UNIVERSITY OF AMSTERDAM
ellis | UNIT AMSTERDAM

Prof. dr. Cees Snoek
University of Amsterdam

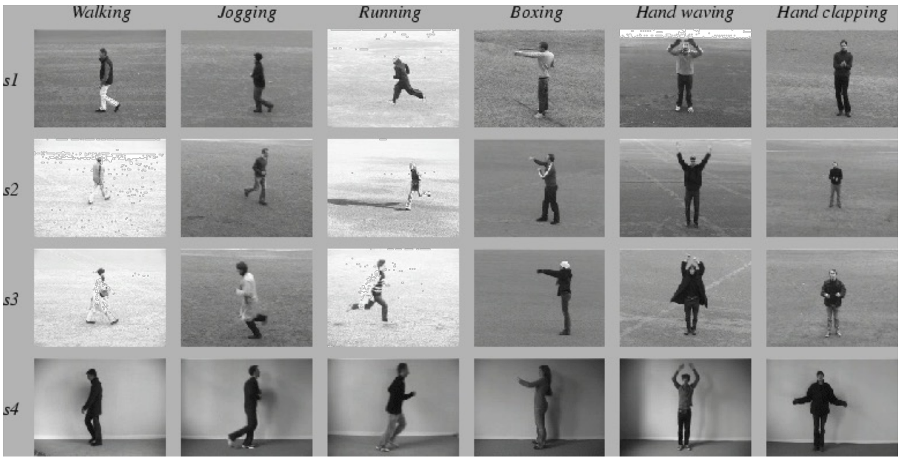
Head of VIS lab
Director QUVA ICAI-lab
Director Atlas ICAI-lab
Director AIM ICAI-lab
CSO, Kepler Vision
Director ELLIS Unit Amsterdam
Scientific Director Amsterdam AI

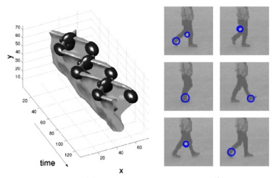
Learning to Generalize in Video Space and Time

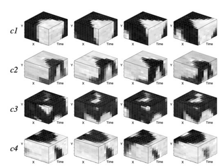
VIS LAB
VIDEO & IMAGE SENSE LAB

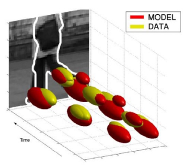
1

How it started…









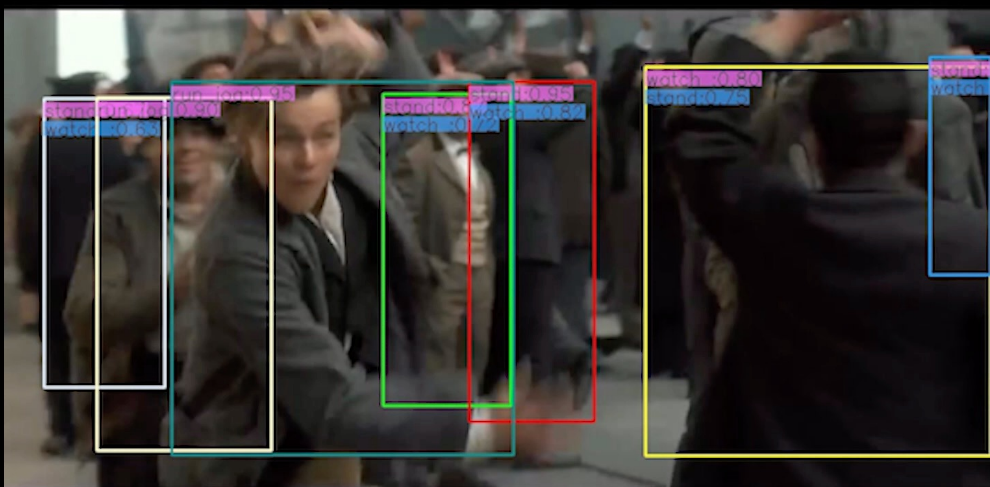
Laptev & Lindeberg, ICCV 2003

2

Du Tran et al., ICCV 2015

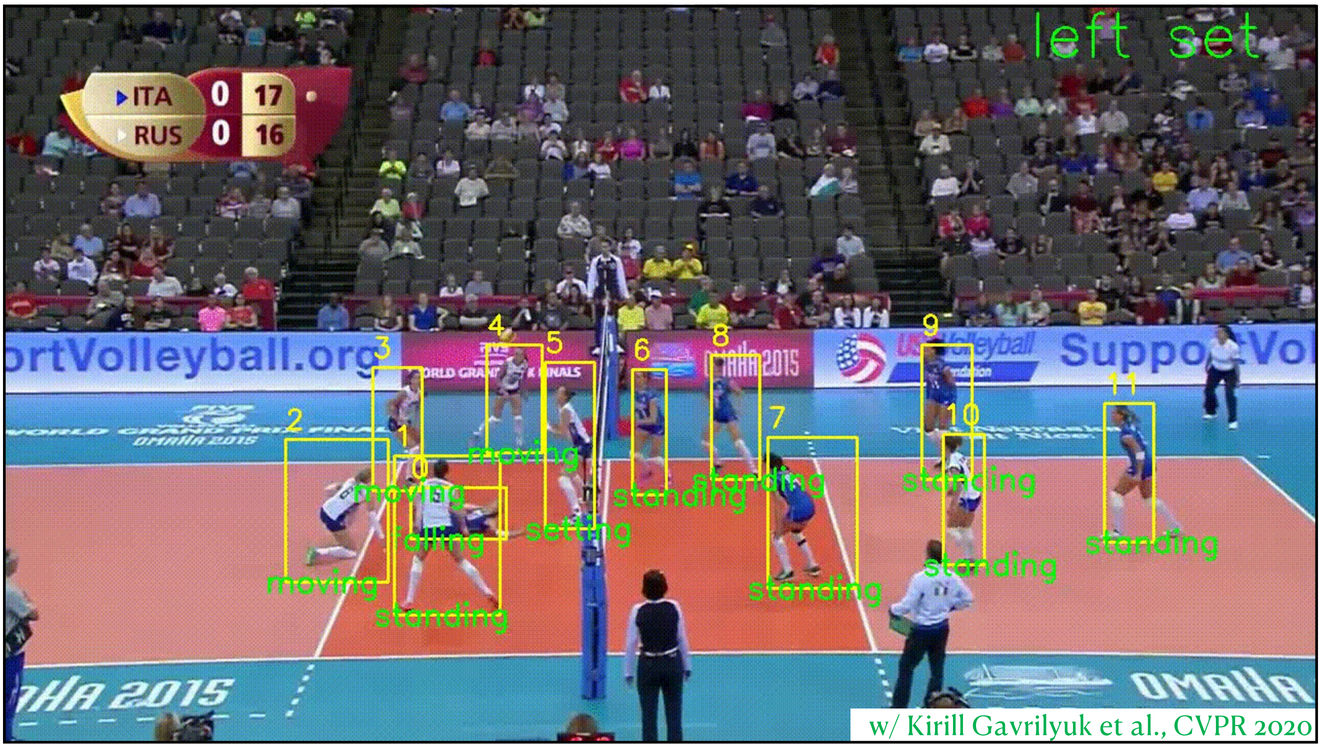
How it's going…

3



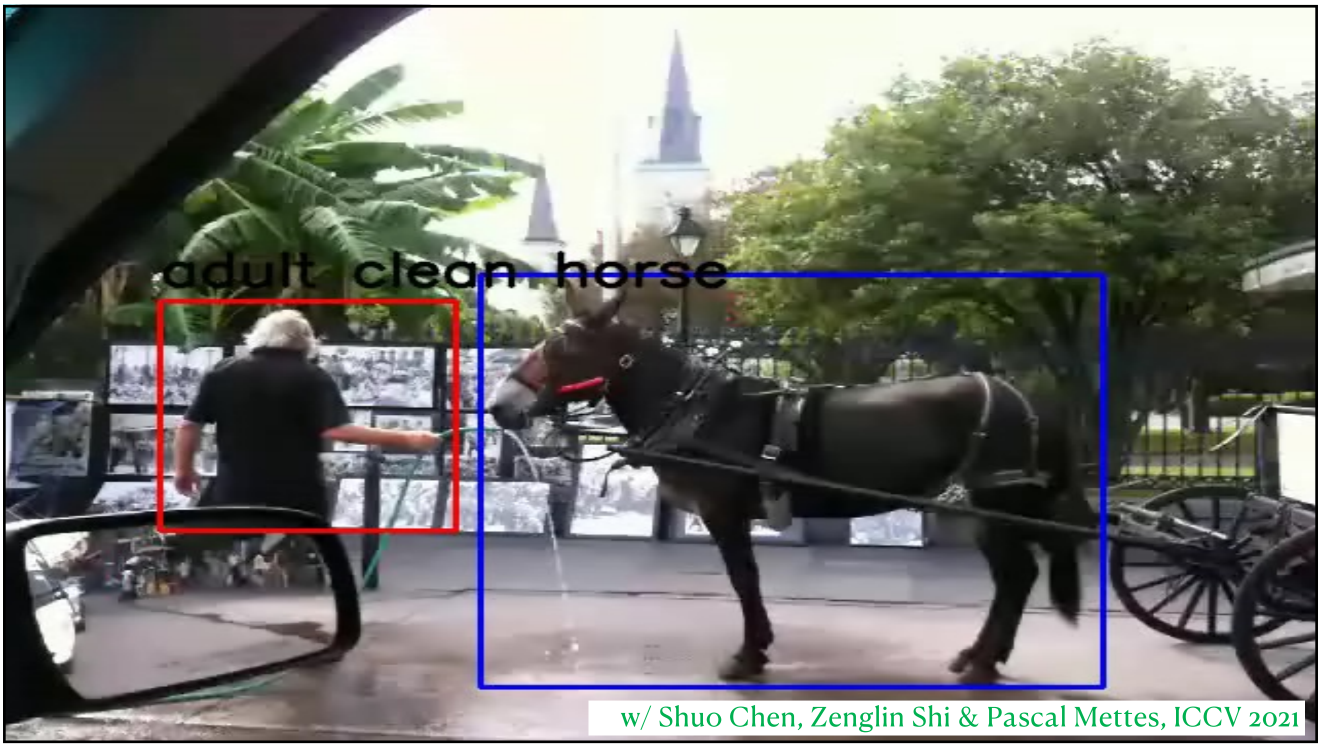
w/ Jiaojiao Zhao et al., CVPR 2022

4



w/ Kirill Gavrilyuk et al., CVPR 2020

5



w/ Shuo Chen, Zenglin Shi & Pascal Mettes, ICCV 2021

6


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

7


w/ Mohammadreza Salehi, Efstratios Gavves & Yuki Asano, ICCV 2023

8

w/ Hazel Doughty, CVPR 2022

Action: **peel**

How is the action done?

evenly, backwards, carefully, quickly, properly

9

Dual-use concerns & responsibility

10

11

12

13

Powerful yet irresponsible

- Mis-alignment with human values
- Hallucination
- Lacking adaptability to social dynamics and cultural context
- Limited transparency and explainability
- Non-inclusive and often closed access
- Unsustainable energy footprint
- Lacking robustness

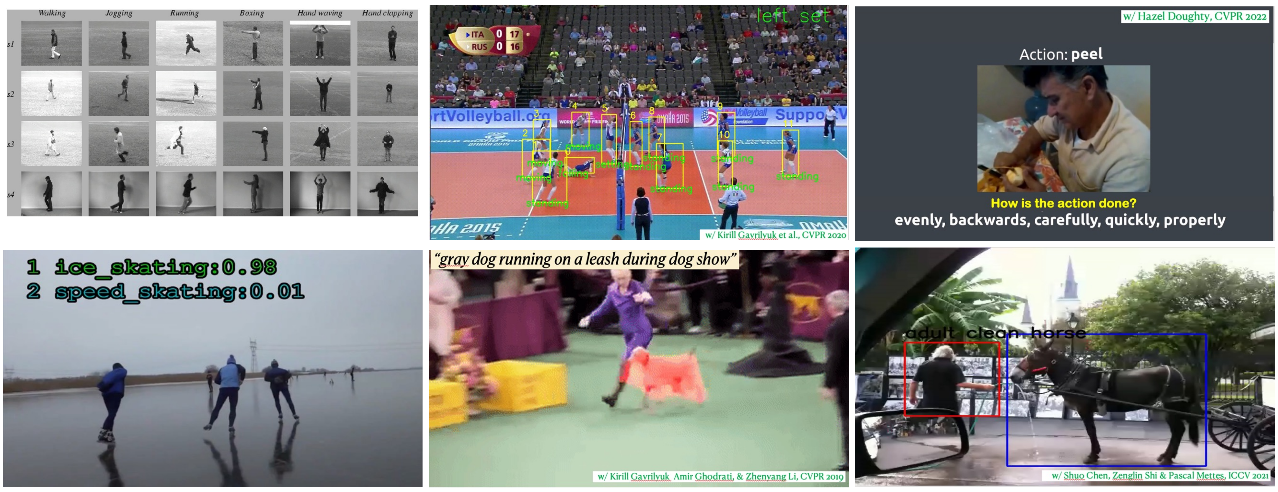
14

Powerful yet irresponsible

- Mis-alignment with human values
- Hallucination
- Lacking adaptability to social dynamics and cultural context
- Limited transparency and explainability
- Non-inclusive and often closed access
- Unsustainable energy footprint
- Lacking robustness

