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: Serge Belongie, University of Copenhagen

http://computervisionbylearning.info

Program

Monday	Fundamentals
Tuesday	Computer vision by learning
Wednesday	Machine learning for computer vision
Thursday	Computer video by learning
Friday	Invited tutorial by Serge Belongie


Serge Belongie

Guest speakers


Subhransu Maji


Martin Oswald


Erik Bekkers


Yuki Asano


Hazel Doughty

Reminder: Where and When

Monday 9th of May to Thursday 12th of May

Lectures	09:30-12:15	CASA – theater room
Lunch	12:15-13:30	*included*
Lab	13:30-17:00	CASA – 3 lab rooms

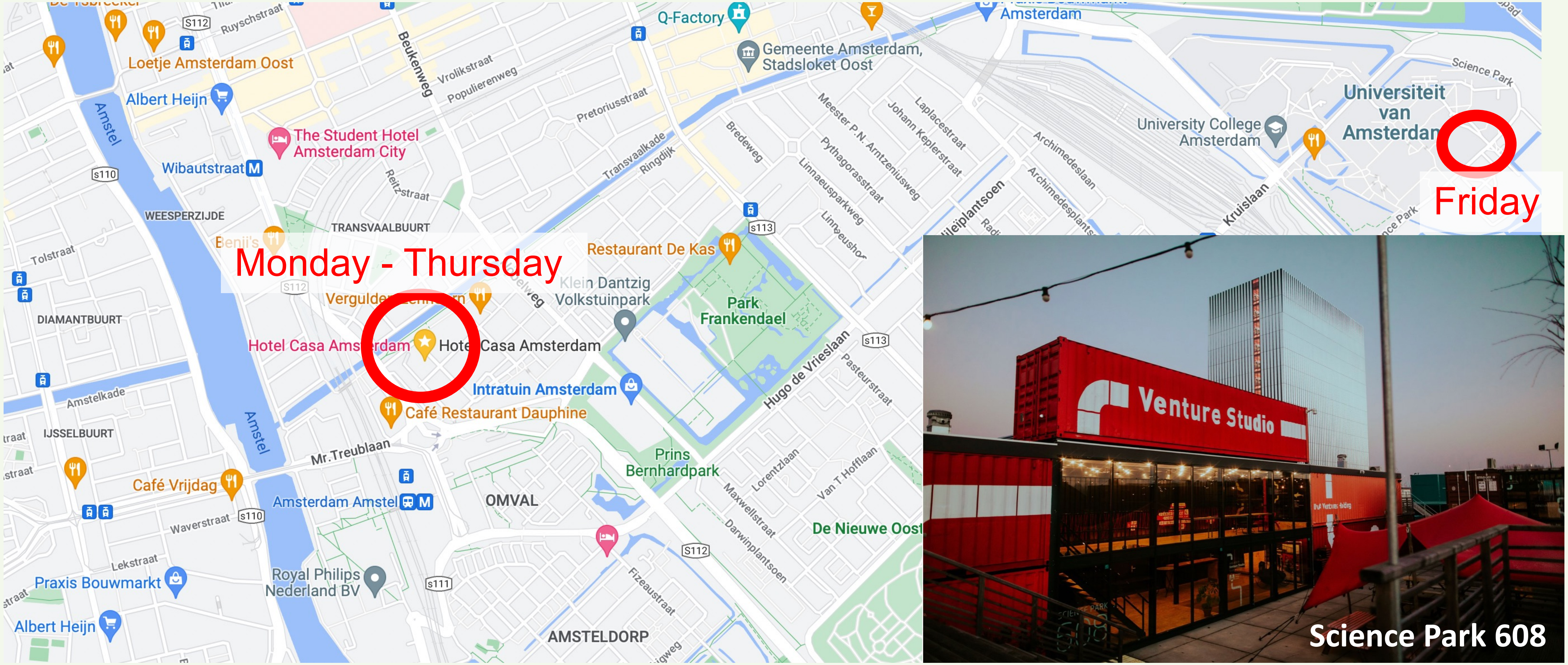
Thursday 12th of May

Borrel	17:00-18:00	CASA

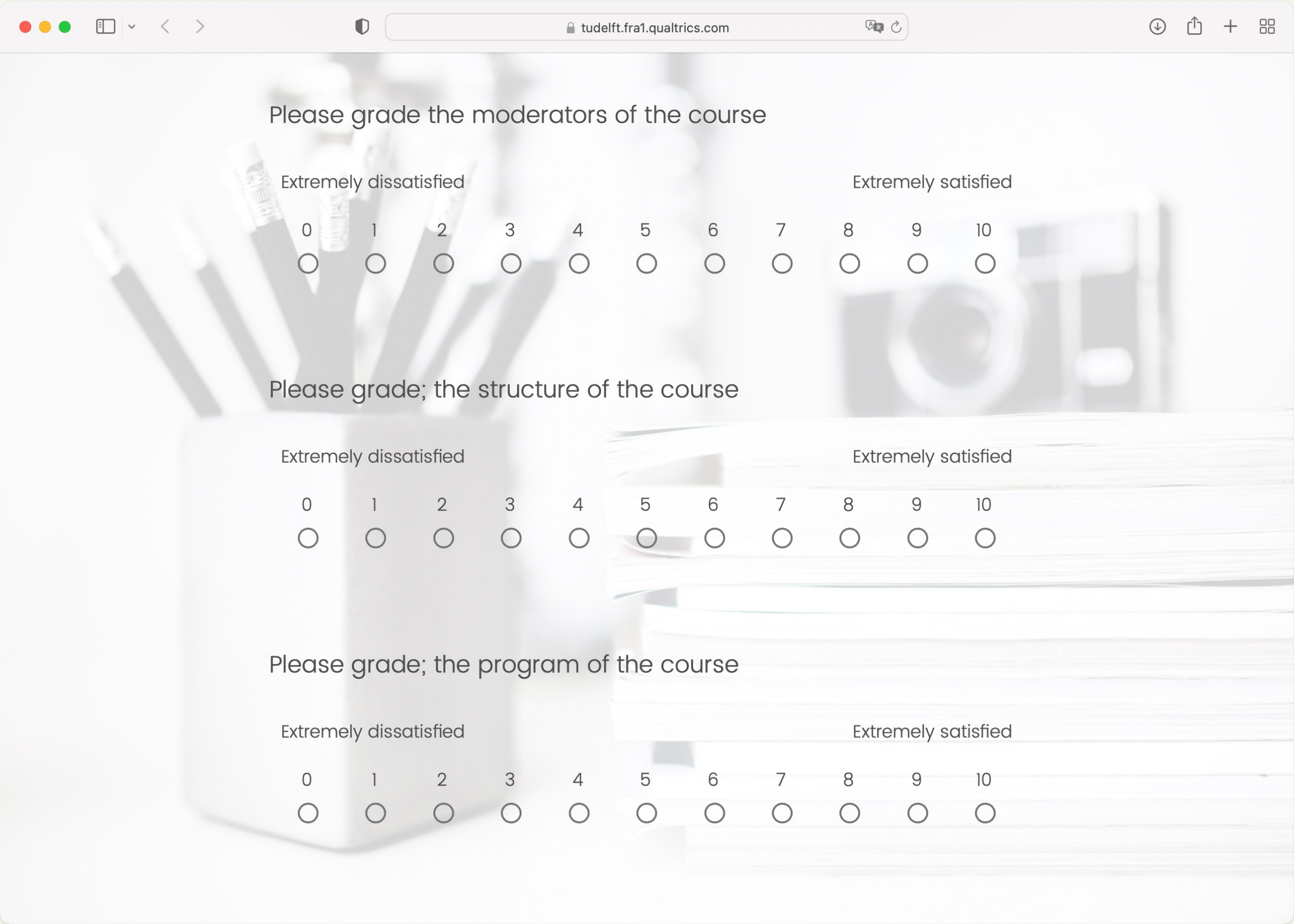
Friday 13th of May

Invited tutorial	09:30-12:15	Startup Village – Venture studio
Closing	12:15-12:30	

Reminder: Map



Your feedback on the course



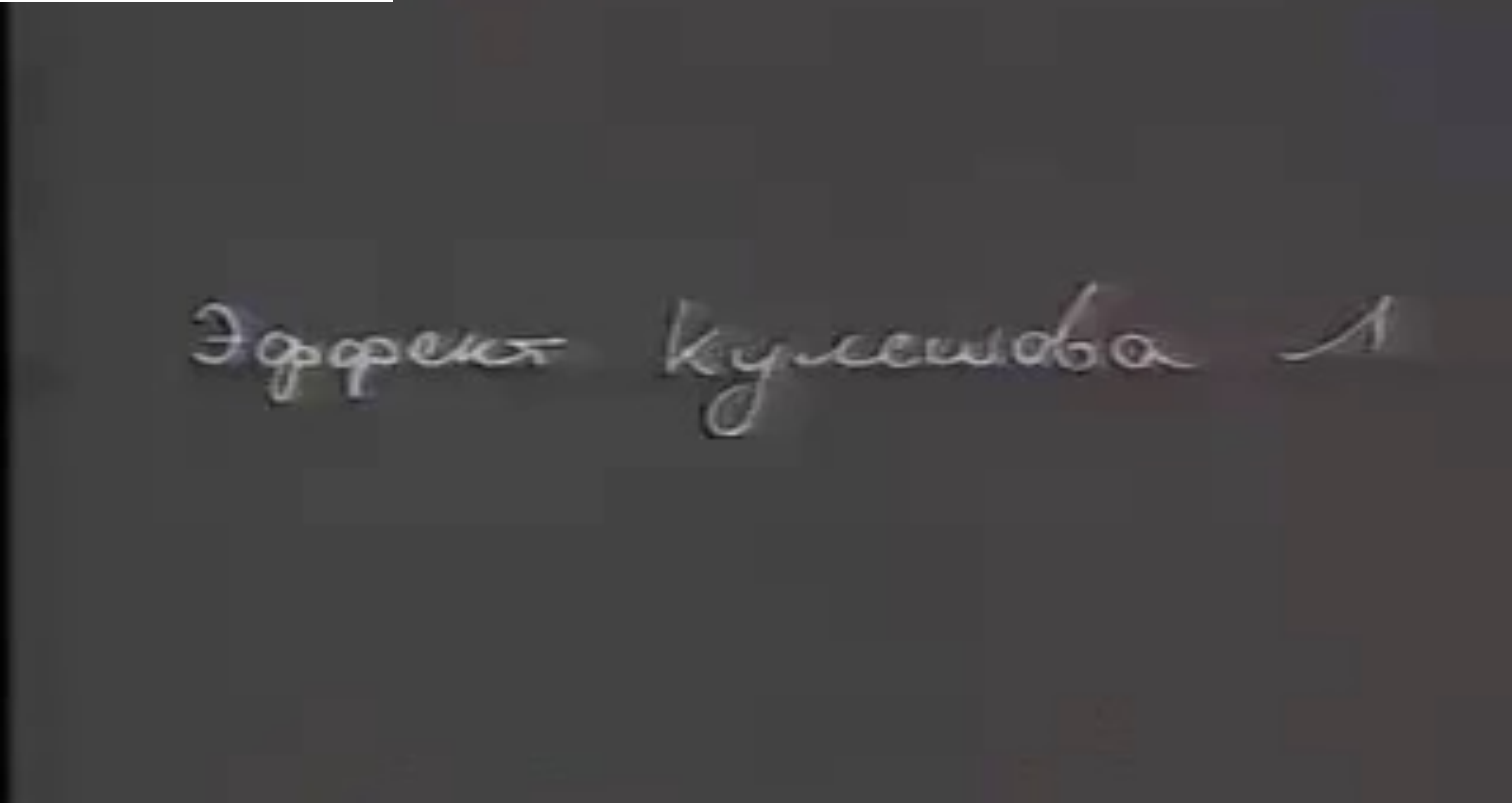
Beyond spatial classification


1 ice_skating:0.98
2 speed_skating:0.01
Tran et al., ICCV 2015

Motivating question for this tutorial

Is video more than the sum of its individual images?

by Lev Kuleshov 1899-1970



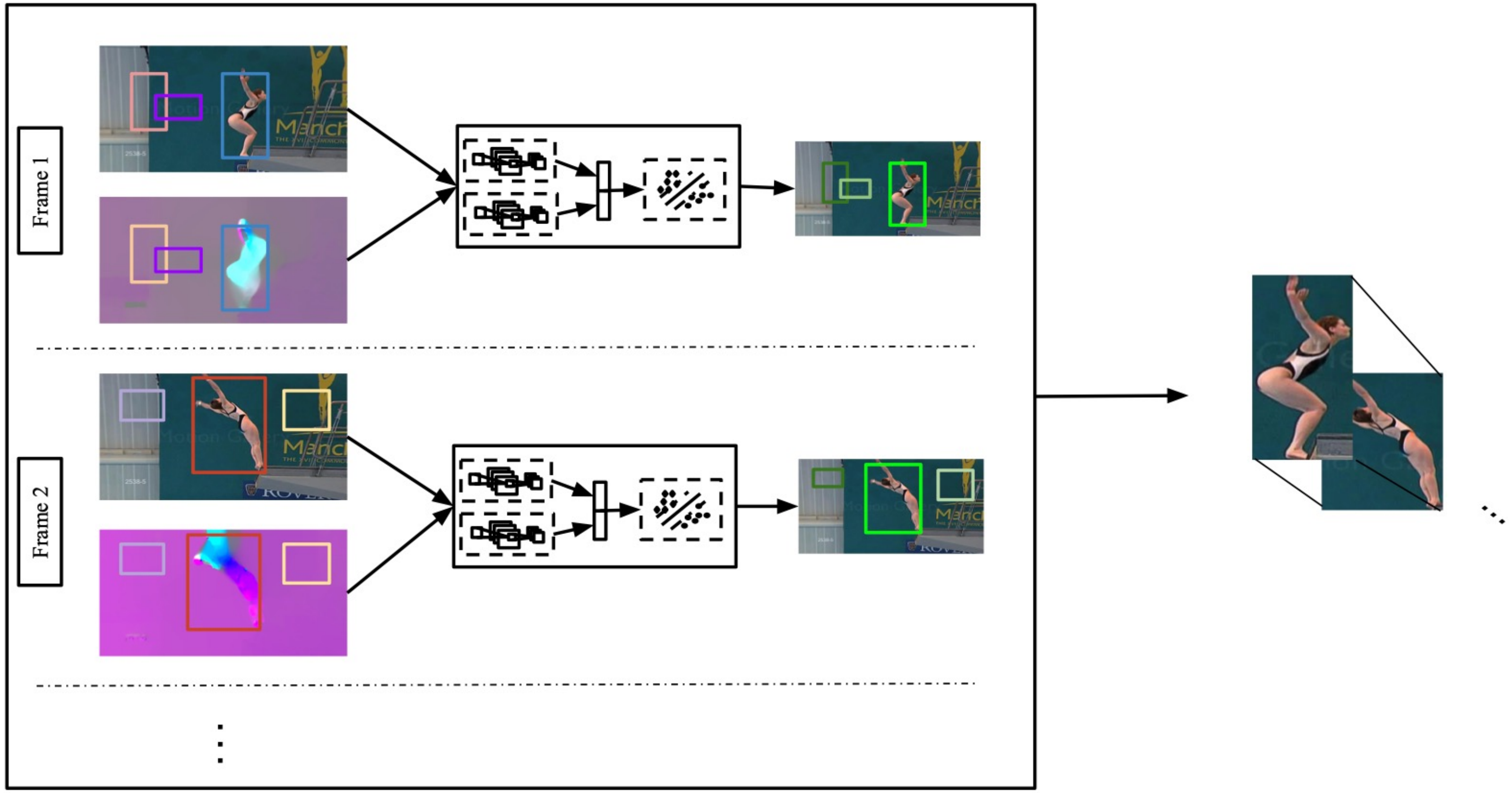
frame 2201 of 9075
vehicle_moving 19
Vehicle_5
activity_walking 61
Interacts 8
Other_5005
Other_5000
Other_5001
Other_5002
Other_5003

Overview

1. **Spatial then temporal**, person detection, linking, attention
2. **Spatial and temporal**, tubelets, convnet, transformer
3. **Spatial and temporal and sound**, repetition count, domain adaptation.

