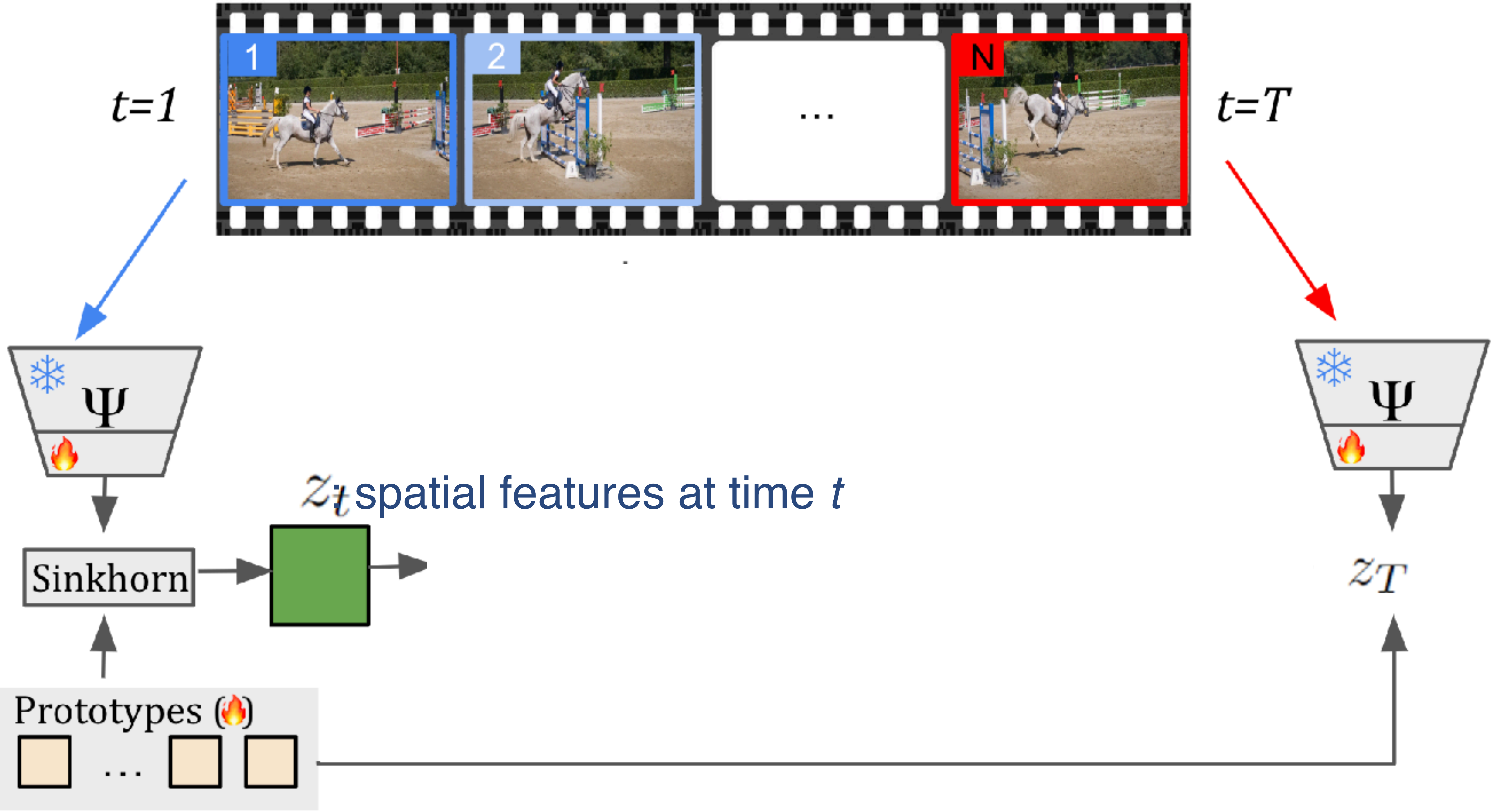
very much not the same!

Solution is obvious



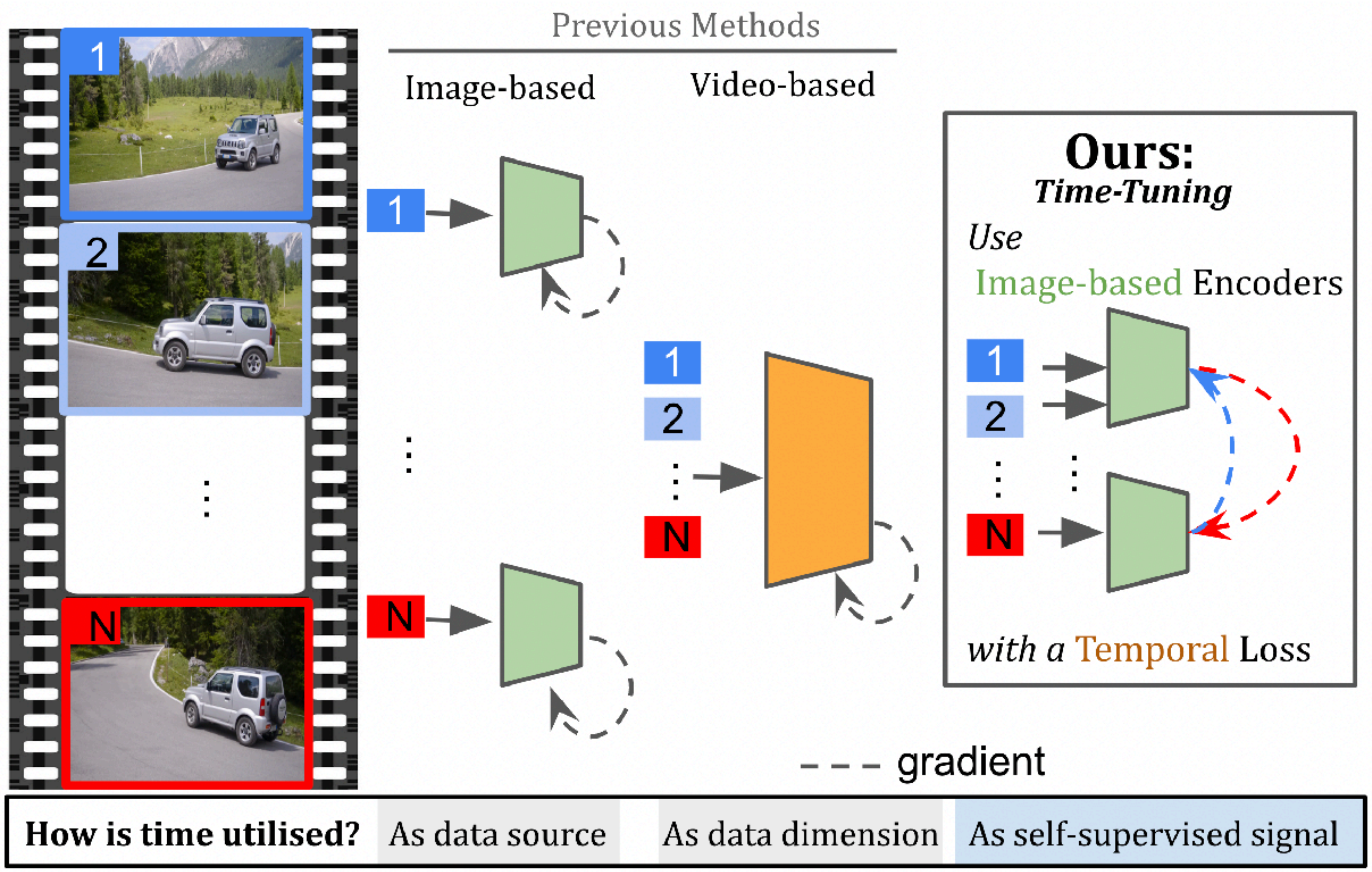
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations.* ICCV 2023