15

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



16

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*

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

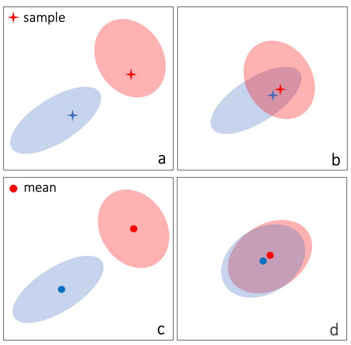
i.i.d. assumption

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

17

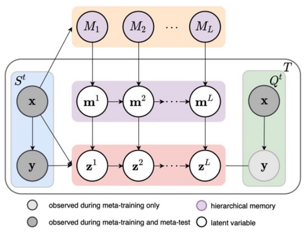
Machine learning inspiration

Domain-invariant learning



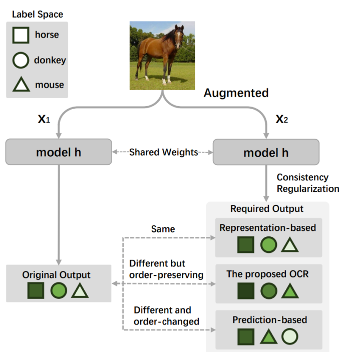
w/ Zehao Xiao et al., ICML 2021

Meta-learning



w/ Yingjun Du et al., ICLR 2022

Source-domain augmentation



w/ Mengmeng Jing et al., ICCV 2023

18

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 discov-

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

19

This lecture

Looks into the generalization abilities of modern video AI

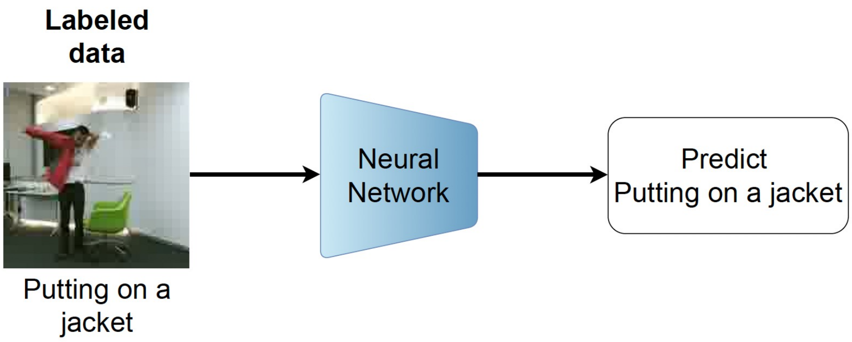
1. Video generalization by pre-training
2. Video generalization by adaptation
3. Video generalization at test-time

20

1. Video generalization by pre-training

21

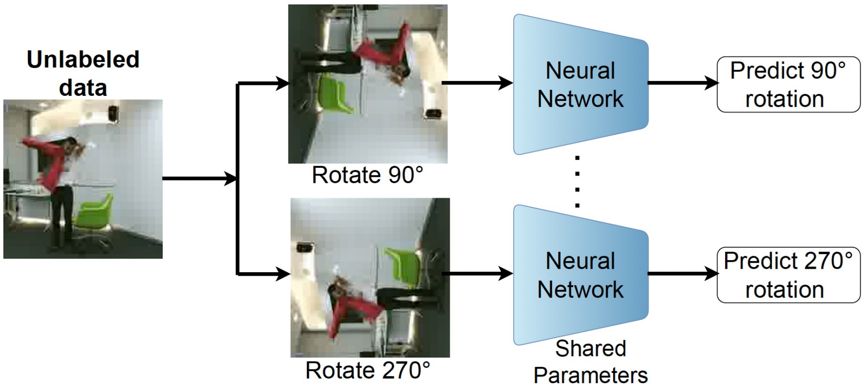
Supervised learning



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

22

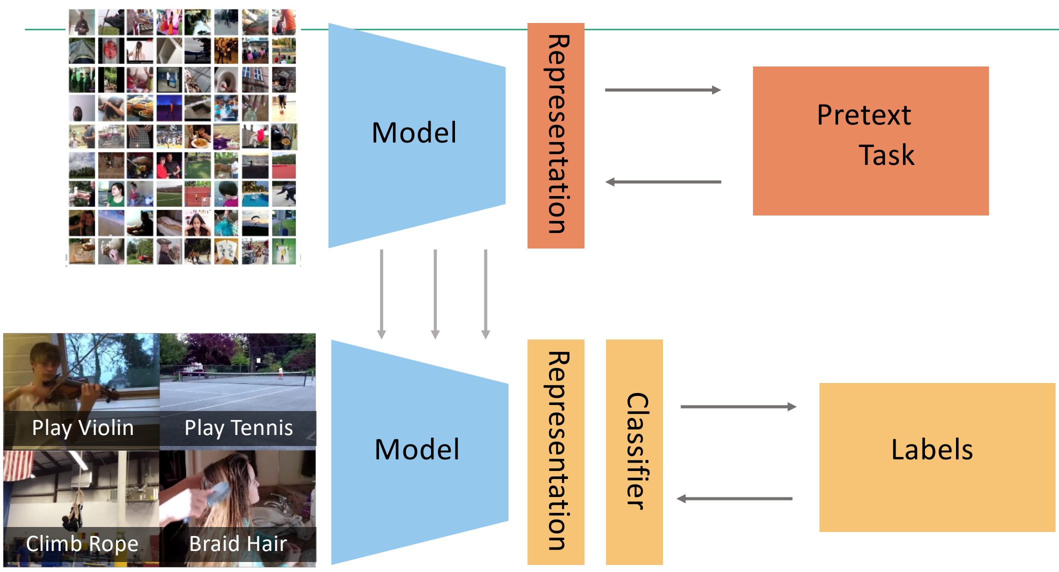
Self-supervised learning using a proxy task



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

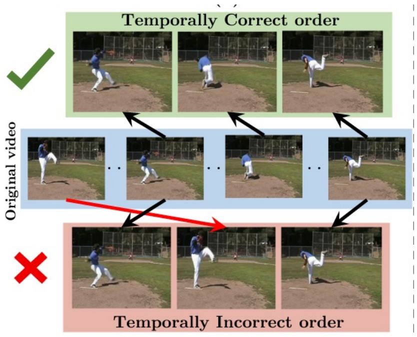
23

Self-Supervision

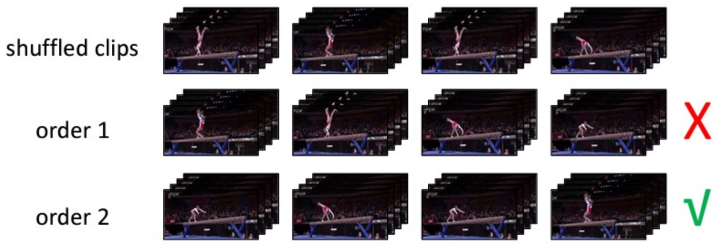


24

Example proxy tasks



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

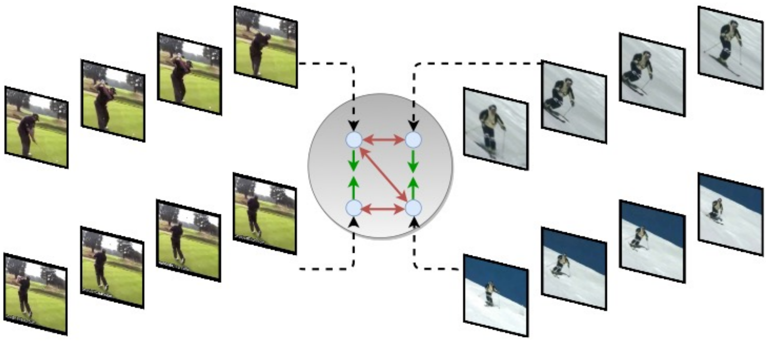


Video Clip Order Prediction, Xu et al., CVPR 2019

25

A more advanced proxy task: contrastive learning

Uses Instance discrimination and enforces augmentation invariance.

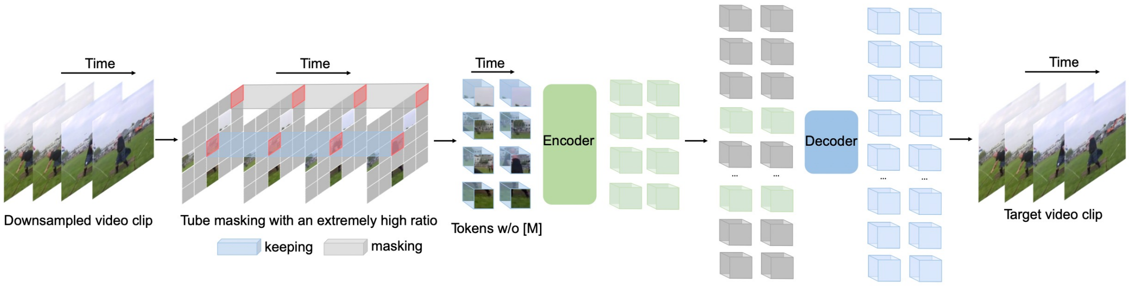


Adaptation of image-based methods like MoCo, SimCLR, to video domain.

26

Masked auto encoding transformers

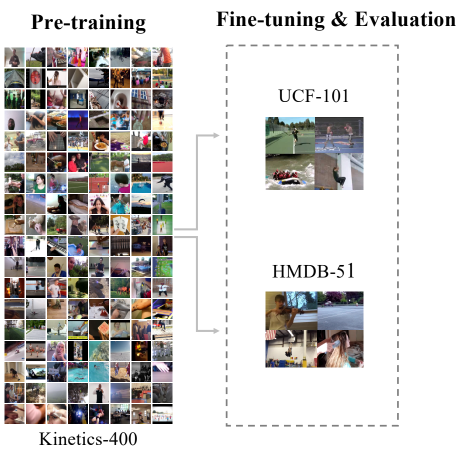
VideoMAE masks random cuboids and reconstructs the missing one



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

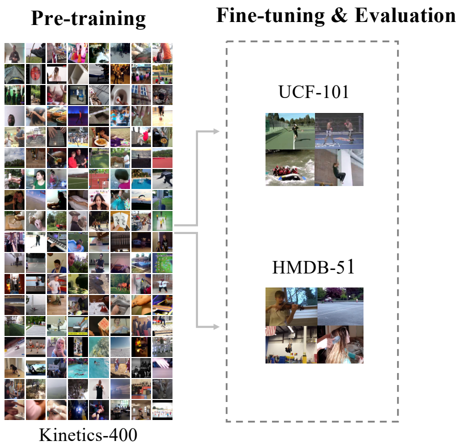
27

Problem: Video self-supervised learning evaluation

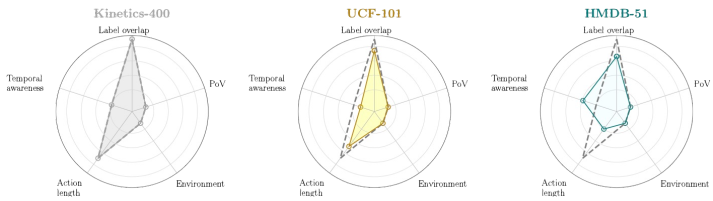


28

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.*

29

1.a How severe is benchmark-sensitivity?

