Advanced School for Computing and Imaging

UNIVERSITY OF AMSTERDAM

Computer Vision by Learning

Cees Snoek, University of Amsterdam

Efstratios Gavves, University of Amsterdam

With an invited tutorial by: Yuki Asano, University of Technology Nuremberg

http://computervisionbylearning.info

amsterdam ai

Practicals: How and Where to Submit?

Lab website referred to the 2022 version, now updated for 2025 version
https://asci-cbl-practicals2025.readthedocs.io/en/latest/

Website contains an excel where you can enter your 2-person team info, please do so today.

Submission email: asci.cbl.practicals2025@googlemail.com

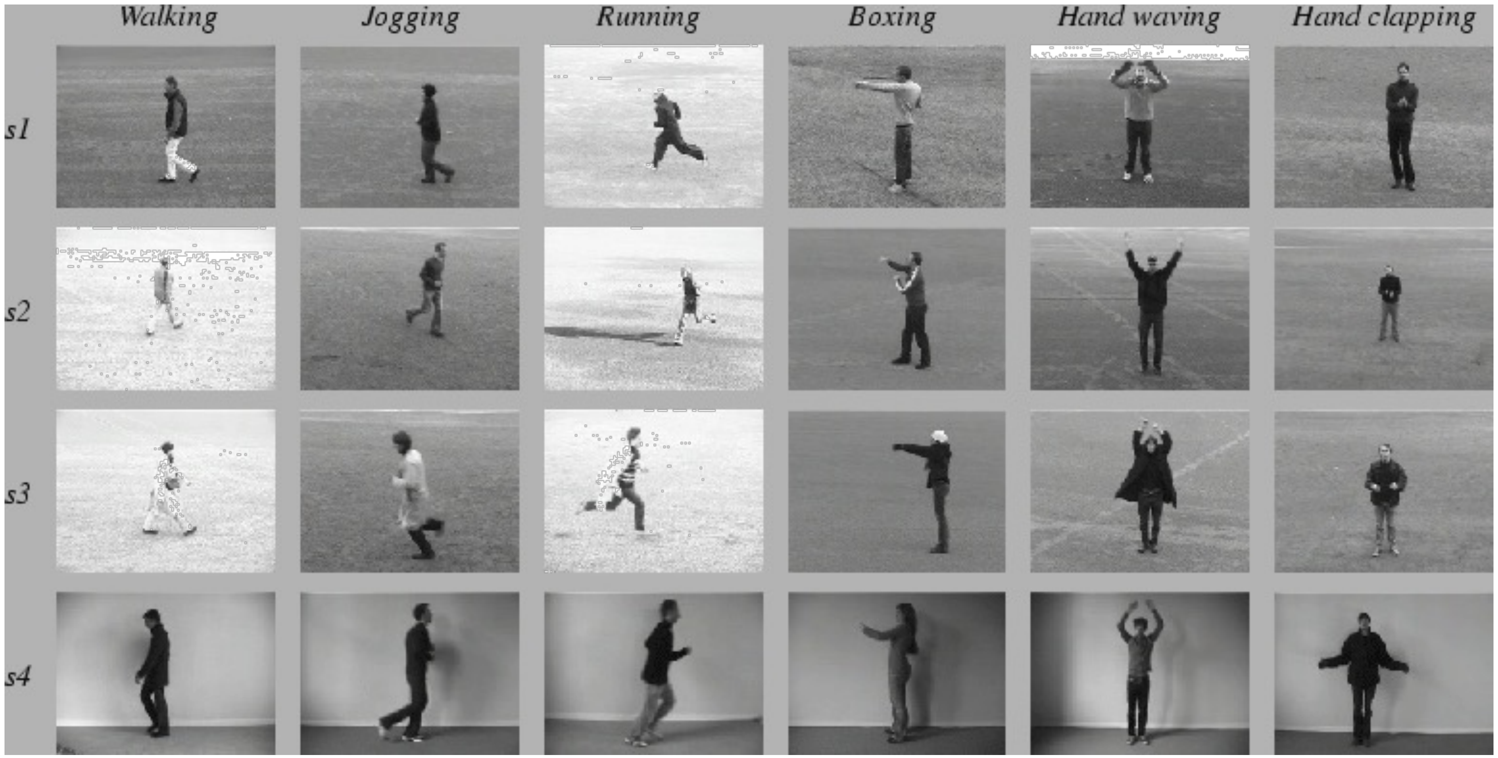
Deadline: **31 January 2025** (23:59 CEST)

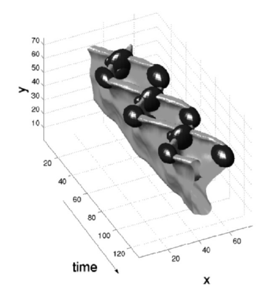
Learning to Generalize in Video Space and Time

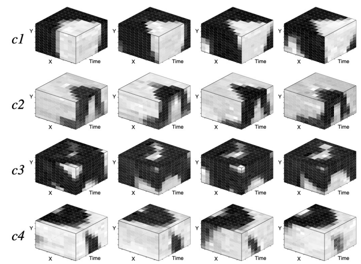
Prof. dr. Cees Snoek
University of Amsterdam

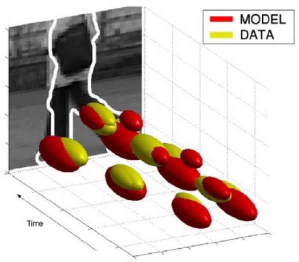
Head of VIS lab, HAVA lab
Scientific Director Amsterdam AI

How it started…





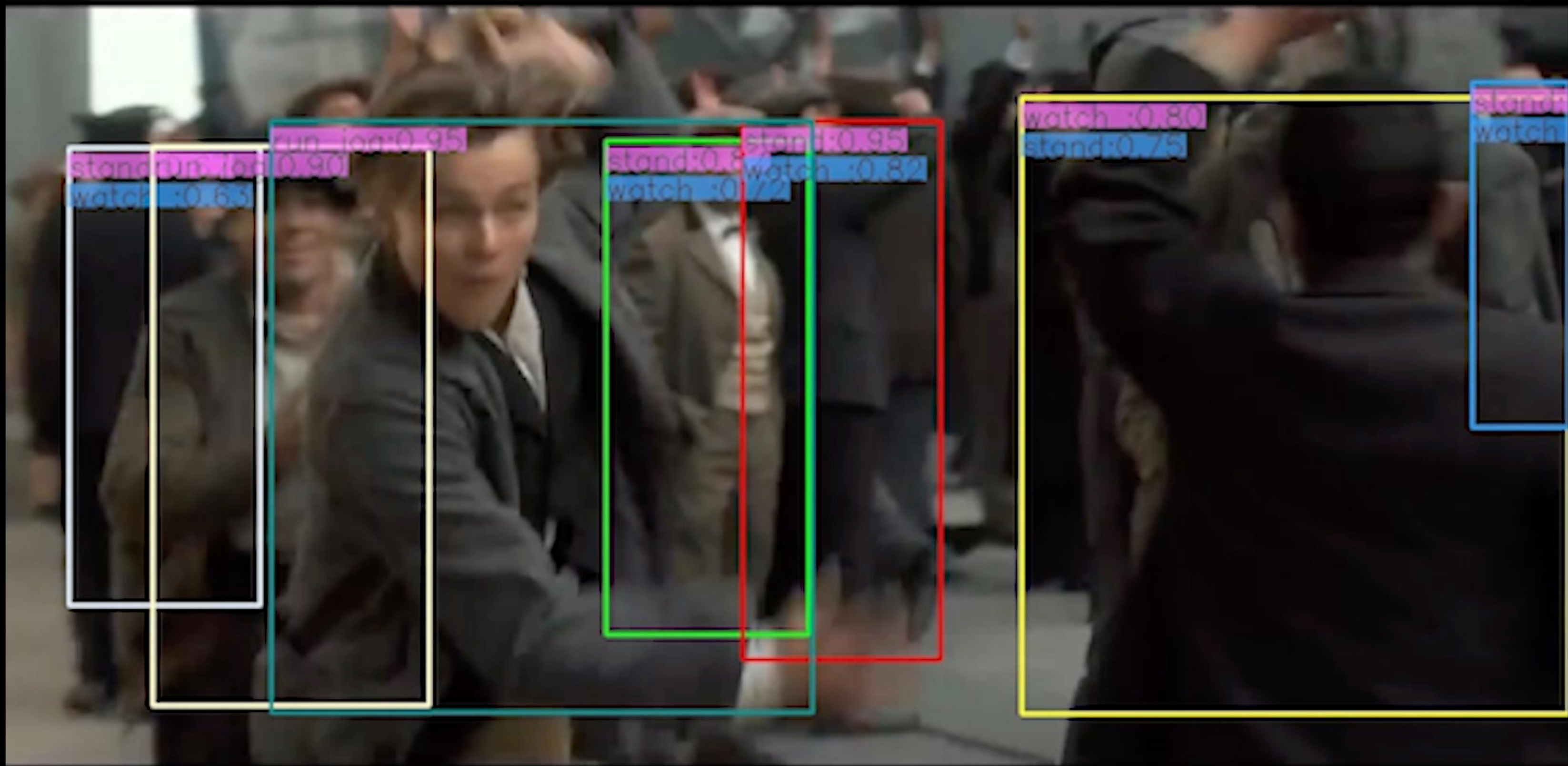




Laptev & Lindeberg, ICCV 2003

Du Tran et al., ICCV 2015

How it's going...



w/ Jiaojiao Zhao et al., CVPR 2022


“gray dog running on a leash during dog show”
w/ Kirill Gavrilyuk Amir Ghodrati, & Zhenyang Li, CVPR 2019

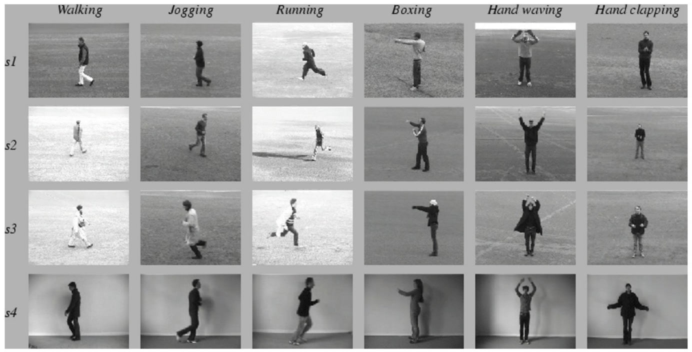
w/ Hazel Doughty, CVPR 2022

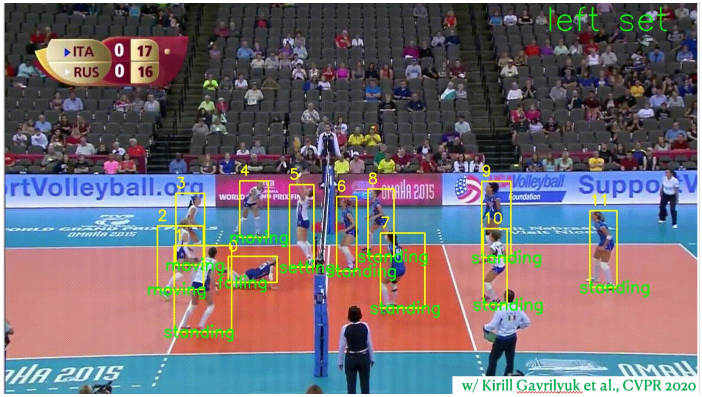
Action: **peel**

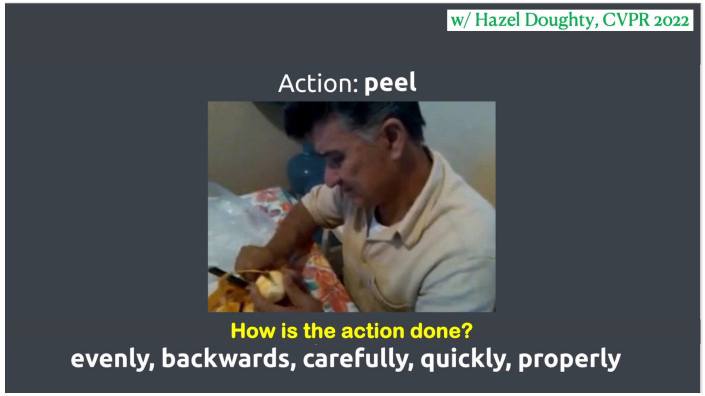
How is the action done?

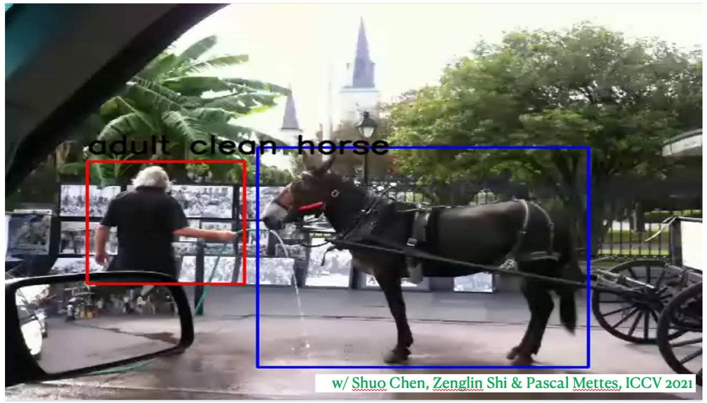
evenly, backwards, carefully, quickly, properly

What assumption do all these works have in common at training time?









Empirical risk minimization and the i.i.d. assumption

Empirical risk minimization

Definition. *Given a set of labeled data points* $S = ((x_1, y_1), \ldots, (x_n, y_n))$*, the* empirical risk *of a predictor* $f : \mathcal{X} \to \mathcal{Y}$ *with respect to the sample* S *is defined as*

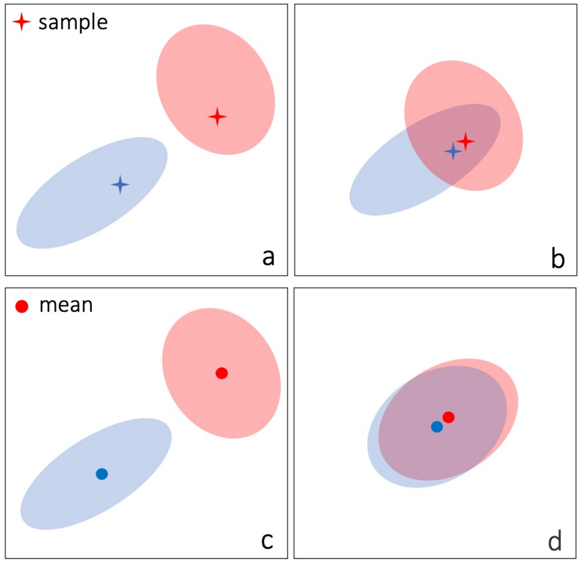
$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} loss(f(x_i), y_i)\,.$$

i.i.d. assumption

It is typically assumed that training, validation and test set are independent and identically distributed.