We *model* a video by tracking image patches, and aligning their clustered features



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

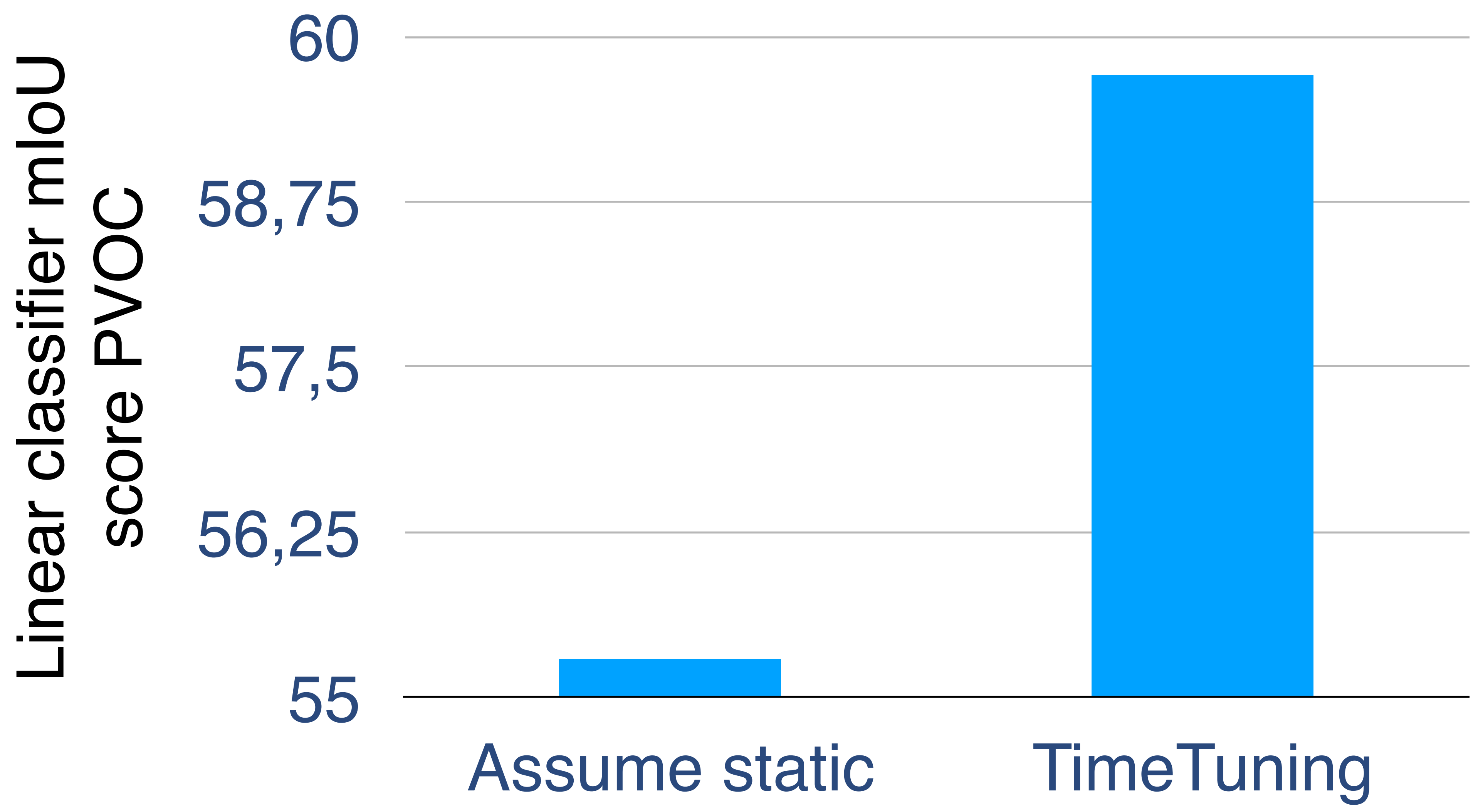
Using videos to learn self-supervised image encoders



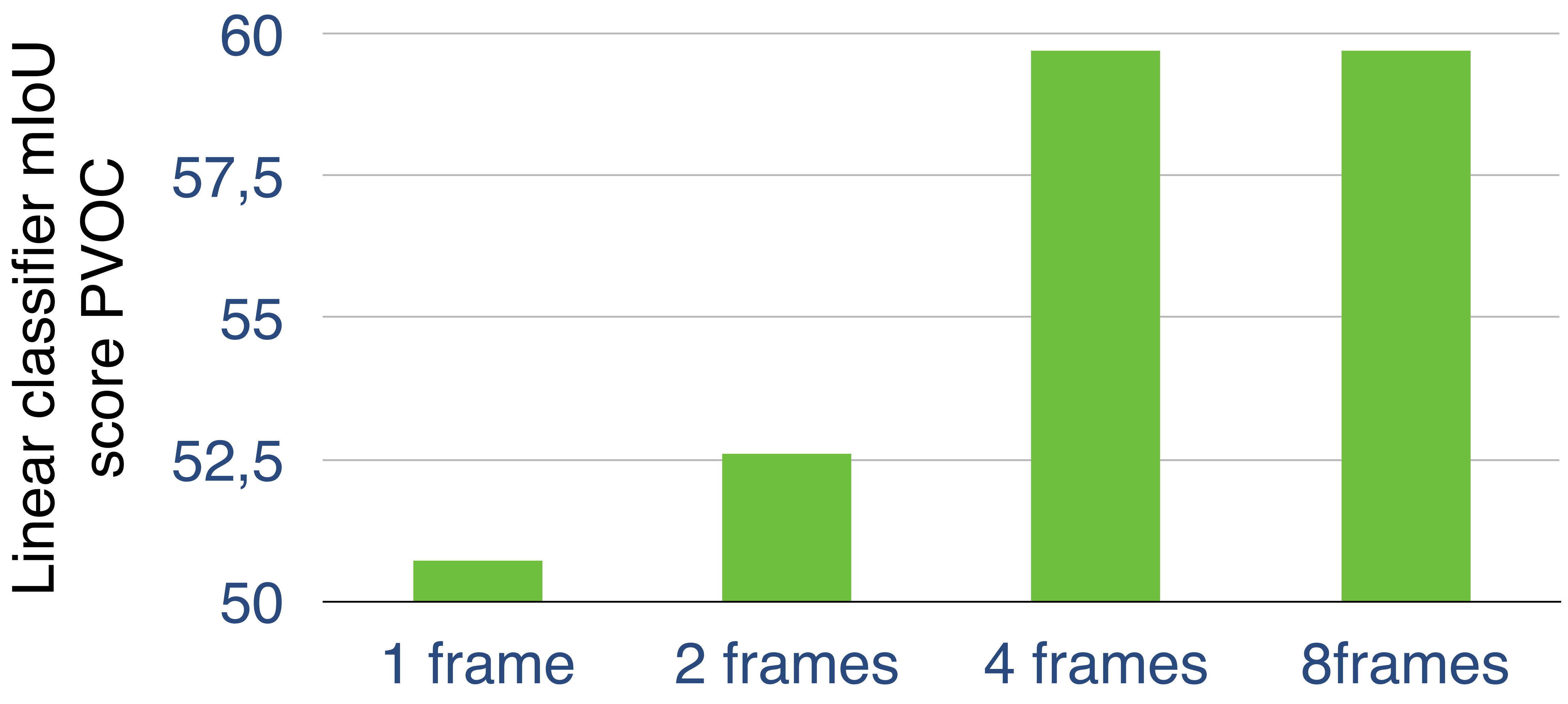
Salehi, Gavves, Snoek, Asano. *Time does tell: self-supervised time-tuning of dense image representations*. ICCV 2023