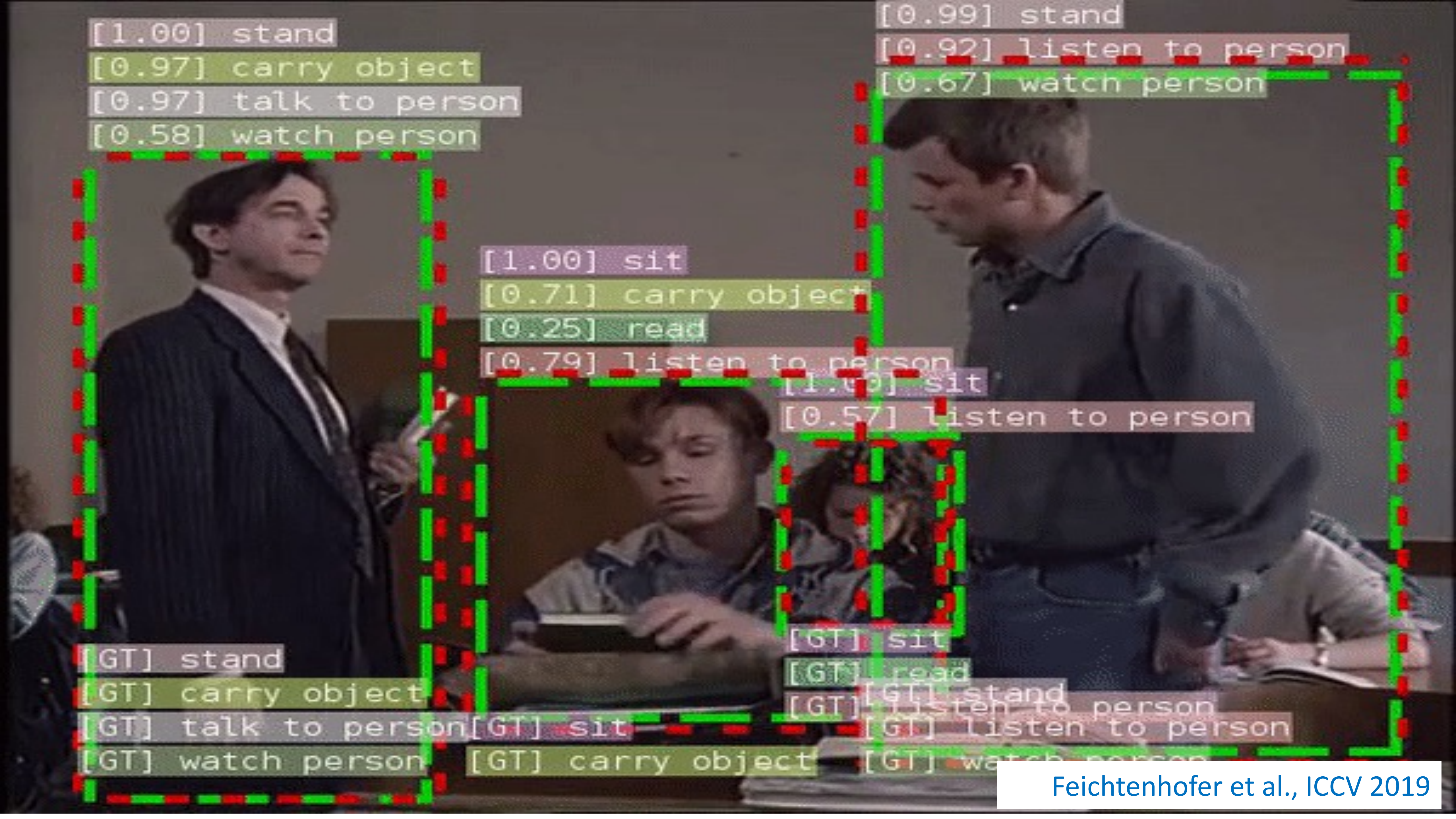
1. Spatial then temporal

Finding action tubes



Detection per video frame

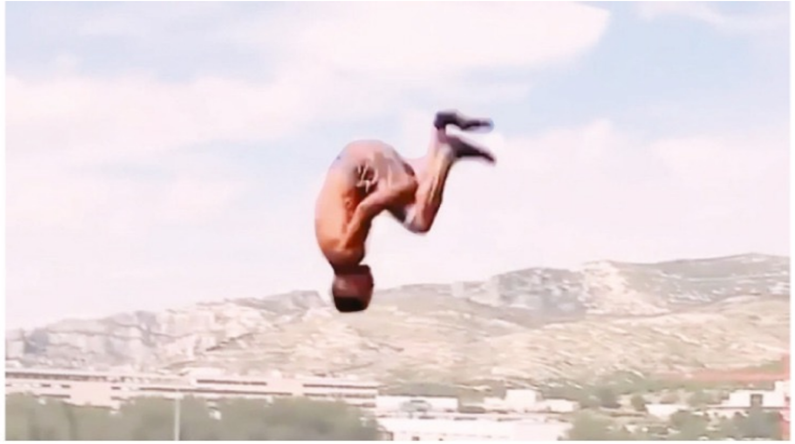
Link detections by dynamic programming



Feichtenhofer et al., ICCV 2019

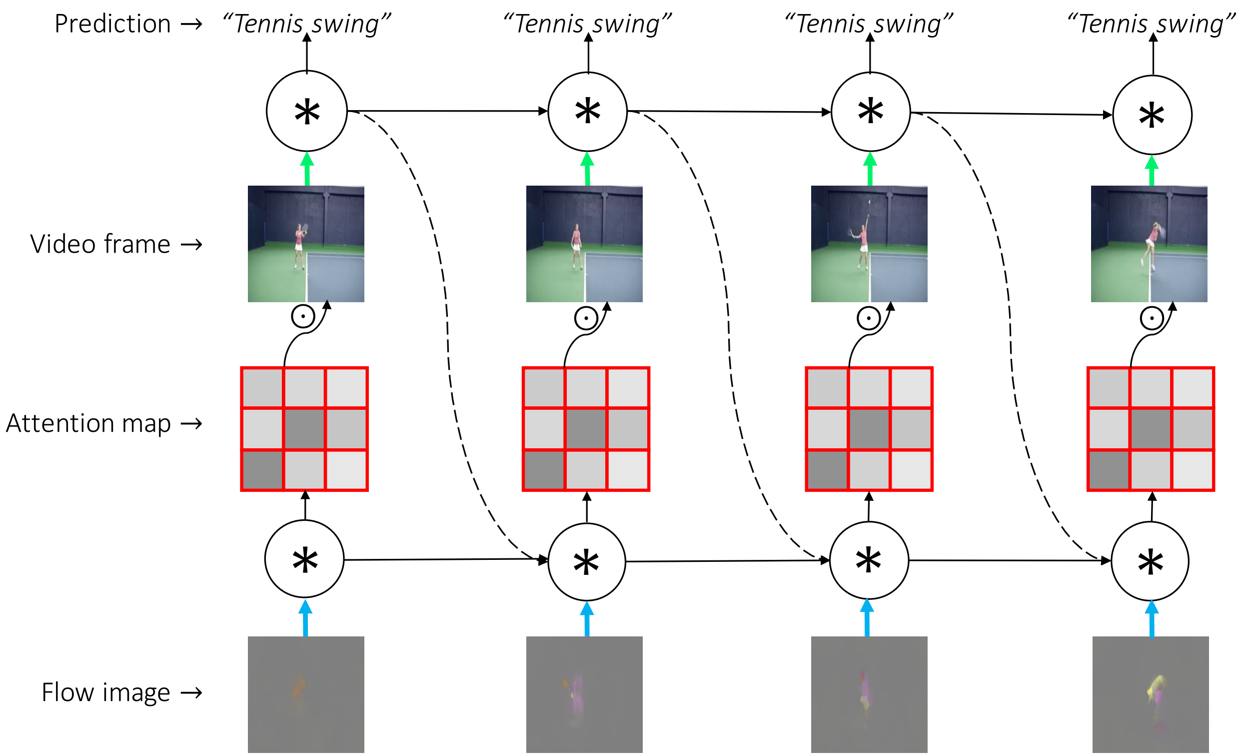
Escorcia *et al.*, CVIU 2020

Siamese linking of spatial detectors



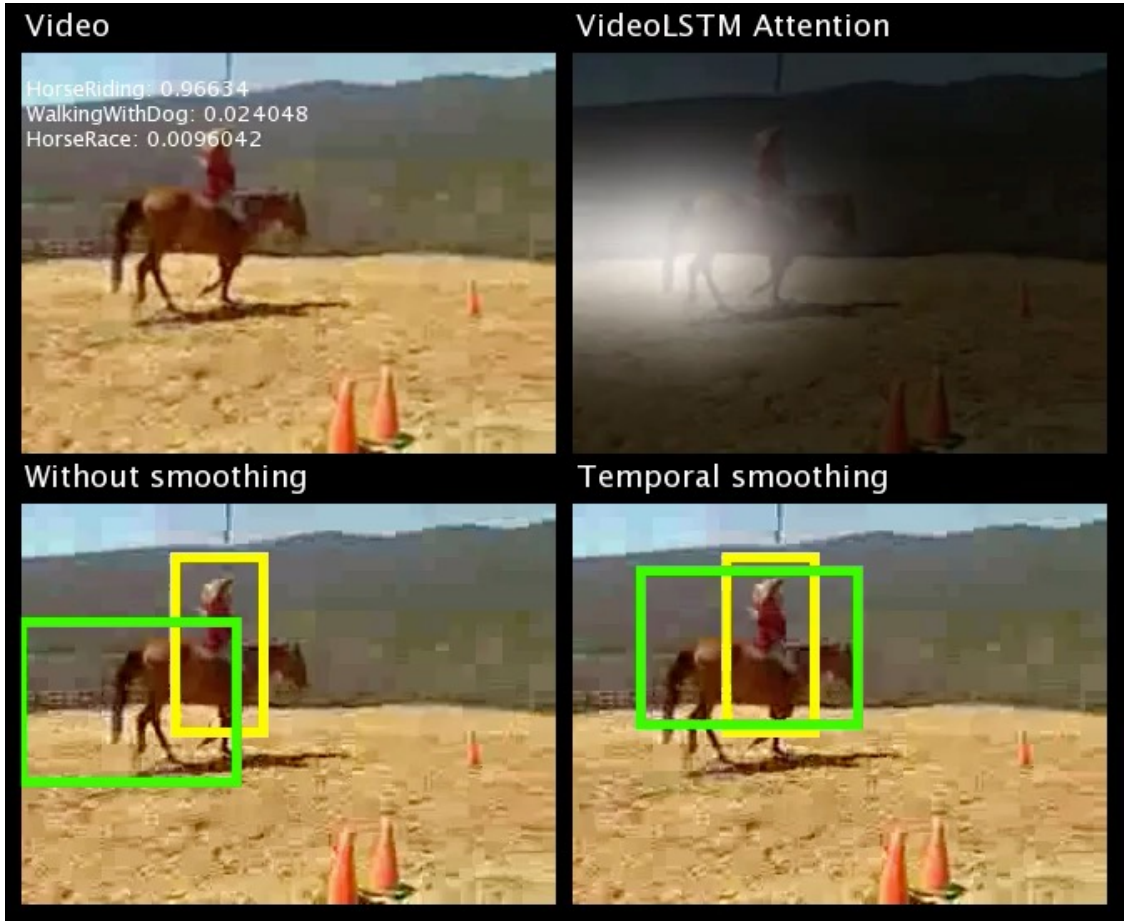
Li *et al.*, CVIU 2018

VideoLSTM convolves, attends and flows



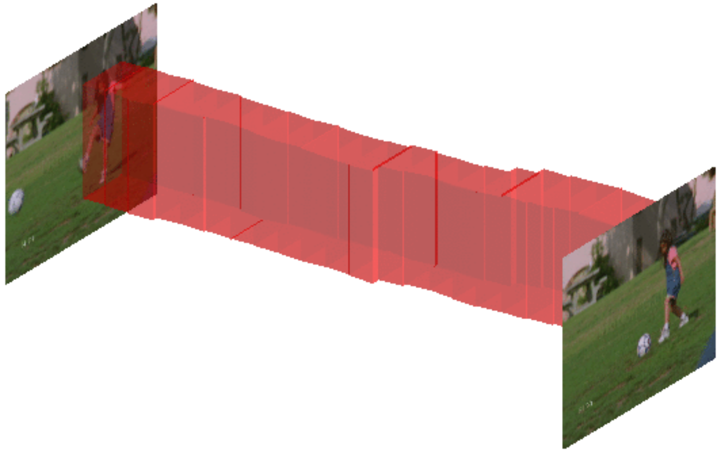
Enable action localization from action class labels only

Temporal smoothing by linear regression

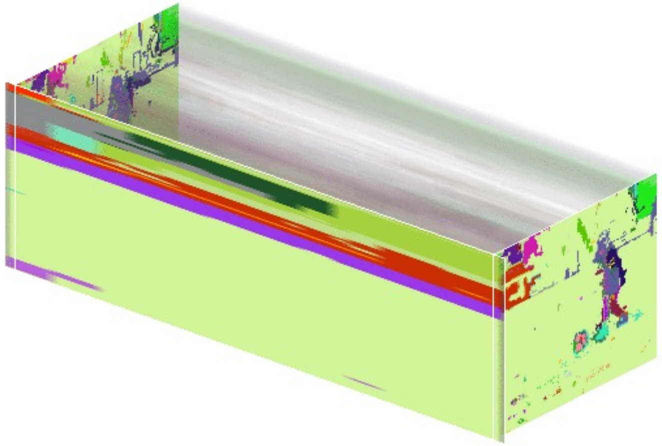


2. Spatial and temporal

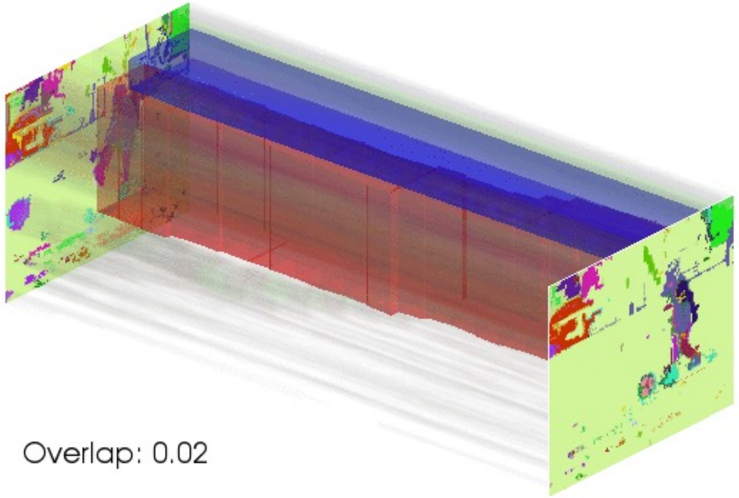
Tubelets: unsupervised activity proposals



Ground truth



Super-voxel segmentation



Proposals from merged voxels

Jain *et al.*, CVPR 2014 / IJCV 2017

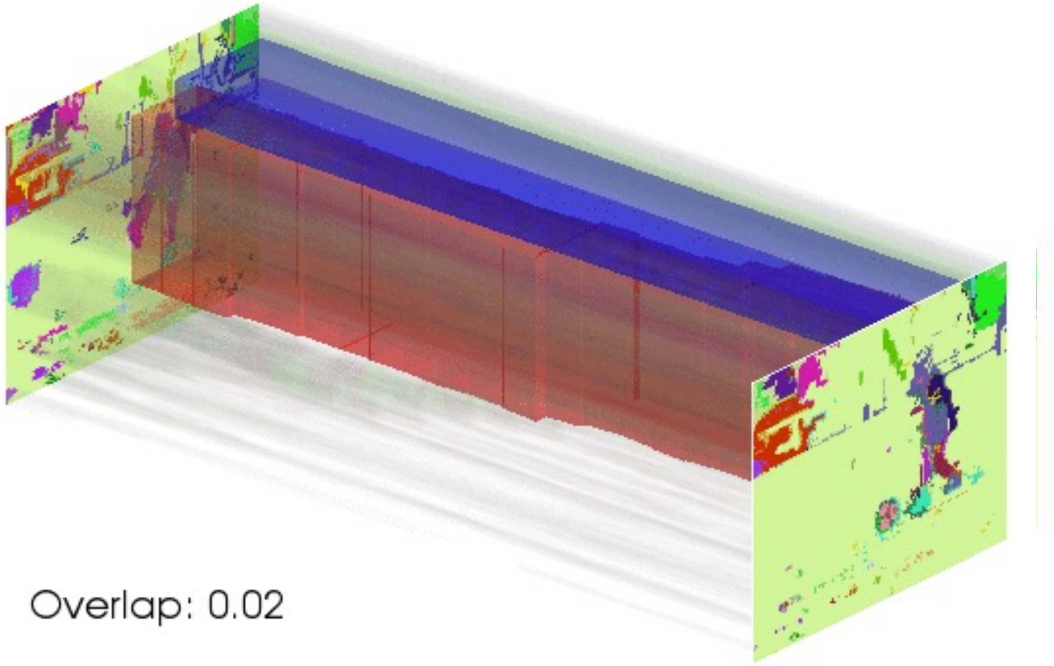
Tubelets: unsupervised activity proposals

Analyze **space and time jointly** to obtain action proposals

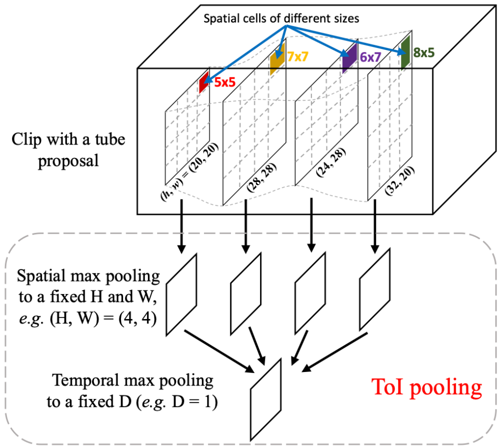
Action-class agnostic, covers variable aspect ratios and temporal lengths