Fida Mohammad Thoker
University of Amsterdam

Hazel Doughty
University of Amsterdam

Piyush Bagad
University of Amsterdam

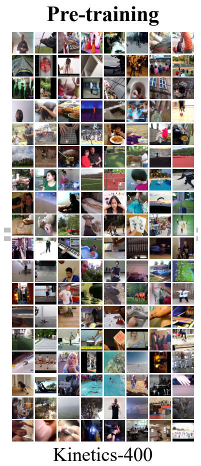
Cees Snoek
University of Amsterdam

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

Project Website

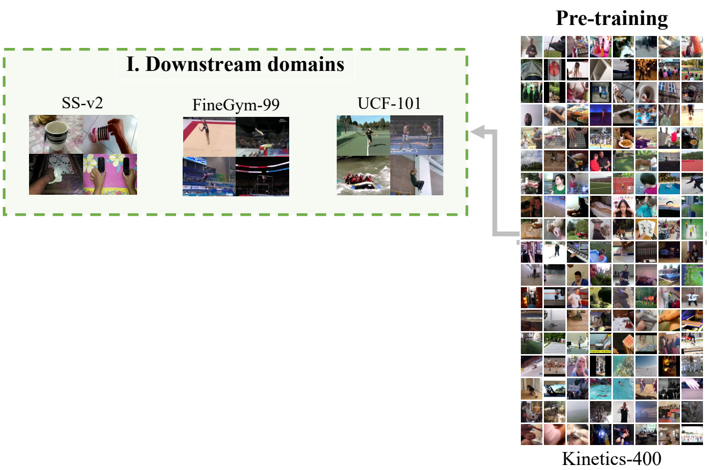
30

Proposed evaluation: four factors of sensitivity



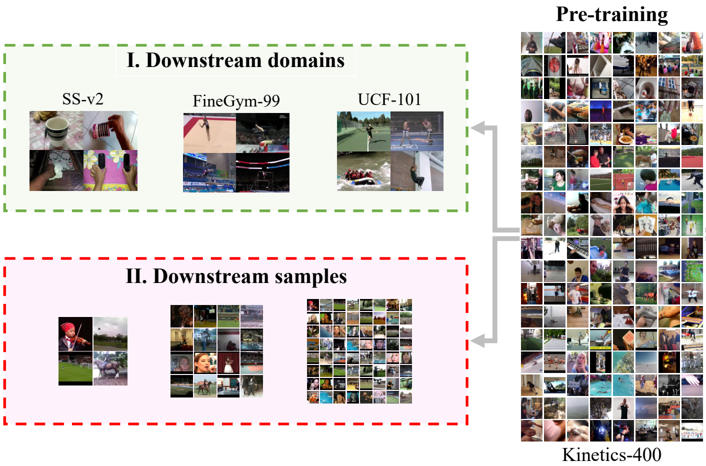
31

Proposed evaluation: four factors of sensitivity



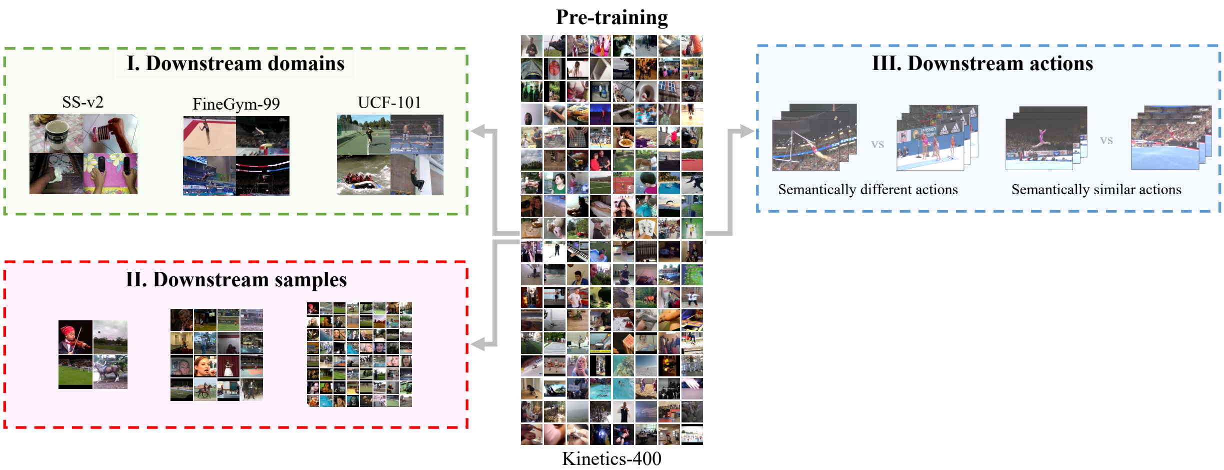
32

Proposed evaluation: four factors of sensitivity



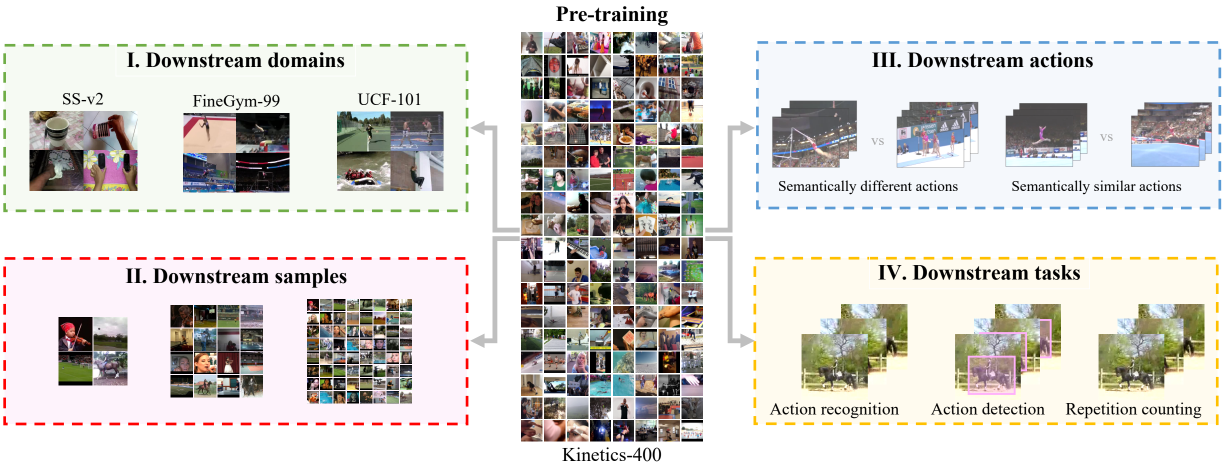
33

Proposed evaluation: four factors of sensitivity



34

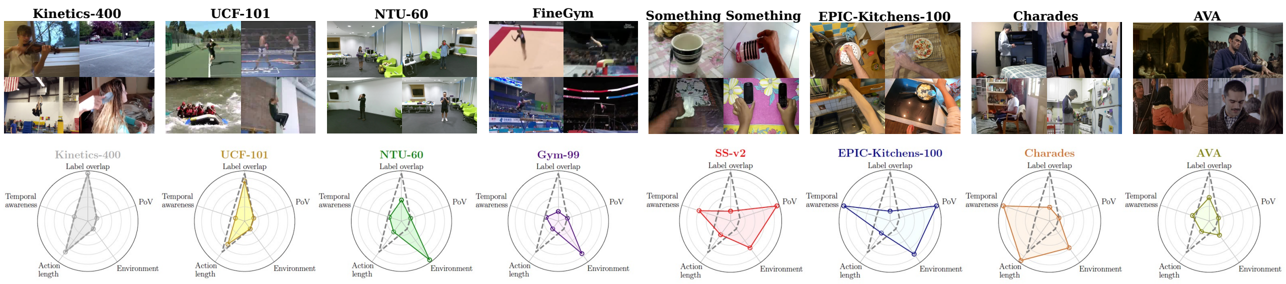
Proposed evaluation: four factors of sensitivity



35

7 datasets / 6 tasks / 500 experiments

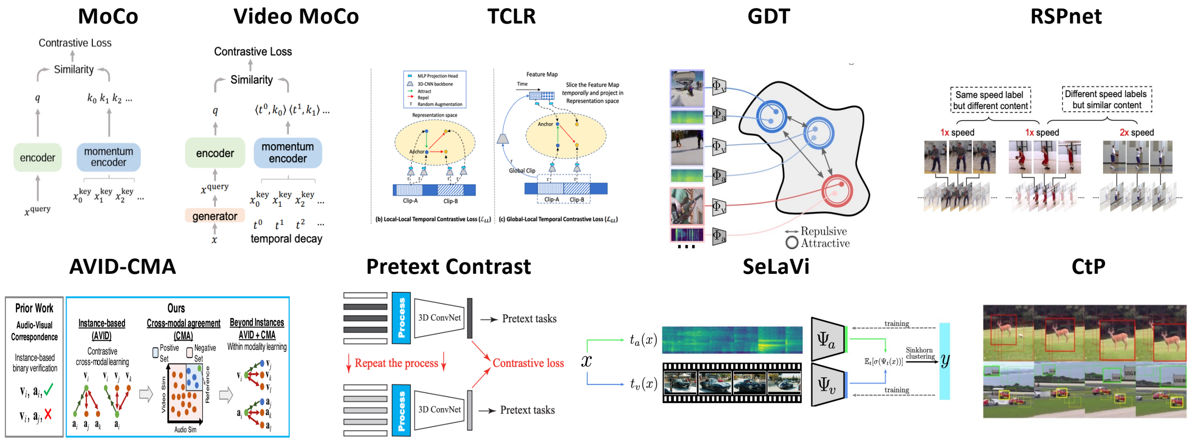
Considerable variety in video domain, the actions and tasks



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

36

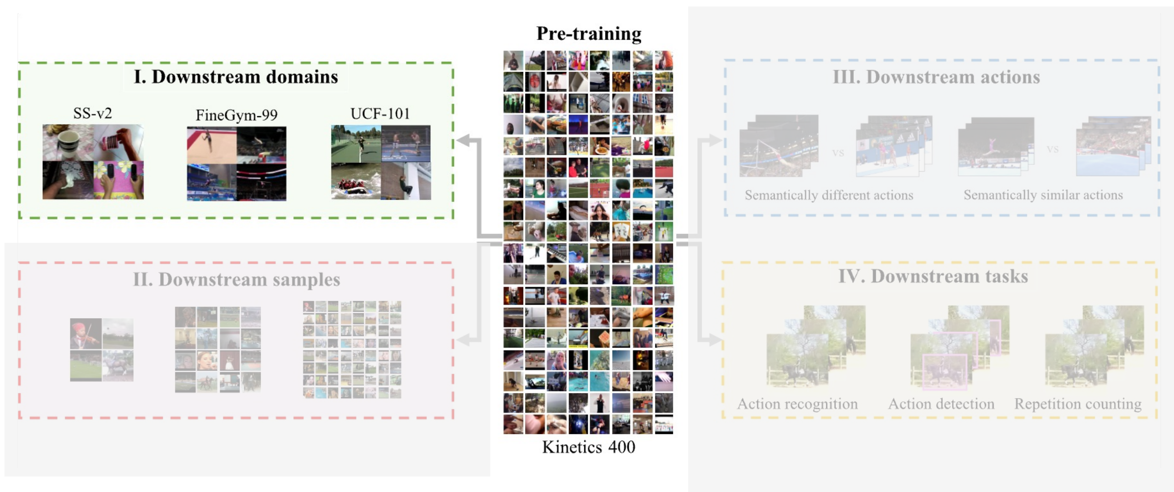
9 video self-supervised learners



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

37

Sensitivity factor I: Downstream domain



38

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 →

39

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 →

40

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 →

41

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 →

42

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 →

43

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 →

Downstream Domains

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

44

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

45

Sensitivity factor I: Downstream domain

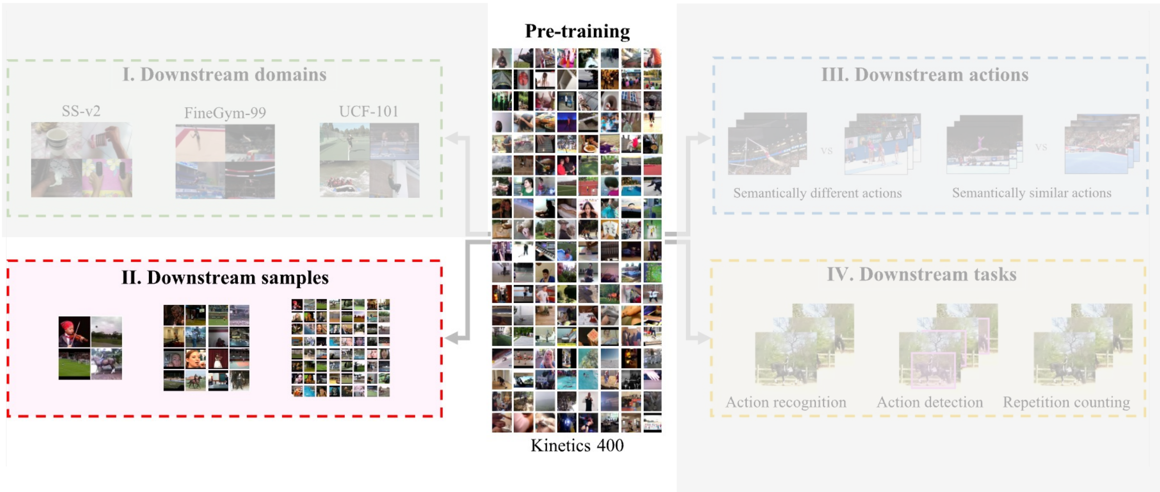
Downstream Domains

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

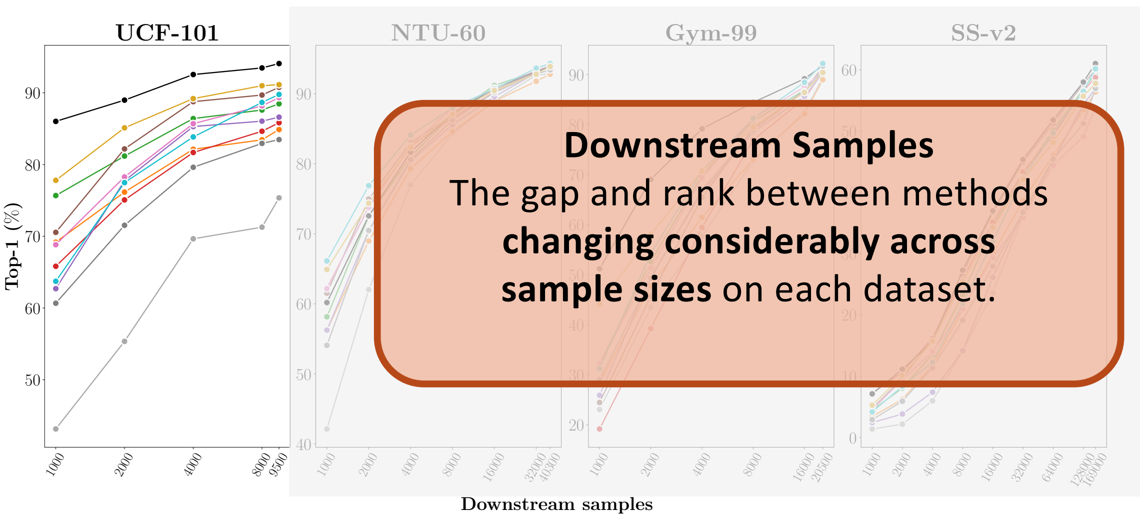
46

Sensitivity factor II: Downstream samples



47

Sensitivity factor II: Downstream samples



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

48

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.

49

Key takeaways

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

Supervised pre-training is dominant across all sensitivity factors.

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

We select a subset of experiments as the **'SEVERE' benchmark**

50

SEVERE benchmark: subset of our experiments

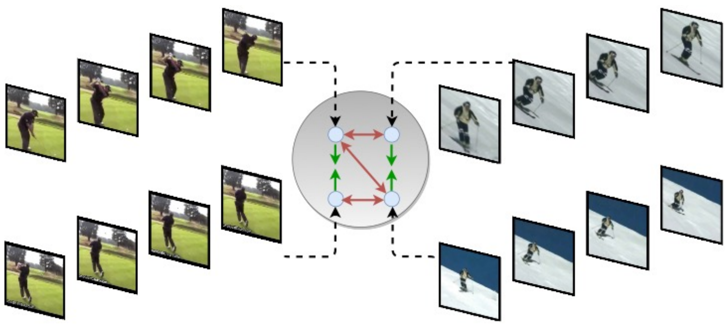
Pre-training	Existing	SEVERE-benchmark							
		Domains		Samples		Actions		Tasks	
	UCF101	SS-v2	Gym-99	UCF (10^3)	Gym-99 (10^3)	FX-S1	UB-S1	UCF-RC	Charades-MLC
None	75.4	56.8	89.4	43.1	23.1	45.0	84.0	0.232	7.9
MoCo	83.5	57.0	90.6	60.7	29.0	65.1	85.0	0.220	8.1
SeLaVi	84.9	56.4	88.9	69.2	28.3	50.2	81.5	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.

51

Problem of holistic contrastive learning

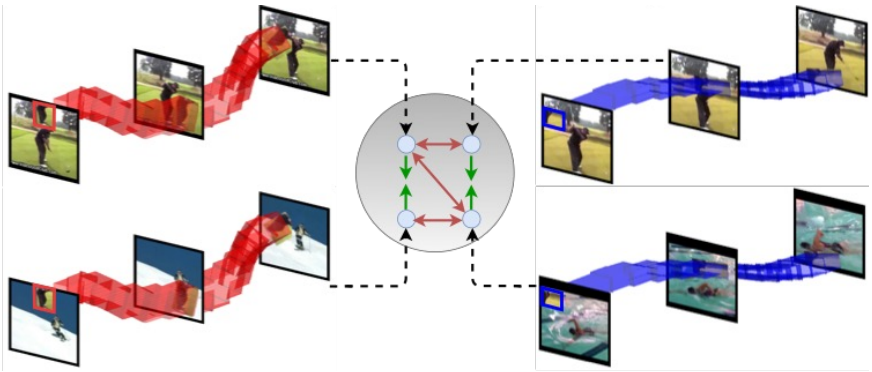
Uses Instance discrimination and enforces augmentation invariance.