Ablations demonstrate using time helps learn better features

Modelling time is esential



Model learns from temporal info



Results

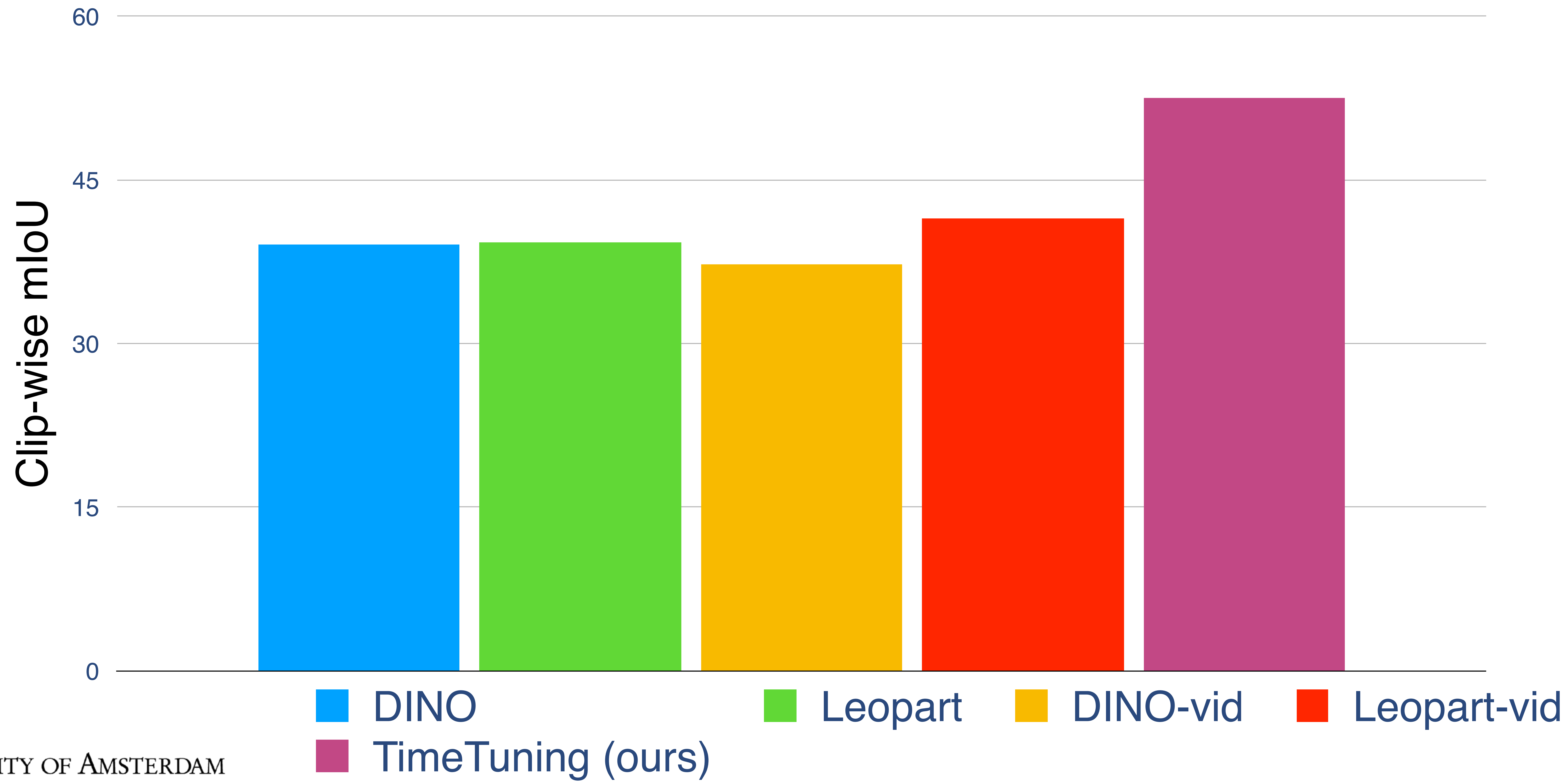
SoTA on unsupervised video segmentation

	Clustering					
	YTVOS			DAVIS		
	F	*C*	*D*	*F*	*C*	*D*
Trained on Images						
Resnet50	44.0	43.4	1.7	39.3	37.4	4.2
SwAV [8]	39.5	38.2	3.2	32.0	29.6	7.3
DINO [9]	39.1	37.9	1.9	30.2	31.0	1.6
Leopart [74]	39.2	37.9	11.7	30.3	30.2	16.5
Trained on Videos						
STEGO*	41.5	40.3	2.0	31.9	31.0	3.2
DINO*	37.2	36.1	1.2	29.3	29.2	2.4
Leopart*	41.5	40.5	7.7	37.5	36.5	12.6
TIMET(ours)	**52.5**	**51.3**	**13.3**	**53.7**	**53.0**	**20.5**

SoTA on unsupervised image segmentation

	Pascal VOC			
	K=21	K=500	LC	FCN
Trained on Images				
ResNet-50	4.5	36.5	53.8	-
DINO [9]	5.5	17.4	50.6	60.6
SwAV [8]	11.6	35.7	50.7	-
MaskContrast [57]	35.0	45.4	49.2	-
DenseCL [61]	-	43.6	49.0	69.4
STEGO [21]	7.0	19.5	59.1	63.5
CrOC [52]	20.6	-	61.6	-
Leopart [74]	**36.6**	50.5	**68.0**	70.1
Trained on Videos				
STEGO*	4.0	15.5	51.1	55.5
Leopart*	14.9	21.2	53.2	63.2
Flowdino† [70]	-	-	59.4	-
TIMET (ours)	34.5	**53.2**	**68.0**	**70.6**

Results on unsupervised video semantic segmentation



Unsupervised Semantic Segmentation on videos

mply running k-means on a couple of videos' spatial features, k=10]