Relies on **supervoxels**

High recall with few proposals

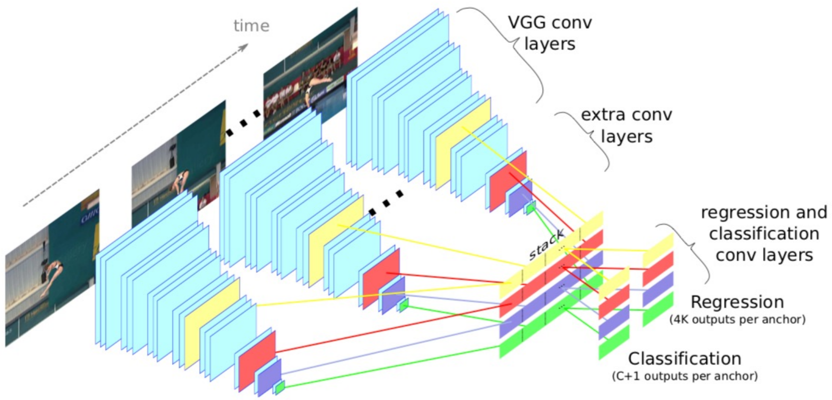


Tube Convolutional Neural Network



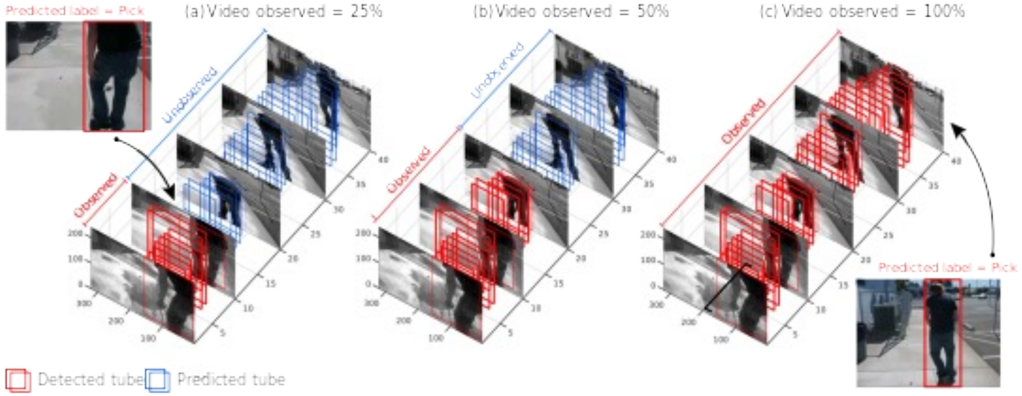
Hou *et al.* ICCV 2017

Action Tubelet Detector



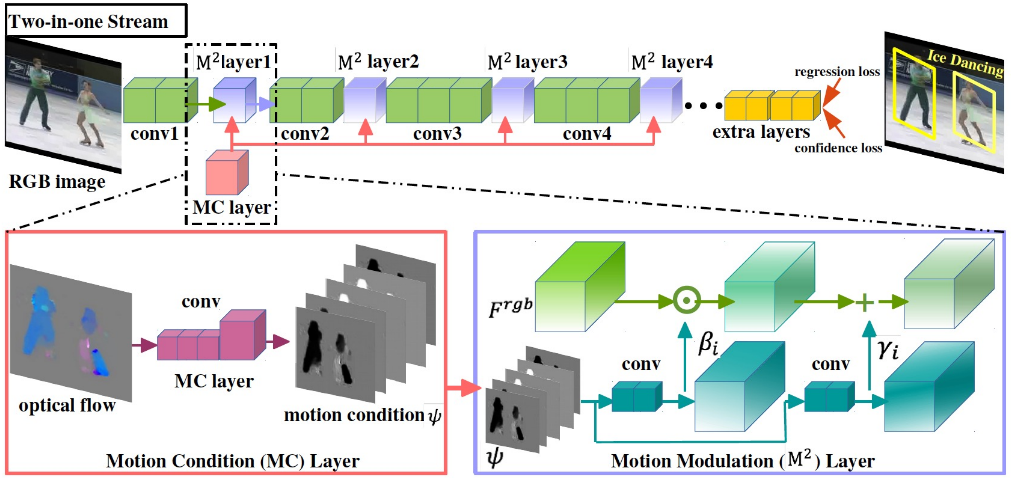
Kalogeiton *et al*, ICCV, 2017

Predicting Action Tubes



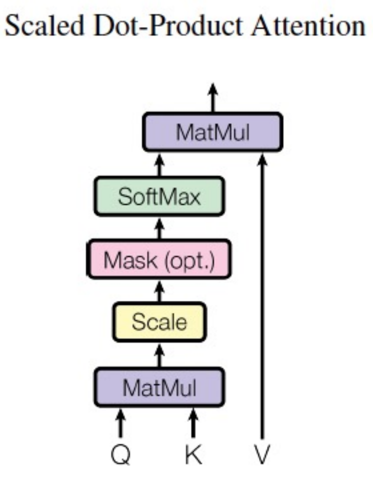
Singh *et al.*, ECCVw 2018

Two-in-One Stream

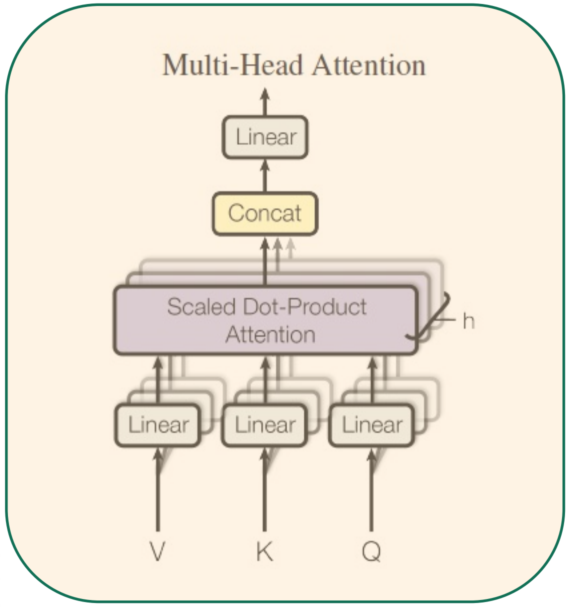


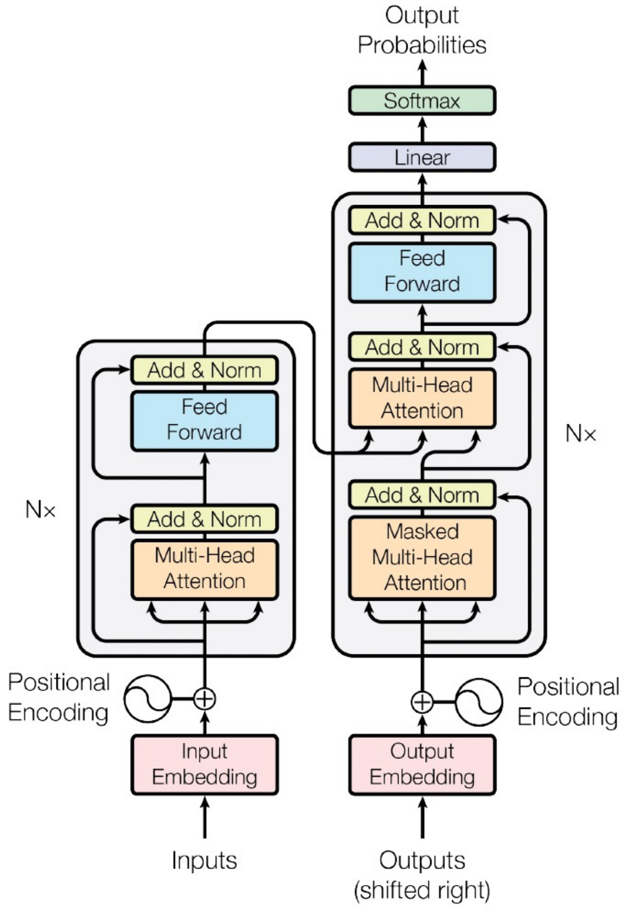
Zhao & Snoek, CVPR 2019

What about transformers?



$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





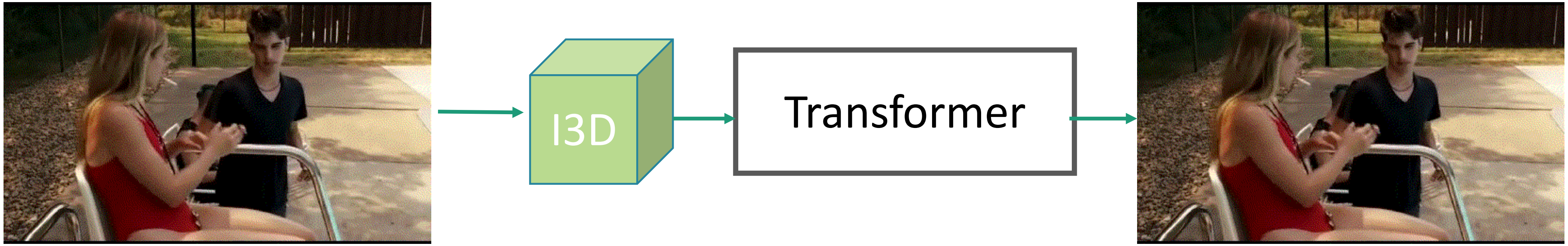
TubeR: Tubelet Transformer for Video Action Detection

Jiaojiao Zhao
University of Amsterdam

Joint work with Yanyi Zhang, Xinyu Li, Hao Chen, Shuai Bing, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G M Snoek, Joseph Tighe, while at Amazon internship.

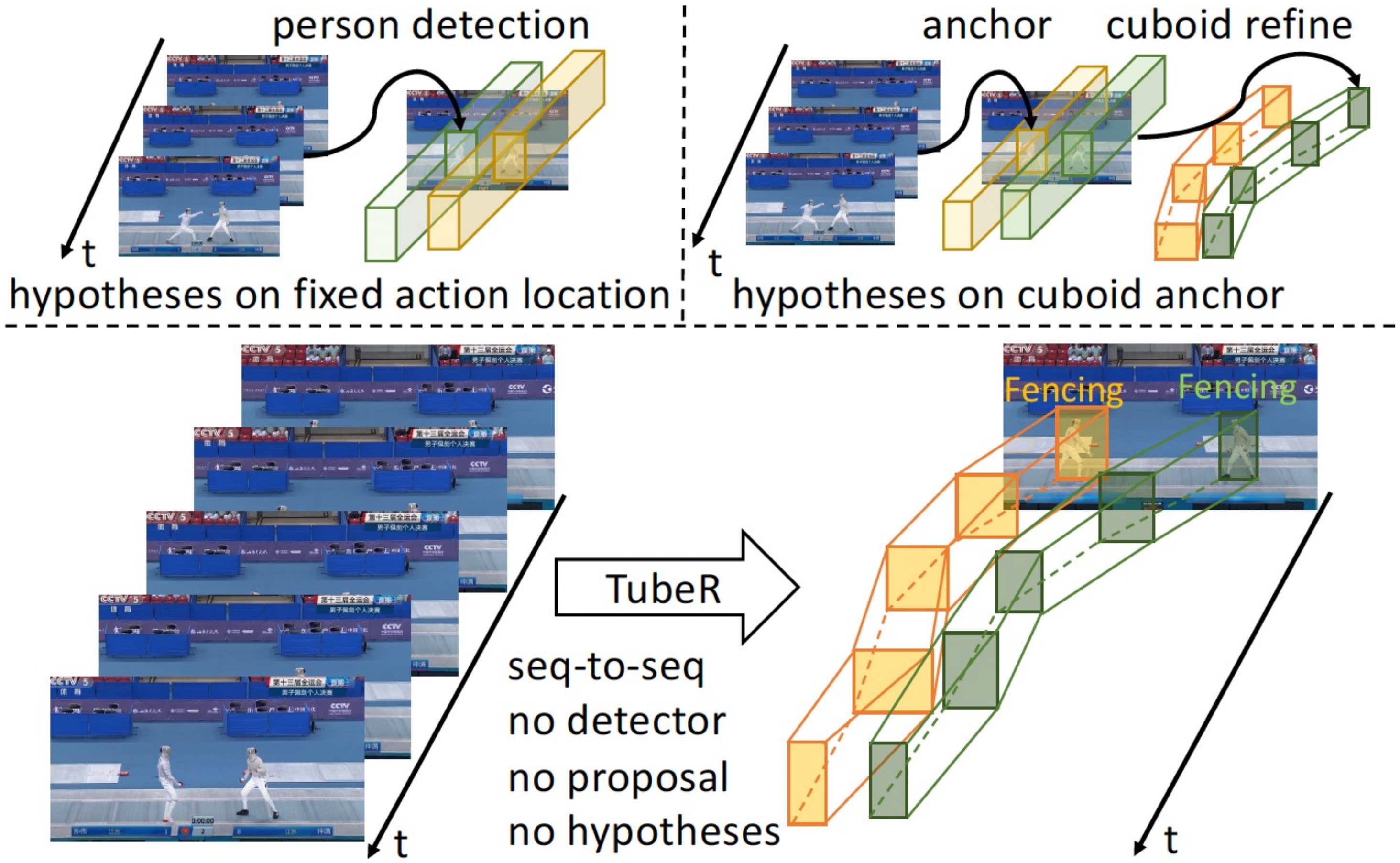
To appear in CVPR 2022 (oral).

TubeR: Tubelet Transformer for Action Detection

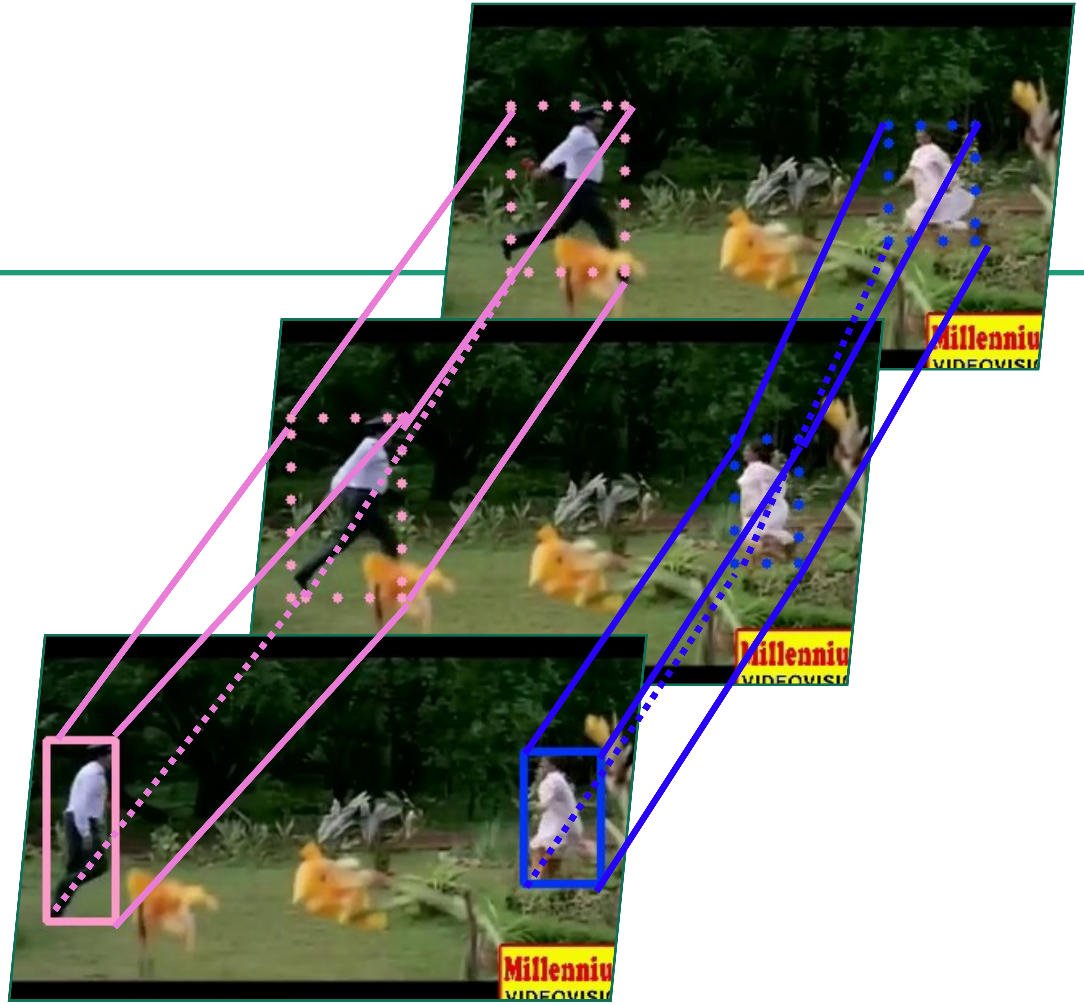


Allows each 2D+t position to attend to all other 2D+t positions in a video clip, which is essential for modeling action relations.

Advantages of transformer



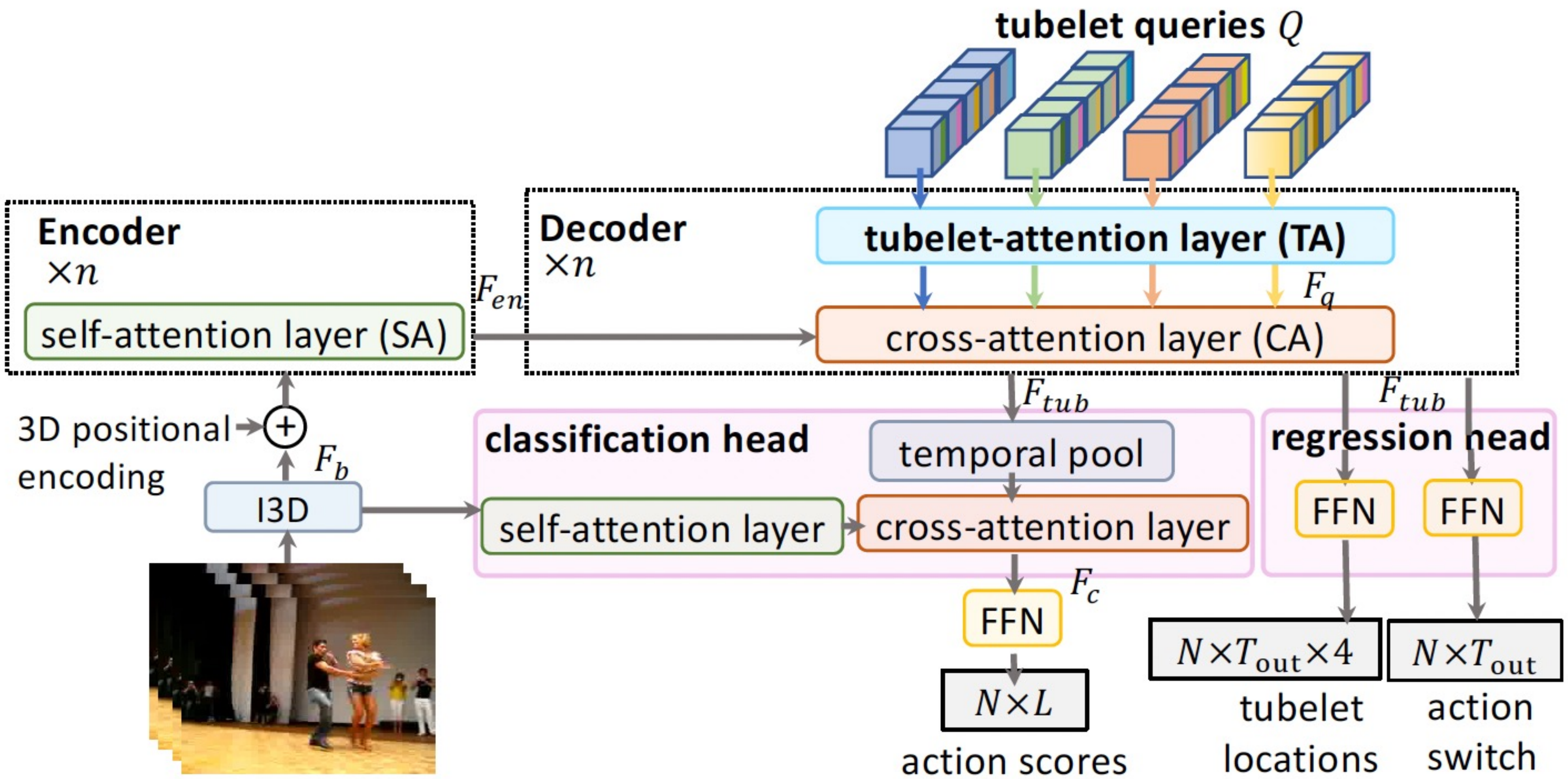
Motivation



The self-attention mechanism facilitates the exchange of boxes between frames, which helps to form action tubelets