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

52

Solution: add synthetic tubelets during pretraining



53

1.b Tubelet-contrastive self-supervision

Fida Mohammad Thoker
University of Amsterdam

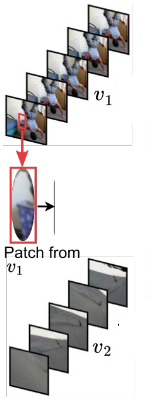
Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

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

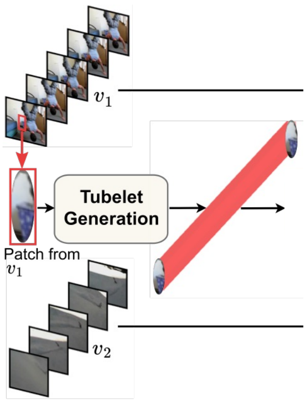
54

Step 0: Crop a random patch from one clip



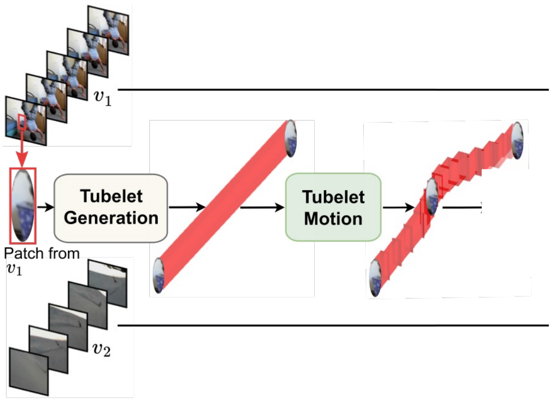
55

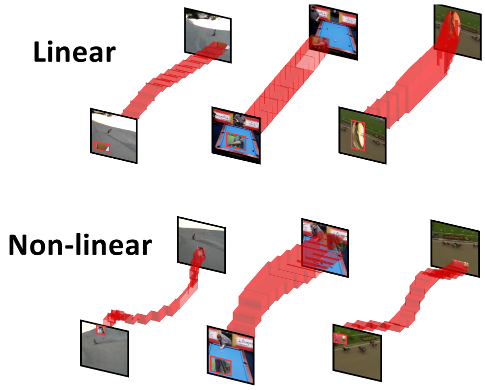
Step 1: Generate a tubelet



56

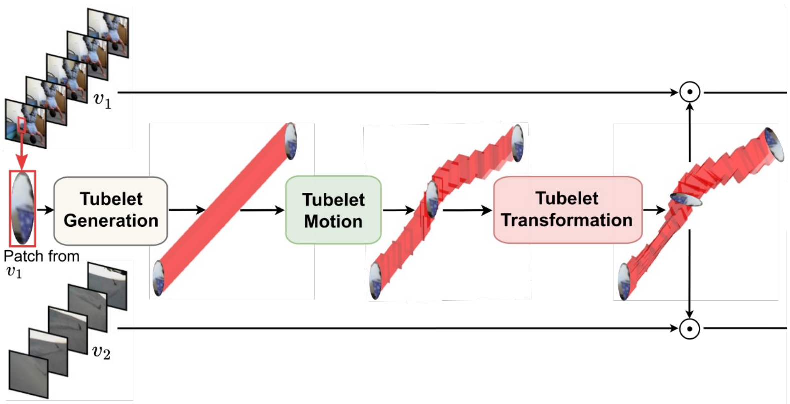
Step 2: Add motion to the patch





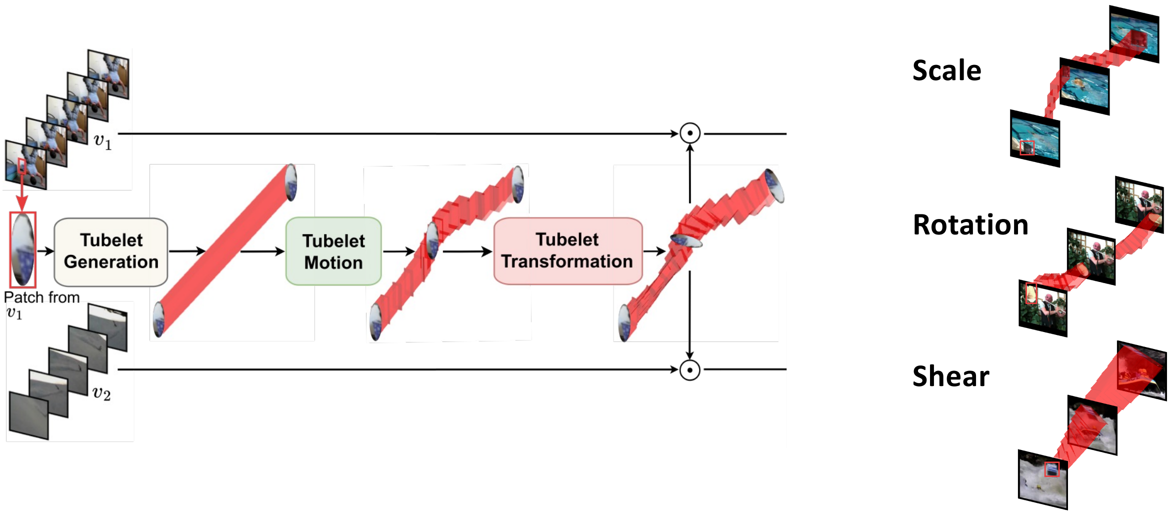
57

Step 3: Add motion complexity by transformations



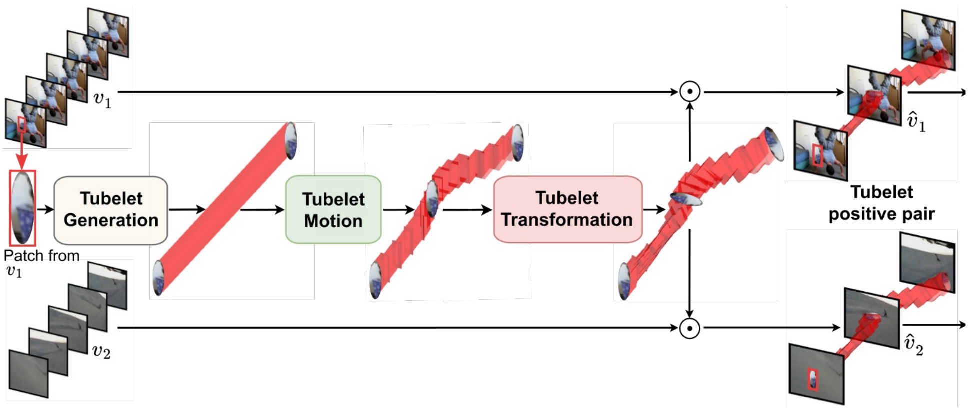
58

Step 3: Add motion complexity by transformations



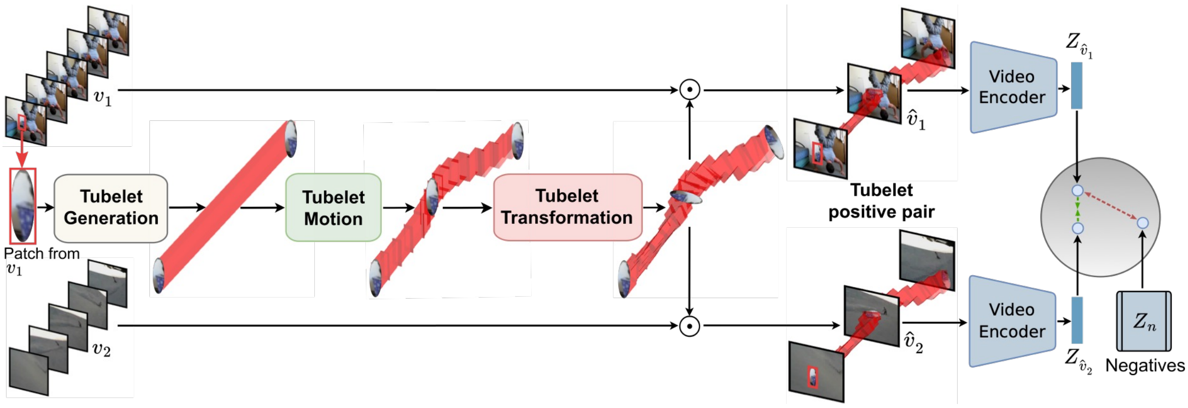
59

Step 4: Overlay identical tubelet on two clips



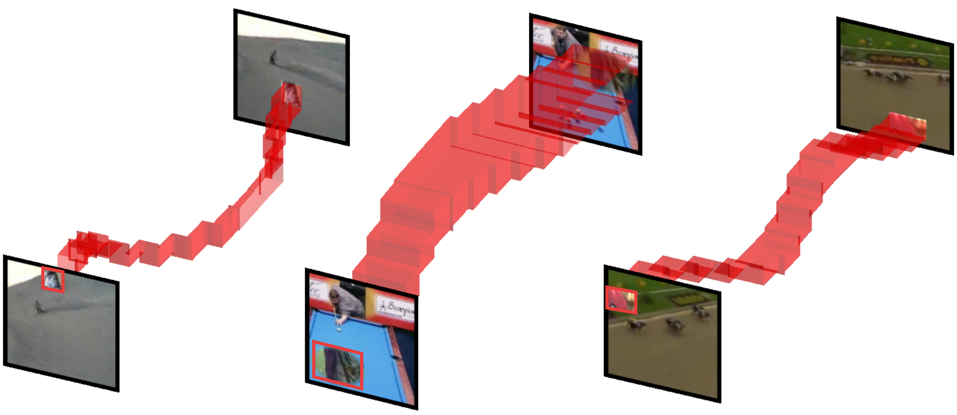
60

Step 5: Tubelet-contrastive learning



61

Examples of synthetically added tubelets



62

Ablations

	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.

63

Ablations

	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.

64

Ablations

	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	47.5	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.

65

Ablations

	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	47.5	47.3	90.9
Rotation	65.5	48.0	47.9	90.9

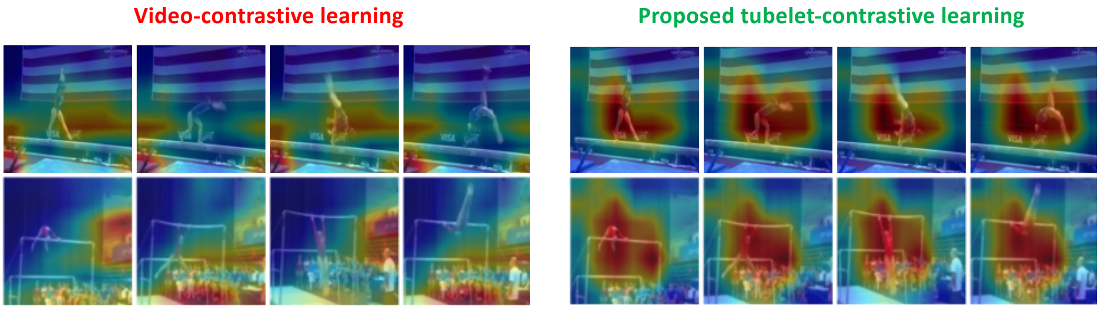
Table 4: **Tubelet Transformation.** Adding motion patterns to tubelet-contrastive learning through transformations improves downstream performance. Best results for rotation.

#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.

66

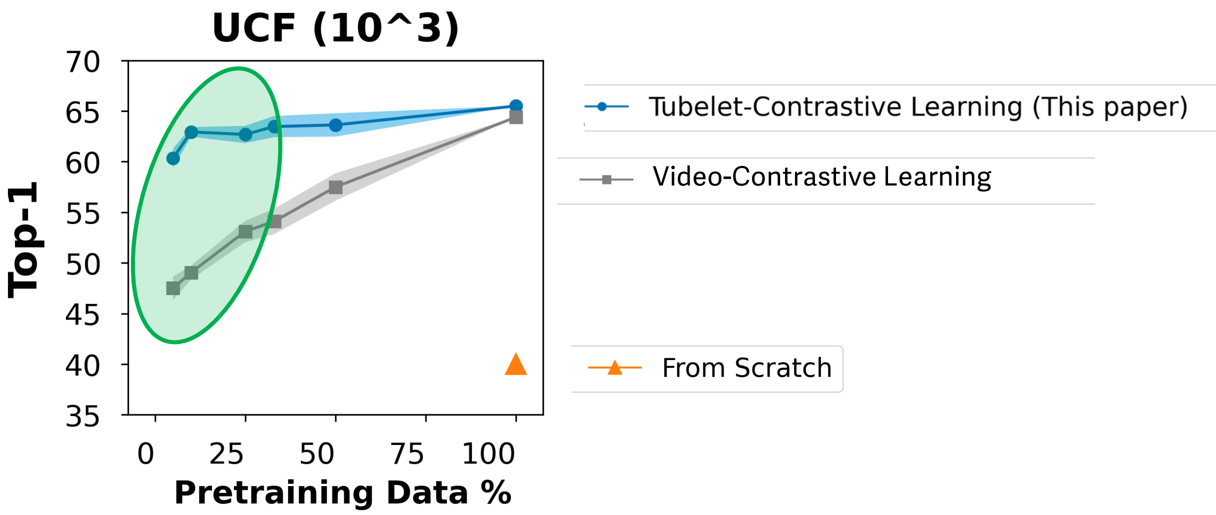
What does the model learn?