DINO

❌ part-centric maps

STEGO

❌ noisy maps

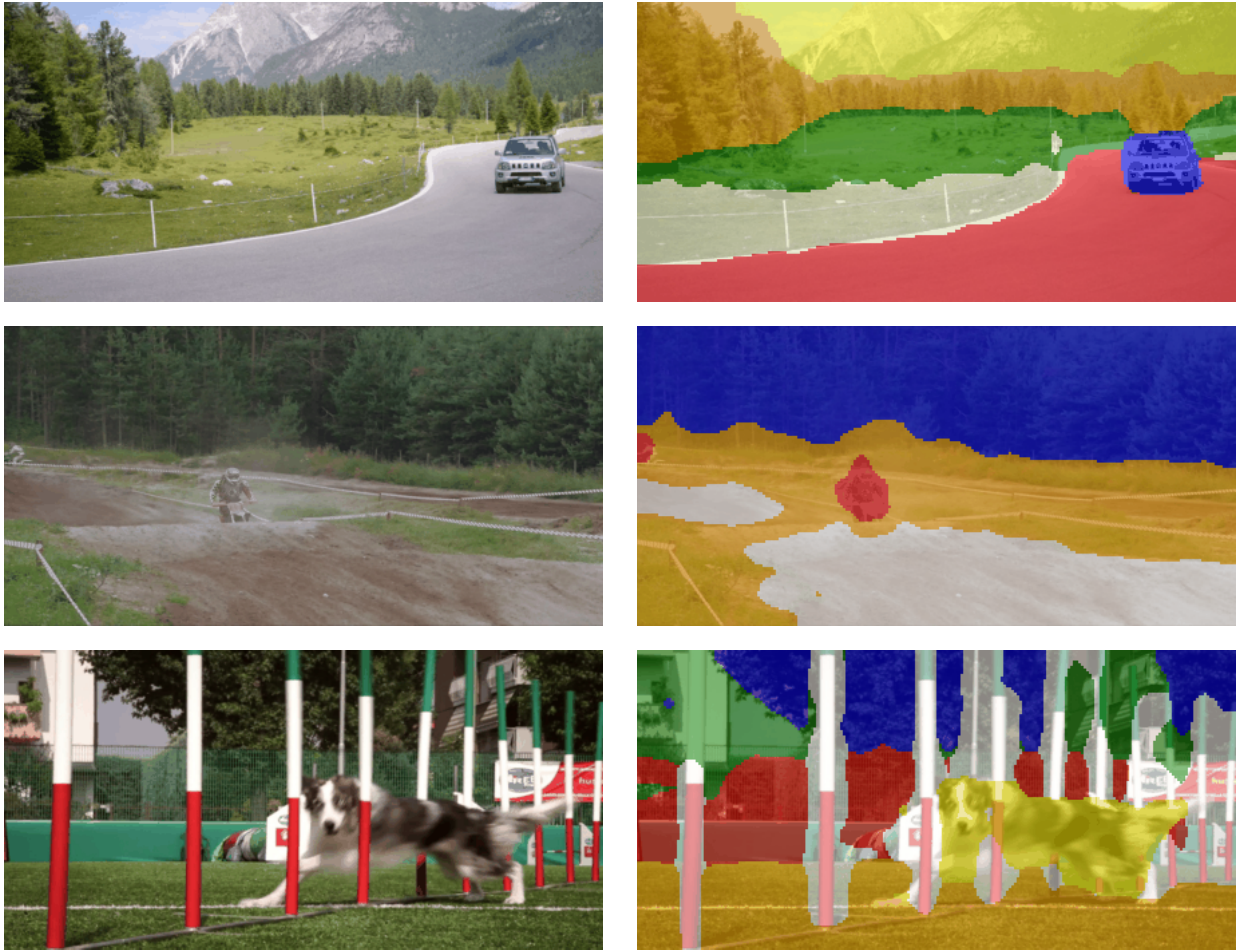
Ours

✅ crisp semantic maps

Unsupervised Semantic Segmentation on videos

[here: running k-means on the whole video's spatial features, k=5]

More examples



UNIVERSITY OF AMSTERDAM
YUKI ASANO

What's the sauce?

- **Videos** provide rich supervision signal
- Don't use frame-wise invariance across time, but instead patch-level invarianc
- We can **improve upon the (strong) DINO** model
- Strongest improvements in unsupervised segmentation

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

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

Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yannis Avrithis*, Yuki M. Asano*

https://www.barilla.com/it-it/ricette/tutte/farfalle-con-fave-e-pesto-ricotta-e-noci

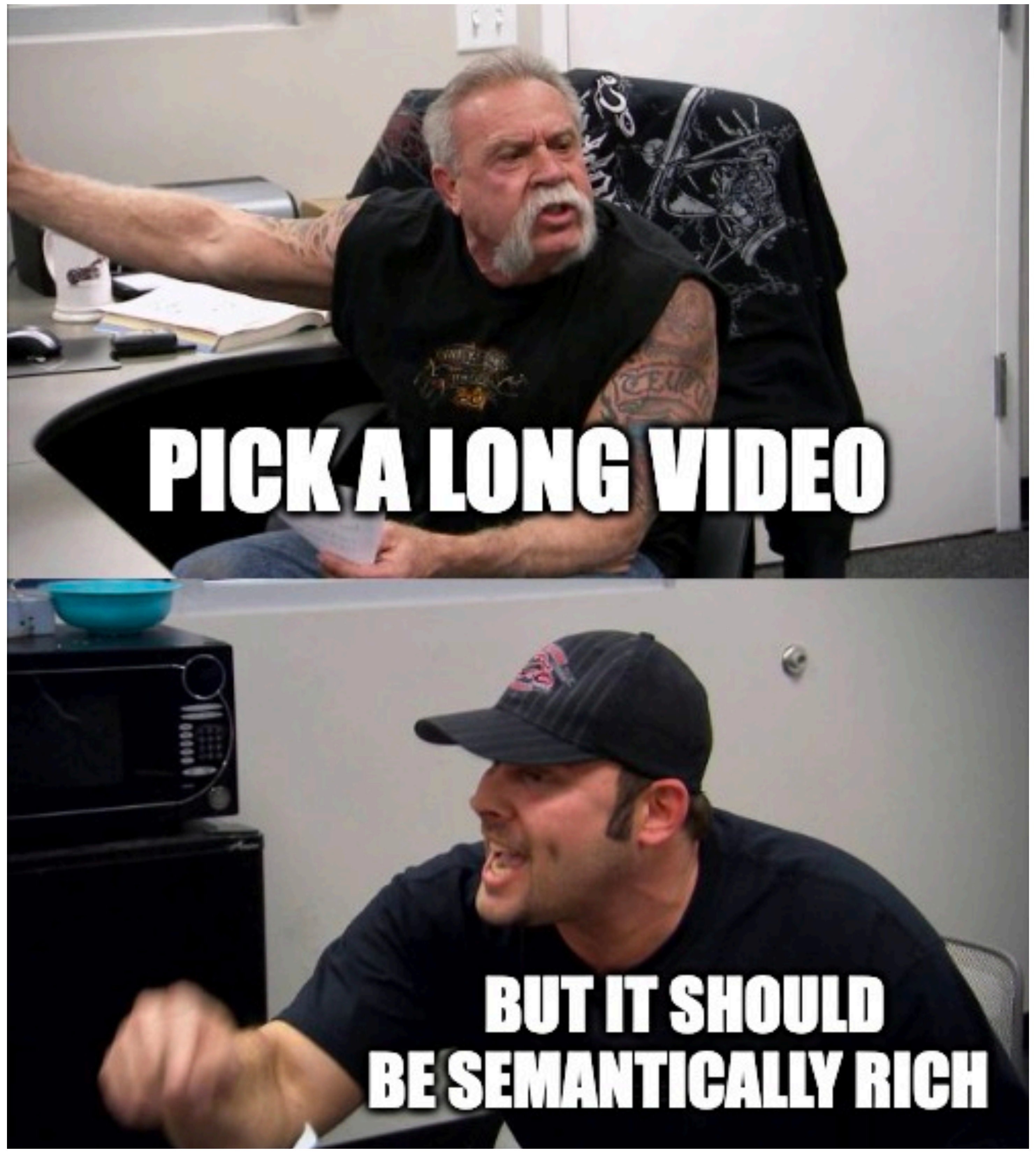
TimeTuning:
DINO as init & use temporal info of videos.