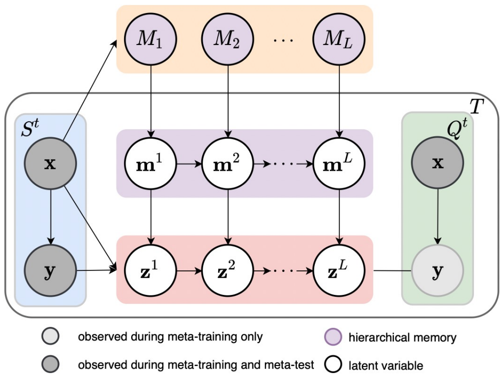
Machine learning inspiration

Domain-invariant learning



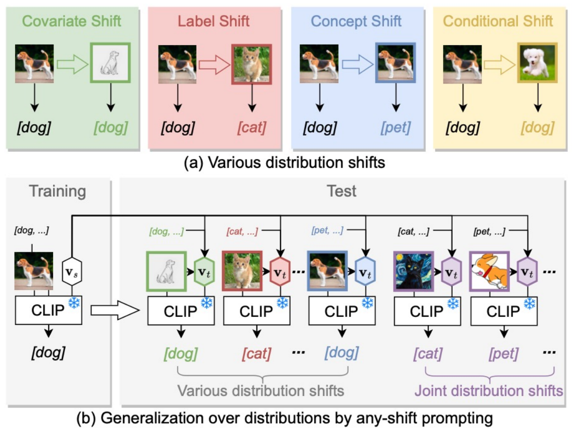
w/ Zehao Xiao et al., ICML 2021

Meta-learning



w/ Yingjun Du et al., ICLR 2022

Prompt learning



w/ Zehao Xiao et al., CVPR 2024

More is different

4 August 1972, Volume 177, Number 4047 SCIENCE

Philip Anderson crystallized the idea of emergence, arguing that “at each level of complexity entirely new properties appear” — that is, although, for example, chemistry is subject to the laws of physics, we cannot infer the field of chemistry from our knowledge of physics.

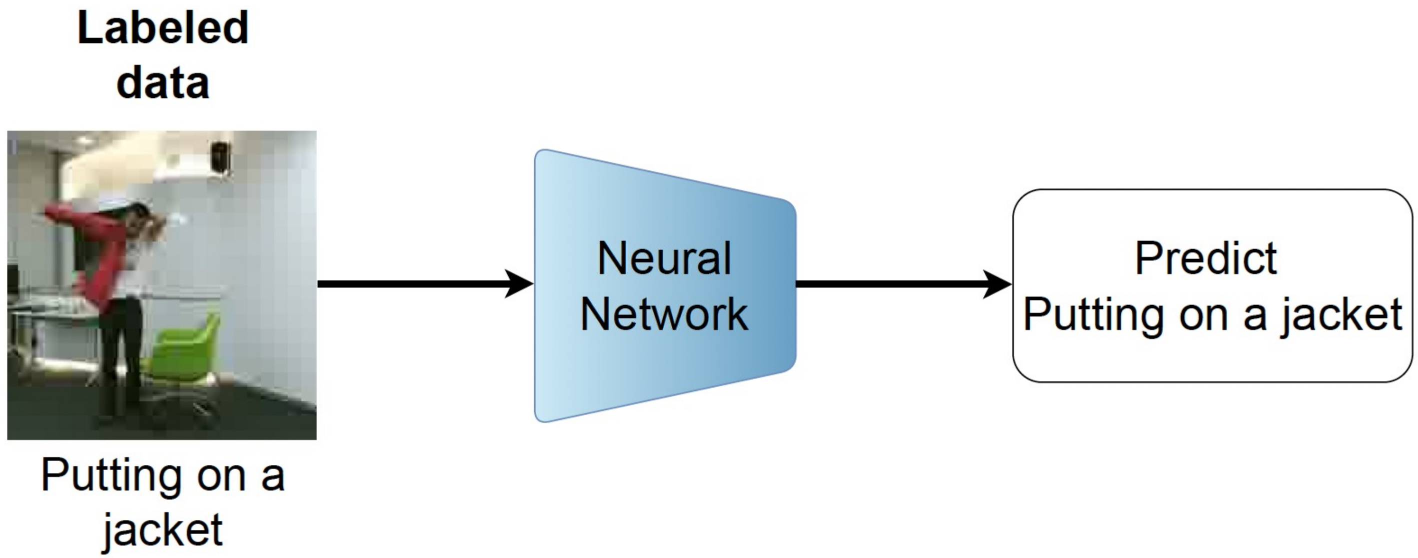
The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we have any detailed knowledge, are assumed to be controlled by the same set

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive. Once new fundamental laws are discovered, a large and ever increasing activity

search which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

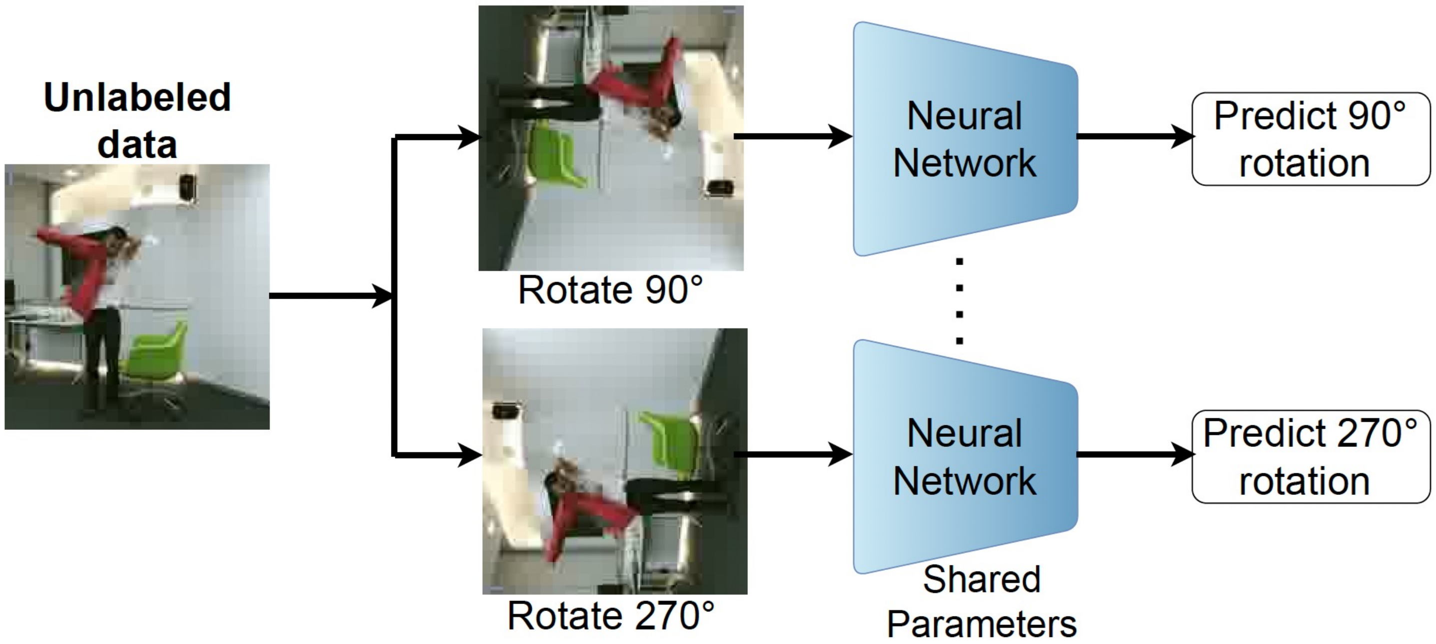
X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry

Supervised learning



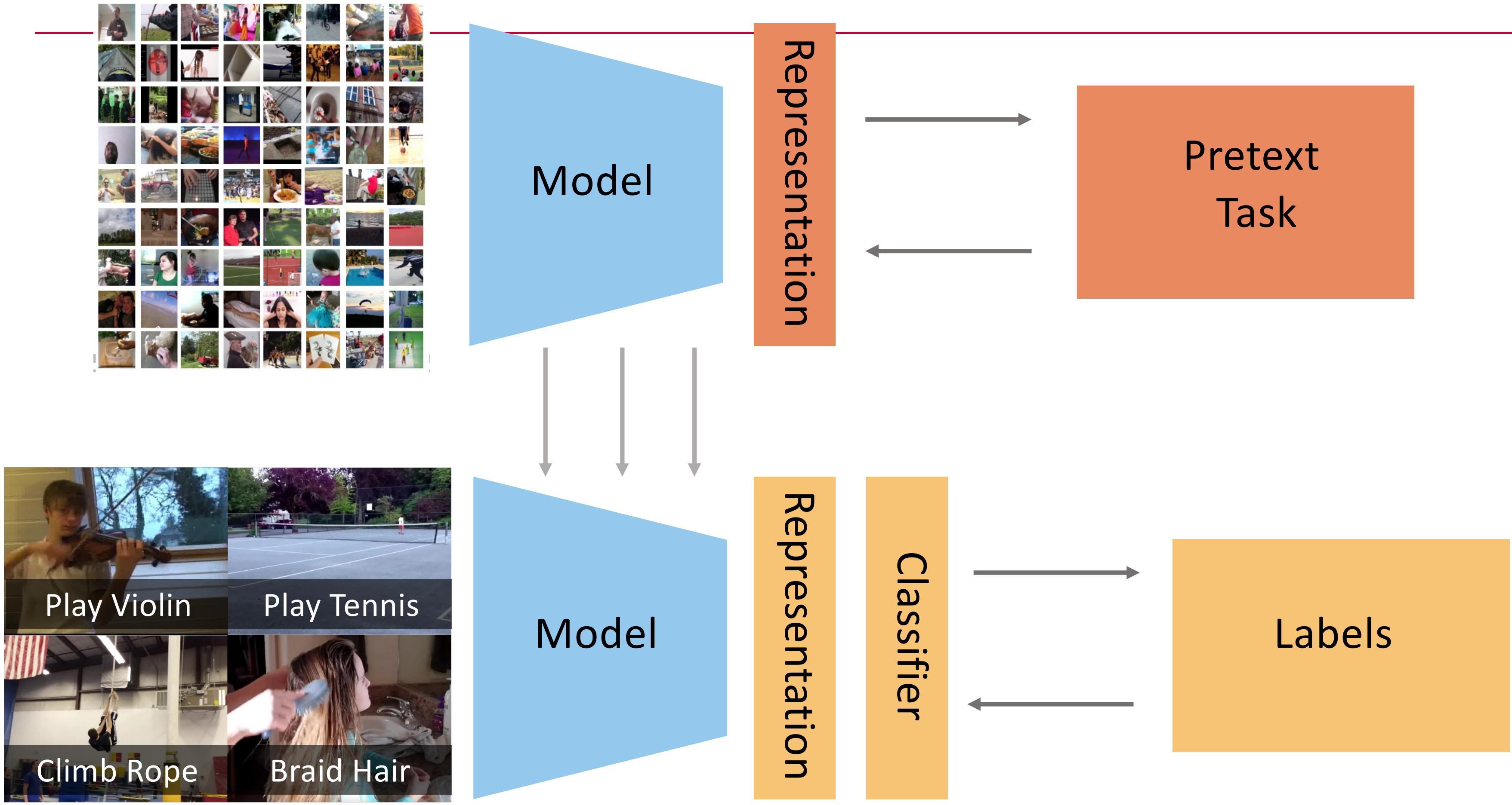
Depends on a manual labeling effort, which is costly, errorprone, and biased

Self-supervised learning using a proxy task

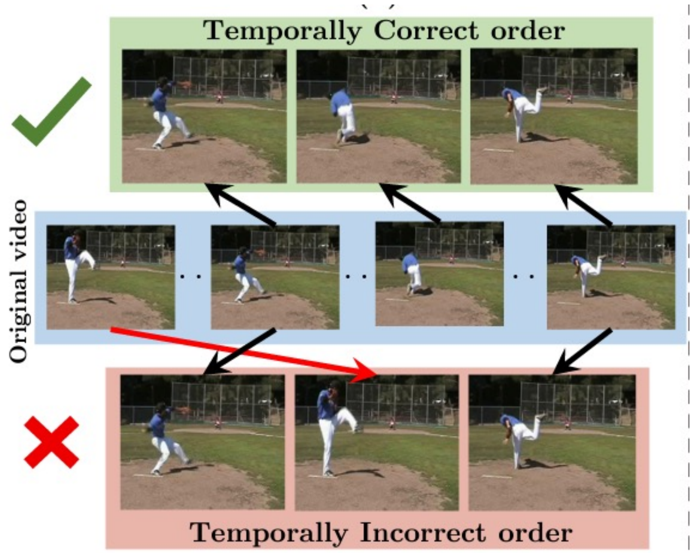


Self-supervised learning exploits (imposed) regularities in the data to learn from.

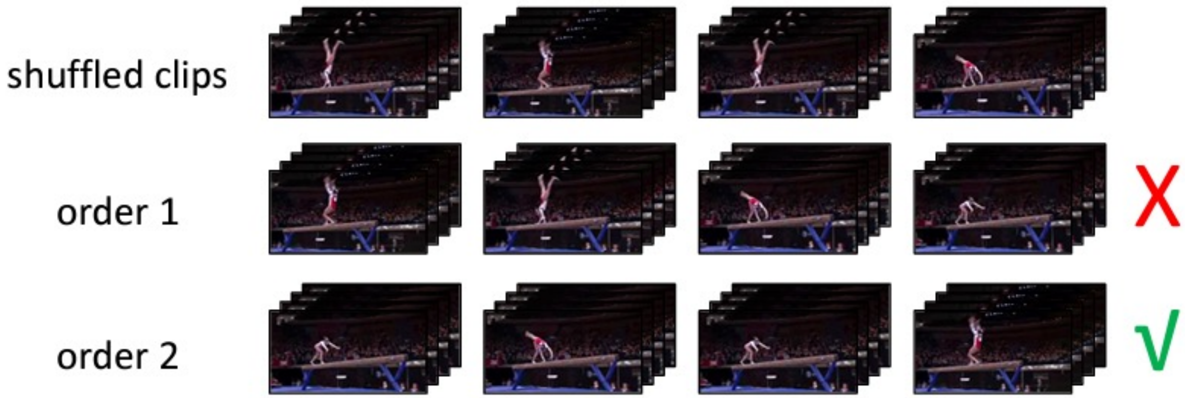
Self-Supervision



Example proxy tasks



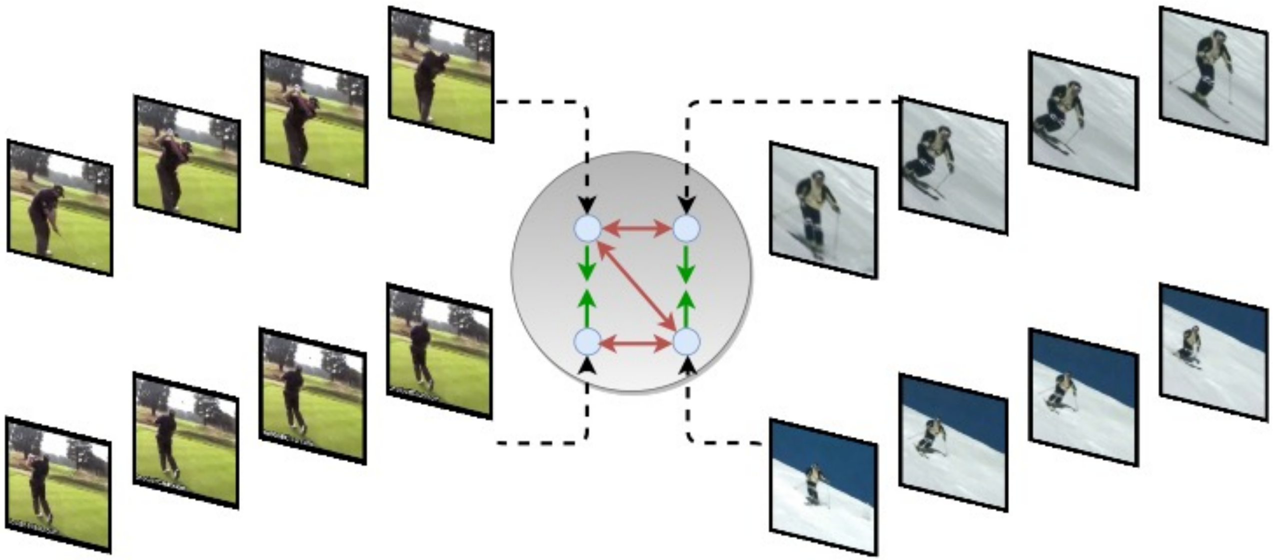
Shuffle and Learn, Mishra et. al., ECCV 2016



Video Clip Order Prediction, Xu et al., CVPR 2019

A more advanced proxy task: contrastive learning

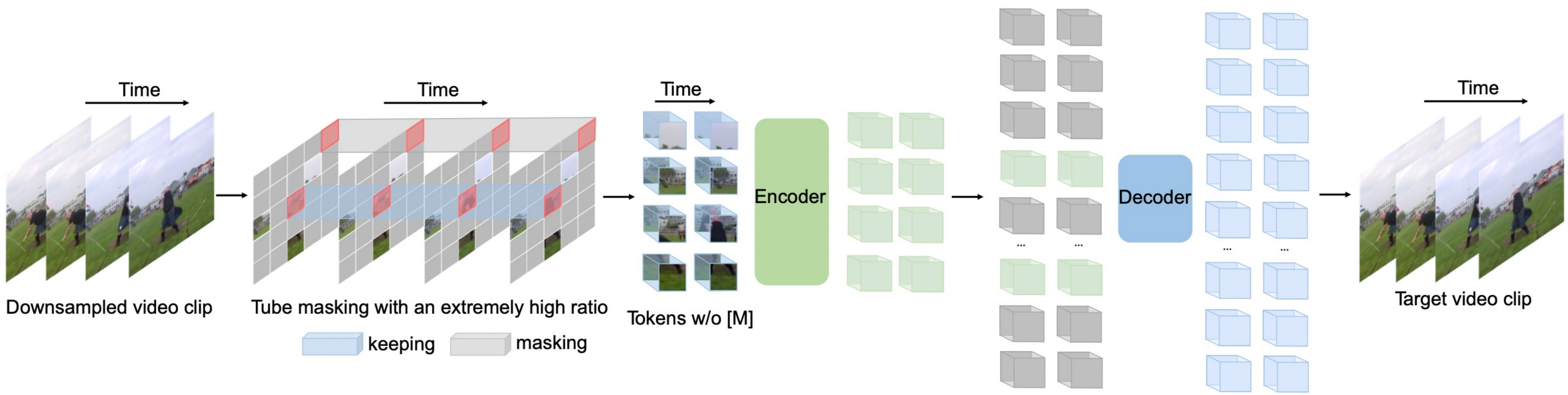
Uses Instance discrimination and enforces augmentation invariance.