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

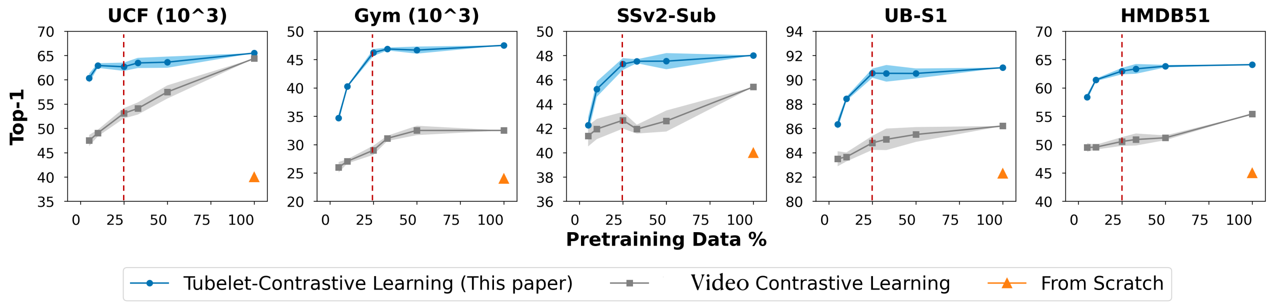
67

Adding synthetic motion improves data efficiency



68

Key benefit: we need 4x less video data



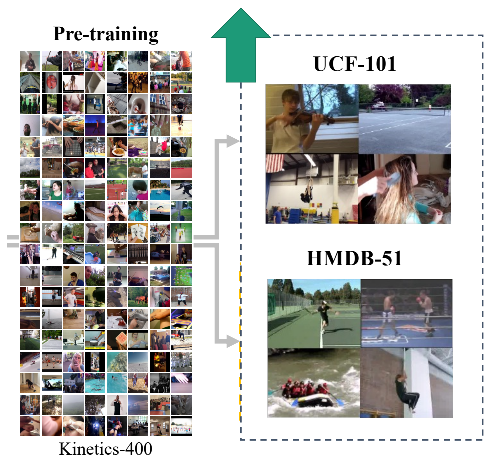
Tubelets simulate a richer variety of fine-grained motion than present in the original video

69

Solid accuracy gain on UCF-101 and HMDB-51

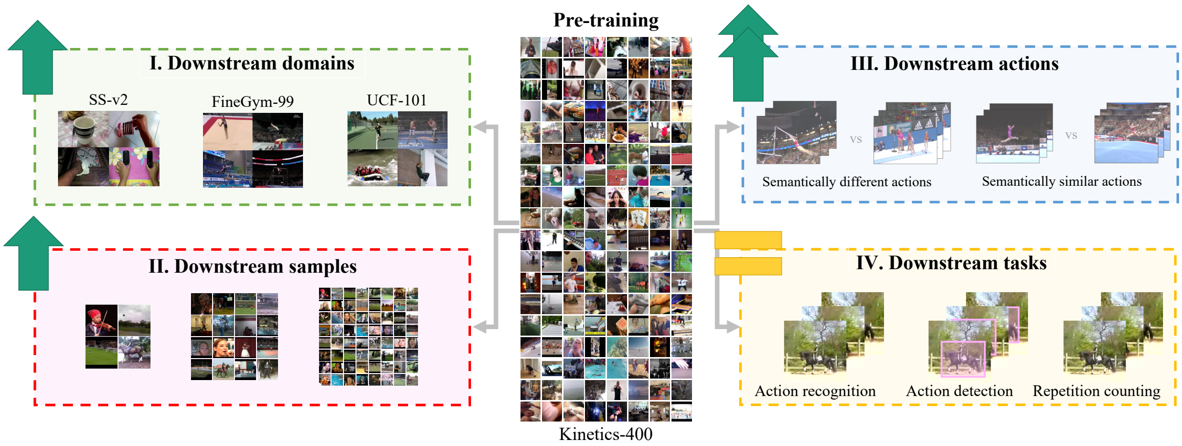
R(2+1)D Backbone pretrained on Kinetics-400

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**



70

Generalization on SEVERE-benchmark



71

Generalization on SEVERE-benchmark

		Domains		Samples		Actions		Tasks			
	Backbone	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.

72

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.

73

2. Video generalization by adaptation

74

Introduction

In this chapter we consider generalization by video representation adaptation, without retrainining from scratch.

We will exploit the (often) multimodal nature of video to aid us in this endeavour.

75

2.a Adaptation by video-language

Piyush Bagad
University of Amsterdam

Makarand Tapaswi
IIIT, Hyderabad

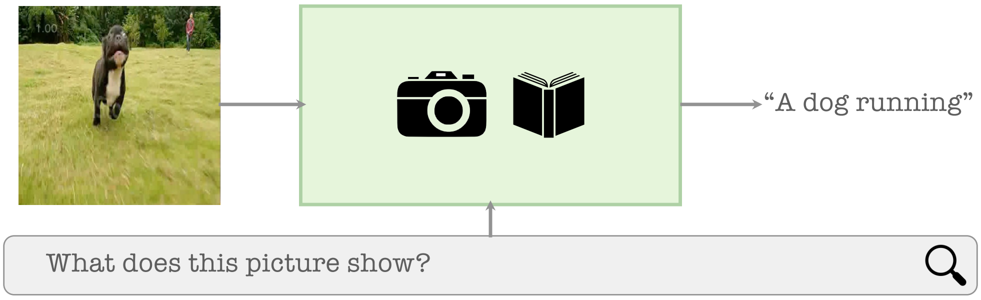
Cees Snoek
University of Amsterdam

Test of Time: Instilling Video-Language Models with a Sense of Time. In C*VPR* 2023.

76

The problem

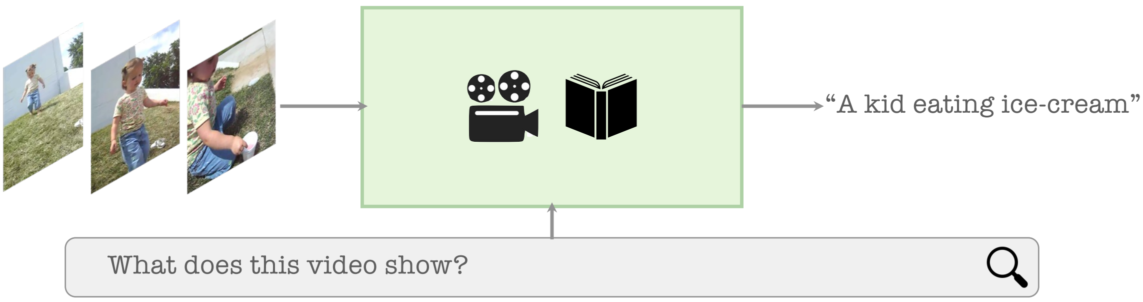
- Foundation models: Language interface + a few (or no) training samples



77

The problem

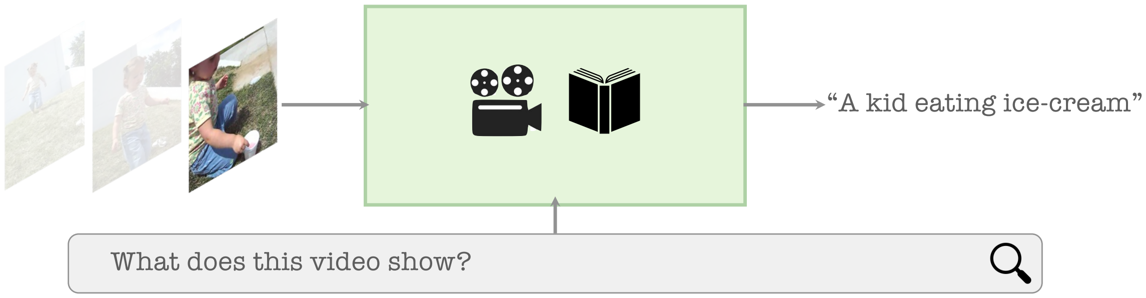
- Foundation models: Language interface + a few (or no) training samples
- Particularly attractive for videos given high cost



78

The problem

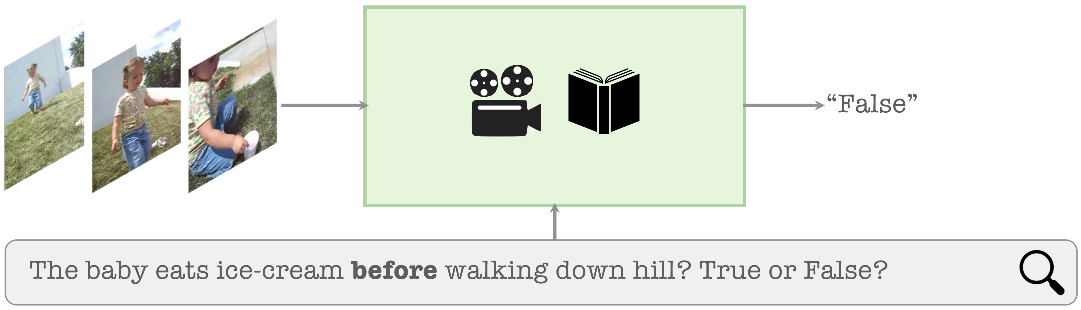
- Do video foundation models truly understand time?



79

The problem

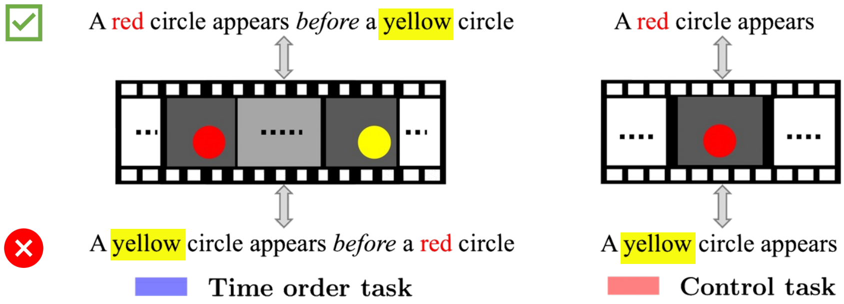
- Do video foundation models truly understand time?
- Our idea for a “test of time”: ask questions that have temporal relations



80

The test of time

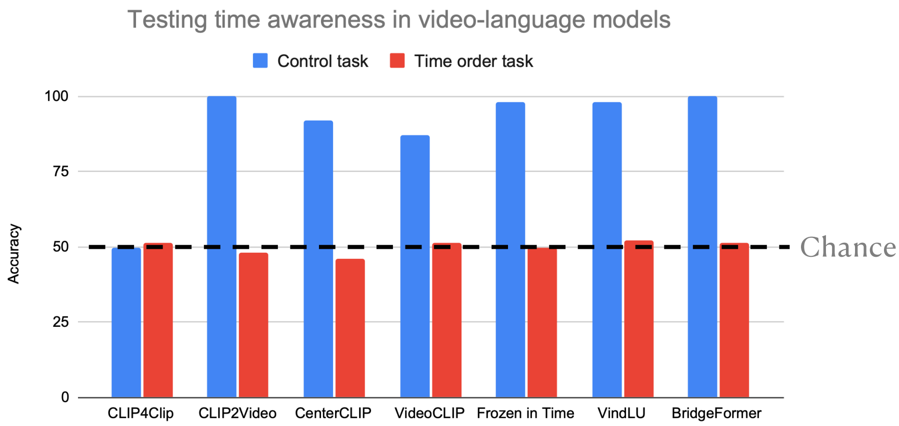
- The spatial image bias in current video benchmarks
- Synthetic benchmark



81

Existing models fail this test of time

- We pick a suite of seven openly available video-language models
- While excelling at the control task, they all fail at the time-order task



82

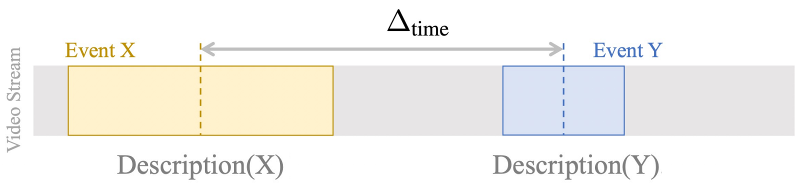
How to instil this sense of time?

- Post-pretraining: instead of training from scratch, we run another round of pre-training

83

How to instil this sense of time?

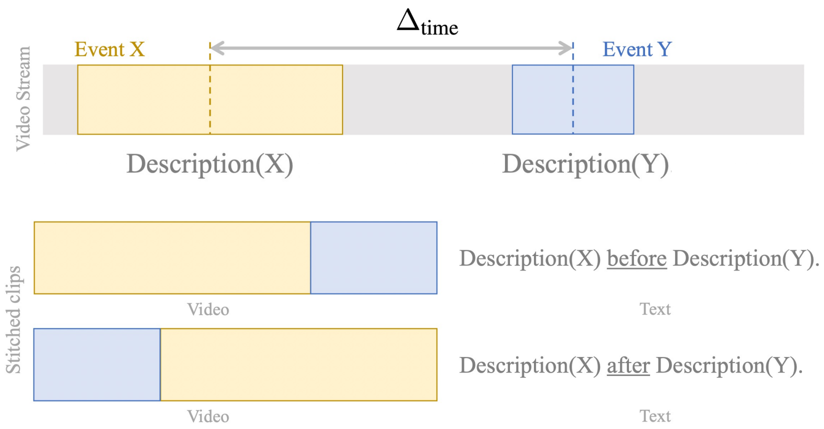
- Data: any dense video-captioning dataset!



84

How to instil this sense of time?

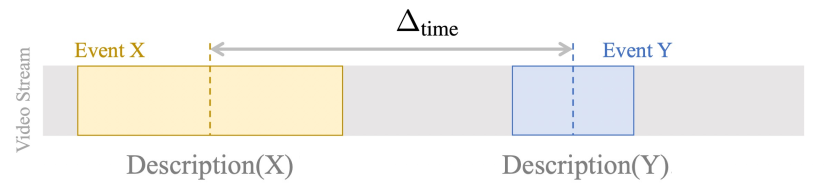
- Data: any dense video-captioning dataset!



85

How to instil this sense of time?

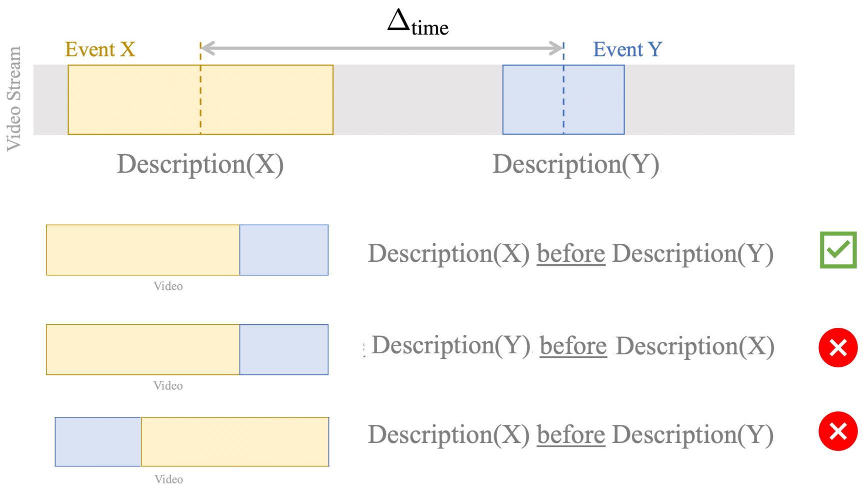
- Data: any dense video-captioning dataset!



86

How to instil this sense of time?