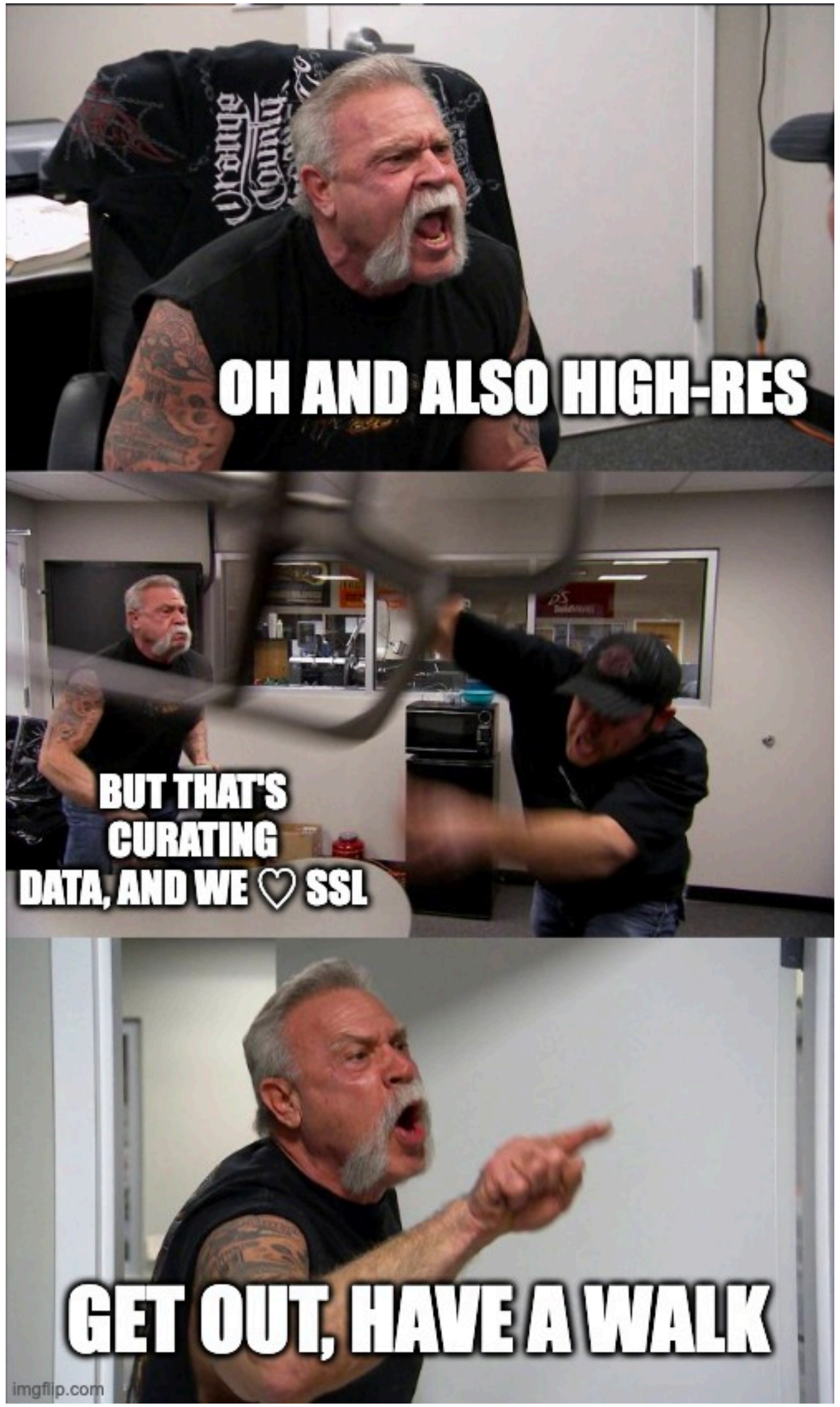
~~How powerful is time~~ without image-

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

us figuring out





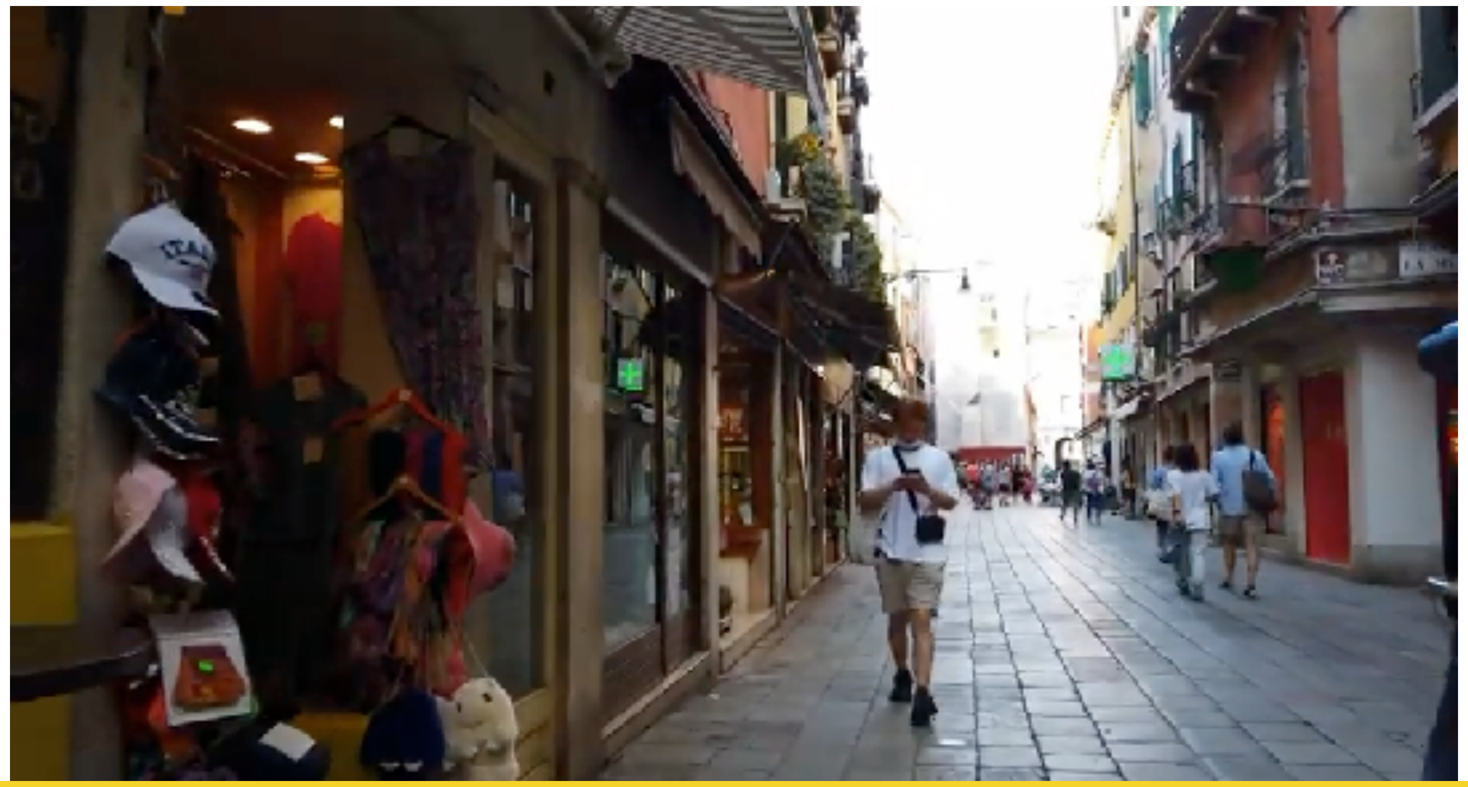
- ✓ Long
- ✓ High-res, smooth
- ✓ Semantically rich
- ✓ Scalable (we ❤️ SSL)