Adaptation of image-based methods like MoCo, SimCLR, to video domain.

Masked auto encoding transformers

VideoMAE masks random cuboids and reconstructs the missing one



Zhan Tong, Yibing Song, Jue Wang, Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In NeurIPS, 2022.

This talk

Looks into the generalization abilities of modern video AI

1. The problem of video evaluation

2. The problem of video contrastive-learning

3. The problem of video masked auto encoding

1. The problem of video evaluation

Fida Mohammad Thoker
University of Amsterdam

Hazel Doughty
University of Amsterdam

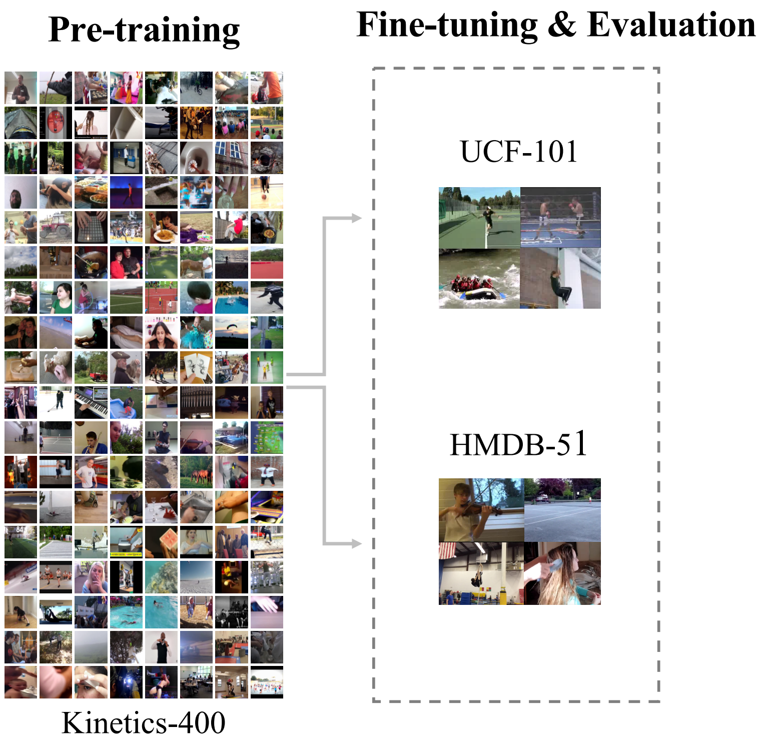
Piyush Bagad
University of Amsterdam

Cees Snoek
University of Amsterdam

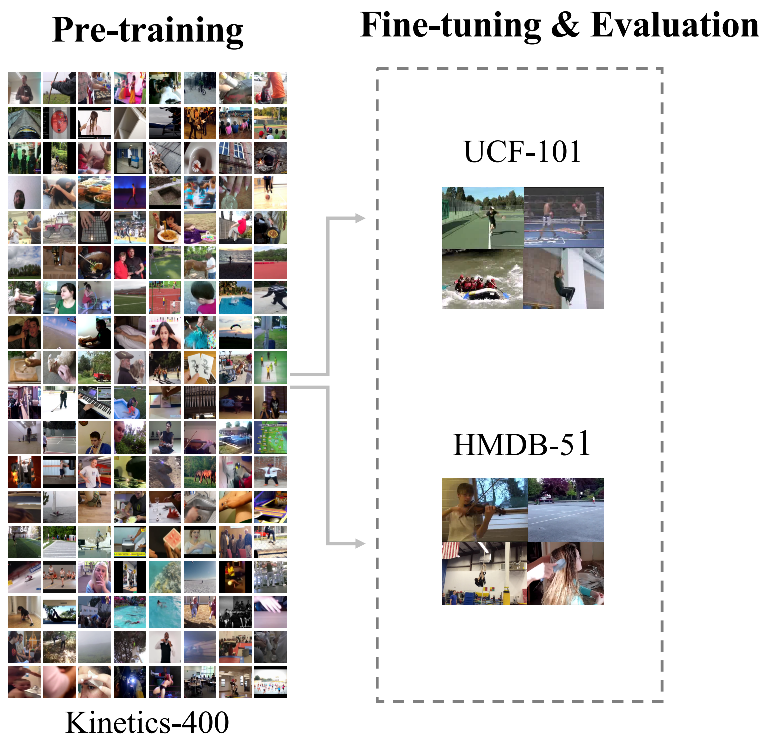
How Severe is Benchmark-Sensitivity in Video Self-Supervised Learning? In *ECCV* 2022.

Project Website

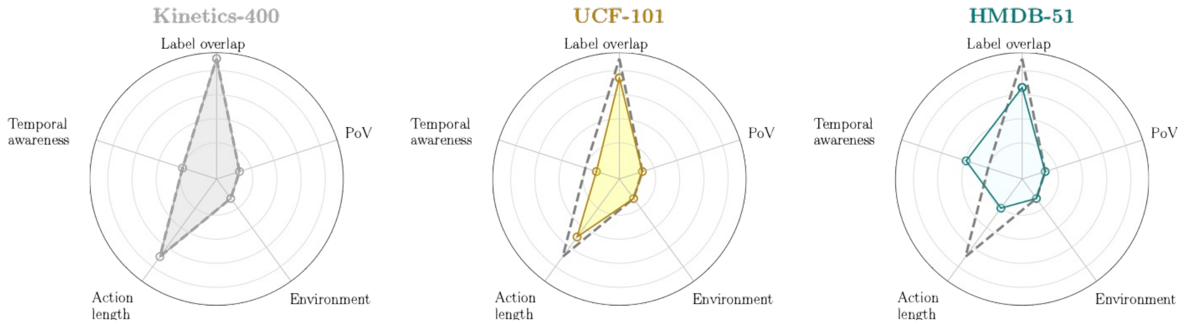
Problem: Video self-supervised learning evaluation



Problem: Video self-supervised learning evaluation



Pre-training and evaluation video too similar?

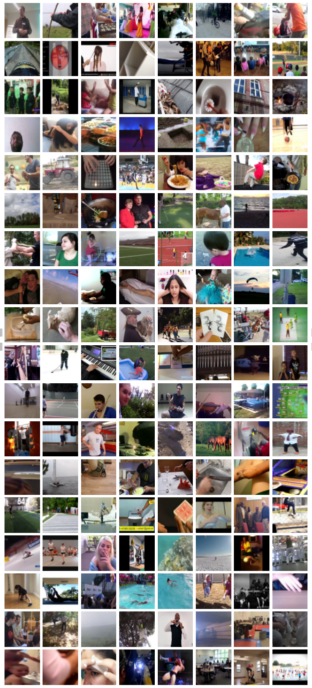


What if downstream video task is different?

Airport, shopping mall, hospital, *etc.*

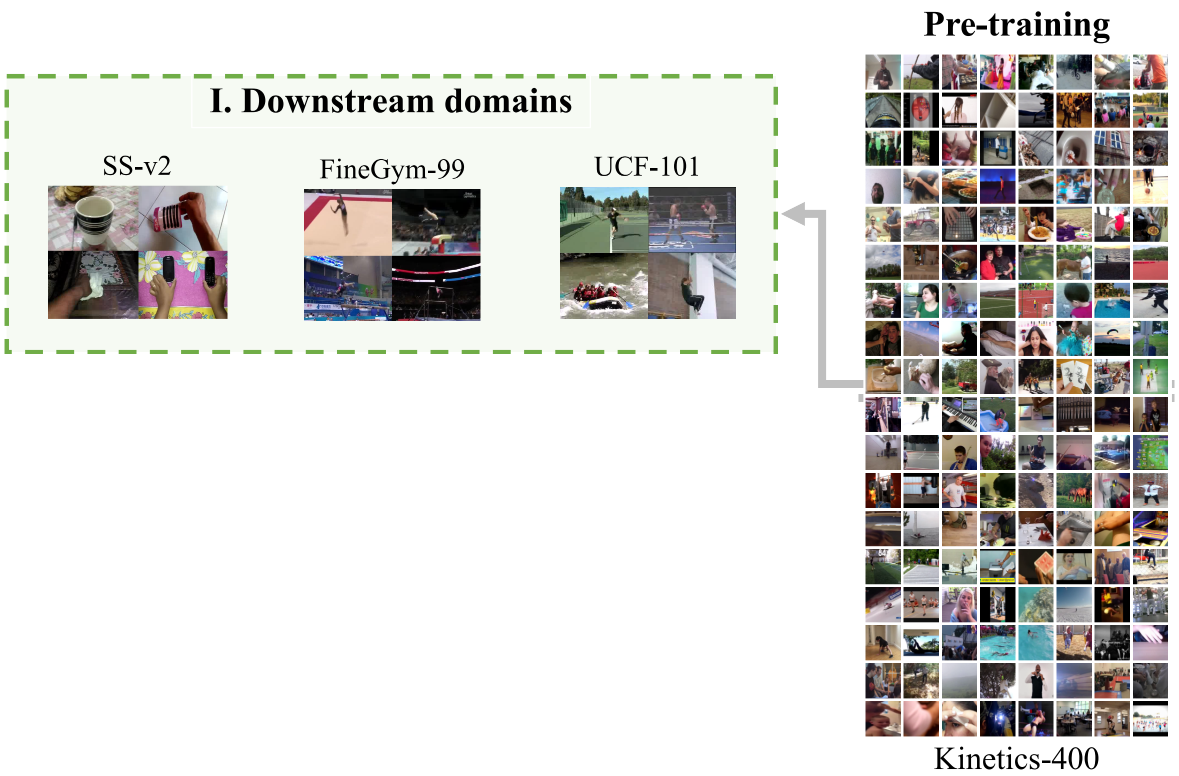
Proposed evaluation: four factors of sensitivity

Pre-training

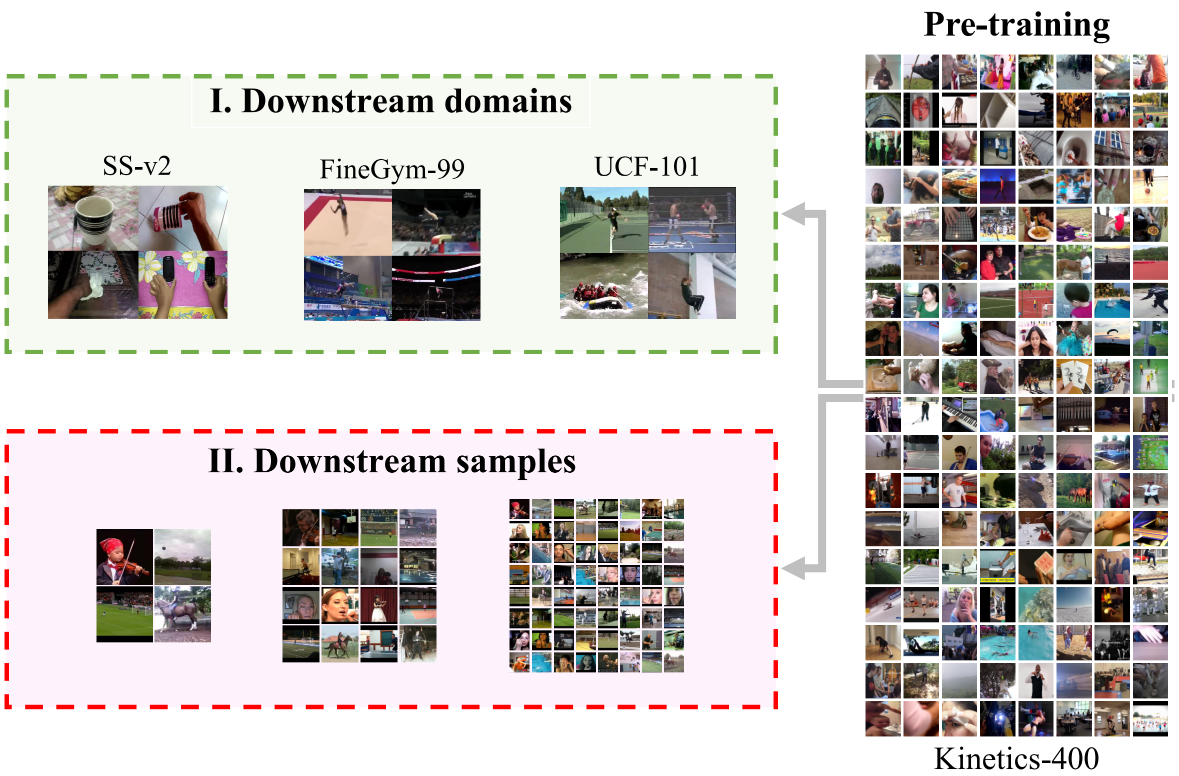


Kinetics-400

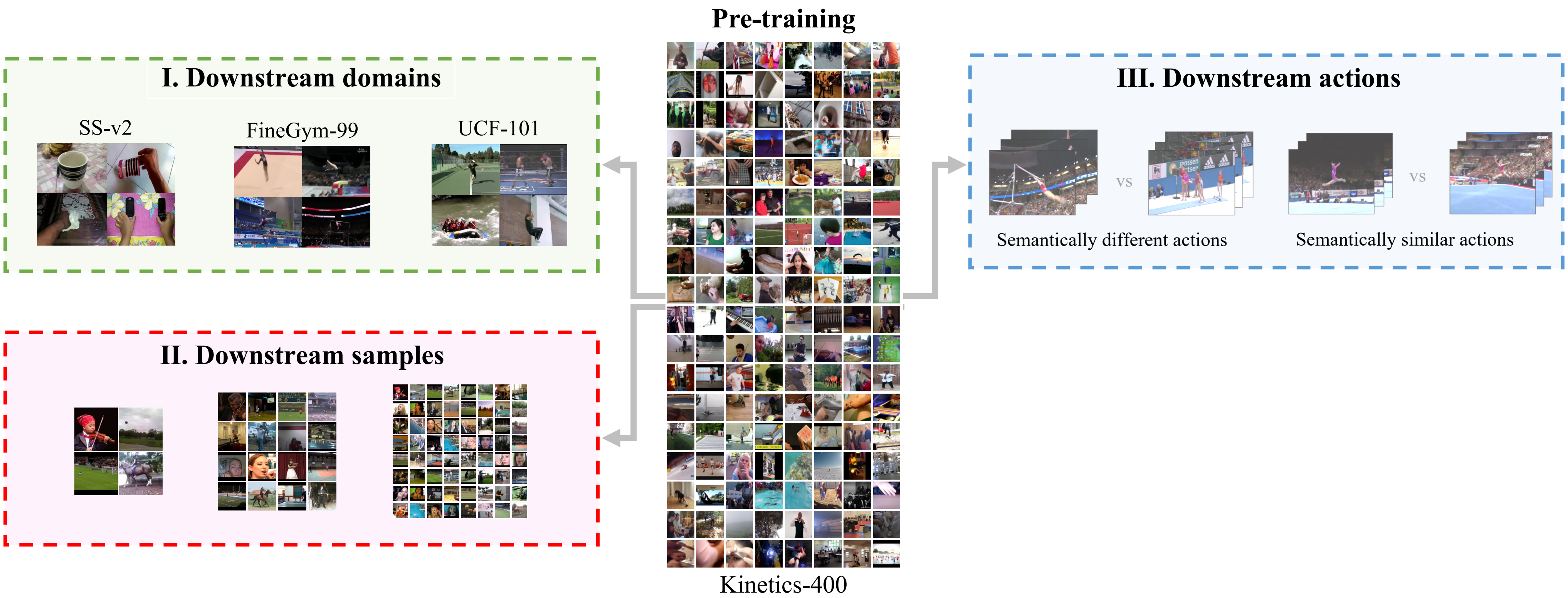
Proposed evaluation: four factors of sensitivity