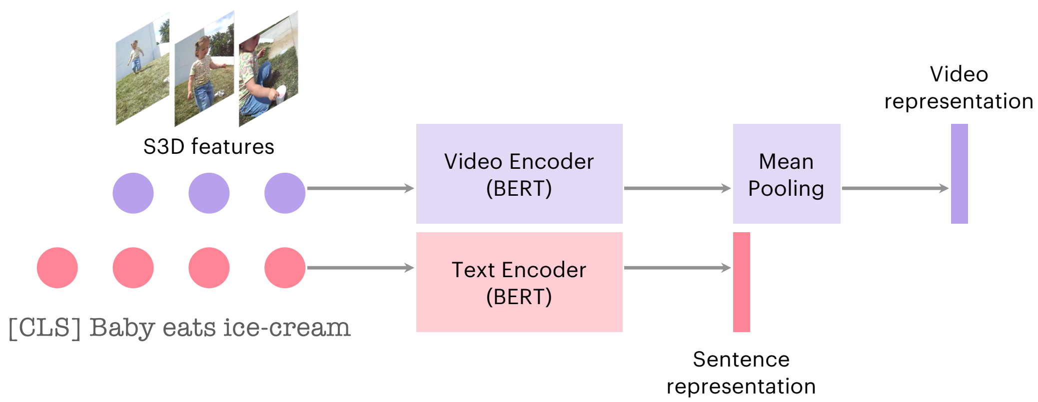
- Data: any dense video-captioning dataset!



87

How to instil this sense of time?

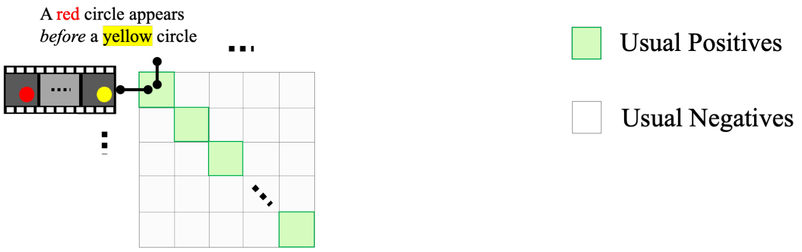
- Base model: We start with a pre-trained model: VideoCLIP [1]



[1] Xu et al, VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021.

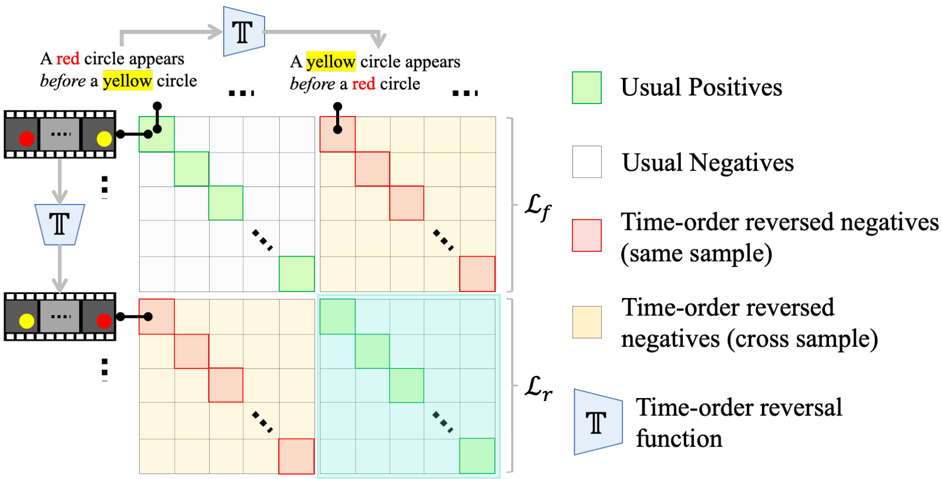
88

How to instil this sense of time?



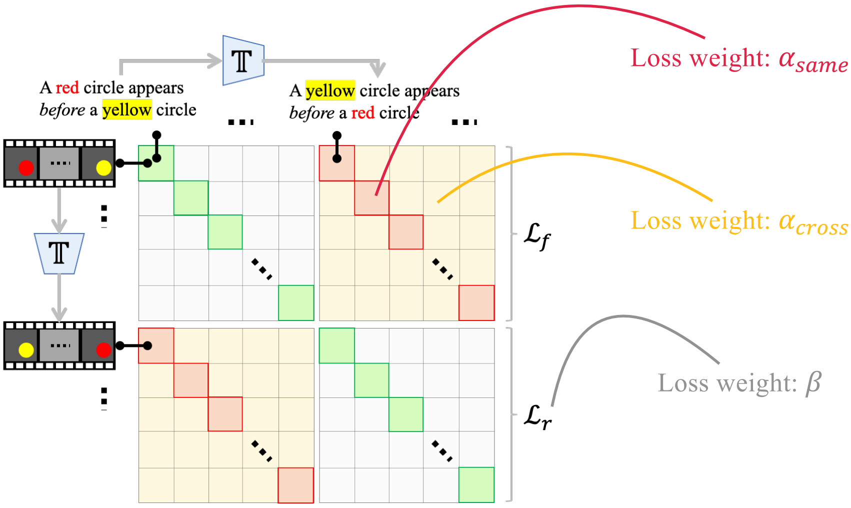
89

How to instil this sense of time?



90

How to instil this sense of time?



TACT: Temporal Adaptation by Consistent Time-ordering

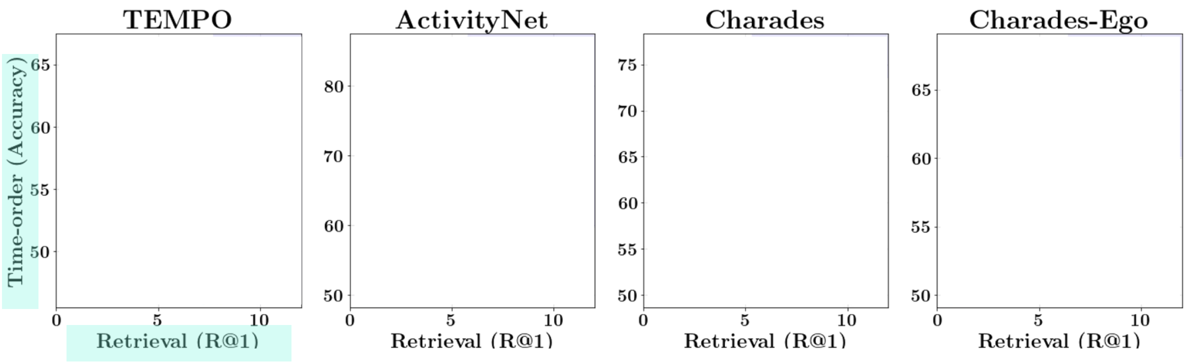
91

Experiments

(a) TEMPO
(b) ActivityNet
(c) Charades
(d) Charades-Ego

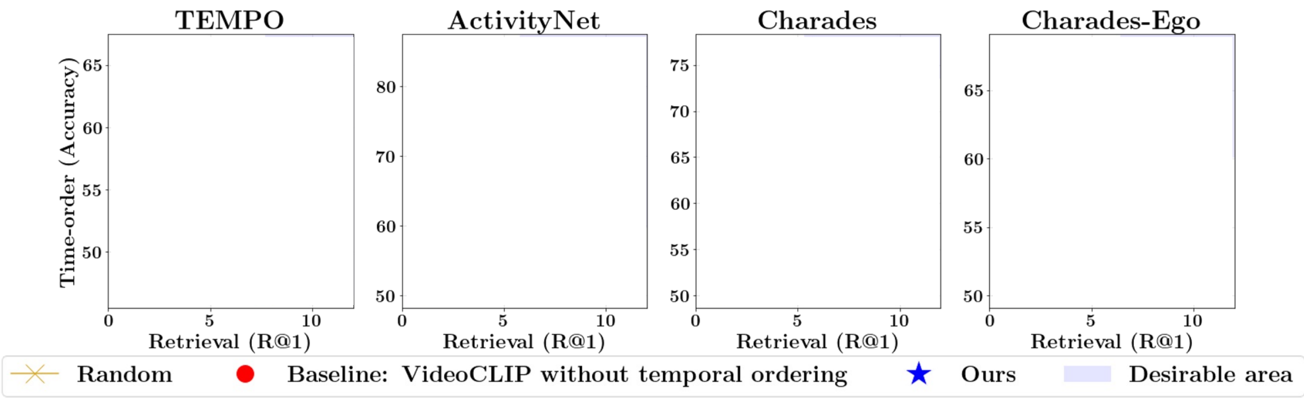
92

Experiments



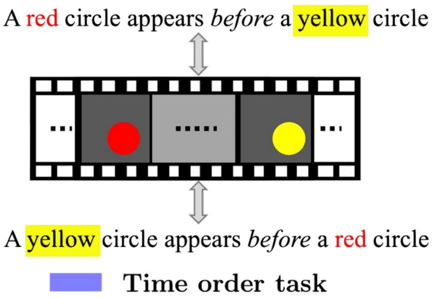
93

Experiments



94

Experiments: Synthetic benchmark

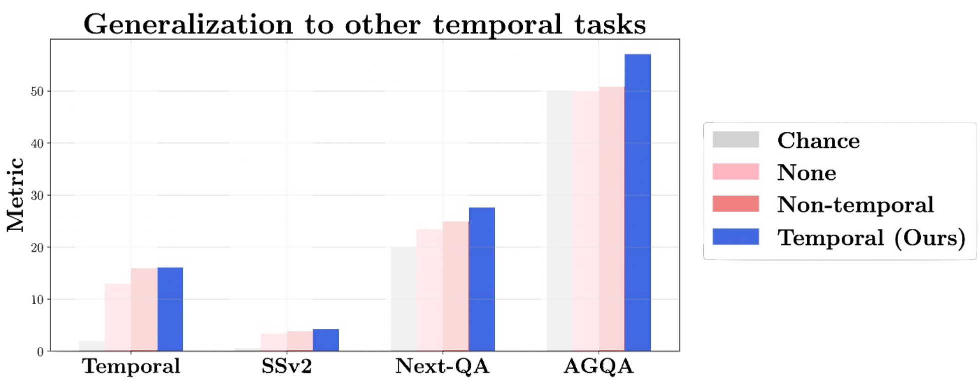


Training Dataset	Accuracy on synthetic data
TEMPO	64.4
ActivityNet	52.5
Charades	65.0
Charades-Ego	85.6

95

Does it work beyond this narrow sense of time?

- Does acquiring this narrow sense of time help other general temporal tasks? We find benefits on several temporal reasoning tasks.



96

Key takeaways

We propose a "test of time" for video-language models. Existing models fail this test.

Our simple TACT recipe instills this sense of time **without re-training from scratch.**

bpiyush.github.io/testoftime-website/

97

2.b Adaptation by video-audio


Yunhua Zhang
University of Amsterdam


Hazel Doughty
University of Amsterdam

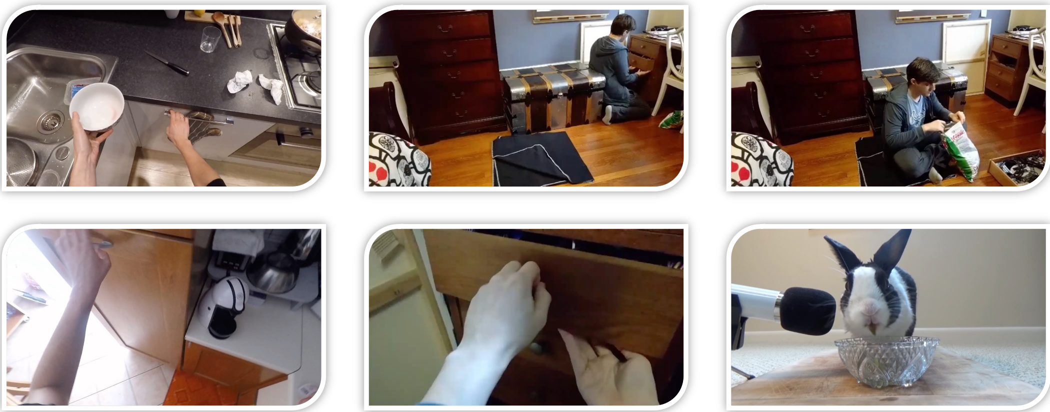

Ling Shao
Inception Institute of AI


Cees Snoek
University of Amsterdam

Audio-Adaptive Activity Recognition Across Video Domains. In *CVPR* 2022.

98

Activity recognition under domain shift

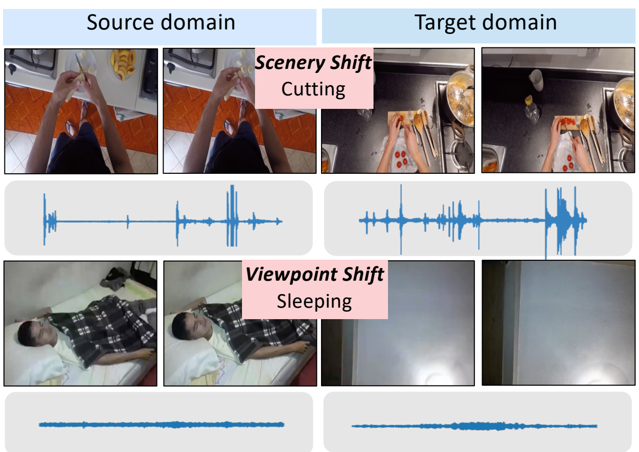


Scenery shift | Camera viewpoint shift | Actor shift

99

Proposed solution

*We deal with the vision distribution shift with the aid of **activity sounds**.*



Characteristic sound signals of audible activities

(Playing piano, playing guitar, ...)

Environmental sounds of silent activities

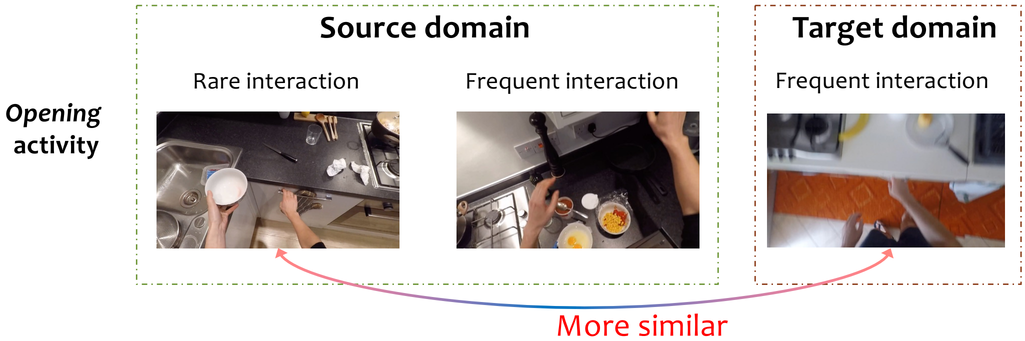
Situp (Sounds in the gym), Camping (Outdoor sounds)

100

Audio-balanced learning

Motivation: videos from **different domains** often have **different label distributions**, not only in terms of activity classes but also their interactions with objects or the environment.