↓

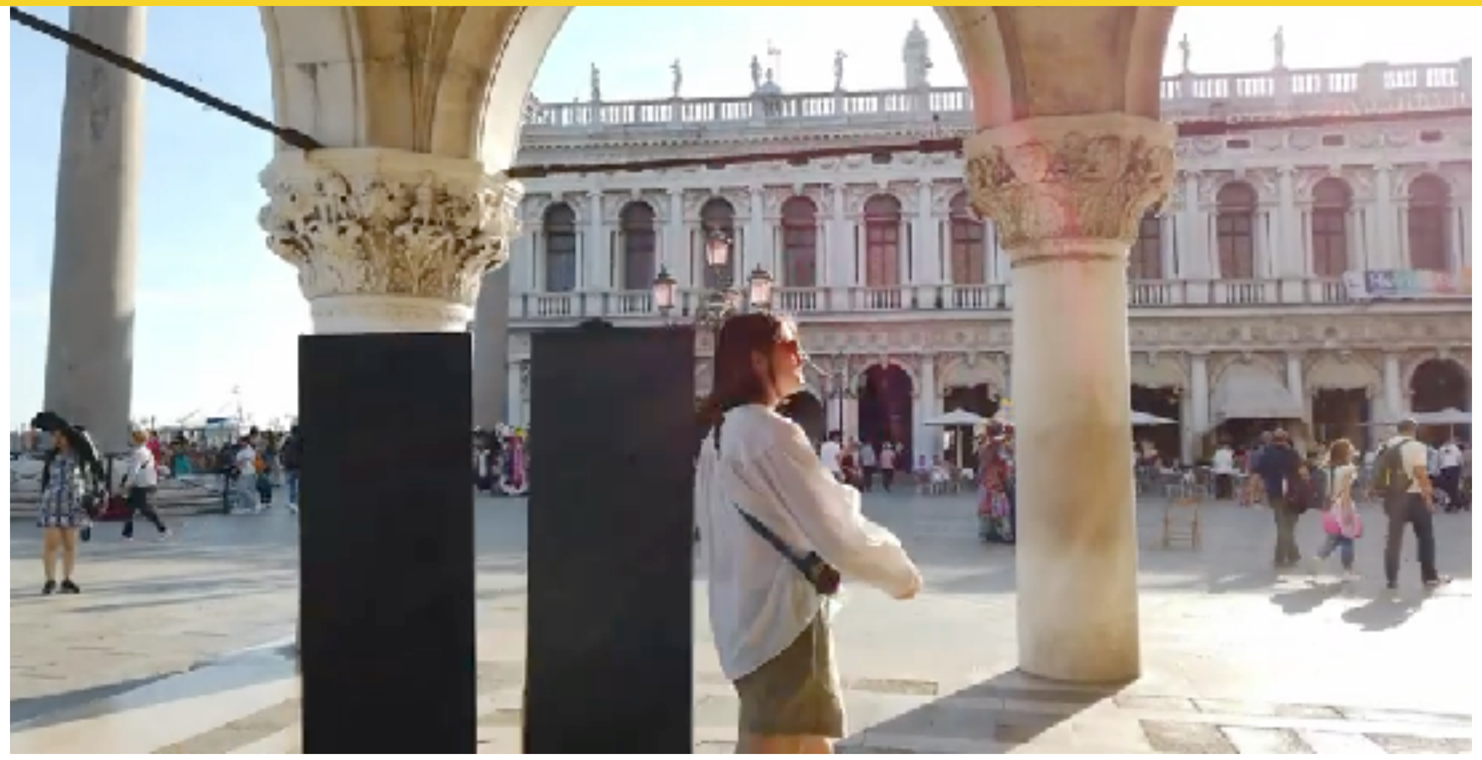
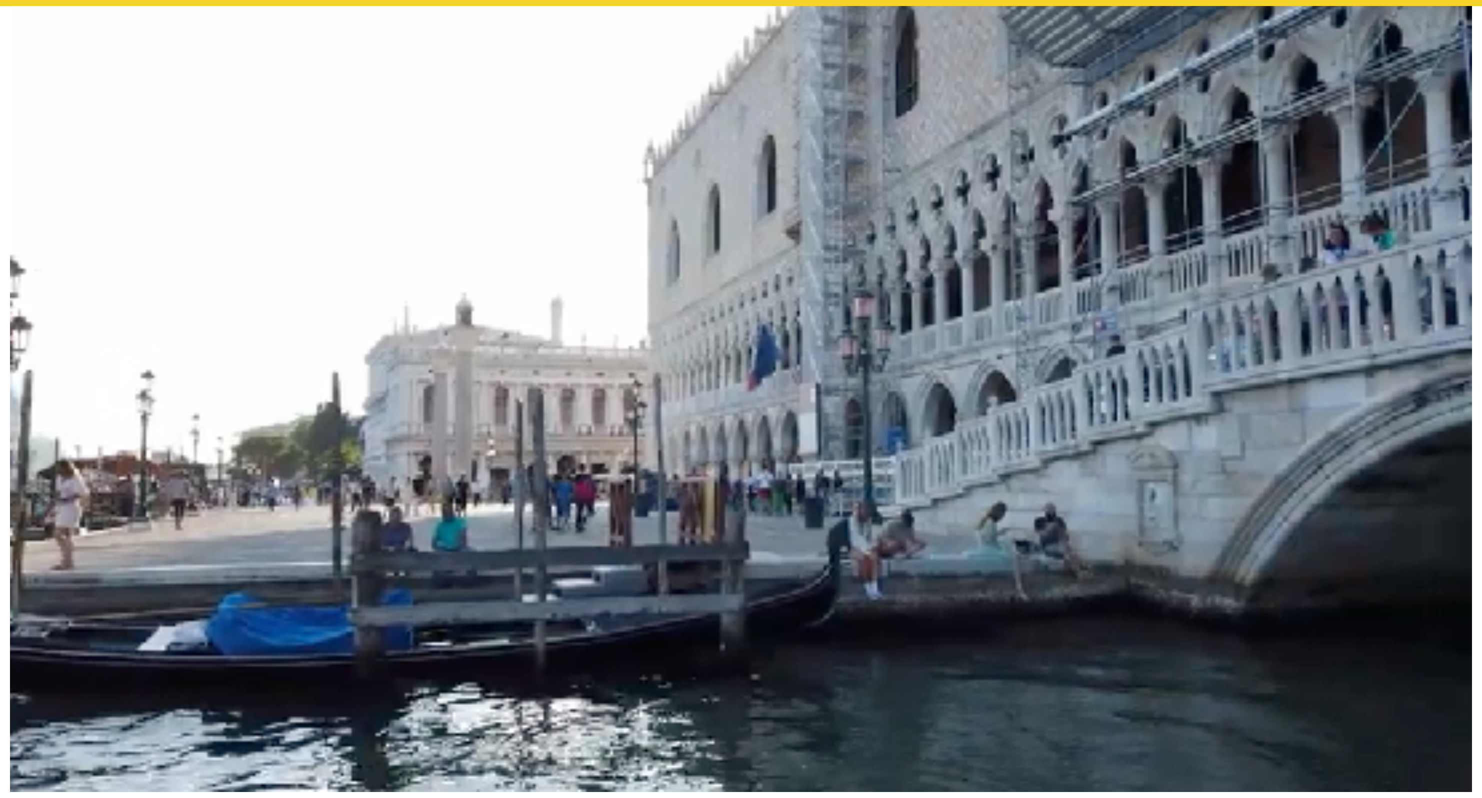
Walking Tours