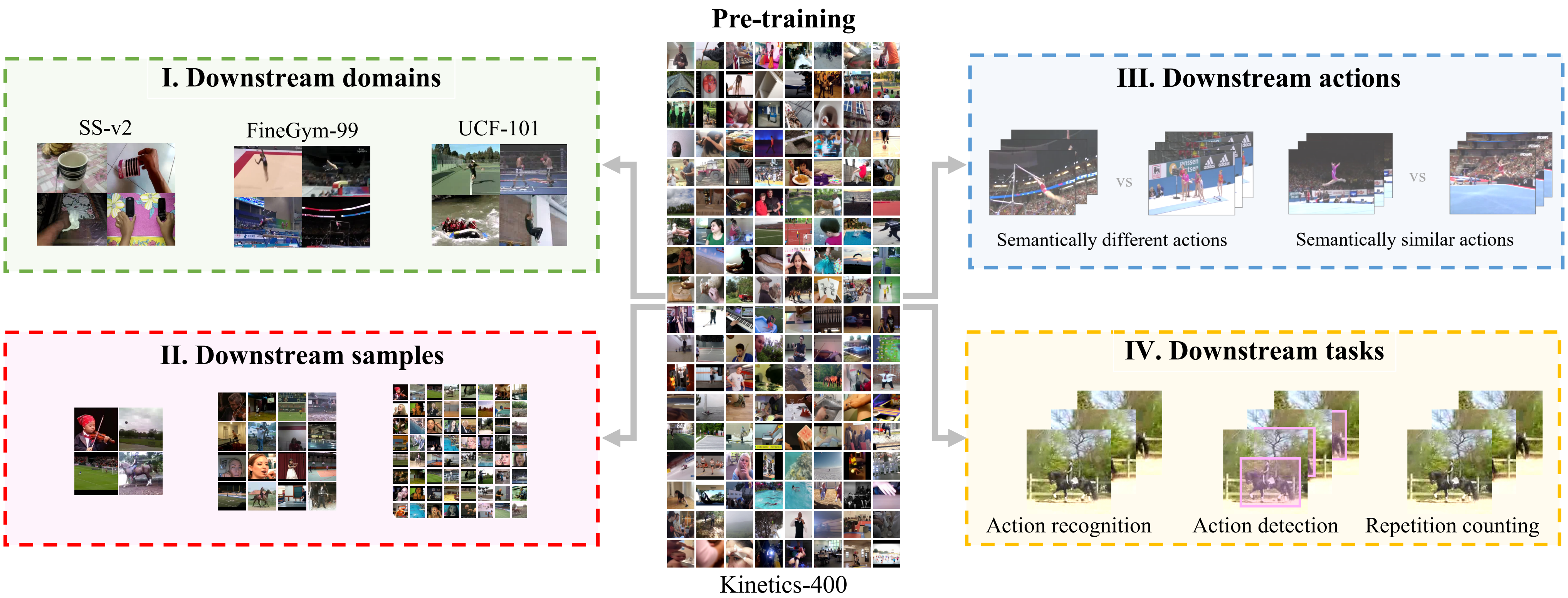
Proposed evaluation: four factors of sensitivity



Proposed evaluation: four factors of sensitivity

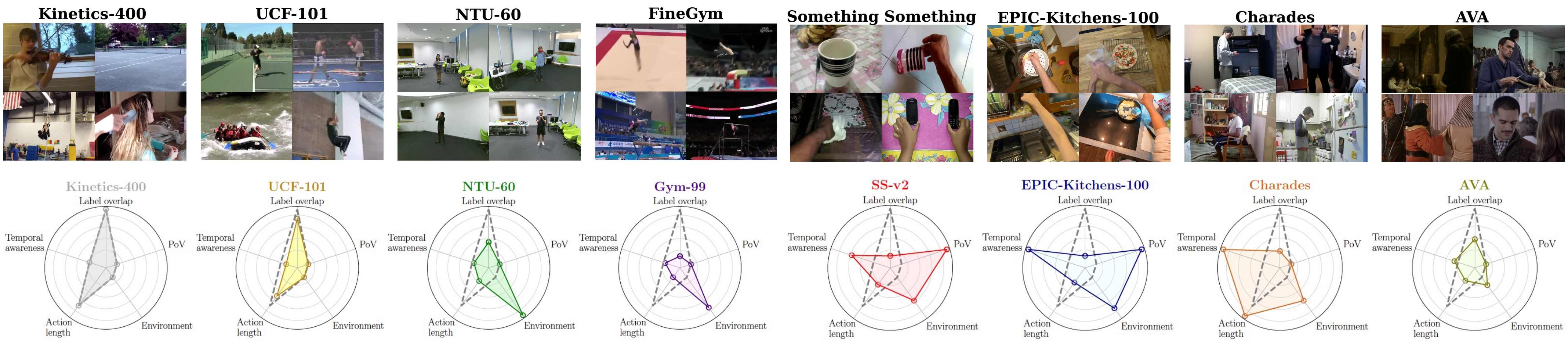


Proposed evaluation: four factors of sensitivity



7 datasets / 6 tasks / 500 experiments

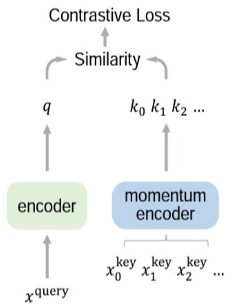
Considerable variety in video domain, the actions and tasks



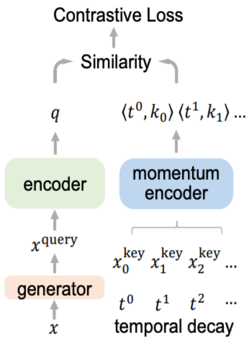
Tasks: *Action classification, Action detection, Repetition counting, Arrow of time prediction, Spatio-temporal detection, Multi-label classification*

9 video self-supervised learners

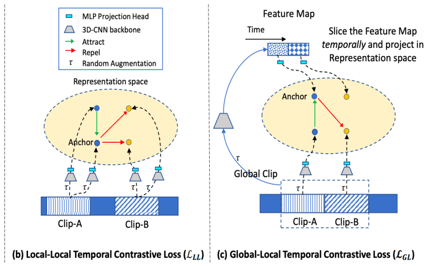
MoCo



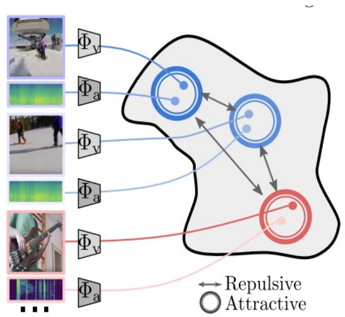
Video MoCo



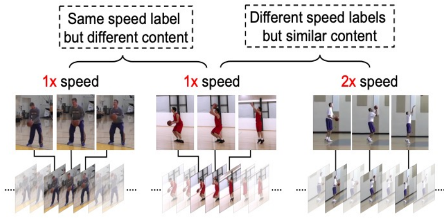
TCLR



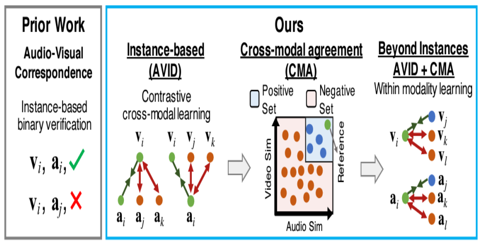
GDT



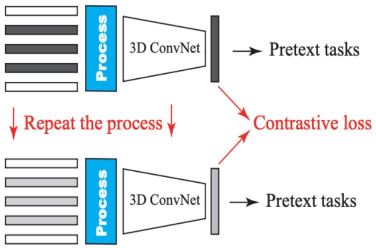
RSPnet



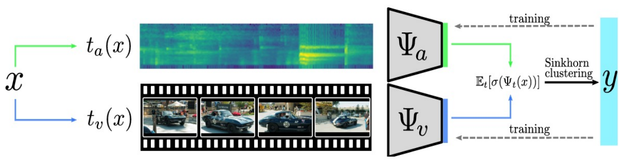
AVID-CMA



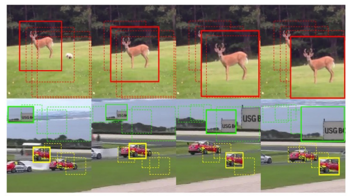
Pretext Contrast



SeLaVi

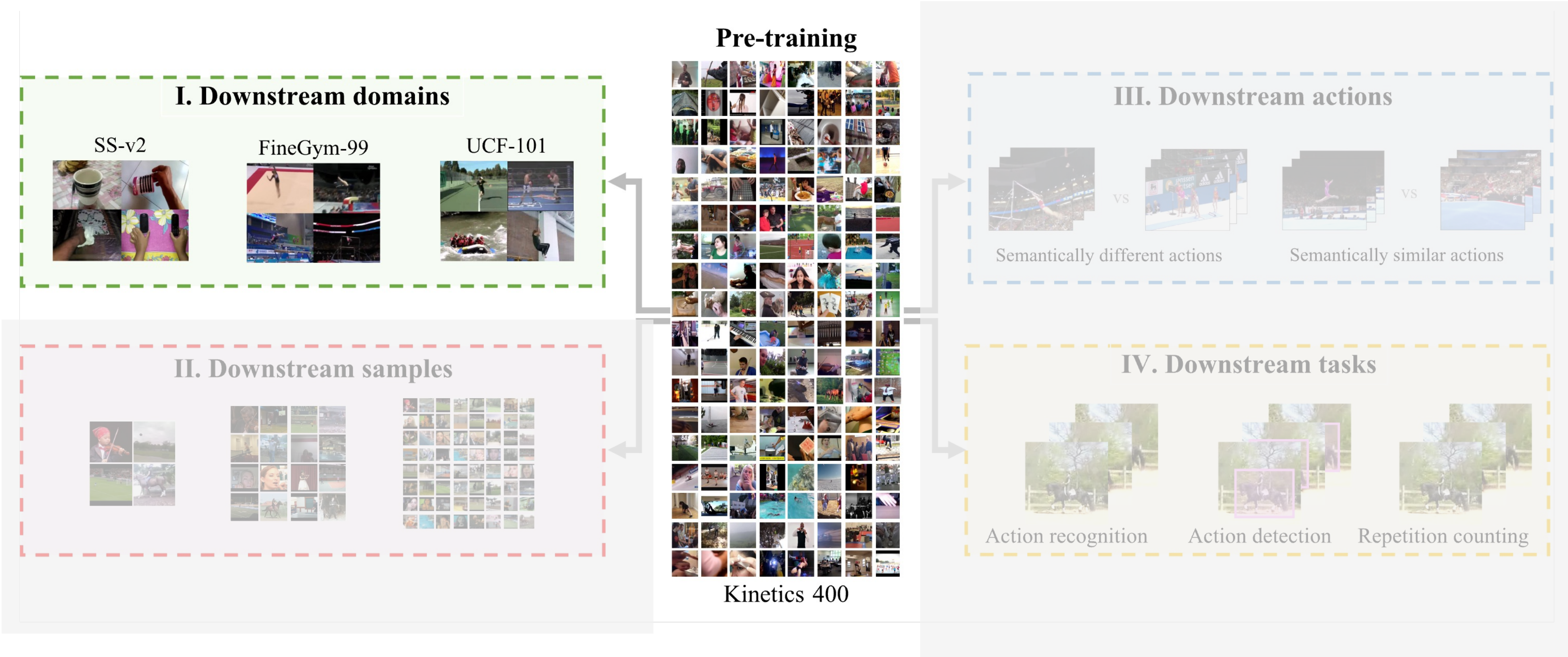


CtP



All methods come with weights for a R(2+1)D-18 network pre-trained on Kinetics-400

Sensitivity factor I: Downstream domain



Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift →

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	Finetuning				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift

Sensitivity factor I: Downstream domain

Downstream Domains

Pre-training	**Finetuning**				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	83.5	93.4	90.6	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	86.6	93.9	90.3	57.0	34.3
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	89.3	94.0	90.6	53.8	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift

Sensitivity factor I: Downstream domain

Downstream Domains

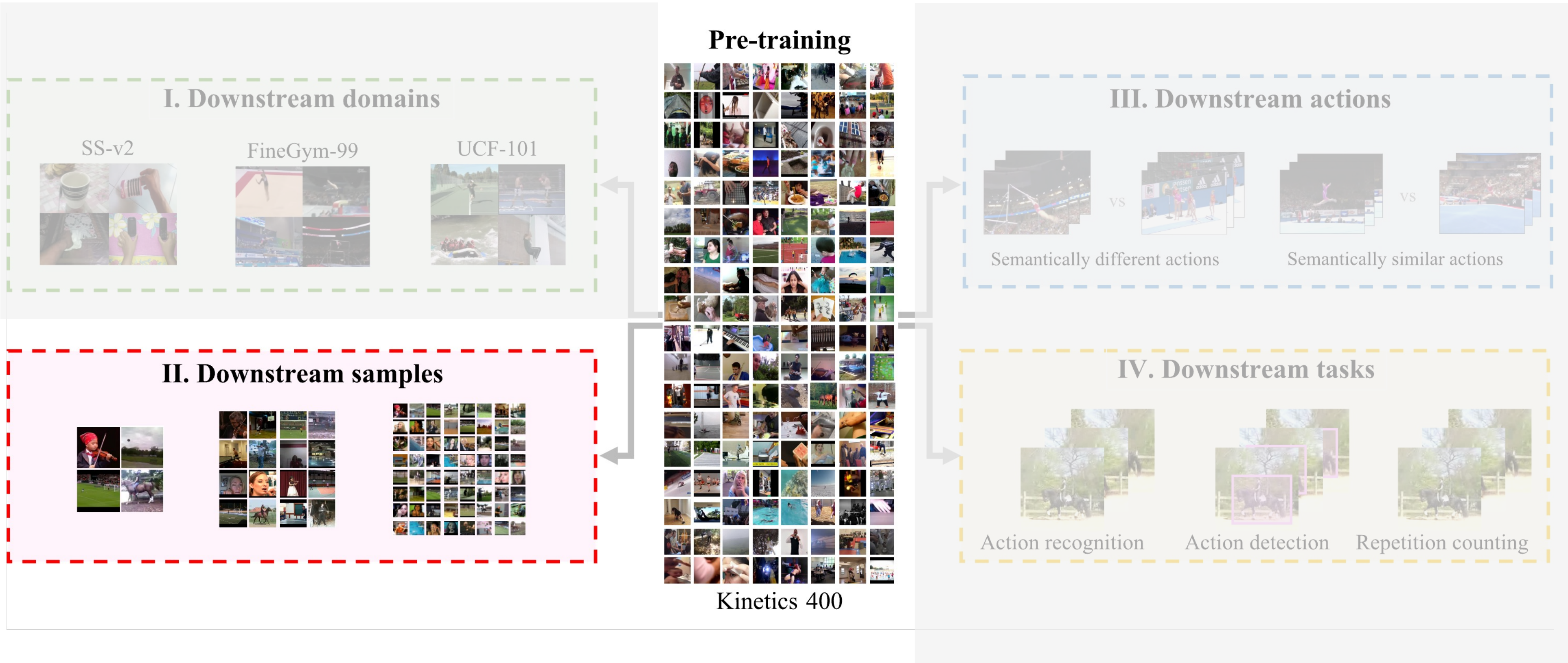
Pre-training	**Finetuning**				
	UCF101	NTU60	Gym99	SSv2	EK 100
None	75.4	92.9	89.4	56.8	25.7
MoCo	[illegible]	[illegible]	[illegible]	57.0	26.4
SeLaVi	84.9	92.8	88.9	56.4	33.8
VideoMoCo	85.8	94.1	90.5	58.8	43.6
Pretext-Contrast	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
RSPNet	88.5	93.9	91.3	59.4	42.7
AVID-CMA	[illegible]	[illegible]	[illegible]	[illegible]	29.9
CtP	89.8	94.3	92.2	60.2	42.8
TCLR	90.8	94.1	91.5	60.0	36.2
GDT	91.1	93.9	90.4	57.8	37.3
Supervised	94.1	93.9	91.8	61.0	47.7