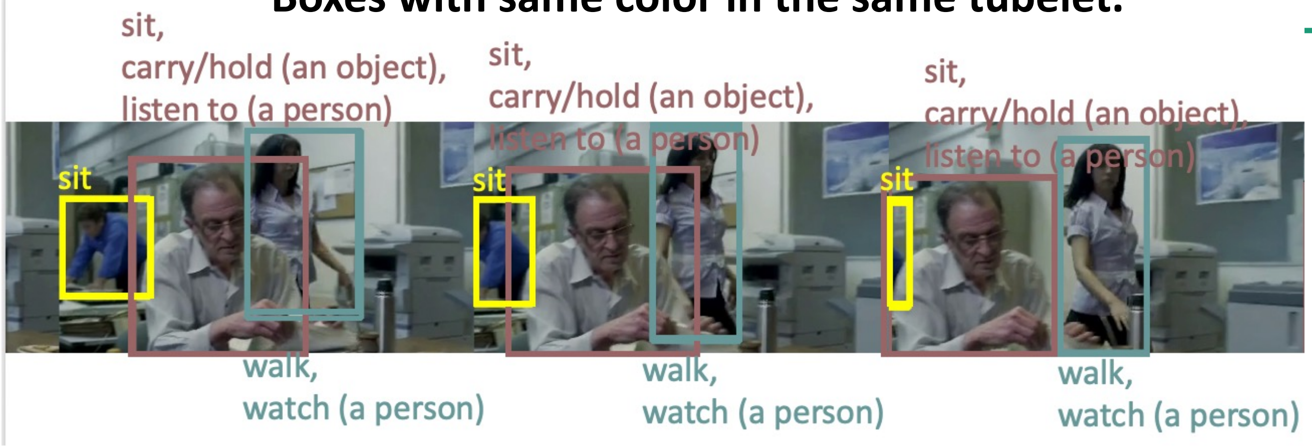
Big picture

Three contributions: **Tubelet query**, **tubelet attention** layer and **task-specific heads**

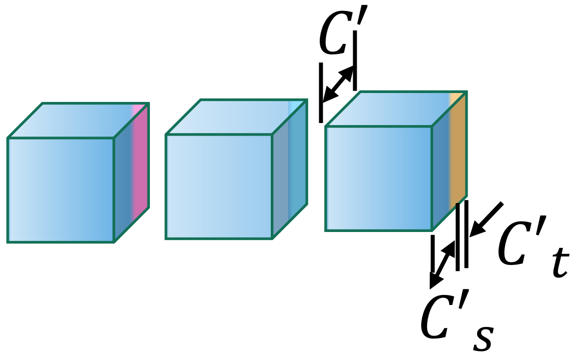


i. Tubelet query

Boxes with same color in the same tubelet.



Tubelet query

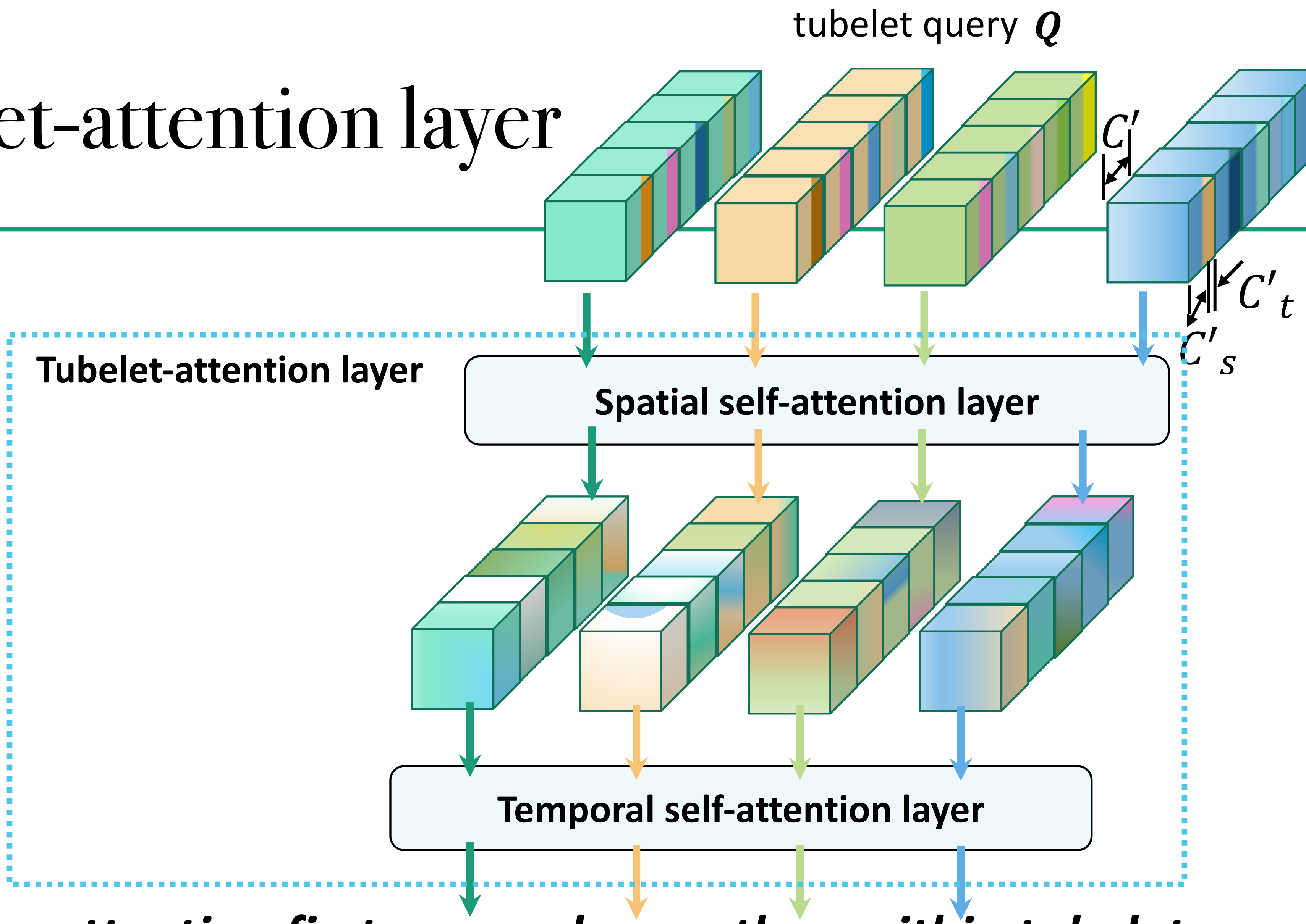


Each tubelet query consists of T box queries.

Box queries share the identity feature C_s' for the visual similarity and have independent features C'_t to capture changes over time.

Without the identity feature, a tubelet is not automatically formed.

ii. Tubelet-attention layer



Perform attention first among boxes, then within tubelets.

iii. Task-specific heads

Context-aware classification

Short-term: query action-specific feature with short-term (global) context

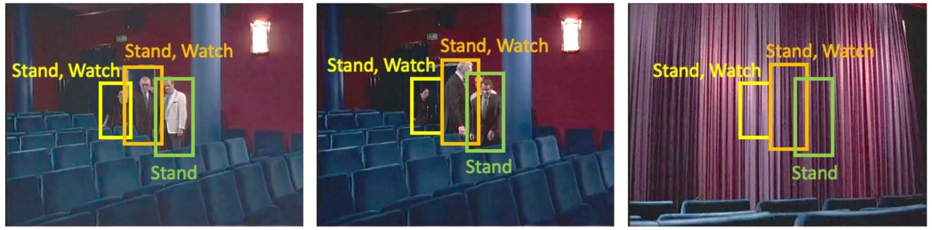
Long-term: buffer containing the backbone feature extracted from a long clip

Action switch regression

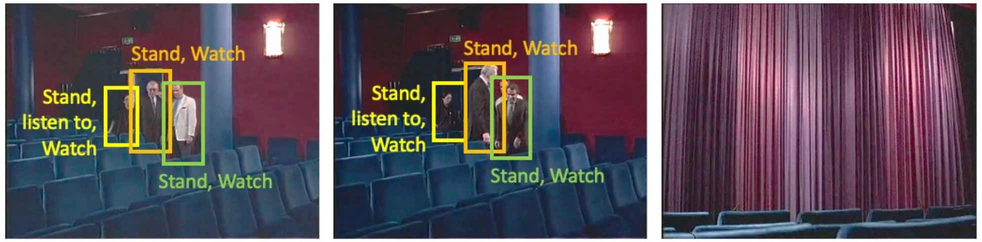
FC layer to decide whether a box prediction depicts an action

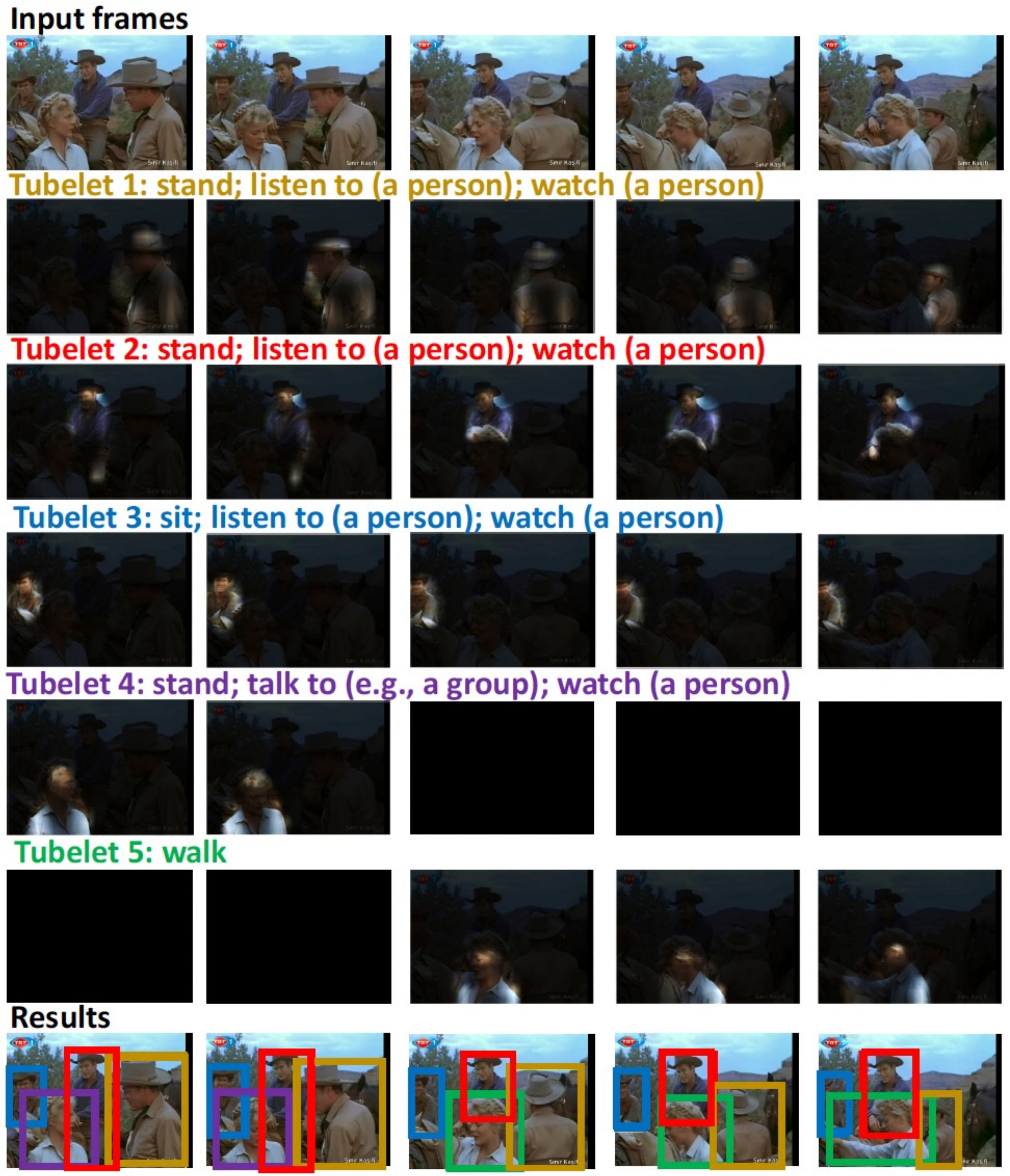
Allows to generate action tubelets with a more precise temporal extent.