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



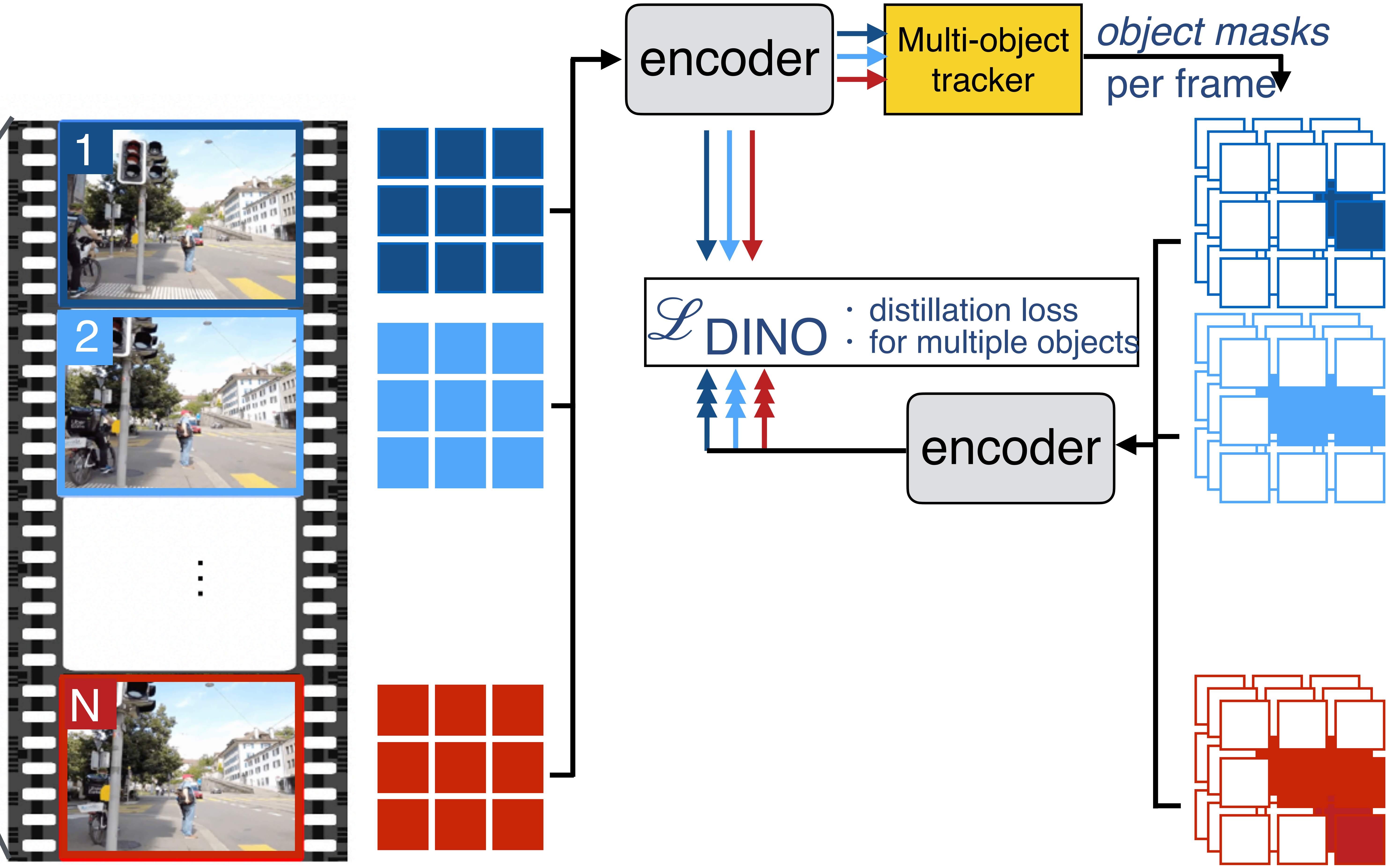
The dataset consists of 10x 4K videos of different cities' Walking Tours.


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

Dora: Discover and Track

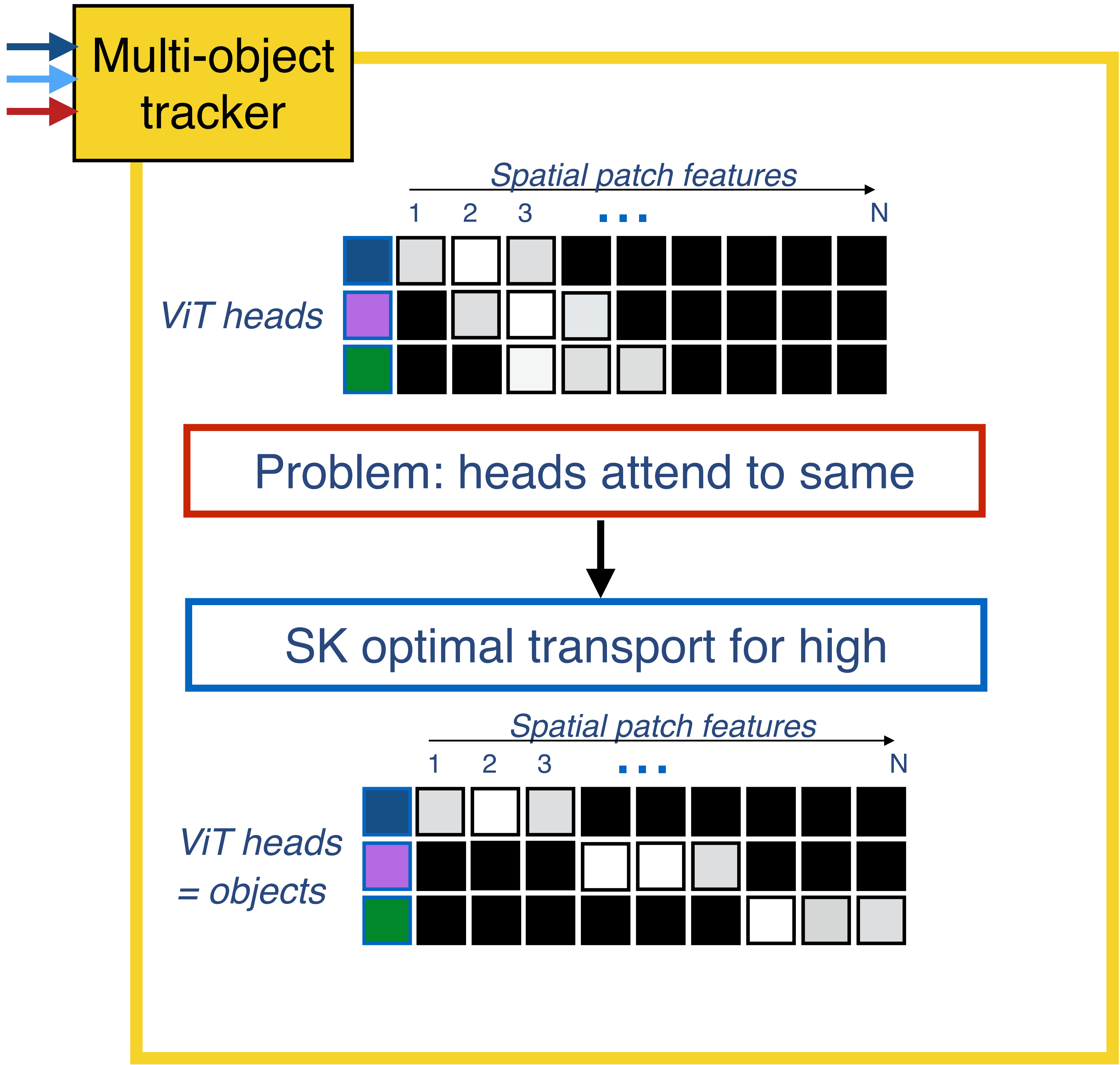
High-level idea:
1) track multiple objects across time
2) enforce invariance of features across