Solution: learn each class and each type of interaction equally



101

Absent-activity learning

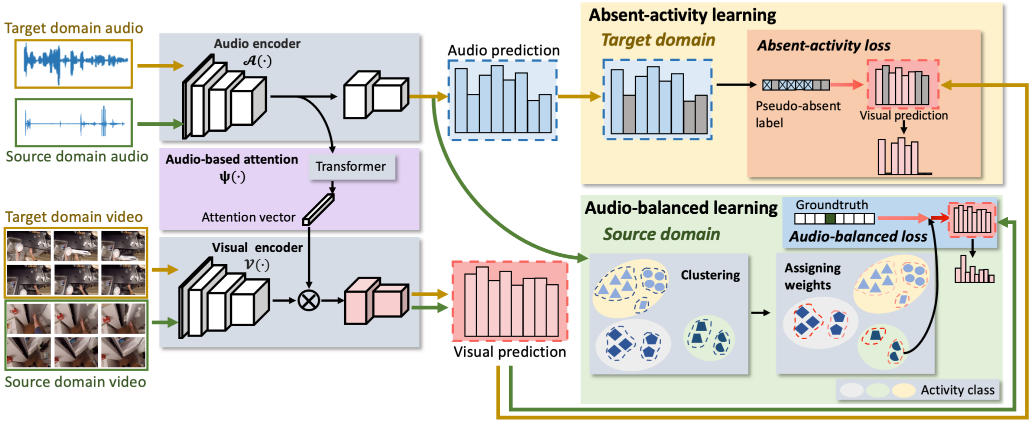
Groundtruth activity:
pour

Absent activities predicted by audio:
wash
close
open

102

Audio-adaptive approach

Supervised by **audio-balanced learning** and **absent-activity learning**



103

Results

Model	Scenery shift EPIC-Kitchens Top-1 (%)	Viewpoint shift CharadesEgo mAP (%)
Visual-only	48.0	23.1
Ours (no audio in testing)	50.7	24.5
Ours	59.2	26.3

104

Actor-shift: success case

Source domain

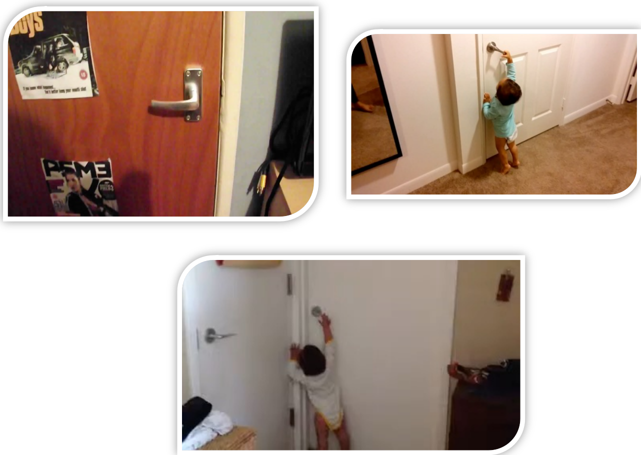
Target domain

Groundtruth: *sleeping*
Prediction: *sleeping*
Confidence: 0.76

105

Actor-shift: success case

Source domain



Target domain

Groundtruth: *opening door*
Prediction: *opening door*
Confidence: 0.85

106

Actor-shift: failure case

Source domain

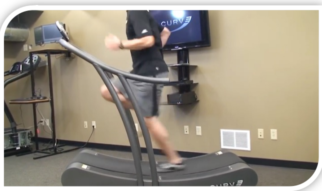
Target domain

Groundtruth: *drinking*
Prediction: *eating*
Confidence: 0.35

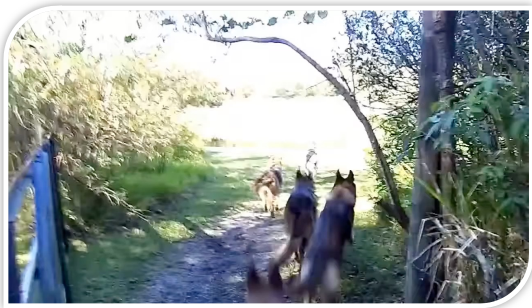
107

Actor-shift: failure case

Source domain



Target domain



Groundtruth: *running*
Prediction: *swimming*
Confidence: 0.48

108

Key takeaways

Showed invariant properties of **sound to reduce visual domain gap**.

Better adaptation ability than visual-only solutions

Benefits from audio more than alternative audiovisual fusion methods

Generalize models to new **environments, viewpoints and actors**

109

2.c Adaptation at night

Yunhua Zhang
University of Amsterdam

Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

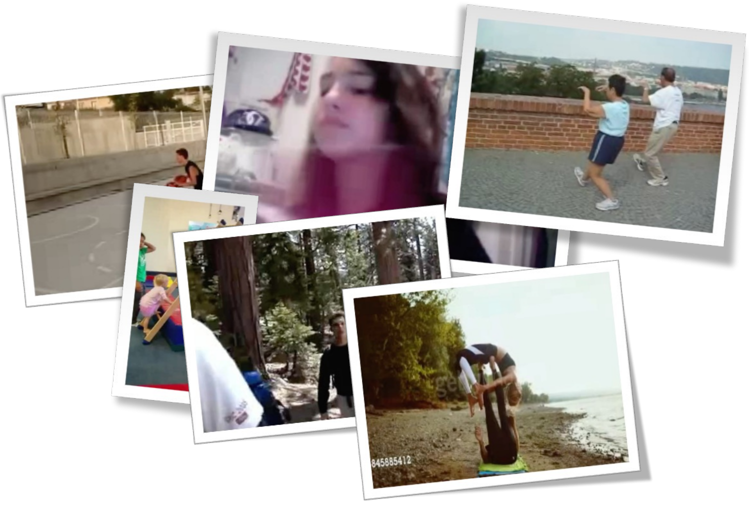
Day2Dark: Pseudo-Supervised Activity Recognition beyond Silent Daylight. *Submitted*.

110

Video datasets are biased to daylight conditions

Video dataset
EPIC-Kitchens
ActivityNet
Charades
Kinetics-400
Moments-in-Time
Kinetics-Sound

$$Y = \frac{\sum_{j=1}^{H_v \times W_v} (0.299 R_j + 0.587 G_j + 0.144 B_j)}{H_v \times W_v}.$$

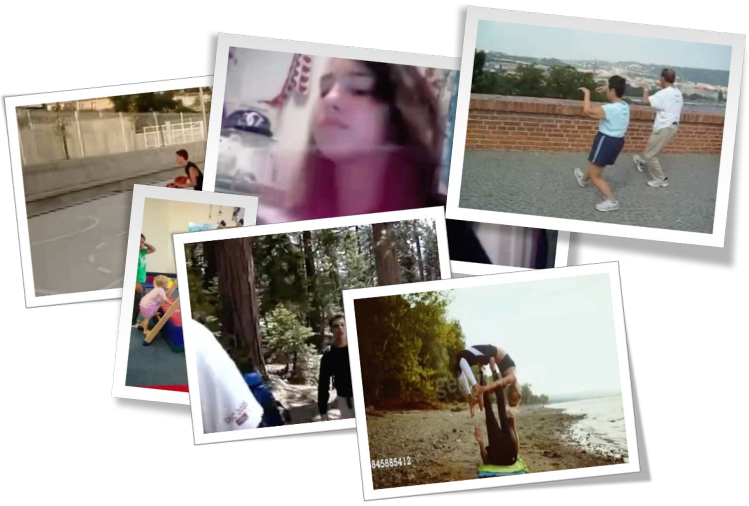


111

Video datasets are biased to daylight conditions

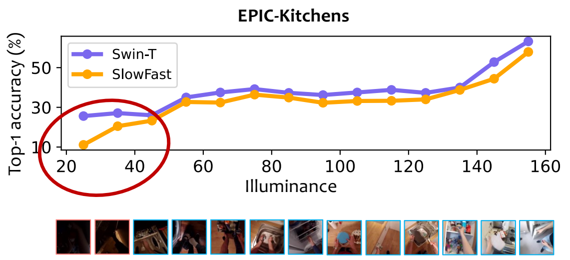
Video dataset	Dark videos (Y<=40)
EPIC-Kitchens	1.9%
ActivityNet	3.2%
Charades	3.6%
Kinetics-400	4.4%
Moments-in-Time	4.9%
Kinetics-Sound	8.3%

$$Y = \frac{\sum_{j=1}^{H_v \times W_v} (0.299 R_j + 0.587 G_j + 0.144 B_j)}{H_v \times W_v}.$$



112

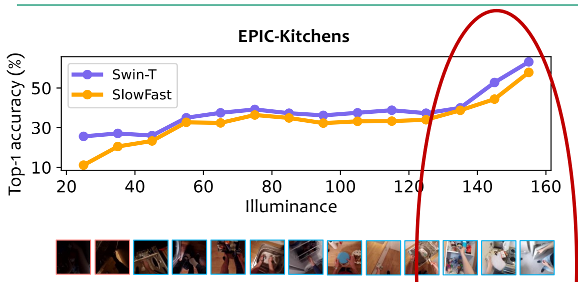
Problem statement: Day2Dark gap



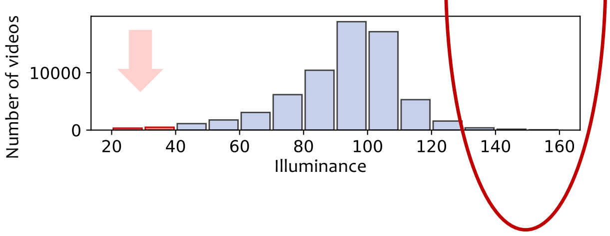
Activity recognition models suffer from **performance drops** in low-illumination.

113

Problem statement: Day2Dark gap



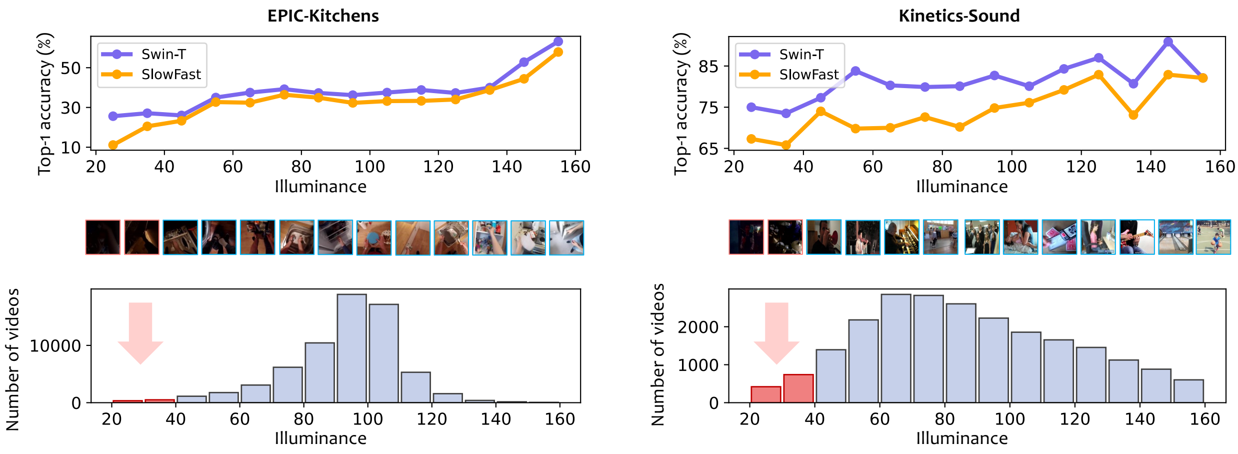
Activity recognition models suffer from **performance drops** in low-illumination.



Caused by lack of **training data** and **distribution shift** by lower color contrast

114

Problem statement: Day2Dark gap



115

Technical contributions

I. A **pseudo-supervised learning** strategy that utilizes **unlabeled dark videos**, which do not contain target activities.

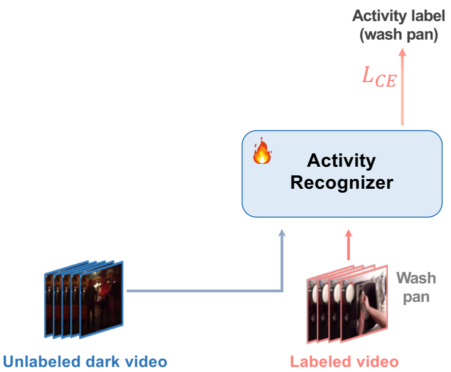
II. **Darkness-aware** audio-visual recognition to **reduce the distribution** shift and find better **cross-modal correspondences** in the dark.