Increasing domain shift

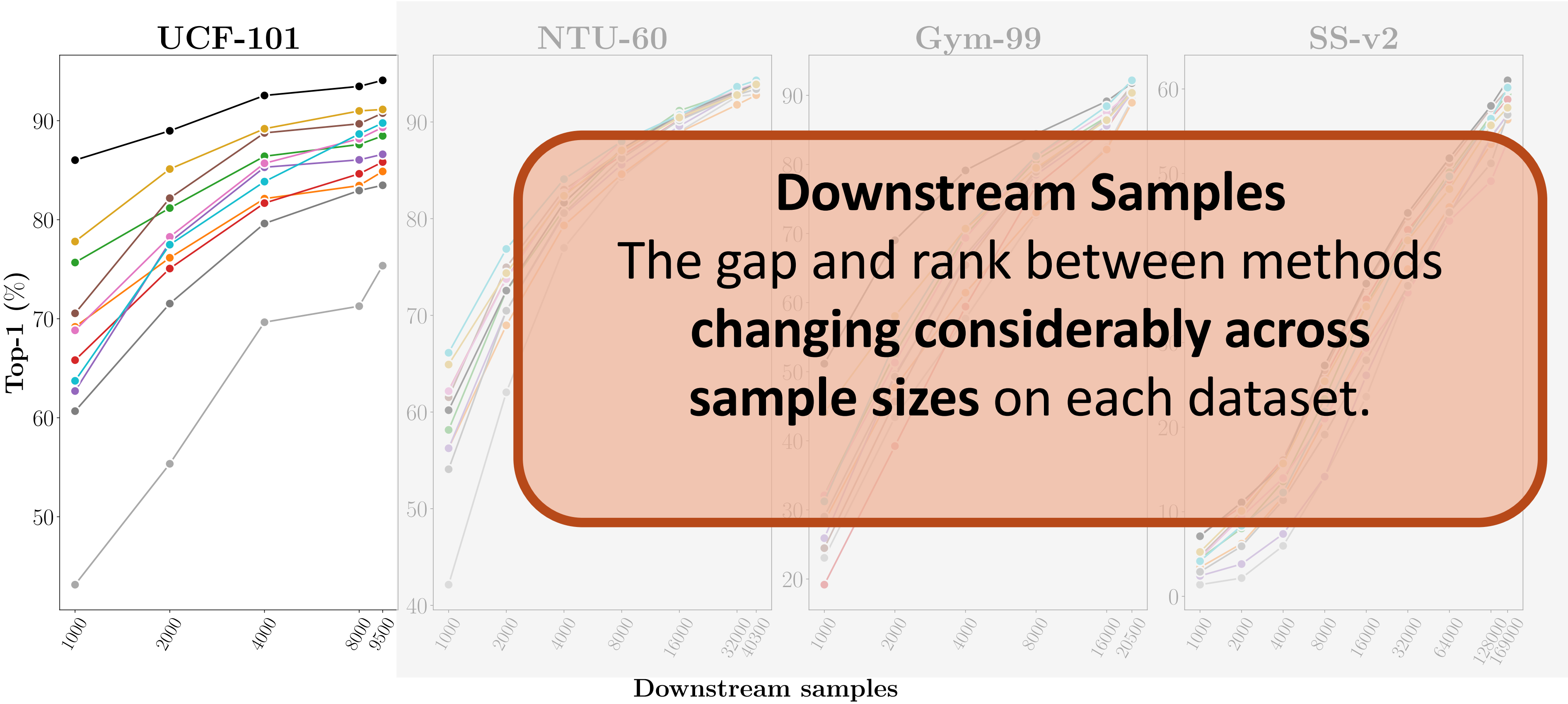
Downstream Domains

UCF-101 finetuning performance **does not generalize** to other target domains.

Sensitivity factor II: Downstream samples



Sensitivity factor II: Downstream samples



Downstream Samples
The gap and rank between methods **changing considerably across sample sizes** on each dataset.

Sensitivity factor III & IV: Downstream actions & tasks

Downstream Actions

Most self-supervised methods are **sensitive to action granularity** in downstream dataset.

Downstream Tasks

UCF-101 action classification performance is **mildly indicative** on other tasks.

Key takeaways

No clear winner, different methods standing out in different settings.

Contrastive methods encouraging **temporal distinctiveness** transfer well.

We select a subset of experiments as the **‘SEVERE’ benchmark**

SEVERE benchmark: subset of our experiments

Pre-training	Existing	SEVERE-benchmark							
		Domains		Samples		Actions		Tasks	
	UCF101	SS-v2	Gym-99	UCF (10^3)	Gym-99 (10^3)	FX-S1	UB-S1	UCF-RC	Charades-MLC
None	75.4	56.8	89.4	43.1	23.1	45.0	84.0	0.232	7.9
MoCo	83.5	57.0	90.6	60.7	29.0	65.1	85.0	0.220	8.1
SeLaVi	84.9	56.4	88.9	69.2	28.3	50.2	81.5	0.171	8.2
VideoMoCo	85.8	58.8	90.5	65.8	19.2	60.4	82.1	0.171	10.5
Pretext-Contrast	86.6	57.0	90.3	62.7	25.9	65.8	86.2	0.168	8.9
RSPNet	88.5	59.4	91.3	75.7	32.2	63.5	85.1	0.151	9.1
AVID-CMA	89.3	53.8	90.6	68.8	32.1	67.2	88.4	0.162	8.4
CtP	89.8	60.2	92.2	63.7	31.2	79.7	88.4	0.178	9.6
TCLR	90.8	60.0	91.5	70.6	24.5	61.0	85.3	0.149	11.1
GDT	91.1	57.8	90.4	77.8	44.1	65.7	81.6	0.137	8.5
Supervised	94.1	61.0	91.8	86.0	51.2	81.0	86.9	0.137	23.6

Enables future video self-supervised methods to evaluate generalization along 4 factors.

2. The problem of video-contrastive learning

Fida Mohammad Thoker
University of Amsterdam

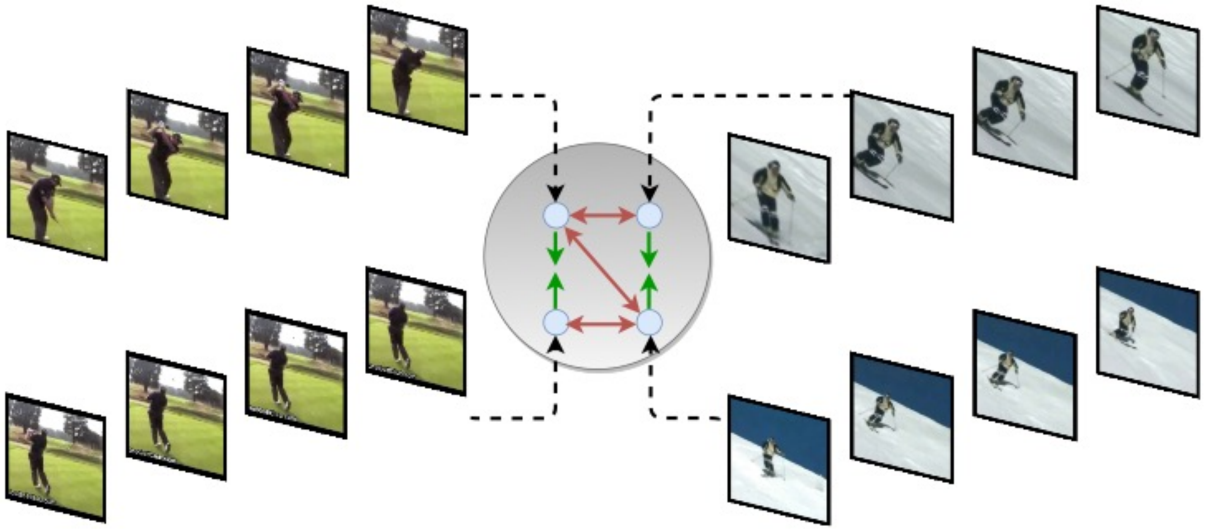
Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization. In *ICCV* 2023.

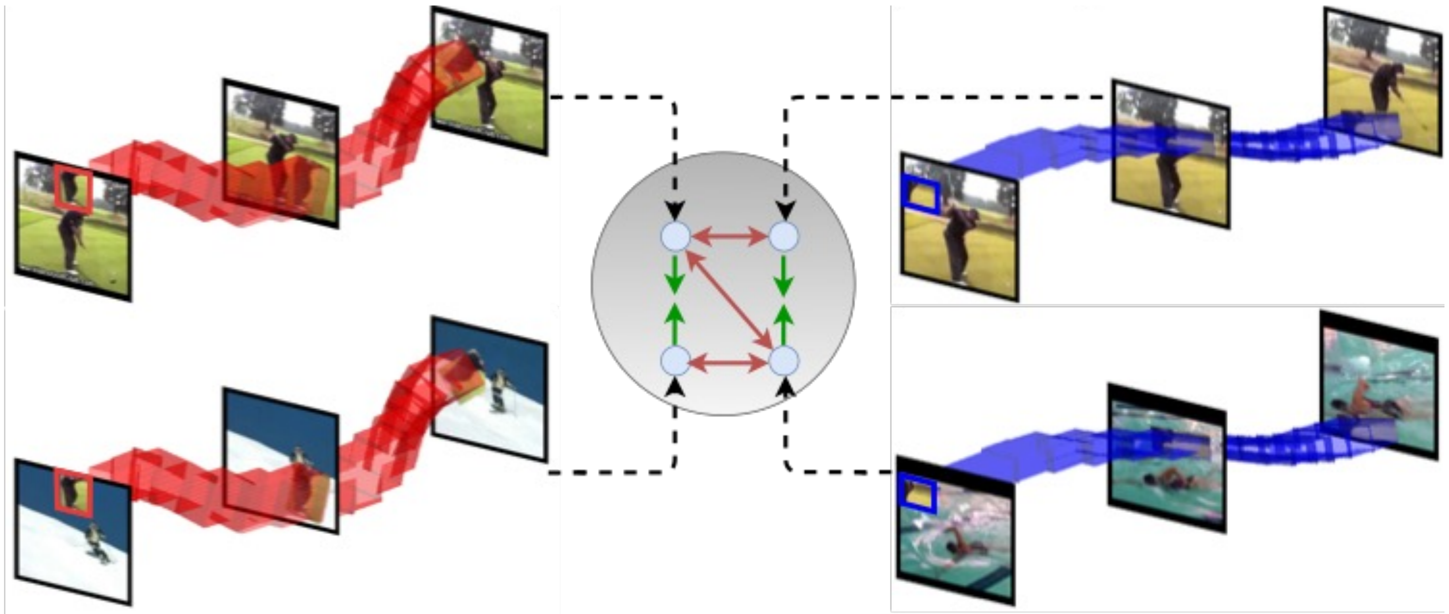
Problem of holistic contrastive learning

Uses Instance discrimination and enforces augmentation invariance.

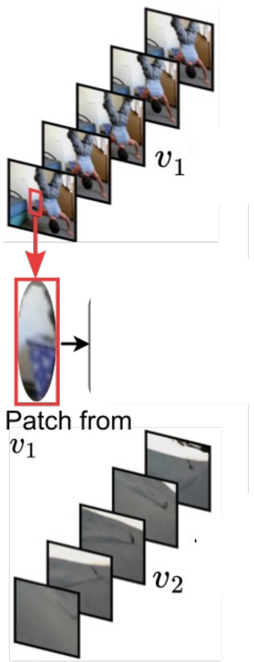


- 👎 Favours **coarse-grained** features
- 👎 Exploits background **shortcut**
- 👎 Limits **generalizability**
- 👎 Motion-variety constraints cause **data hunger**

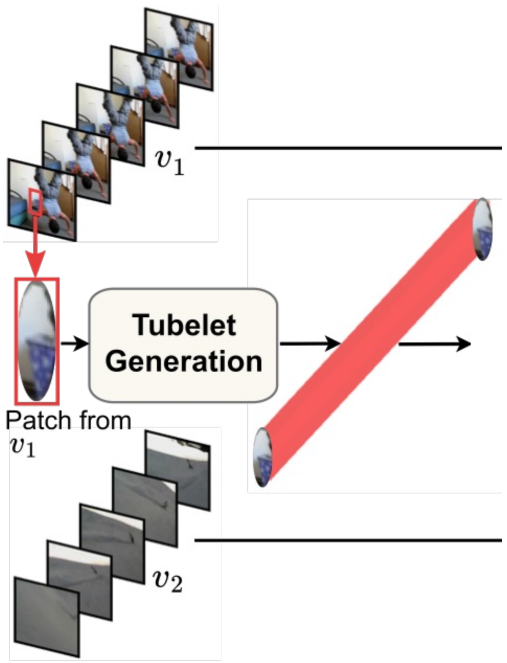
Solution: add synthetic tubelets during pretraining



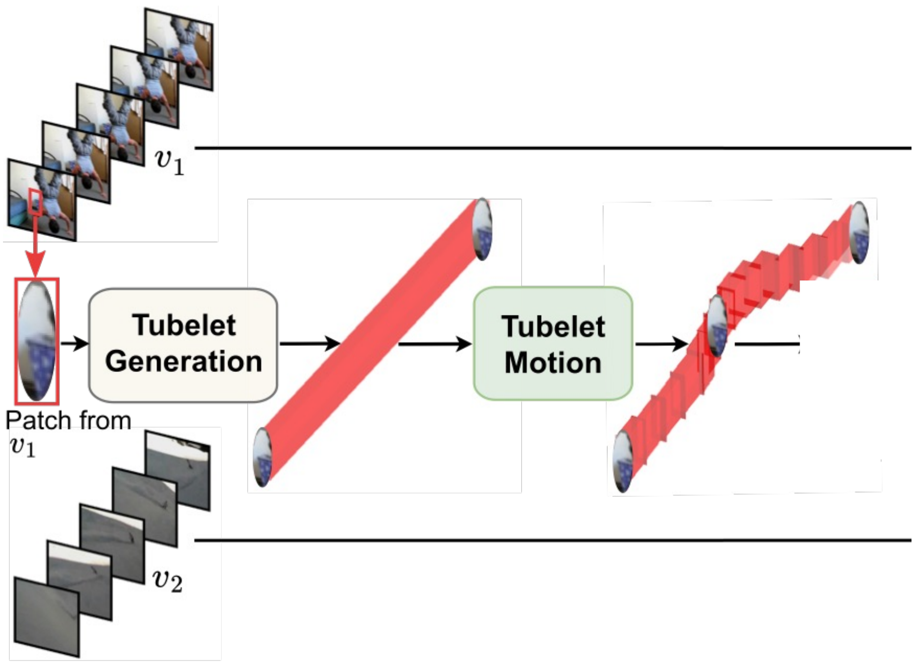
Step 0: Crop a random patch from one clip