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

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

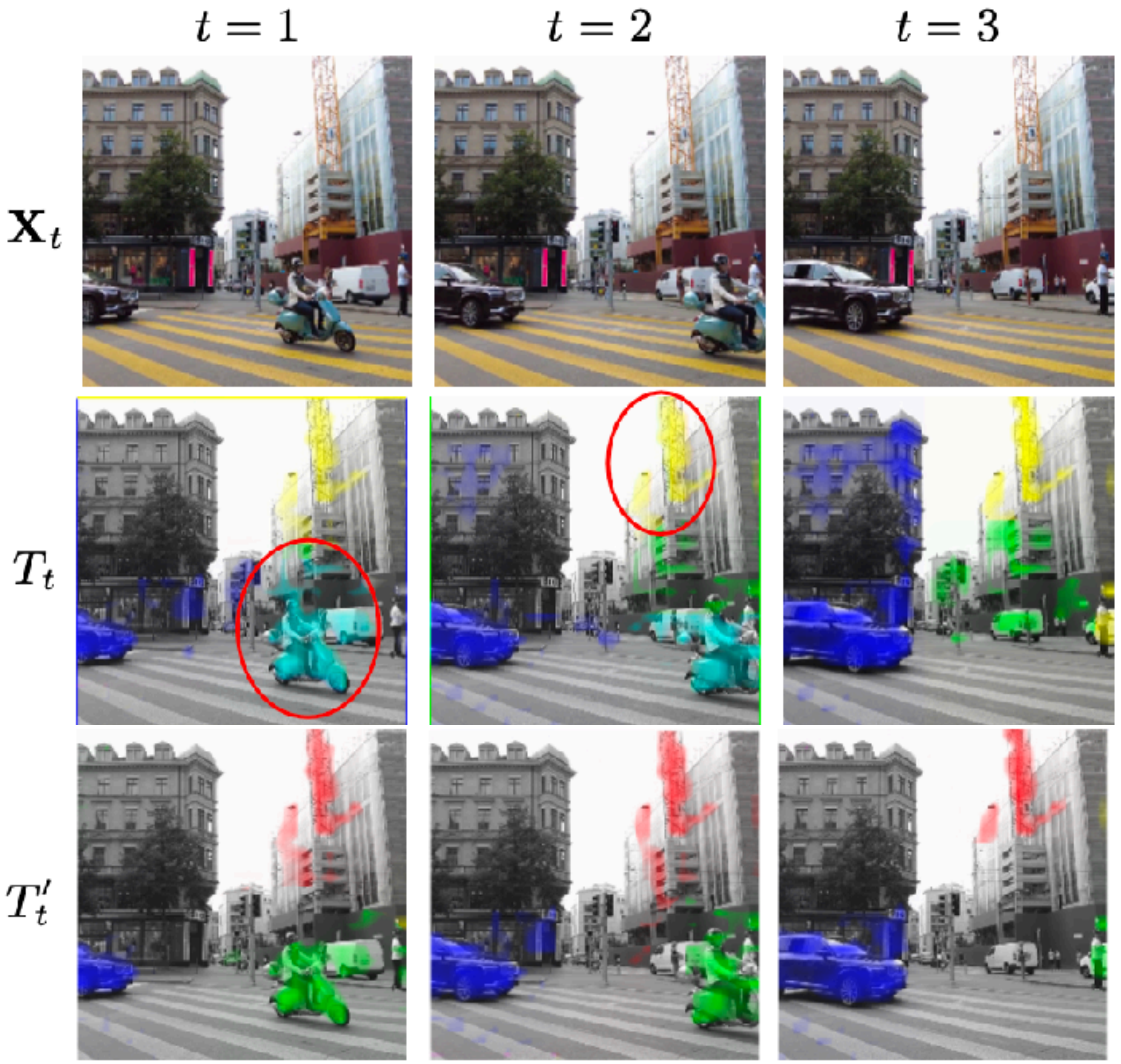
Spreading attention with Sinkhorn-Knopp



Visualise attention of 3 heads with colors R,G,B

without

with SK

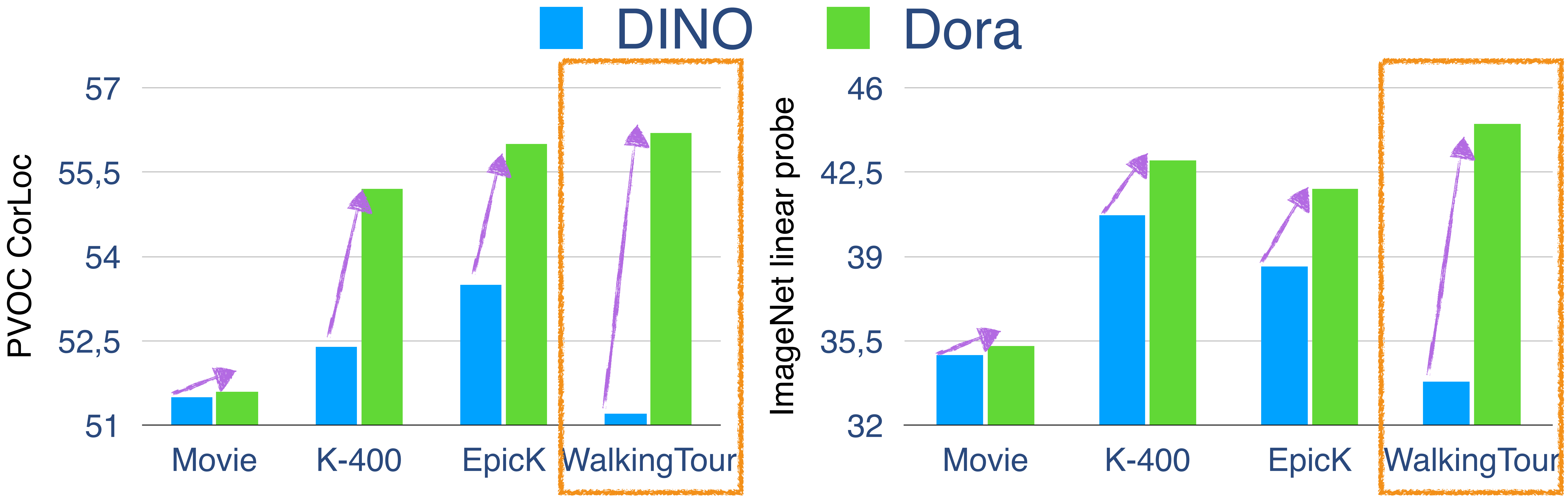


Venkataramanan, Rizve, Carreira, Asano*, Avrithis*. *Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video.* ICLR 2024

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

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

Dora better than DINO
WT+ Dora: great match



University of Amsterdam
YUKI ASANO

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

But how does it compare against ImageNet pretraining?

Dora (1WT) ~ on par with DINO (IN-1k)

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

What's the sauce?

- Training strong encoders **from scratch** with 1 video is possible
- Models match DINO (trained on ImageNet) in terms of performance
- The training loss is **spatially dense** and leverages **time**
- **Multi-object tracking** emerges
- **Walking videos** are great for training vision models

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

Future Foundation Models will be massively pretrained with videos.

Post-pretraining has a large potential that we're only beginning to exploit

Who gave the talk again?

Oh, hi, I'm Yuki

@y_m_asano

- Currently Assistant Professor with Video & Image Sense (VIS) Lab
- In two days:
  Full Professor and head of Fundamental AI Lab, University of Technology Nuremberg

Fundamental AI Lab

- Our focus:
  - Self-supervised Learning
  - Multimodal Learning
  - Large Language Models

- Let's collaborate!