Without switch



With switch

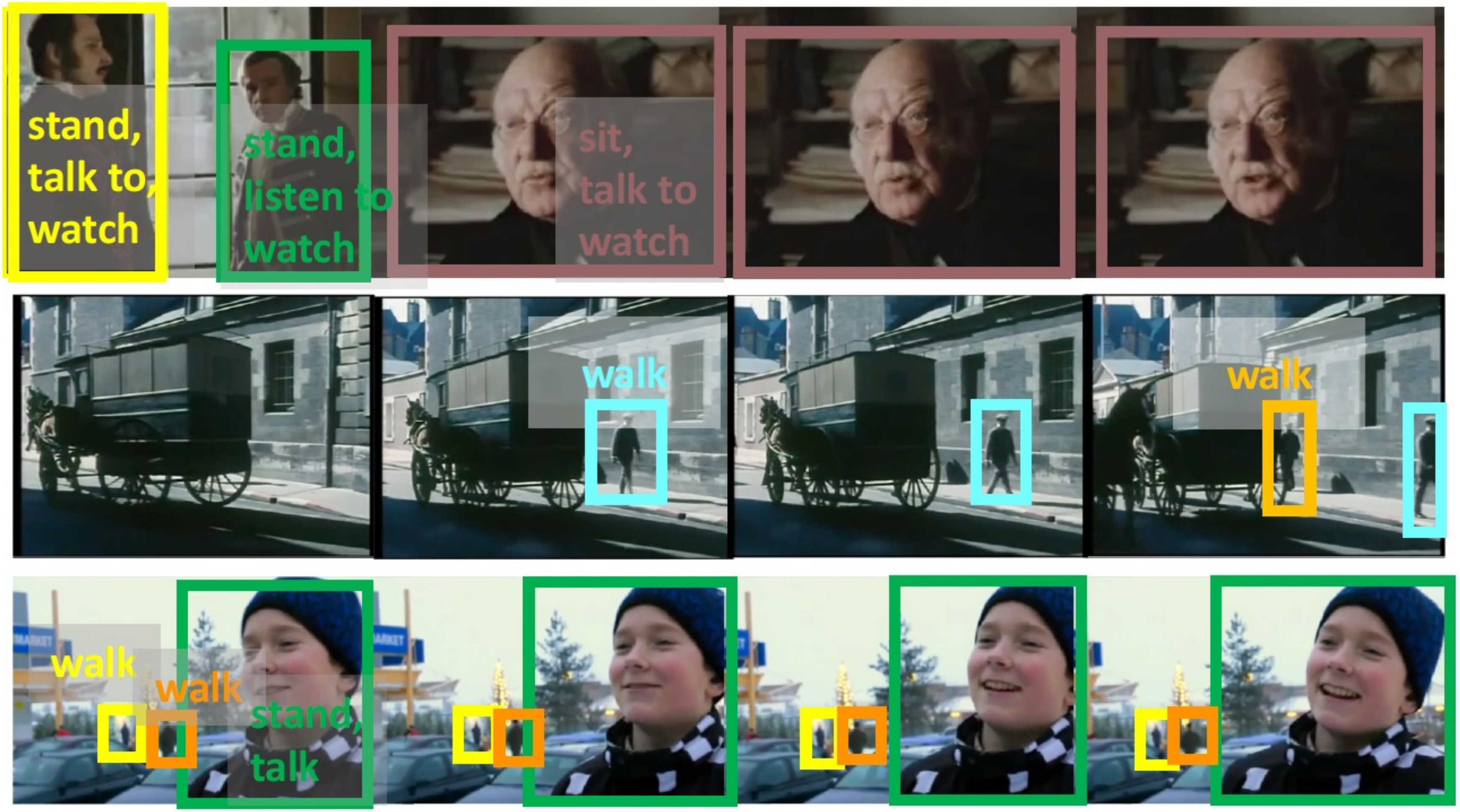




TubeR-behavior

Each tubelet covers a separated action instance

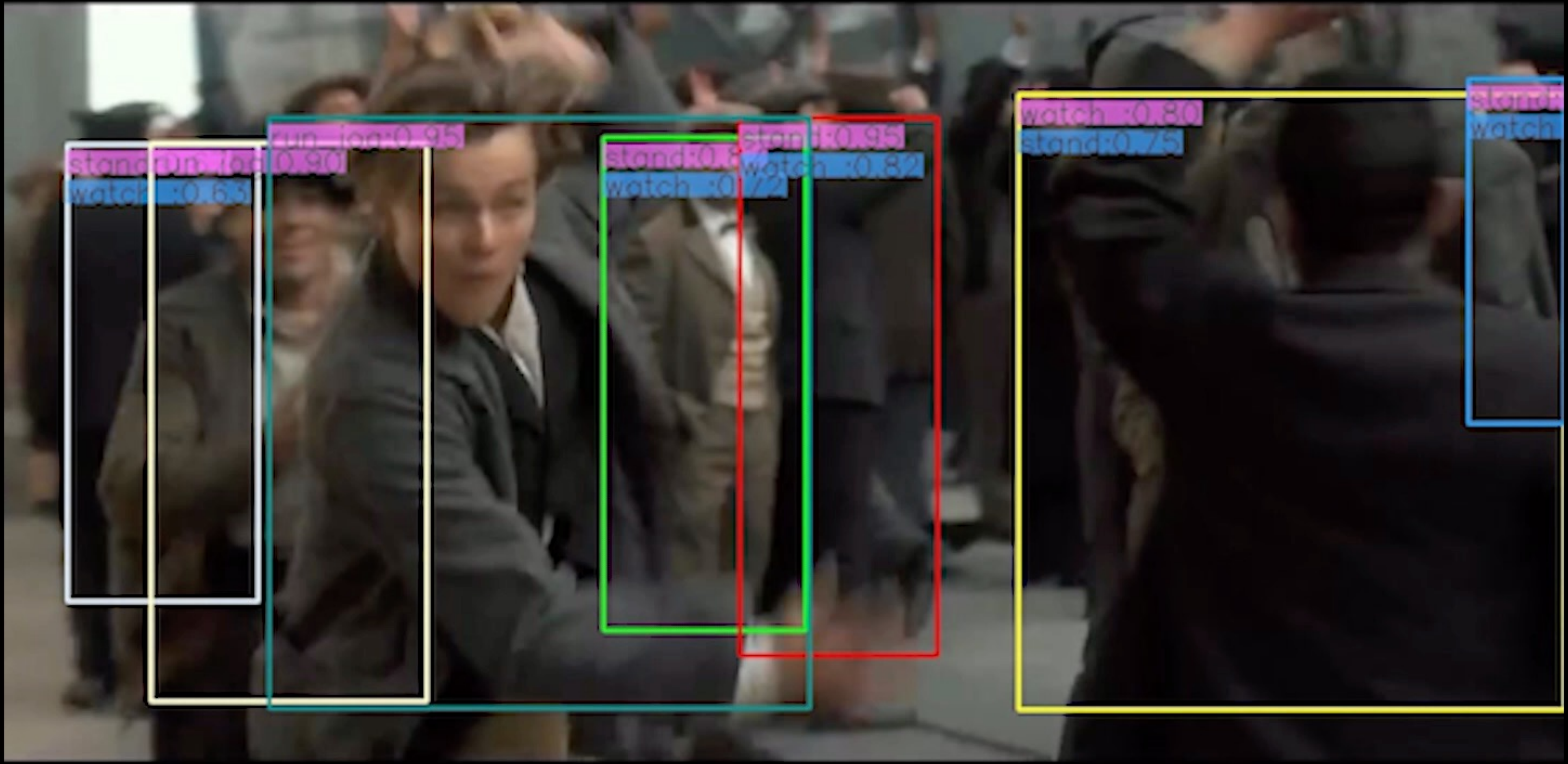
Qualitative results



Shot changes

Occlusions

Scale changes



Jiaojiao Zhao et al., CVPR 2022

3. Spatial and temporal and sound

Repetitive Activity Counting by Sight and Sound

Yunhua Zhang
University of Amsterdam

Ling Shao
Inception Institute of AI

Cees Snoek
University of Amsterdam

In *CVPR* 2021.

Repetitive motion

Sports

Music

Urban

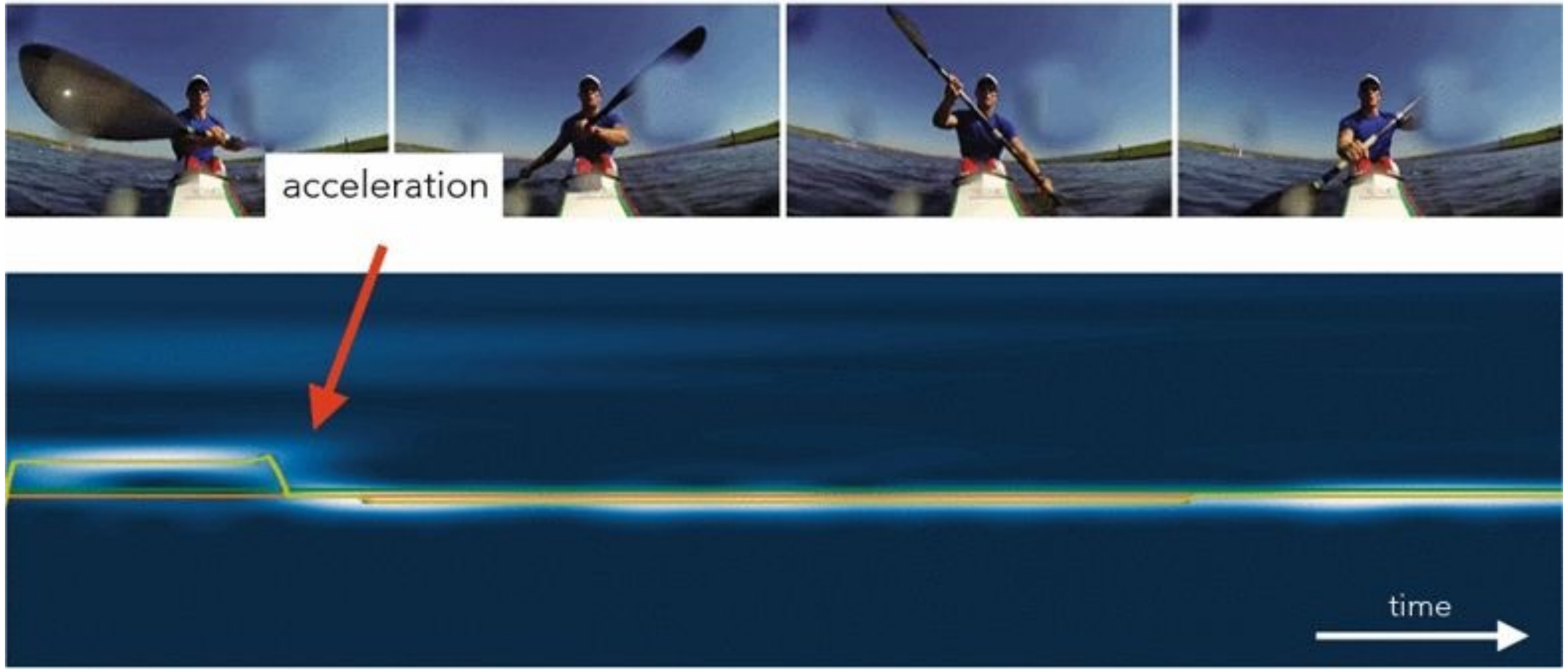
Natural environments

Cutler & Davis, PAMI 2000
Pogalin et al. CVPR 2008

Stationary world

Represent video as one-dimensional fixed-period Fourier signal that preserves repetitive motion structure

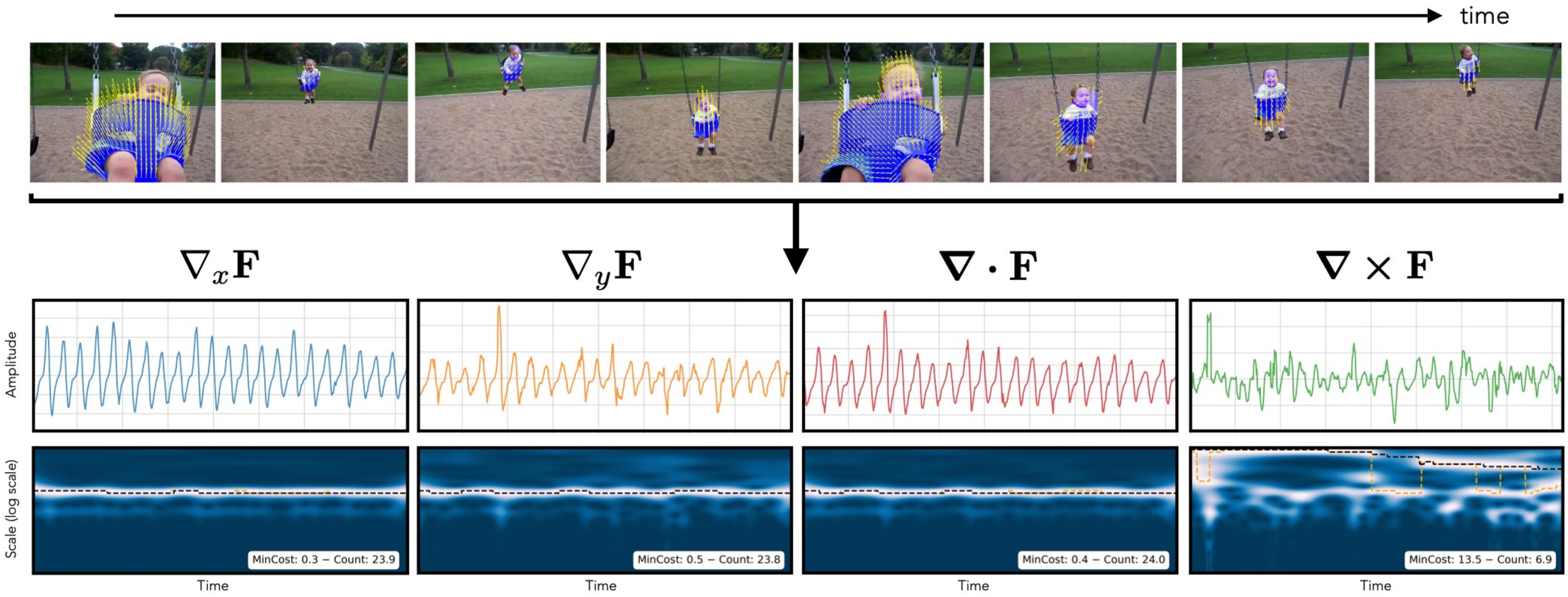
Had to assume **static and stationary** video, inapt for real world



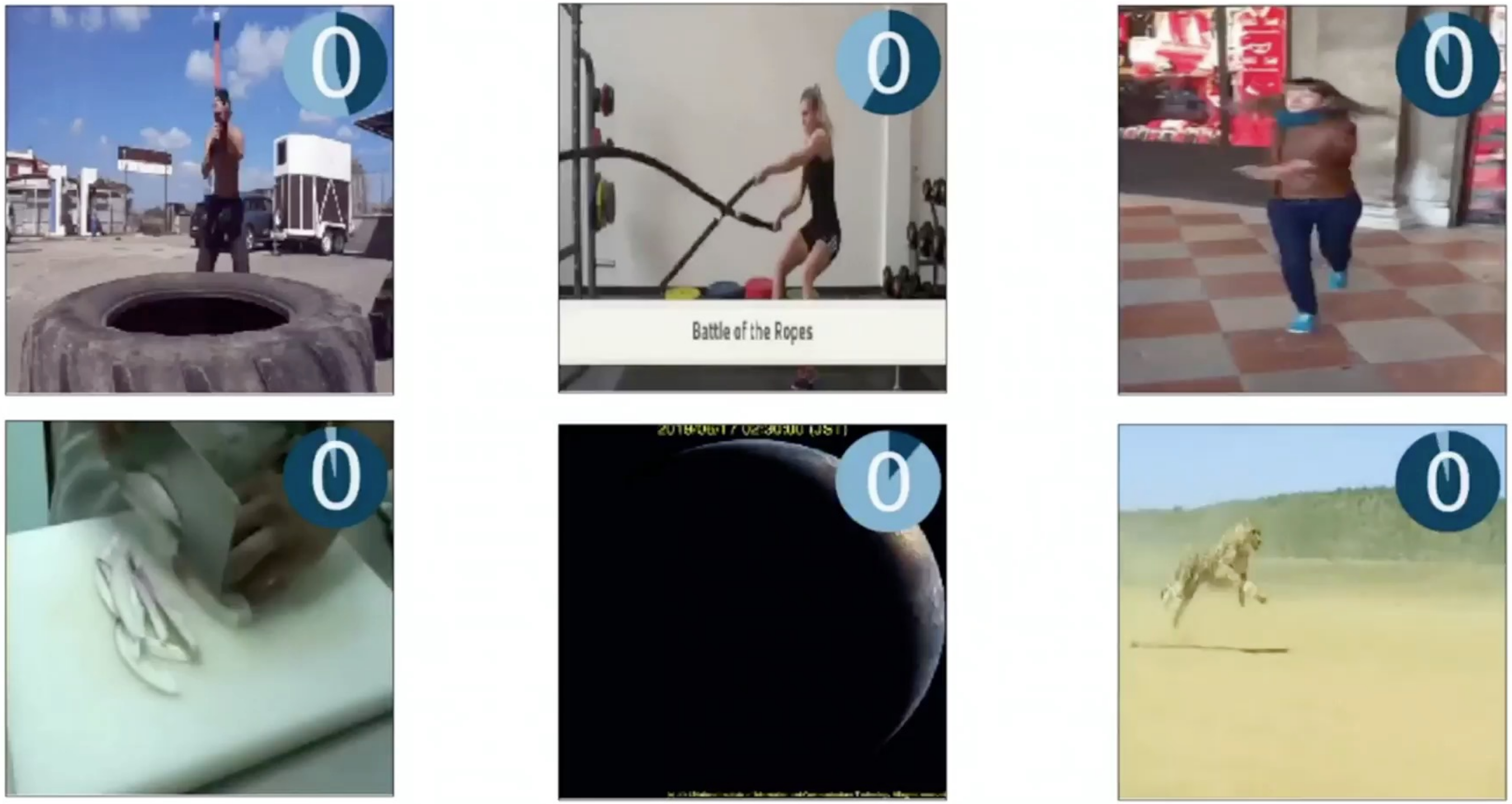
w/ Tom Runia, et al. CVPR 2018 & IJCV 2019

Non-stationary world

Wavelet transform of optical flow features

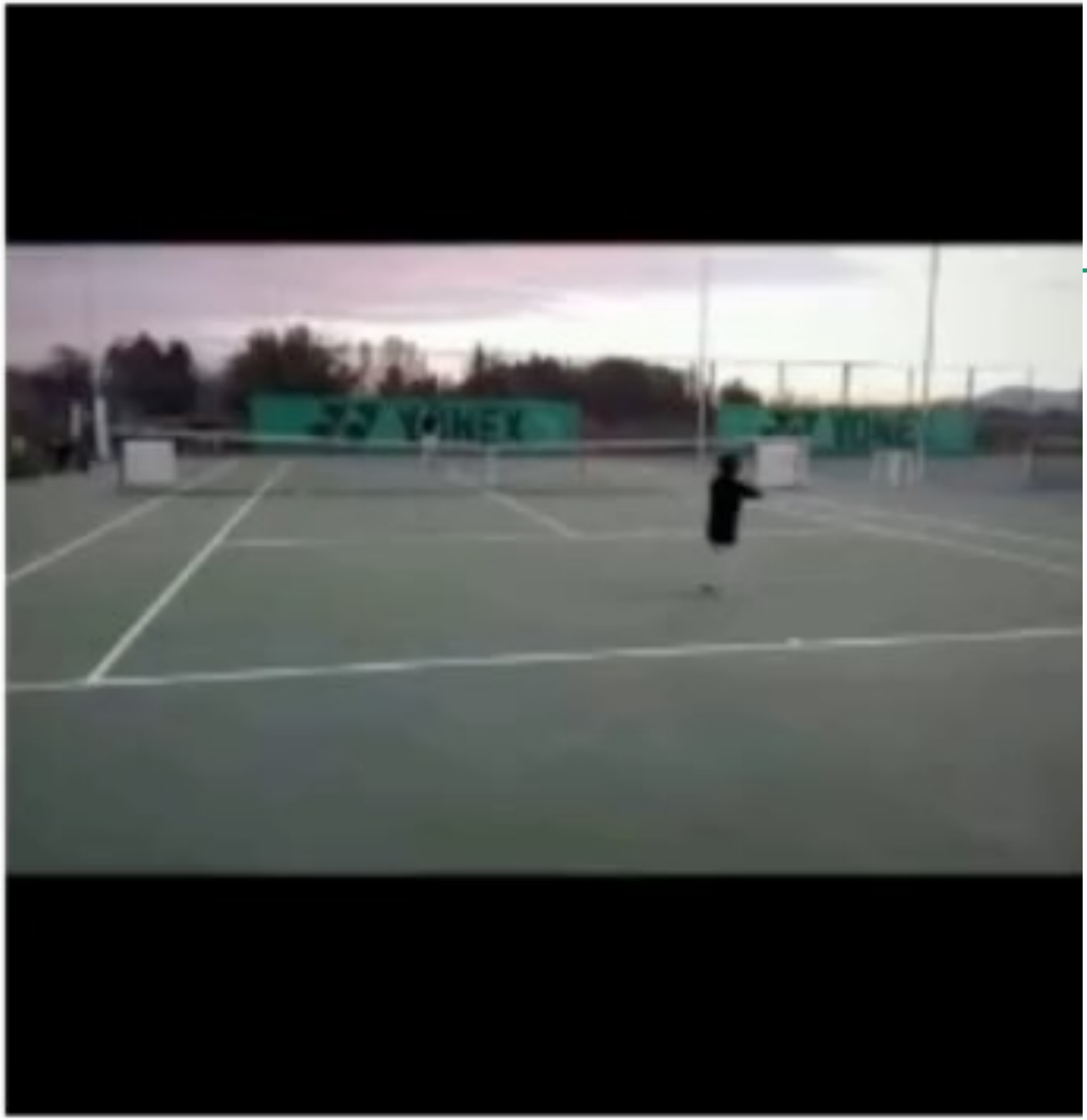


Dataset world



Dwibedi et al., introduce Countix at CVPR 2020
Zhang et al. introduce UCFRep at CVPR 2020

https://github.com/xiaobai1217/Awesome-Video-Datasets



Real world challenges unseen during training

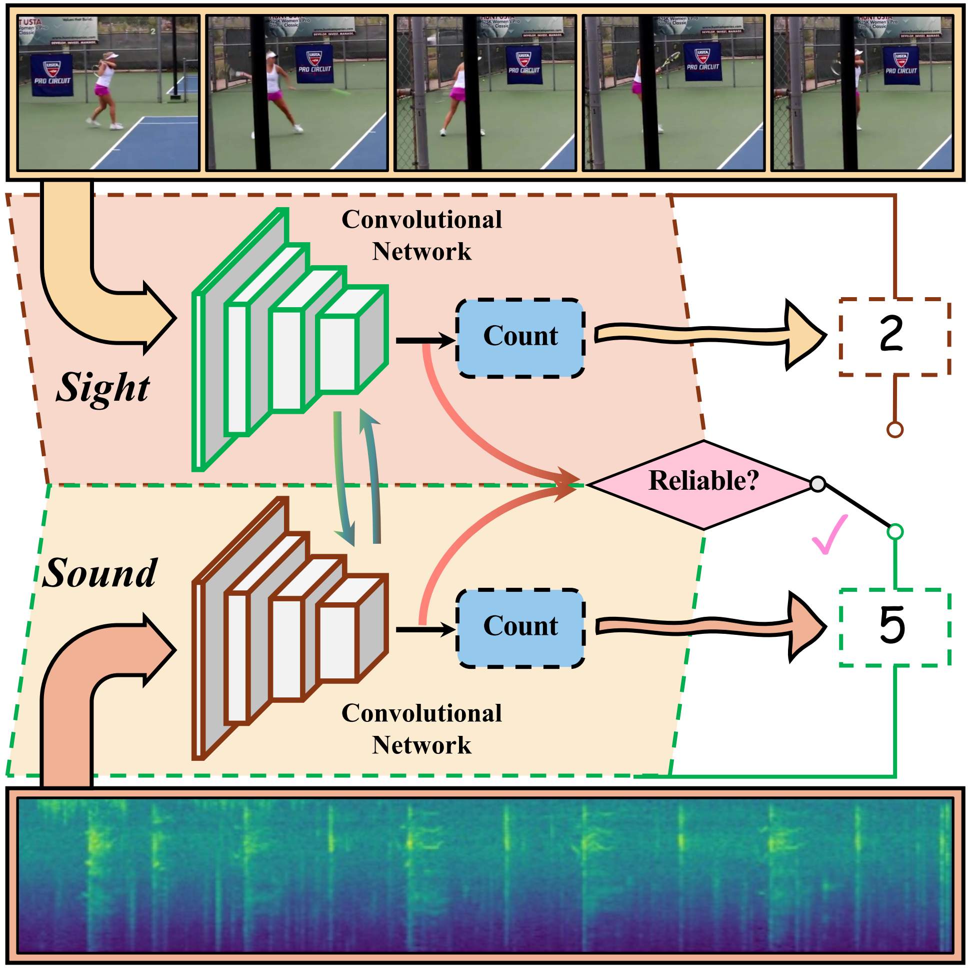
Contributions

Video repetition estimation from a new perspective based on not only the sight but also the sound signal