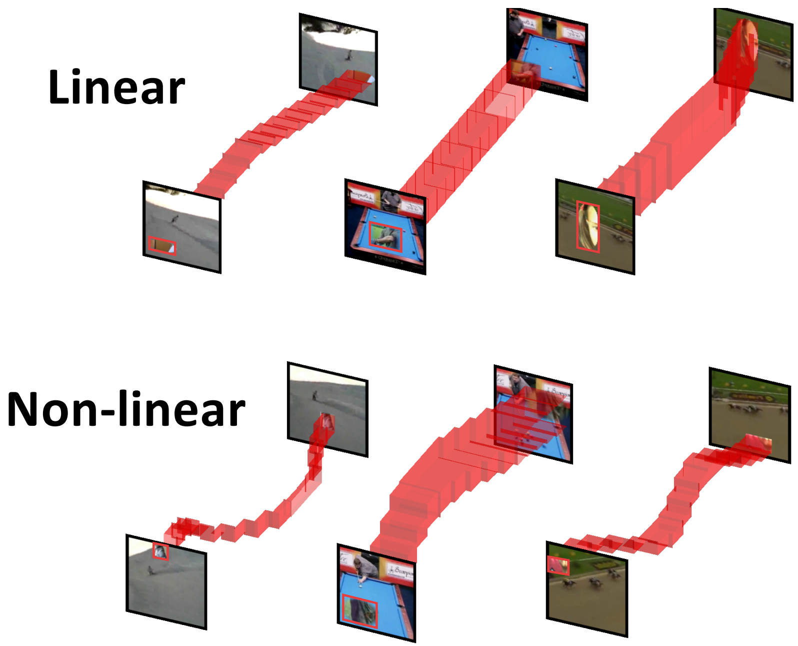


Step 1: Generate a tubelet

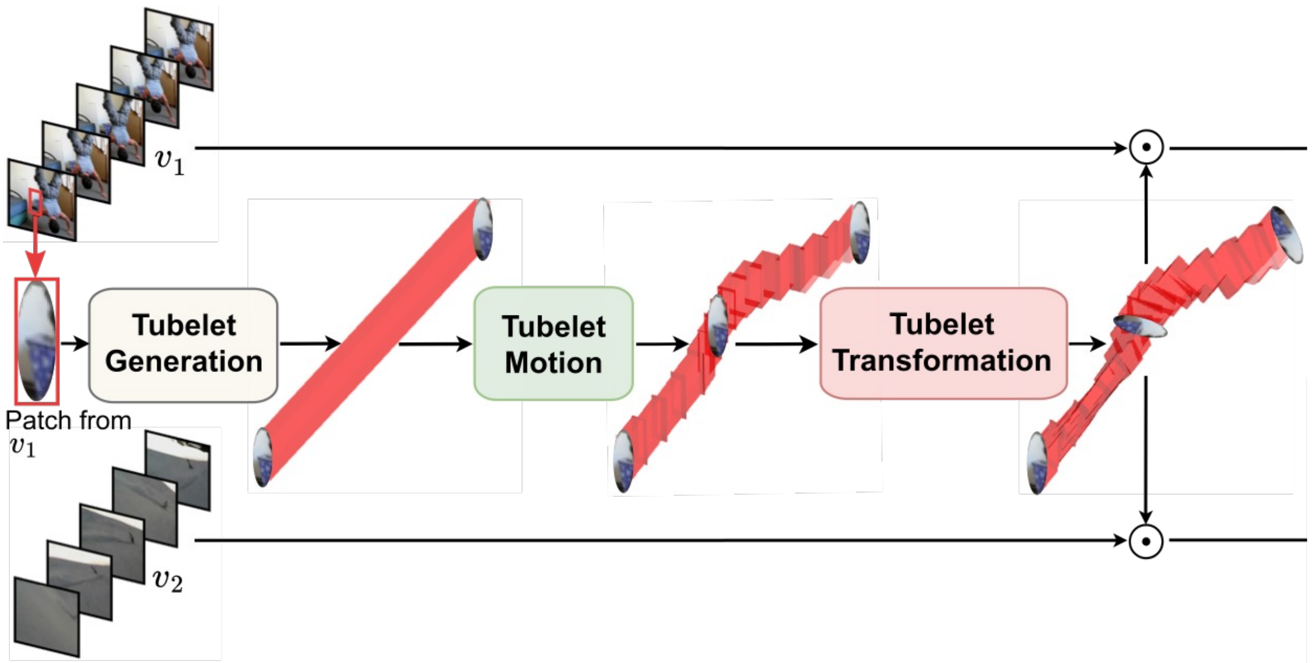


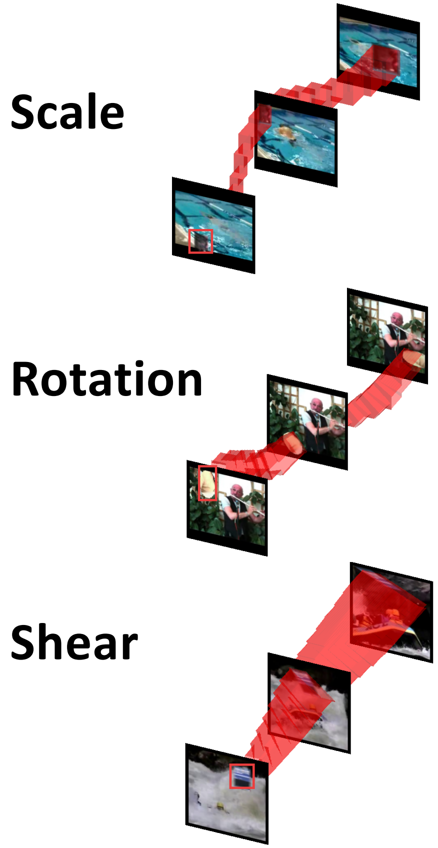
Step 2: Add motion to the patch



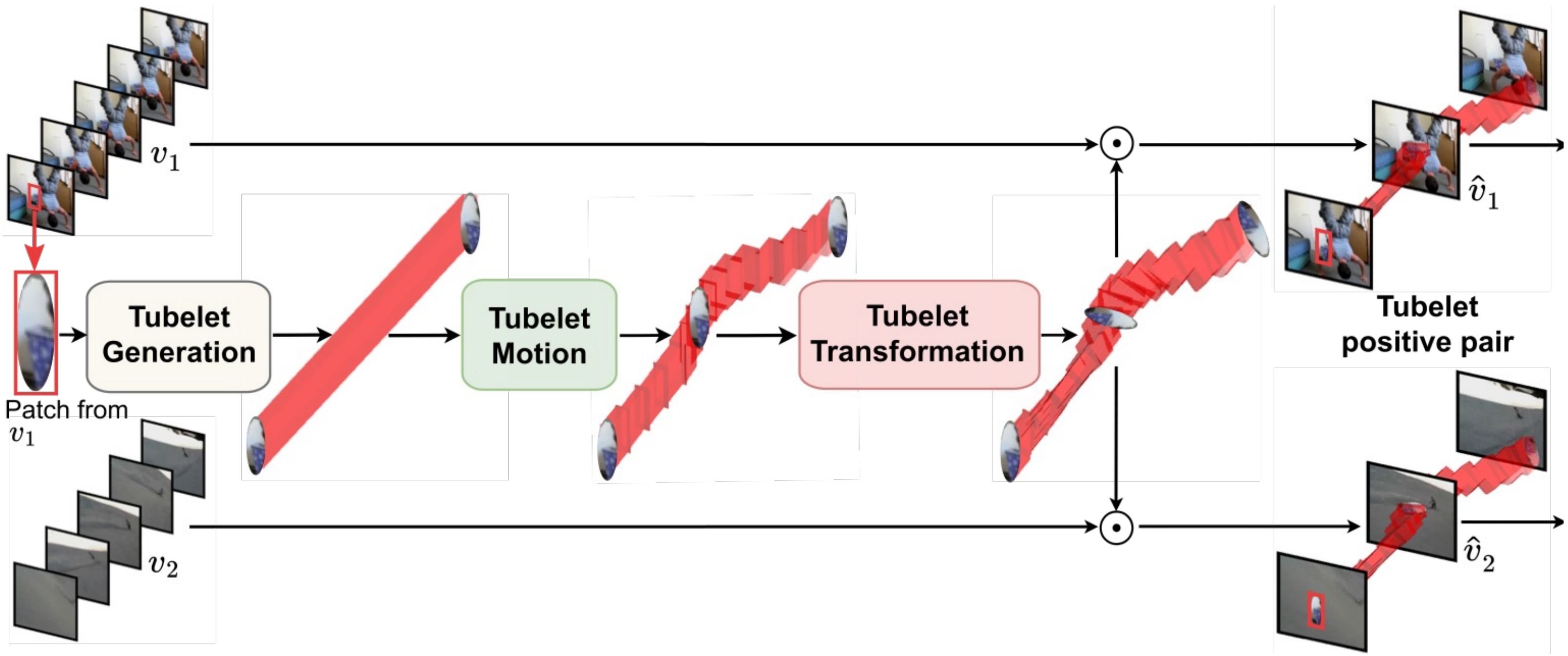


Step 3: Add motion complexity by transformations

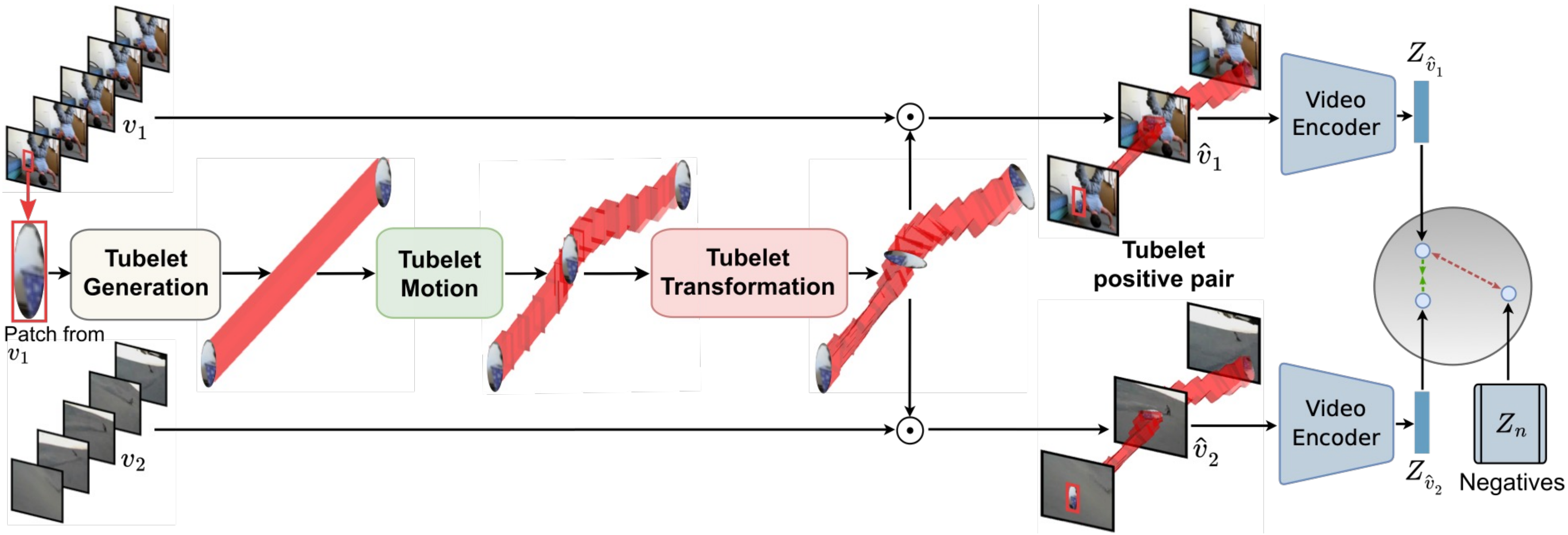




Step 4: Overlay identical tubelet on two clips



Step 5: Tubelet-contrastive learning



Ablations

Ablations confirm intuitions

	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
Video Contrast				
Baseline	57.5	29.5	44.2	84.8
Tubelet Contrast				
Tubelet Generation	48.2	28.2	40.1	84.1
Tubelet Motion	63.0	45.6	47.5	90.3
Tubelet Transformation	65.5	48.0	47.9	90.9

Table 2: **Tubelet-Contrastive Learning** considerably outperforms video contrast on multiple downstream settings. Tubelet motion and transformations are key.

Tubelet Motion	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
No motion	48.2	28.2	40.1	84.1
Linear	55.5	34.6	45.3	88.5
Non-Linear	63.0	45.6	47.5	90.3

Table 3: **Tubelet Motions.** Learning from tubelets with non-linear motion benefits multiple downstream settings.

Transformation	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
None	63.0	45.6	47.5	90.5
Scale	65.1	46.5	47.0	90.5
Shear	65.2	[illegible]	47.3	90.9
Rotation	65.5	48.0	47.9	90.9

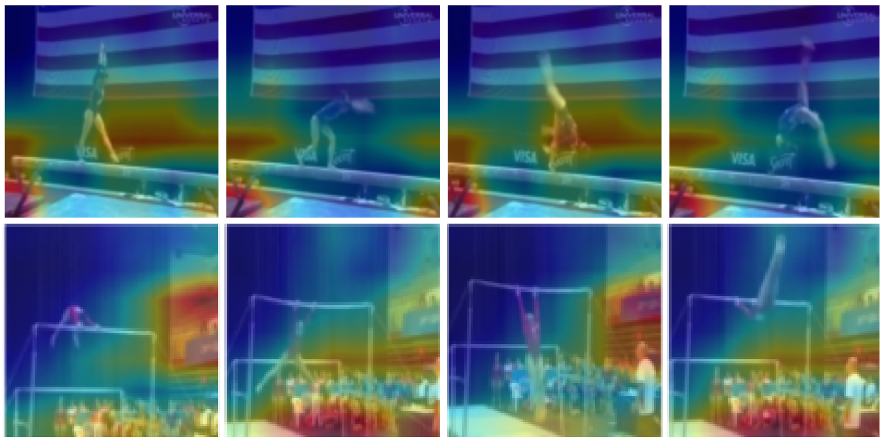
Table 4: **Tubelet Transformation.** Adding motion patterns to tubelet-contrastive learning through transformations improves downstream performance. Best results for rotation.

#Tubelets	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
1	62.0	39.5	47.1	89.5
2	65.5	48.0	47.9	90.9
3	66.5	46.0	47.5	90.9

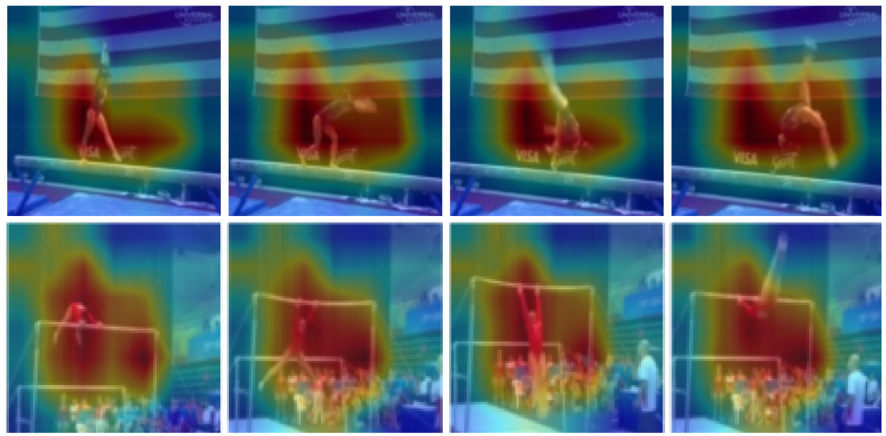
Table 5: **Number of Tubelets.** Overlaying two tubelets in positive pairs improves downstream performance.