116

Unlabeled dark video examples

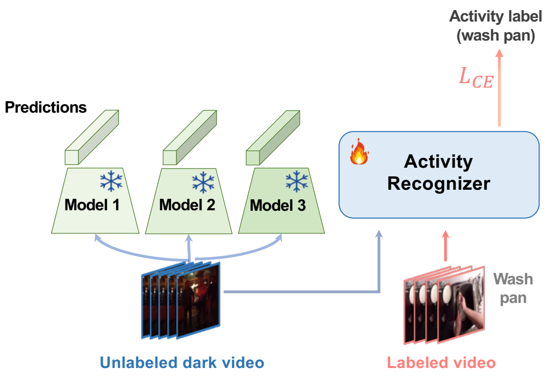
117

I. Supervision beyond daylight



118

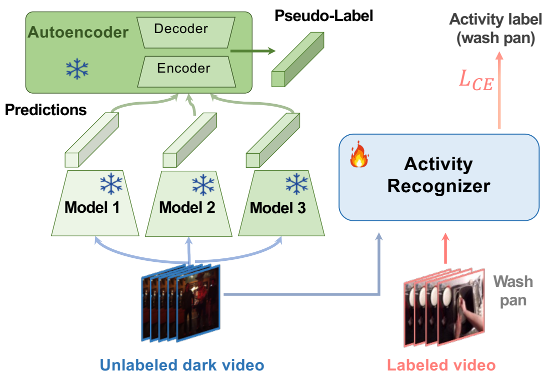
I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

119

I. Supervision beyond daylight

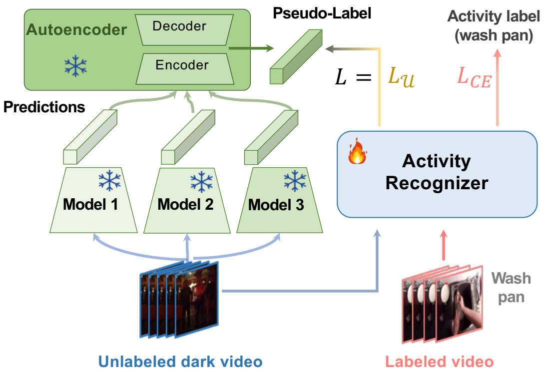


Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

120

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

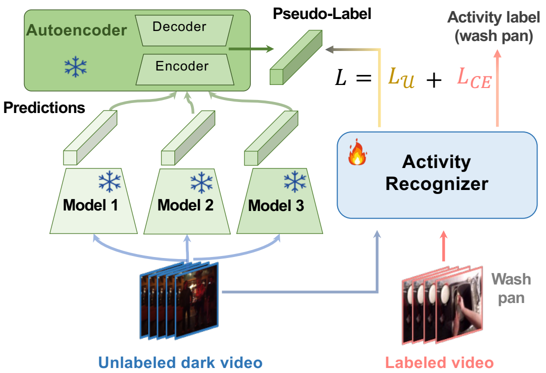
Single **distance** function as the loss

$$L_U = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

Model output
Pseudo-label

121

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

Single **distance** function as the loss

$$L_U = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

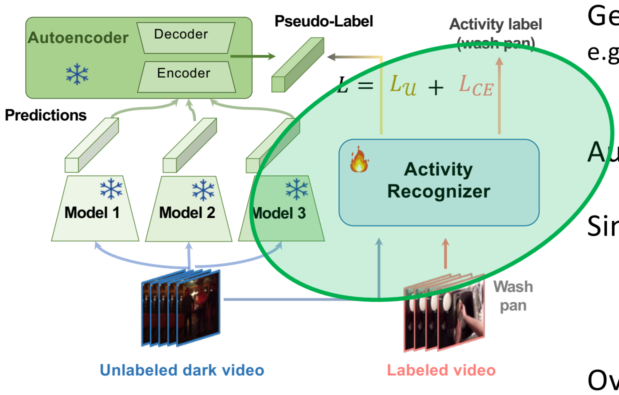
Model output
Pseudo-label

Overall training **objective:**

$$L = L_{CE} + \lambda L_U$$

122

I. Supervision beyond daylight



Generate **pseudo-labels** by auxiliary models
e.g. Video-text retrieval, sound source localization *etc.*

Autoencode predictions into **latent** pseudo-label

Single **distance** function as the loss

$$L_U = \sum_{j=1}^{U} dist(\hat{q}^j, q^j)$$

Model output

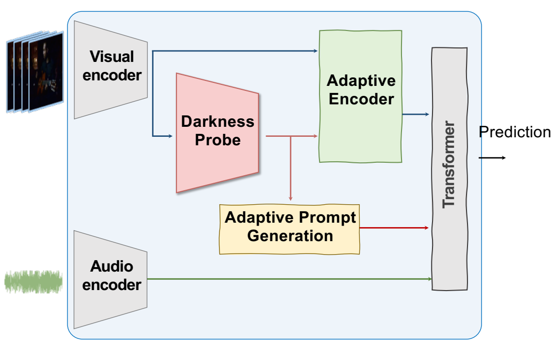
Pseudo-label

Overall training **objective:**

$$L = L_{CE} + \lambda L_U$$

123

II. Darkness-aware audio-visual recognition



Darkness probe produces an *n*-way branch attention to adapt to the current light condition

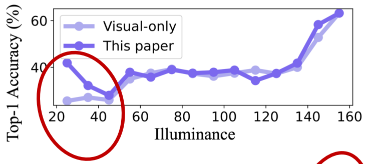
Adaptive encoder encodes the visual features according to perceived darkness

Adaptive prompt generation treats different light conditions as different tasks

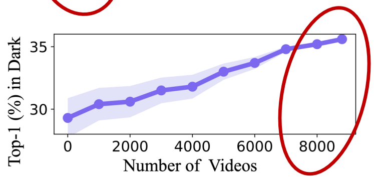
Transformer fuses adapted visual features, prompts and audio features

124

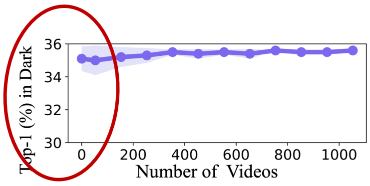
Properties of our proposal



Largest improvement for lowest illuminance.



The more unlabeled dark videos the better.

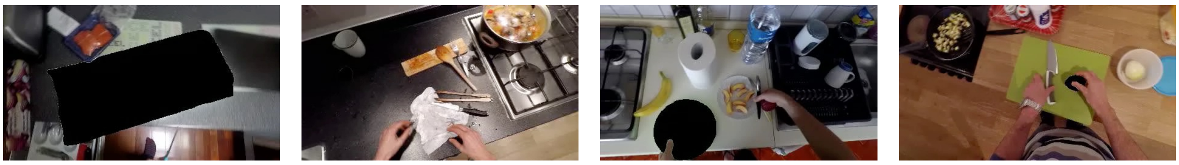


Even successful without labeled dark videos.

125

Bonus: also effective for occlusions

Tested on 182 EPIC-Kitchens videos with segmentation masks from Darkhalil et al.
We simulate occlusions by setting the pixel intensity of object regions to zero.



Visual encoder	26.4%
Vanilla multi-modal transformer	27.7%
This paper	***29.8%***

126

Comparison with image enhancement

Model	Venues	EPIC-Kitchens	
		Dark↑	GFLOPs↓
Vanilla multi-modal transformer		29.8	1.4
KinD	MM 2019	20.3	932.2
SCI	CVPR 2022	24.1	3.4
Unsupervised enhancement	ECCV 2022	26.4	108.8
LEDNet	ECCV 2022	27.8	312.0
This paper		**35.6**	**1.6**

We are superior to image enhancement for both accuracy and computation time.

127

Qualitative result for 'take box'

Illumination for dark frames improve, but color distortions harm activity recognition.

128

Failure case for 'pick up knife'

Vanilla audio-visual transformer

Verb prediction : ***put*** ✗
Noun prediction: ***leek*** ✗

This paper

Verb prediction: ***put*** ✗
Noun prediction: ***leek*** ✗

The right hand draws more attention than the left hand

129

Failure case: 'slapping'

Vanilla audio-visual transformer

prediction: ***laughing*** ✗
Confidence: 1.0

This paper

Prediction; ***laughing*** ✗
Confidence: 0.86

The environmental sound distracts the model

130

Key takeaways

Day2dark gap is wide-spread for multiple action recognition datasets and backbones.

Unlabeled dark videos and adaptively including **sound reduces the gap**.

Proposed model **outperforms image enhancement and alternative fusion** approaches.

131

3. Video generalization at test-time

132

3.a Generalize over unseen modality combo's



Yunhua Zhang
University of Amsterdam

Hazel Doughty
University of Amsterdam

Cees Snoek
University of Amsterdam

Learning Unseen Modality Interaction. *Submitted*.

133

Problem: Modality-complete assumption

Training

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Validation/Testing

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

134

Others: Robustness for modality-incomplete data

Training

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Validation/Testing

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Device 3
RGB

Device 4
Audio

One or more modalities could be missing during inference

Antoine Miech, et al. "Learning a text-video embedding from incomplete and heterogeneous data." In arXiv preprint 2018.
Mengmeng Ma, et al. "Smil: Multimodal learning with severely missing modality." In AAAI 2021.
Nina Shvetsova, et al. "Everything at once-multi-modal fusion transformer for video retrieval." In CVPR 2022.

135

Our goal: Recognize unseen modality-interactions

Training

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Validation/Testing

Device 1
RGB & Audio

Device 2
Audio & Optical Flow

Device 3
RGB & Optical Flow

136

Challenges

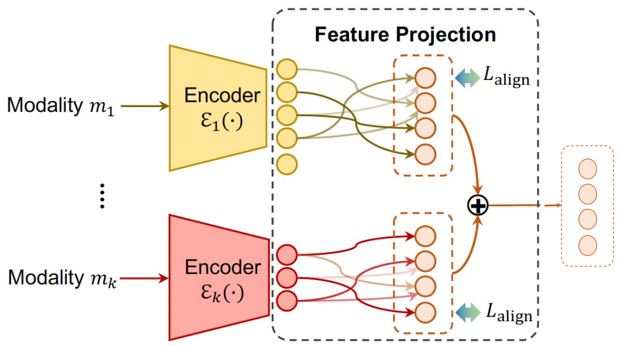
Simple concatenation of unimodal features cannot learn cross-modal correspondences when modality-complete data unavailable.

The accumulation should be agnostic for the order of modalities, and also allow for any modality combinations

Simple addition of unimodal features is hard as modalities come in different feature spaces and dimensionalities

137

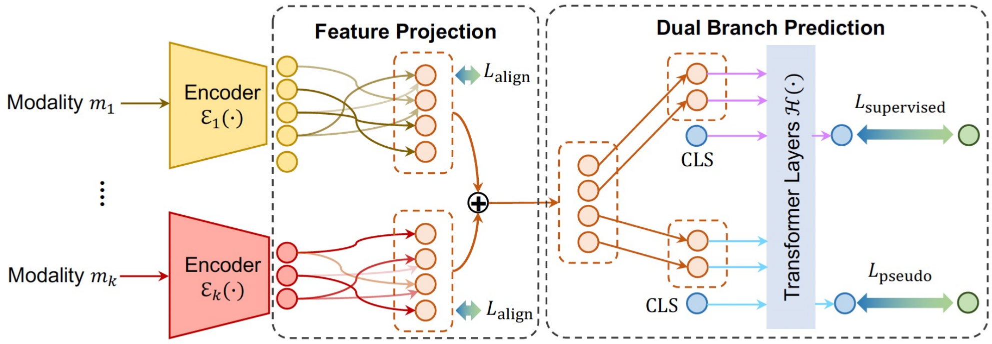
Approach



Project modality-specific features into a common space while maintaining differentiating information.

138

Approach

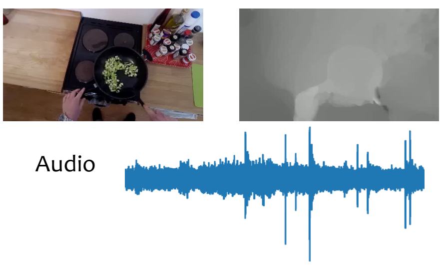


Some modalities are less discriminative than others and cause overfitting.
To mitigate overfitting, we introduce an additional pseudo-supervised branch.

139

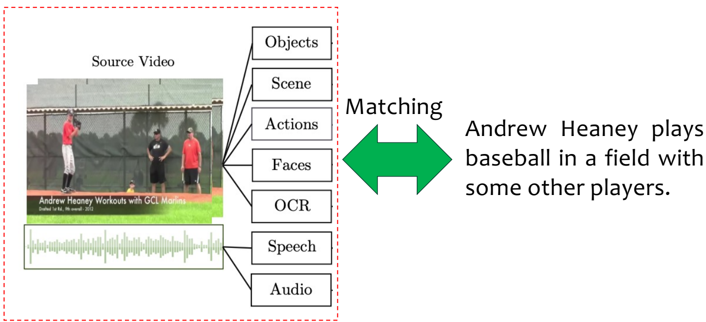
Evaluation

Video Classification



EPIC Kitchens with 3 provided modalities

Video Retrieval



MSR-VTT with 7 provided modalities

Note we define new splits to assure unseen modality interactions at test time

140

Results

	Video Classification	Video Retrieval
	Top-1 (%) ↑	MnR ↓
Late fusion	18.1	72.3
Modality Complete (Nagrani et al.)	17.5	86.2
Modality Incomplete (Recasens et al.)	18.5	72.2
Ours: *unseen modality interaction*	**23.7**	**66.2**

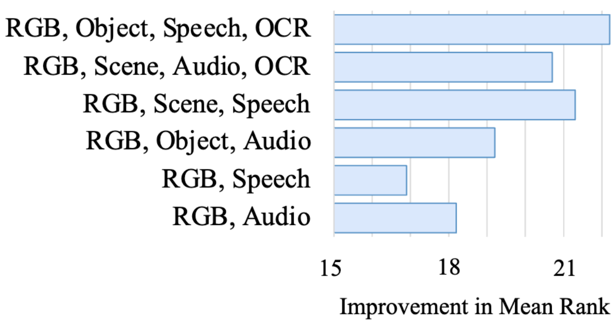
Without the need for modality-complete data, our method learns a more effective cross-modal fusion for unseen modality combinations

141

Robustness for modality incomplete at test-time

We show the improvement over a vanilla multimodal transformer.

Multimedia Retrieval



Our model can handle any input modality

Improves robustness for all unseen combo's

Our model most effective for more modalities

142

Key takeaways

We can effectively make **predictions for unseen modality interactions** by feature projections and pseudo-supervision

Our approach is **suitable for classification, regression and retrieval**, and can handle a wide variety of modality combinations

143

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, ellis.social}

144

Key references

- Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, Cees G M Snoek: **How Severe is Benchmark-Sensitivity in Video Self-Supervised Learning?.** In: ECCV, 2022.
- Fida Mohammad Thoker, Hazel Doughty, Cees G M Snoek: **Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization.** In: ICCV, 2023.
- Piyush Bagad, Makarand Tapaswi, Cees G M Snoek: **Test of Time: Instilling Video-Language Models with a Sense of Time**. In: CVPR, 2023.
- Yunhua Zhang, Hazel Doughty, Ling Shao, Cees G M Snoek: **Audio-Adaptive Activity Recognition Across Video Domains**. In: CVPR, 2022.
- Yunhua Zhang, Hazel Doughty, Cees G M Snoek: **Day2Dark: Pseudo-Supervised Activity Recognition beyond Silent Daylight.** *Submitted.*
- Yunhua Zhang, Hazel Doughty, Cees G M Snoek: **Learning Unseen Modality Interaction**. *Submitted.*

145