Audiovisual model with a sight and sound stream, each stream facilitates each modality to predict the number of repetitions

Two sight and sound datasets for video repetition estimation

Model basics



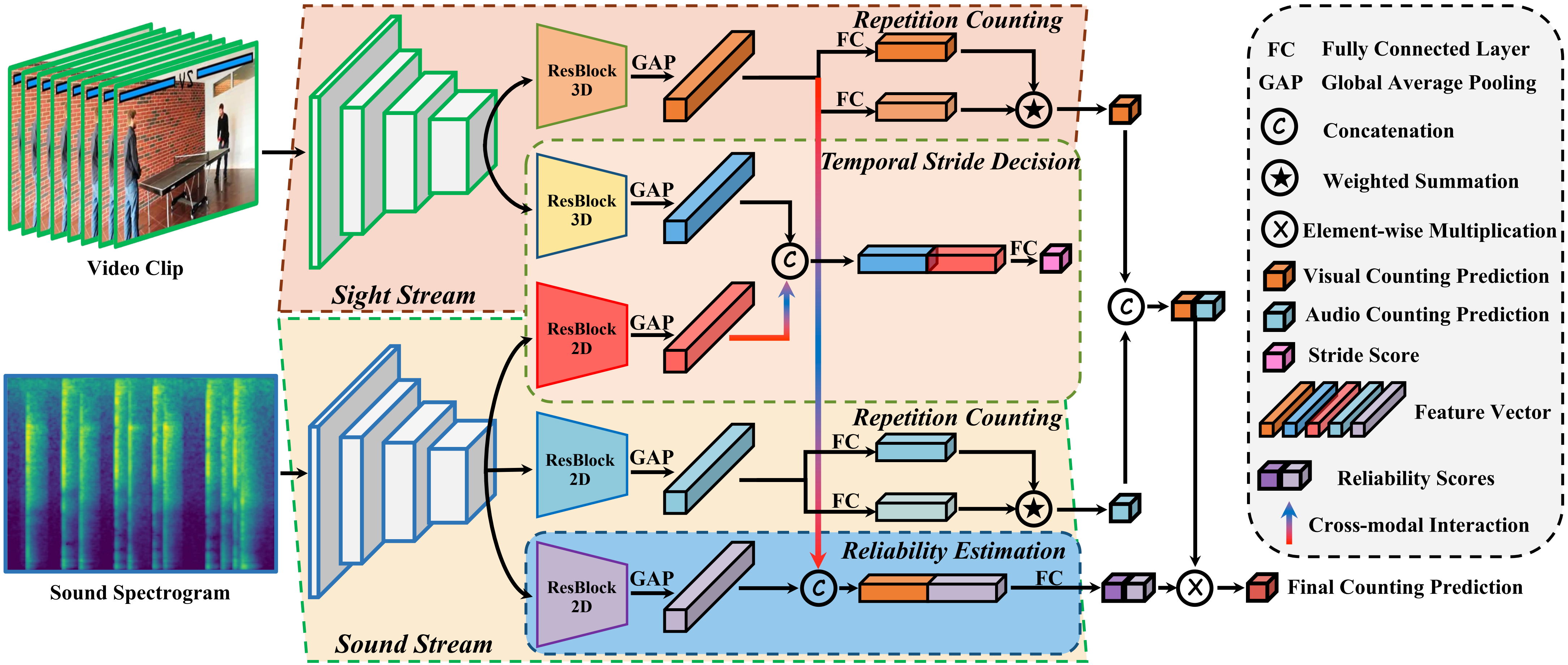
Sight stream: S3D net predicting counting result per input clip and repetition class

Sound stream: Resnet-18 predicting counting result per sound spectrogram and repetition class

Temporal stride: selects best stride per video for the sight stream based on visual and audio features

Reliability: decides what prediction to use

A more detailed view



Repurpose and reorganize Countix dataset

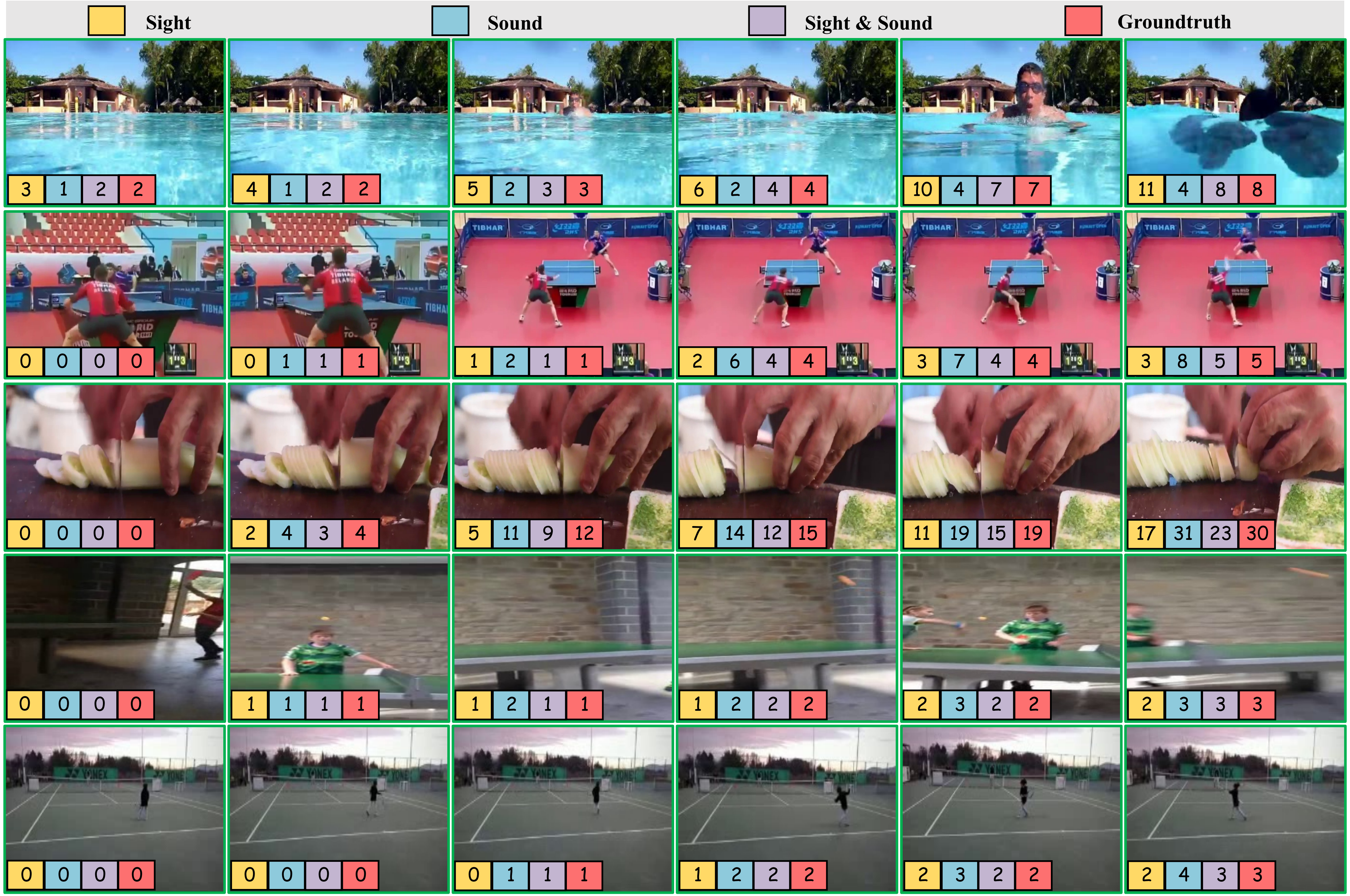
Countix-AV

1,863 videos covering repetitive activity categories with clear sound and without background music, with 987, 311 and 565 for train, val and test.

Extreme Countix-AV

156 videos from Countix-AV and another 58 videos from the VGGSound dataset in which the sight conditions are too poor for counting, for test only,

https://github.com/xiaobai1217/Awesome-Video-Datasets


Sight
Sound
Sight & Sound
Groundtruth
3 1 2 2
4 1 2 2
5 2 3 3
6 2 4 4
10 4 7 7
11 4 8 8
0 0 0 0
0 1 1 1
1 2 1 1
2 6 4 4
3 7 4 4
3 8 5 5
0 0 0 0
2 4 3 4
5 11 9 12
7 14 12 15
11 19 15 19
17 31 23 30
0 0 0 0
1 1 1 1
1 2 1 1
1 2 2 2
2 3 2 2
2 3 3 3
0 0 0 0
0 0 0 0
0 1 1 1
1 2 2 2
2 3 2 2
2 4 3 3

Benefit of model components

Model components	**MAE↓**
Sight stream	0.331
Sound stream	0.375
Sight with temporal stride	0.314
Averaging predictions	0.300
Full sight and sound model	**0.291**

Mean Absolute Error:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{c}_i - l_i|}{l_i}$$

l_i - groundtruth

$\hat{c}_i$ - model prediction

All modules matter, reliability estimation is preferred over simple averaging

Comparison with others

Sight datasets

	UCFRep	**Countix**
	MAE↓	MAE↓
Baseline by Dwibedi et al.	0.474	0.525
Dwibedi et al. CVPR20	-	0.364
Zhang et al. CVPR20	0.147	-
Levy and Wolf ICCV15	0.286	-
Ours: sight only	**0.143**	0.314
Ours: sound only	-	0.793
Ours: sight & sound	-	**0.307**

Sight-only model already good

Comparison with others

	Sight datasets		Sight & Sound datasets	
	UCFRep	**Countix**	**Countix-AV**	**Extreme Countix-AV**
	MAE↓	MAE↓	MAE↓	MAE↓
Baseline by Dwibedi et al.	0.474	0.525	0.503	0.620
Dwibedi et al. CVPR20	-	0.364	-	-
Zhang et al. CVPR20	0.147	-	-	-
Levy and Wolf ICCV15	0.286	-	-	-
Ours: sight only	**0.143**	0.314	0.331	0.392
Ours: sound only	-	0.793	0.375	0.351
Ours: sight & sound	-	**0.307**	**0.291**	**0.329**

Sight-only model already good, adding sound further reduces counting error

Real world video challenges

Real world challenge	Sight	Sound	Sight & Sound
Camera viewpoint changes	0.384	0.376	0.331
Cluttered background	0.342	0.337	0.307
Low illumination	0.325	0.269	0.310
Fast motion	0.528	0.311	0.383
Disappearing activity	0.413	0.373	0.339
Scale variation	0.332	0.386	0.308
Low resolution	0.348	0.303	0.294
Overall	**0.392**	**0.351**	**0.329**

Sound less sensitive than sight, combination always outperforms sight only

Low resolution

Sight	0
Sound	0
Sight & Sound	0
Groundtruth	0

Sound can play a vital role, especially under harsh vision conditions

Audio-Adaptive Activity Recognition Across Video Domains




Yunhua Zhang
University of Amsterdam

Hazel Doughty
University of Amsterdam

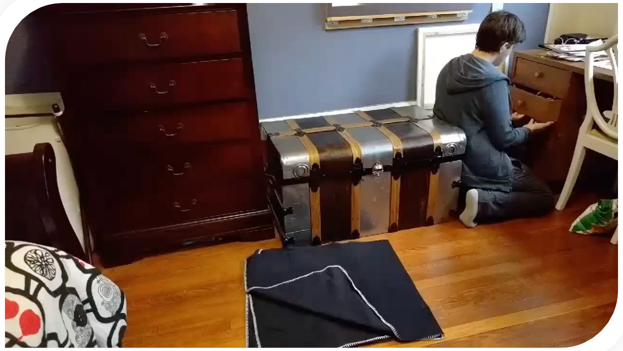
Ling Shao
Inception Institute of AI

Cees Snoek
University of Amsterdam

To appear in CVPR 2022.

Activity recognition under domain shift

Opening **activity**

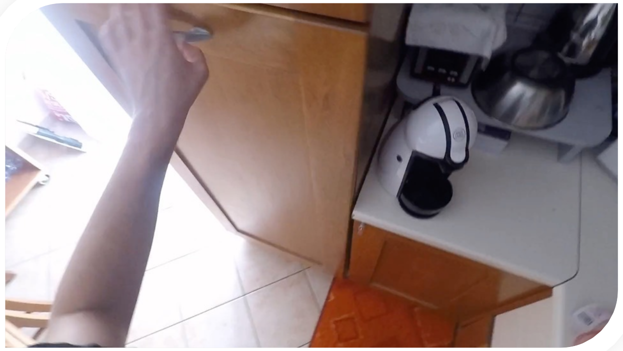


Camera viewpoint shift

Eating **activity**

Actor shift

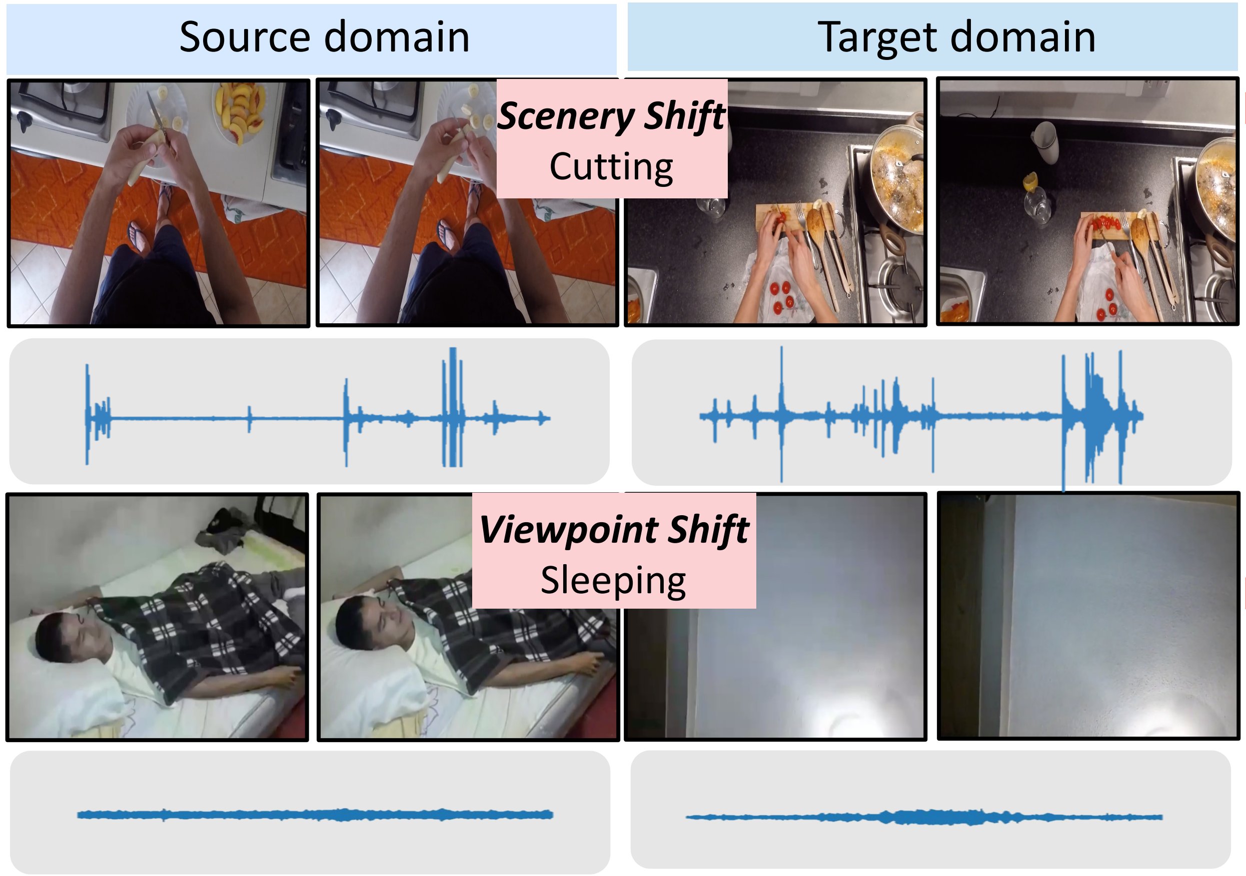
Opening **activity**



Scenery shift

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, …)