What does the model learn?

Video-contrastive learning

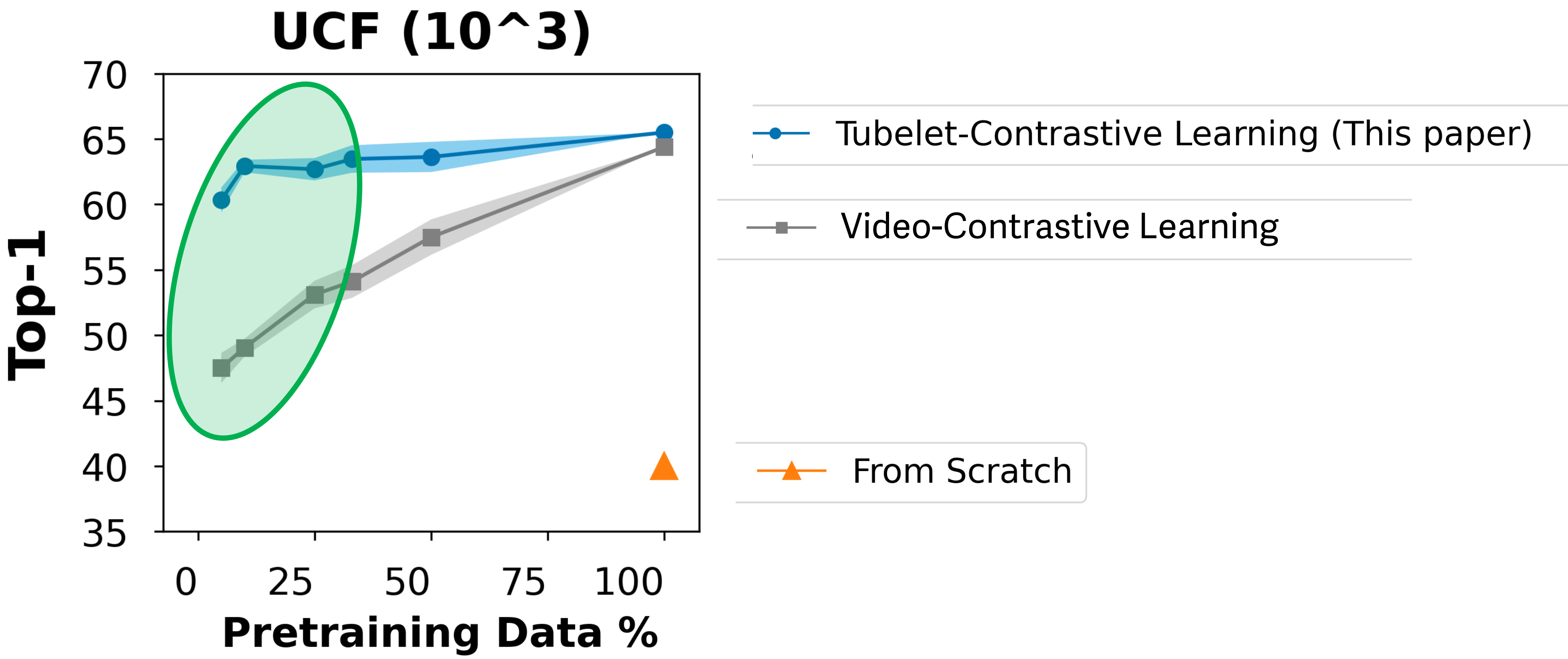


Proposed tubelet-contrastive learning

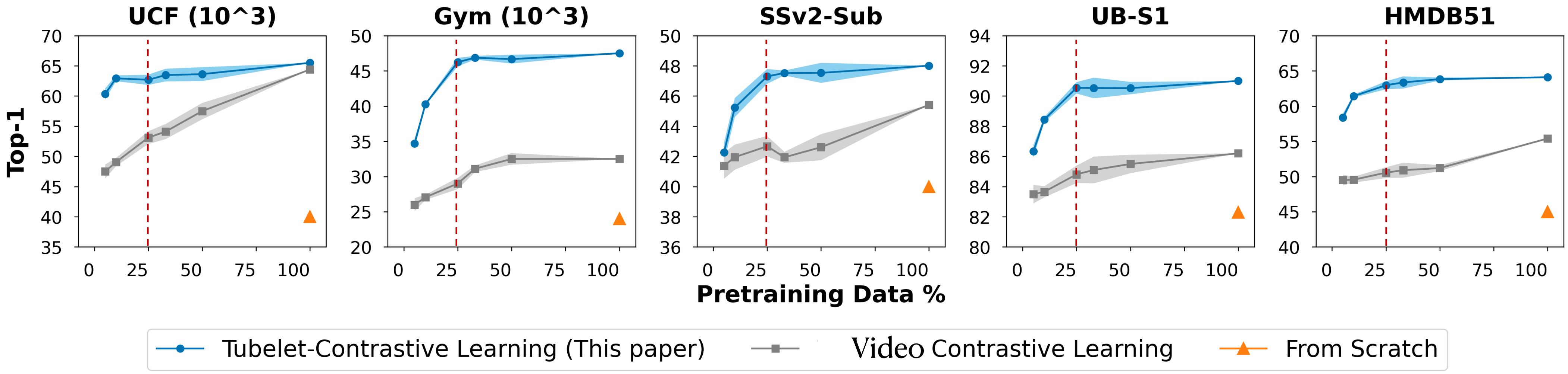


Without seeing any FineGym videos during training, our approach attends to motion

Adding synthetic motion improves data efficiency



Key benefit: we need 4x less video data

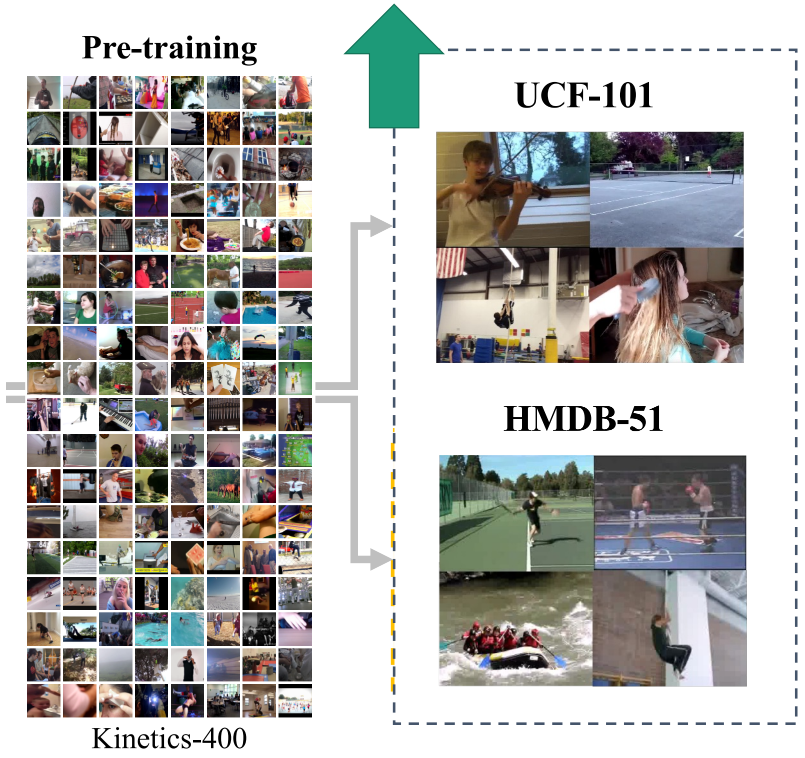


Tubelets simulate a richer variety of fine-grained motion than present in the original video

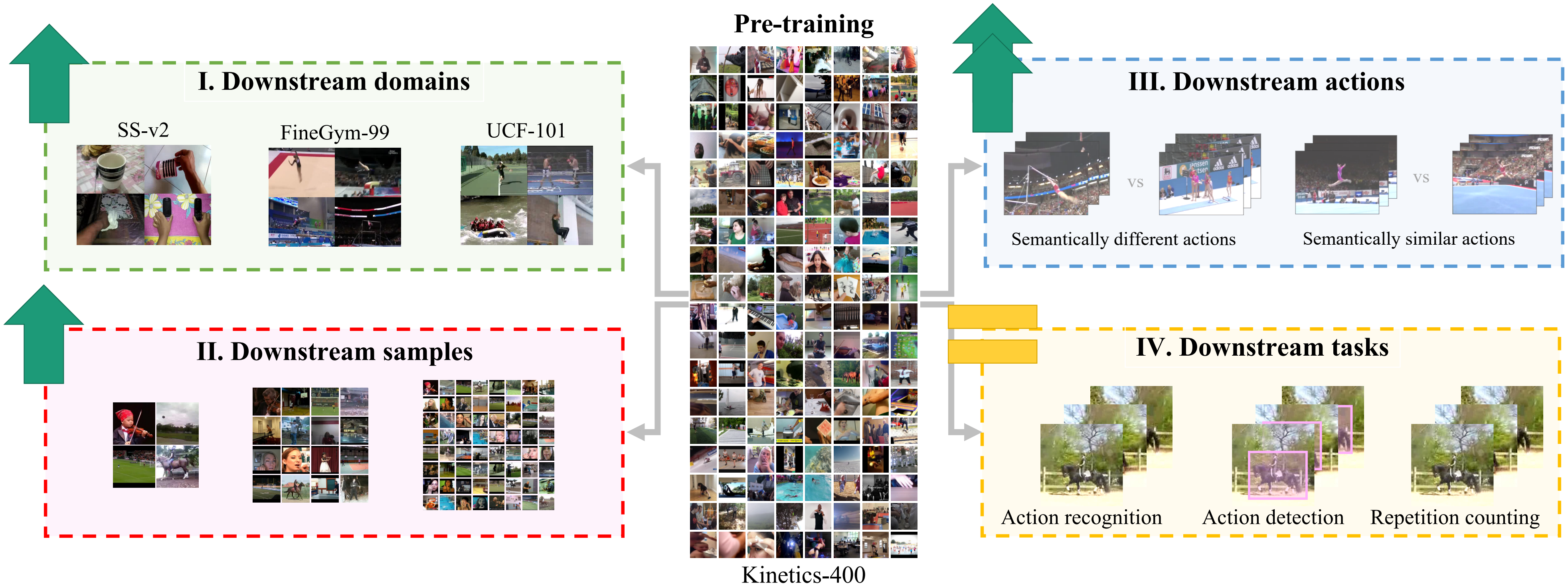
Solid accuracy gain on UCF-101 and HMDB-51

R(2+1)D Backbone pretrained on Kinetics-400

Method	Modality	UCF101	HMDB51
Pace Prediction [76]	RGB	77.1	36.6
VideoMoCo [56]	RGB	78.7	49.2
RSPNet [58]	RGB	81.1	44.6
SRTC [46]	RGB	82.0	51.2
FAME [10]	RGB	84.8	53.5
MCN [45]	RGB	84.8	54.5
AVID-CMA [52]	RGB+Audio	87.5	60.8
TCLR [9]	RGB	88.2	60.0
TE [31]	RGB	88.2	62.2
CtP [74]	RGB	88.4	61.7
MotionFit [20]	RGB+Flow	88.9	61.4
GDT [57]	RGB+Audio	89.3	60.0
Ours w/ mini-Kinetics	RGB	**90.7**	**65.0**
Ours w/ Kinetics	RGB	**91.0**	**64.1**



Generalization on SEVERE-benchmark



Generalization on SEVERE-benchmark

	Backbone	Domains		Samples		Actions		Tasks		Mean	Rank↓
		SSv2	Gym99	UCF (10^3)	Gym (10^3)	FX-S1	UB-S1	UCF-RC↓	Charades	**Mean**	**Rank↓**
SVT [61]	ViT-B	59.2	62.3	83.9	18.5	35.4	55.1	0.421	35.5	51.0	8.9
VideoMAE [71]	ViT-B	69.7	85.1	77.2	27.5	37.0	78.5	0.172	12.6	58.1	8.3
Supervised [72]	R(2+1)D-18	60.8	92.1	86.6	51.3	79.0	87.1	0.132	23.5	70.9	3.9
None	R(2+1)D-18	57.1	89.8	38.3	22.7	46.6	82.3	0.217	7.9	52.9	11.6
SeLaVi [2]	R(2+1)D-18	56.2	88.9	69.0	30.2	51.3	80.9	0.162	8.4	58.6	11.0
MoCo [23]	R(2+1)D-18	57.1	90.7	60.4	30.9	65.0	84.5	0.208	8.3	59.5	9.1
VideoMoCo [56]	R(2+1)D-18	59.0	90.3	65.4	20.6	57.3	83.9	0.185	10.5	58.6	9.1
Pre-Contrast [69]	R(2+1)D-18	56.9	90.5	64.6	27.5	66.1	86.1	0.164	8.9	60.5	9.0
AVID-CMA [51]	R(2+1)D-18	52.0	90.4	68.2	33.4	68.0	87.3	0.148	8.2	61.6	9.0
GDT [57]	R(2+1)D-18	58.0	90.5	**78.4**	45.6	66.0	83.4	**0.123**	8.5	64.8	8.6
RSPNet [58]	R(2+1)D-18	59.0	91.1	74.7	32.2	65.4	83.6	0.145	9.0	62.6	8.0
TCLR [8]	R(2+1)D-18	59.8	91.6	72.6	26.3	60.7	84.7	0.142	**12.2**	61.7	7.6
CtP [74]	R(2+1)D-18	59.6	92.0	61.0	32.9	79.1	88.8	0.178	9.6	63.2	5.6
Ours w/ mini-Kinetics	R(2+1)D-18	59.4	92.2	65.5	**48.0**	78.3	90.9	0.150	9.0	66.0	5.4
Ours w/ Kinetics	R(2+1)D-18	**60.2**	**92.8**	65.7	47.0	**80.1**	**91.0**	0.150	10.3	**66.5**	**4.1**

Better generalization, even when using the 3x smaller Mini-Kinetics for pretraining.

Key takeaways

Contrastive learning with **synthetic tubelets** provides:

Simple and effective self-supervised video representation learning.

Data-efficient pretraining with less unlabelled video data.

Better generalization to diverse video domains and fine-grained tasks.

3. The problem of video masked auto encoding

Fida Mohammad Thoker
University of Amsterdam

Michael Dorkenwald
University of Amsterdam

Fida Mohammad Thoker
KAUST

Efstratios Gavves
University of Amsterdam

Cees Snoek
University of Amsterdam

Yuki Asano
University of Amsterdam

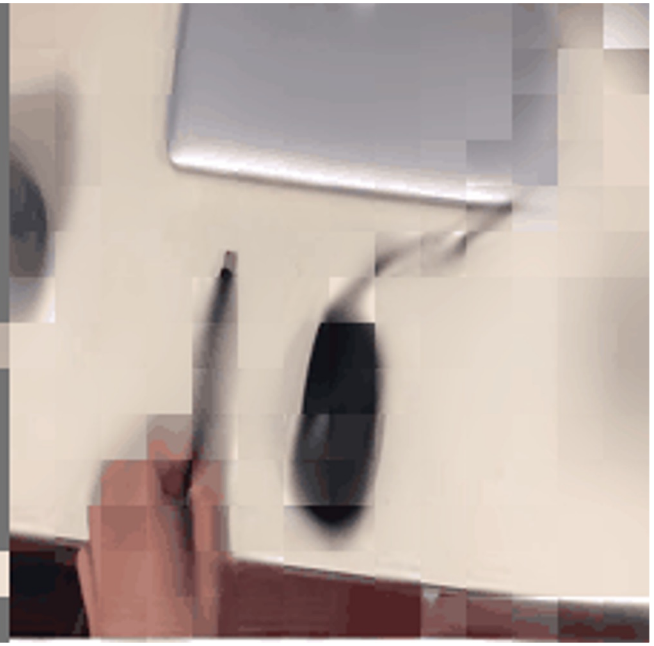
SIGMA: Sinkhorn-Guided Masked Video Modeling. In *ECCV* 2024.