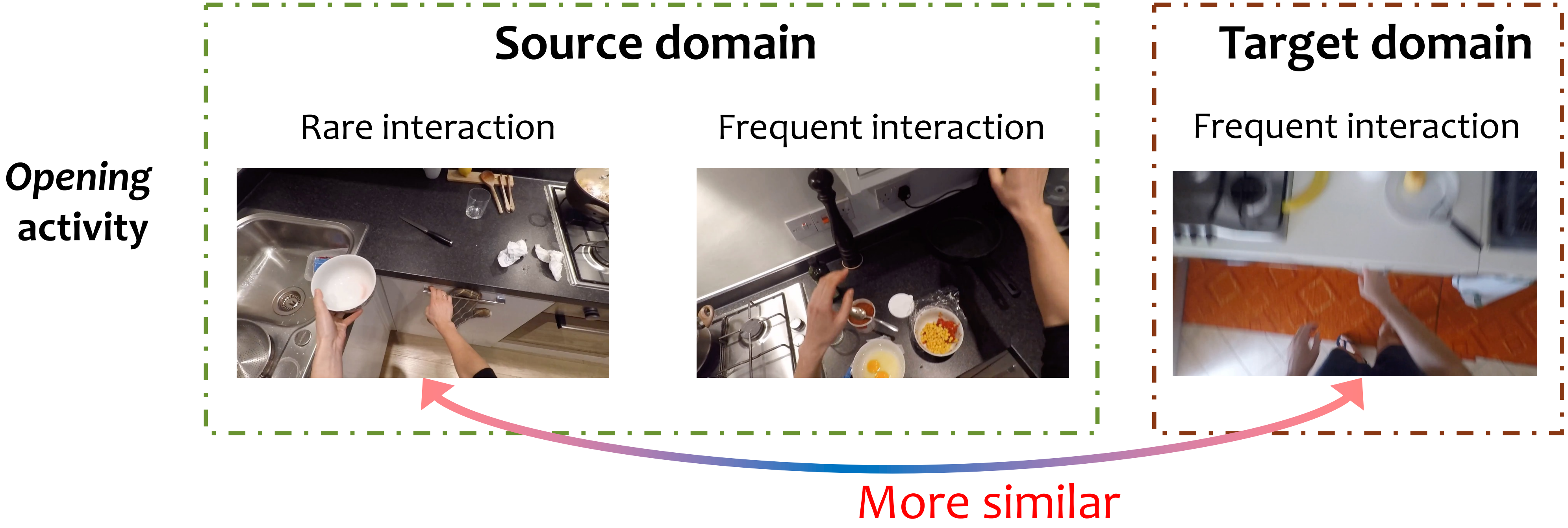
Environmental sounds of silent activities

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

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.

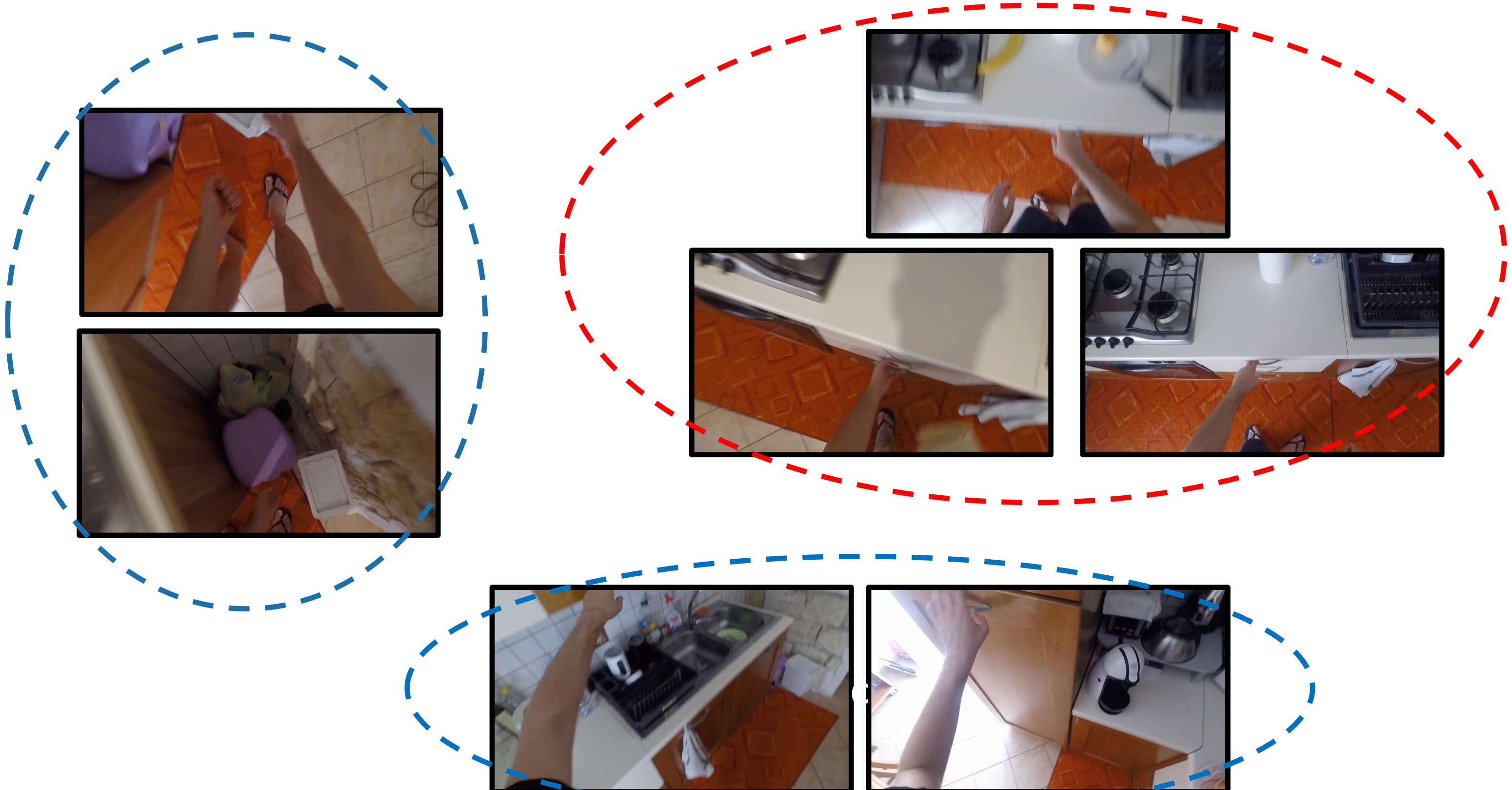
Solution: learn each class and each type of interaction equally



Audio-balanced learning

For source domain data, we use audio to **cluster** the samples inside each class.

Each cluster is treated as one type of interaction



K-means Clustering

Absent-activity learning

Observation: Most activities are silent ⟶ Audio predictions are unreliable

Solution: activities with the lowest audio-based probablities

⟶ unlikely happening inside the video

Example: silent environment ⟶ “playing piano” ✗

Forcing the model to predict low probabilities towards these absent activities.

Absent-activity learning

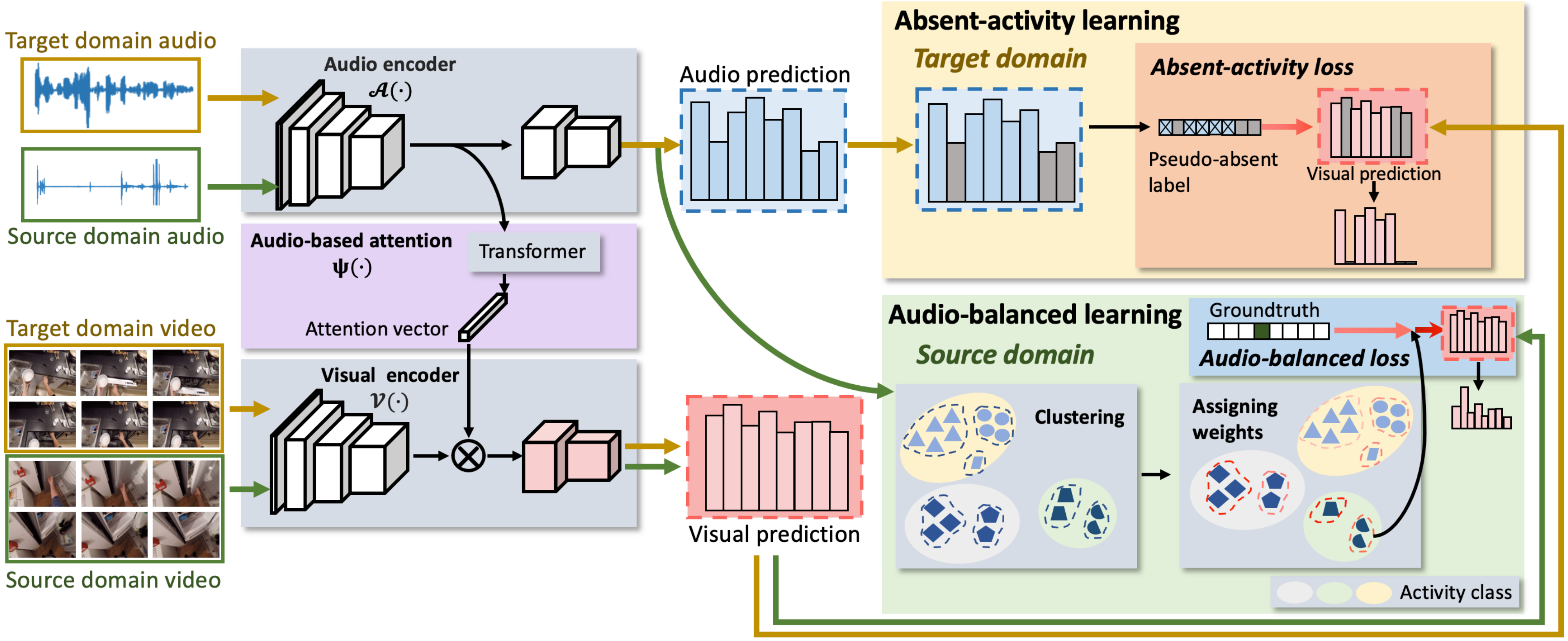
EPIC-Kitchens (scenery shift)
Single-label classification

Groundtruth activity:
pour

Absent activities predicted by audio:
wash
close
open

An audio-adaptive visual encoder

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

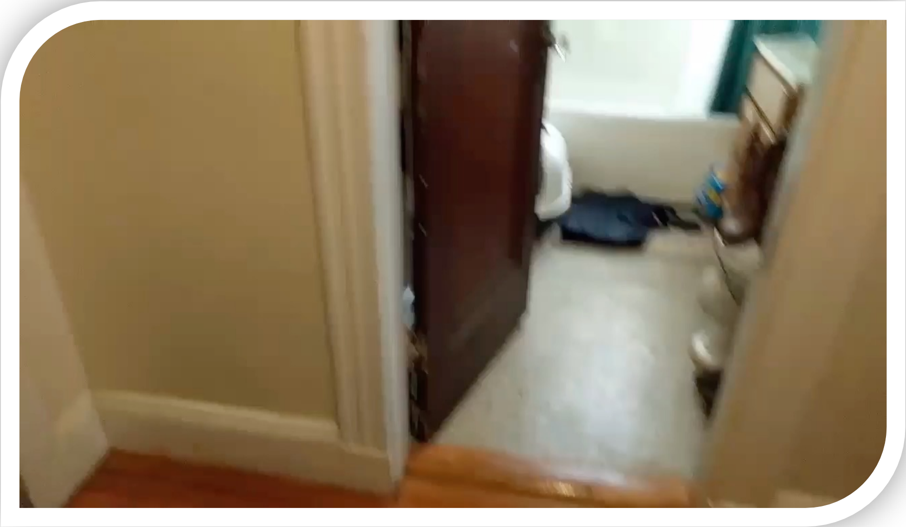


Activity sounds provide out-of-sight information

3rd person view

We can see a person (domain-specific visual feature)