QUVA Deep Vision Lab

Video MAE

Input video

Masked input (80%)



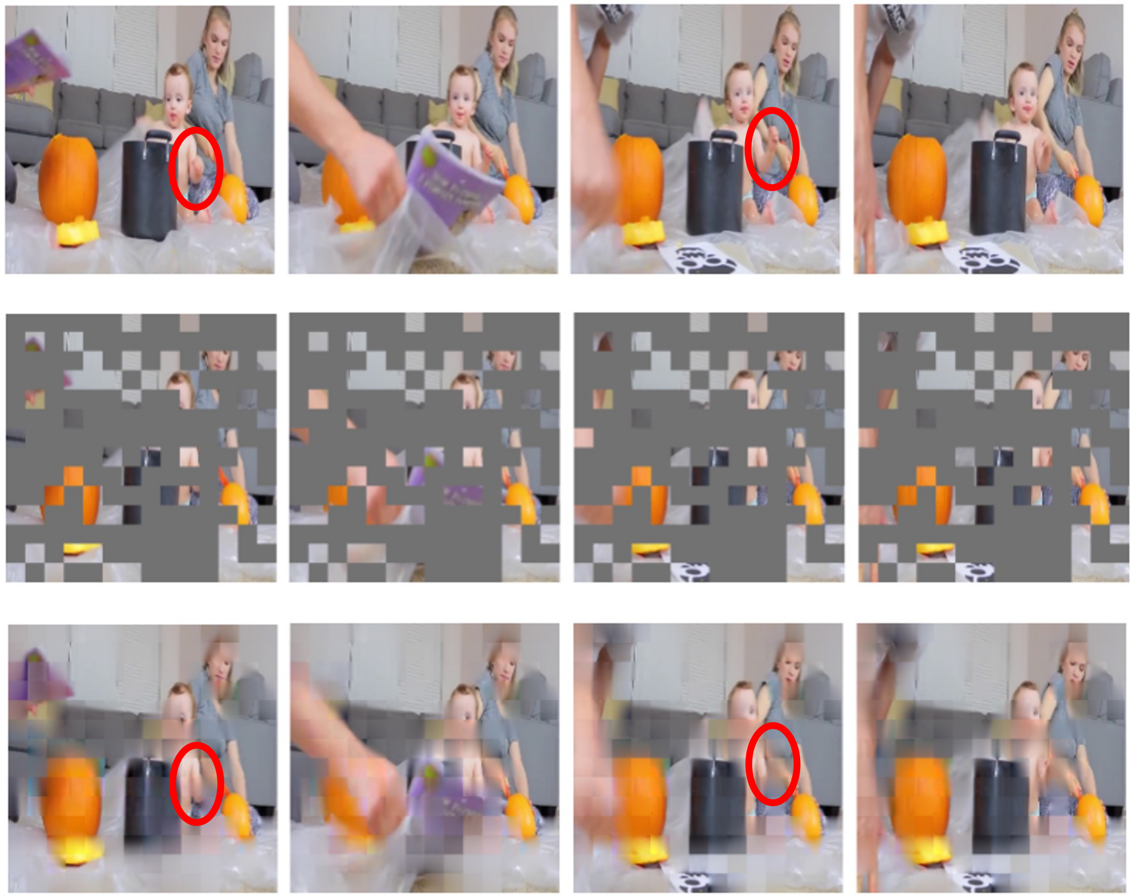
Reconstructed output video

Video MAE Challenge: Poor motion modeling

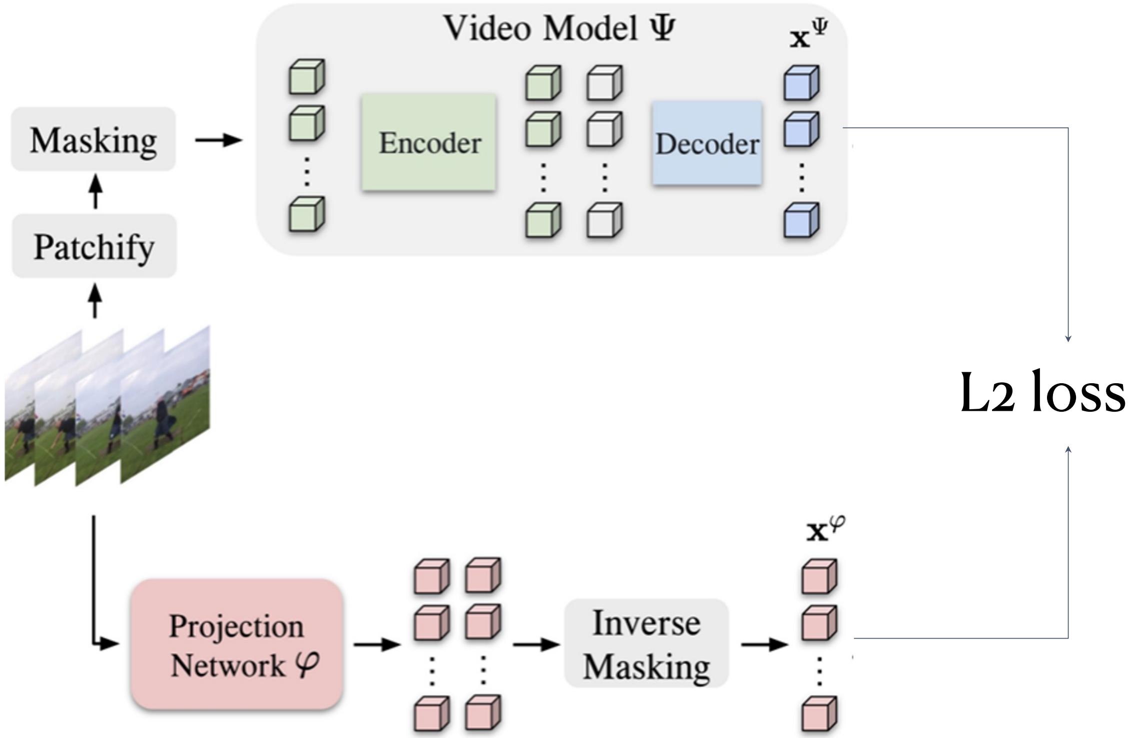
Input video

Masked input

Reconstructed video



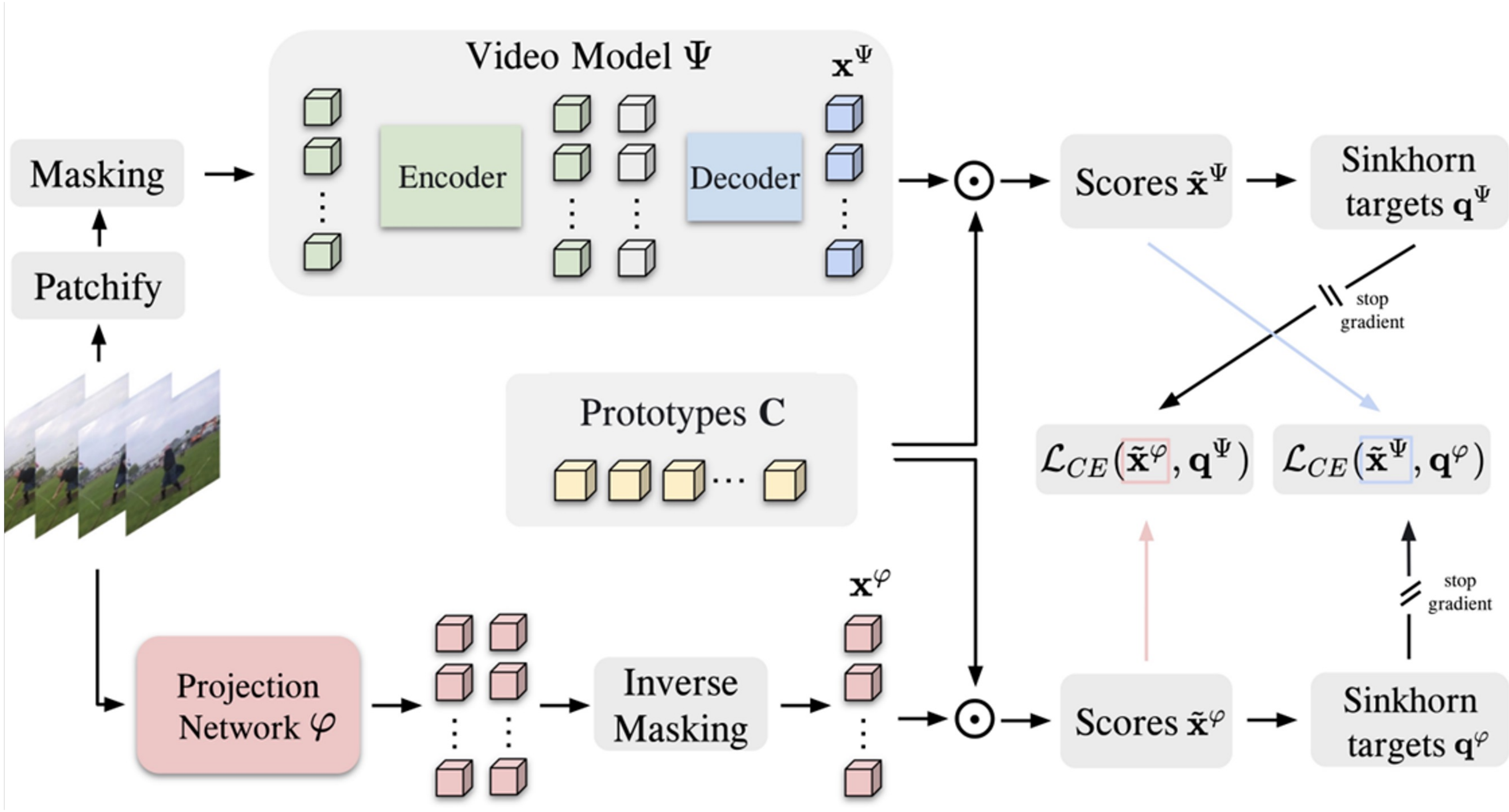
From pixel to feature reconstruction



Trivial Solution! Mode Collapse.

Online clustering methods can help

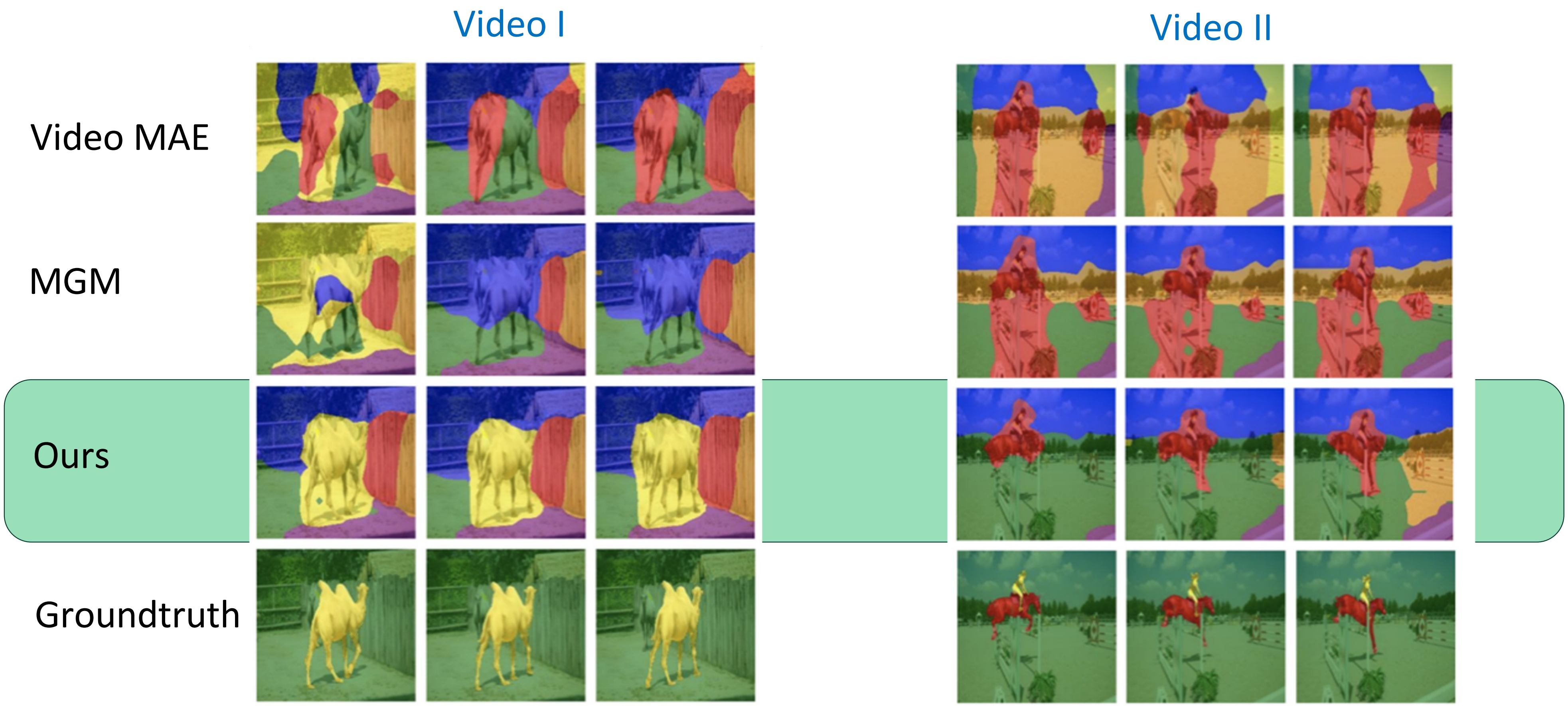
SIGMA: Sinkhorn-Guided Masked Video Modeling



Generalization on SEVERE-benchmark

	Domains	Samples (10^3)		Actions		Tasks		Mean
	Gym99	UCF	Gym	FX-S1	UB-S1	UCF-RC↓	Charades	
SVT	62.3	<u>83.9</u>	18.5	35.4	55.1	0.421	**35.5**	49.8
MVD	79.1	70.2	25.5	35.0	71.5	0.184	16.1	54.2
VideoMAE	85.1	77.2	27.5	37.0	78.5	<u>0.172</u>	12.6	57.3
MGM	86.5	75.1	27.0	41.0	84.4	0.181	17.9	59.1
SIGMA-MLP (ours)	<u>88.6</u>	81.2	<u>33.6</u>	<u>51.0</u>	<u>85.2</u>	0.178	20.1	<u>63.1</u>
SIGMA-DINO (ours)	**90.3**	**86.0**	**35.0**	**64.8**	**87.5**	**0.169**	<u>23.3</u>	**67.1**

Unsupervised video object segmentation on DAVIS



Key takeaways

Sinkhorn-clustering leads to more abstract mask reconstruction

Alleviates **training collapse,** profits from **pretrained image models**

Better generalization to video domains, samples and fine-grained actions.

Concluding encouragement

Learning to generalize in video space and time, and across modalities and tasks, is an **open research challenge**.

First ideas have started to appear, **much more research is needed**.

VIS LAB
VIDEO & IMAGE SENSE LAB

Prof. dr. Cees Snoek

https://ivi.fnwi.uva.nl/vislab/

@cgmsnoek {x, bsky.social}