Ego-centric view

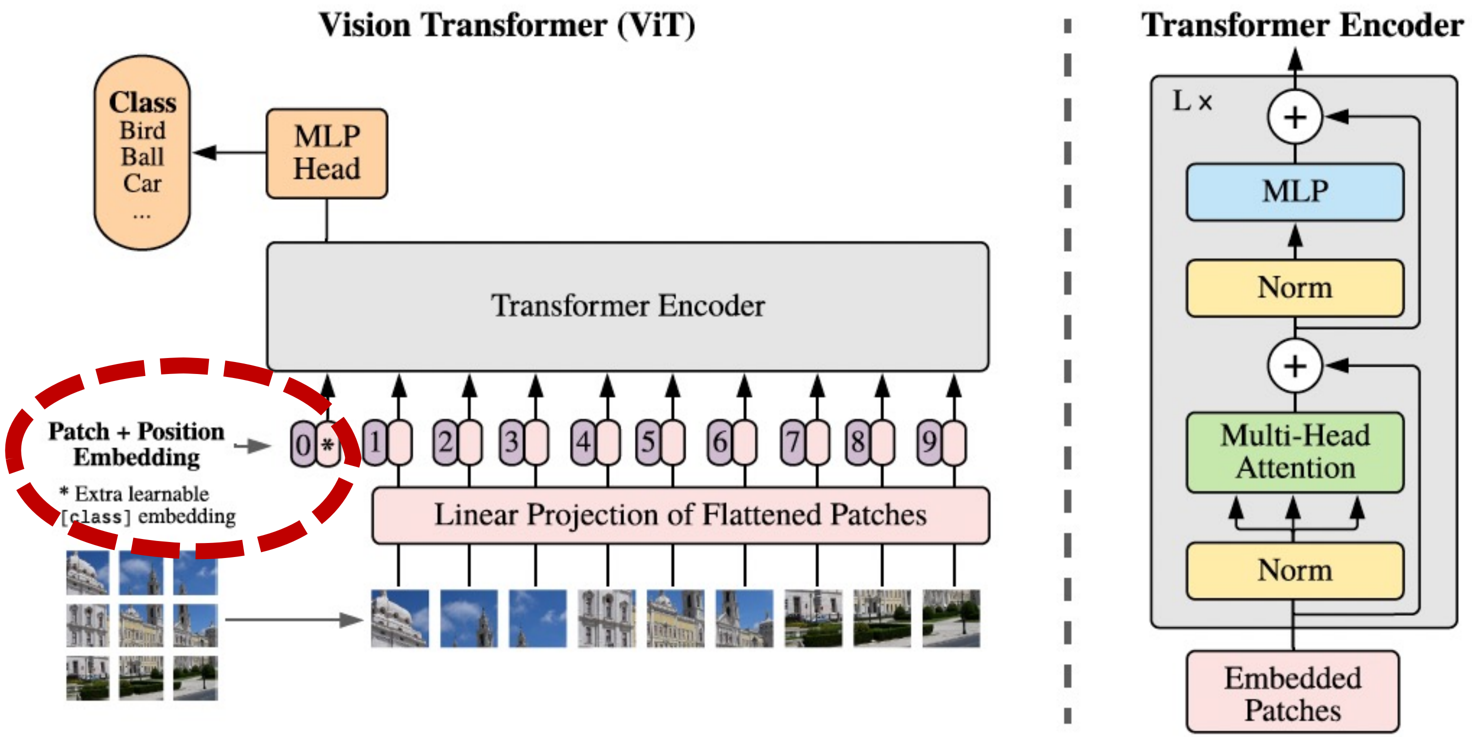


No person can be observed
But the sound can be heard

Remaining problem remove domain-specific visual features

Recap: vision transformer

Follows standard transformer encoder, adds learnable classification token

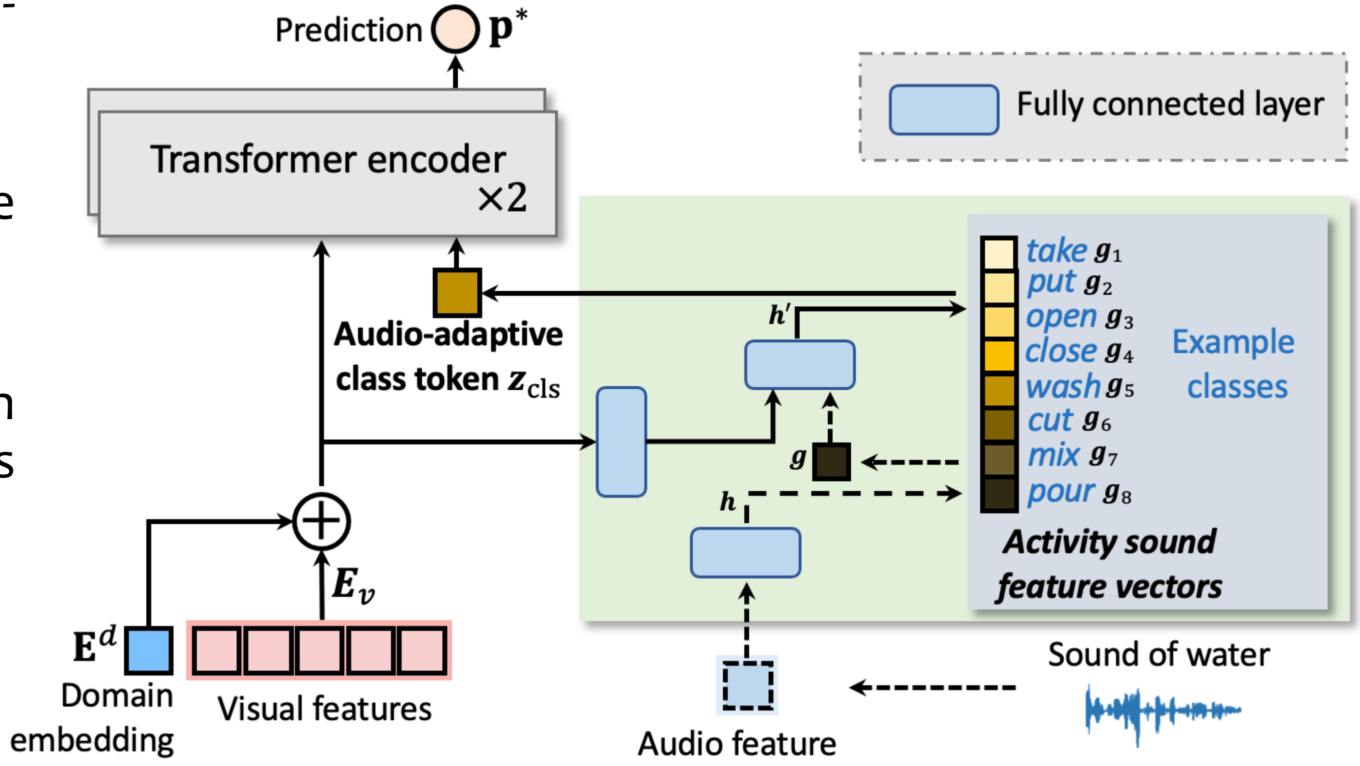


Audio-infused transformer

Domain embedding: remove domain-specific visual features

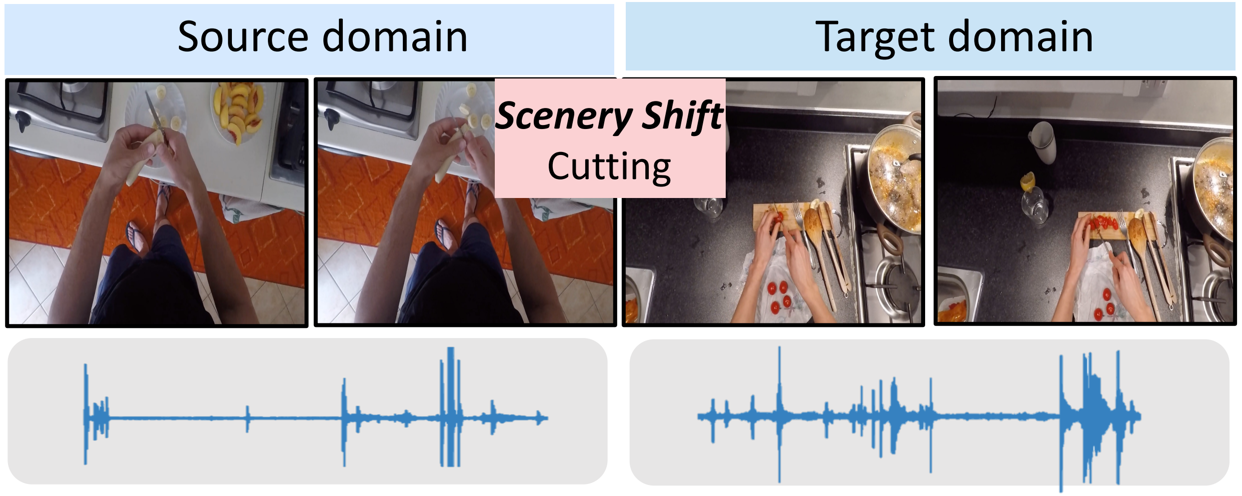
Audio-adaptive class token: incorporate the activity information from sound

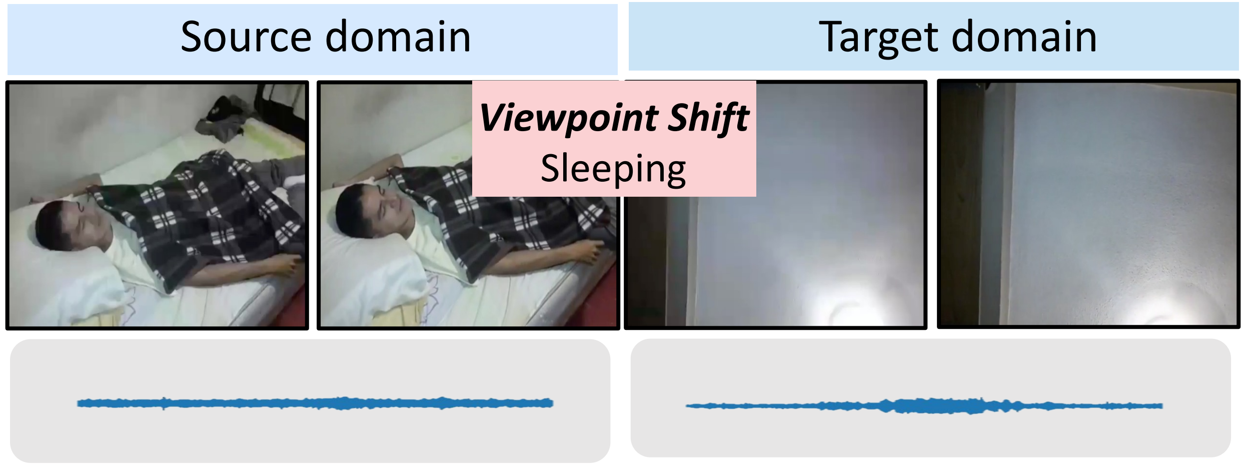
Activity sound feature vectors: chosen by the audio features, which provides regularization for model learning.



Ablation

	Scenery-shift ↑ (EPIC-Kitches, top-1)	Viewpoint-shift↑ (CharadesEgo, mAP)
Stage 1: Audio-adaptive encoder		
Visual encoder (SlowFast)	48.0	23.1
+Audio-based attention	51.2	23.5





Ablation

	Scenery-shift ↑ (EPIC-Kitches, top-1)	**Viewpoint-shift↑** (CharadesEgo, mAP)
Stage 1: Audio-adaptive encoder		
Visual encoder (SlowFast)	48.0	23.1
+Audio-based attention	51.2	23.5

Ablation

	Scenery-shift ↑ (EPIC-Kitches, top-1)	**Viewpoint-shift↑** (CharadesEgo, mAP)
Stage 1: Audio-adaptive encoder		
Visual encoder (SlowFast)	48.0	23.1
+Audio-based attention	51.2	23.5
+Absent-activity learning	53.7	24.4
+Audio-balanced learning	55.7	25.0

Ablation

	Scenery-shift ↑ (EPIC-Kitches, top-1)	Viewpoint-shift↑ (CharadesEgo, mAP)
Stage 1: Audio-adaptive encoder		
Visual encoder (SlowFast)	48.0	23.1
+Audio-based attention	51.2	23.5
+Absent-activity learning	53.7	24.4
+Audio-balanced learning	55.7	25.0
Stage 2: Audio-infused transformer		
+Vanilla multi-modal transformer	56.1	25.0
+Domain embedding	57.2	25.4
+Audio-adaptive class token	59.2	26.3

Scenery-shift on EPIC-Kitchens

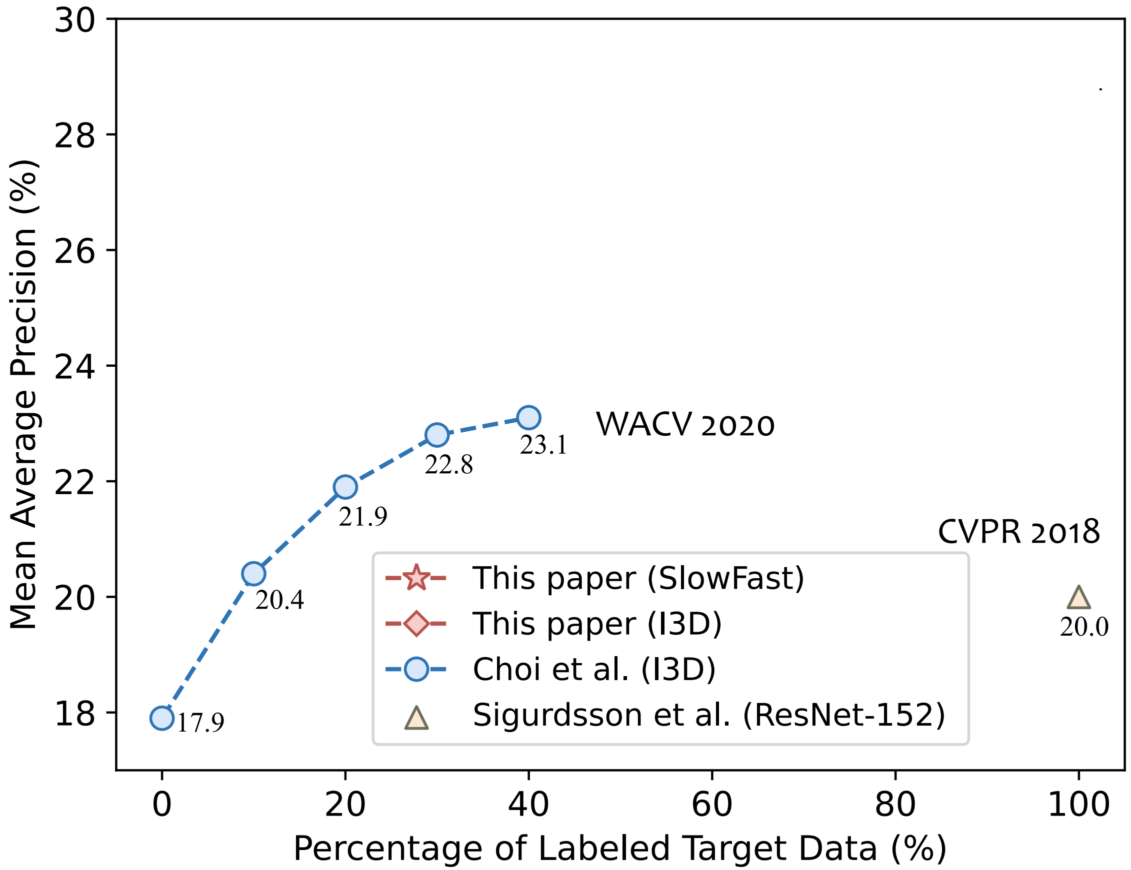
		RGB	Flow	Audio	Mean
I3D Architecture					
Sahoo et al.	NeurIPS 2021	✓			43.2
Munro & Damen	CVPR 2020	✓	✓		50.3
Song et al.	CVPR 2021	✓	✓		51.2
Kim et al.	ICCV 2021	✓	✓		51.0
This paper		✓	✓	✓	**54.1**

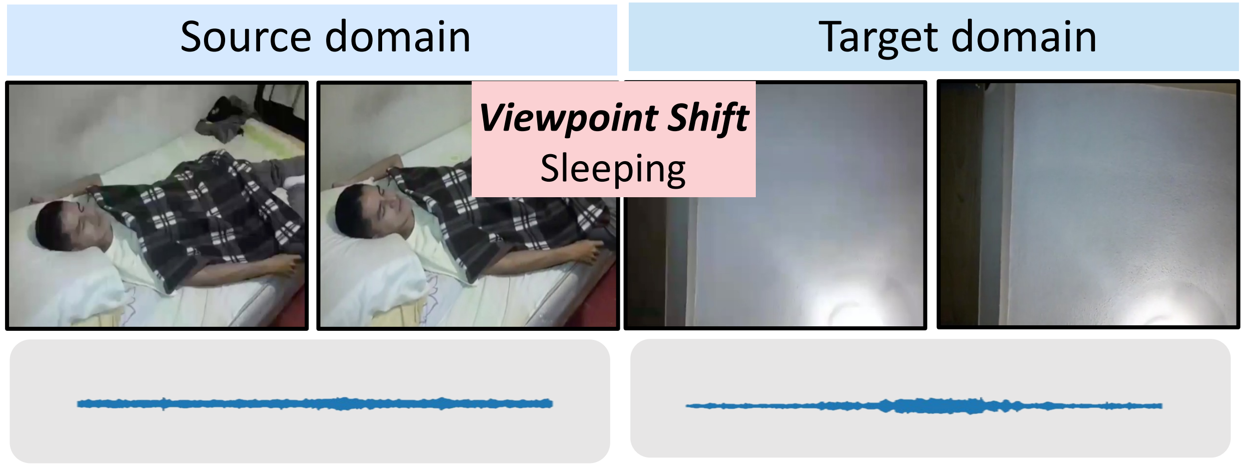
Scenery-shift on EPIC-Kitchens

		RGB	Flow	Audio	Mean
I3D Architecture					
Sahoo et al.	NeurIPS 2021	✓			43.2
Munro & Damen	CVPR 2020	✓	✓		50.3
Song et al.	CVPR 2021	✓	✓		51.2
Kim et al.	ICCV 2021	✓	✓		51.0
This paper		✓	✓	✓	**54.1**
SlowFast Architecture					
This paper		✓	✓	✓	**61.0**

Viewpoint-shift on CharadesEgo

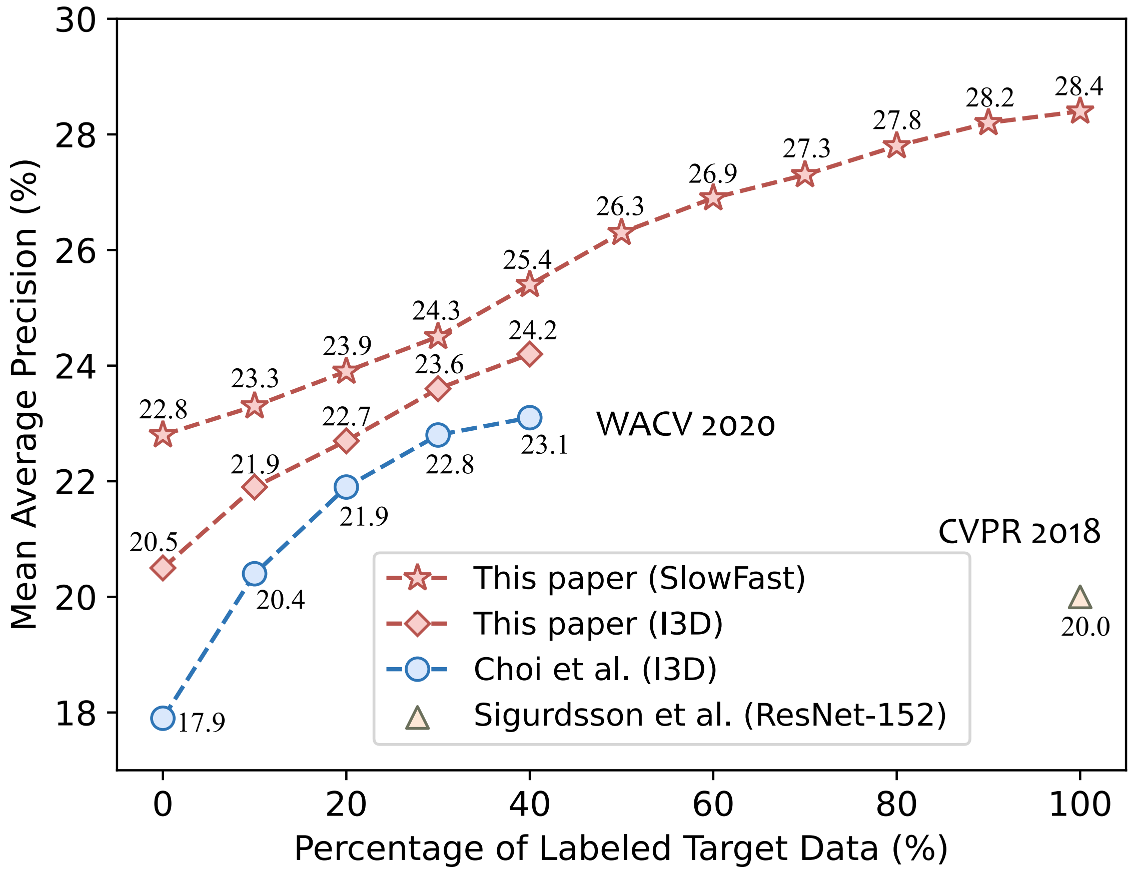
semi-supervised domain adaptation

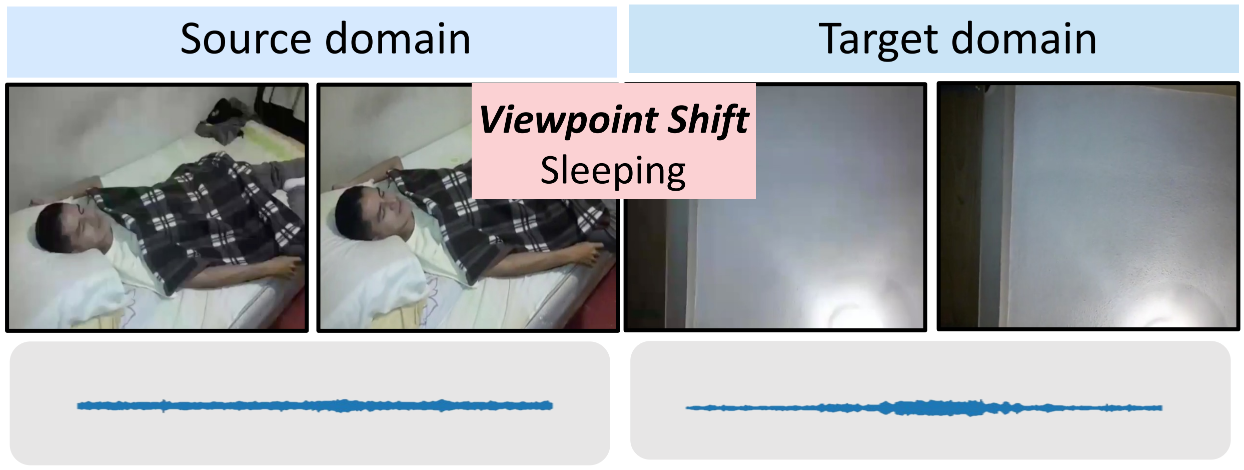




Viewpoint-shift on CharadesEgo

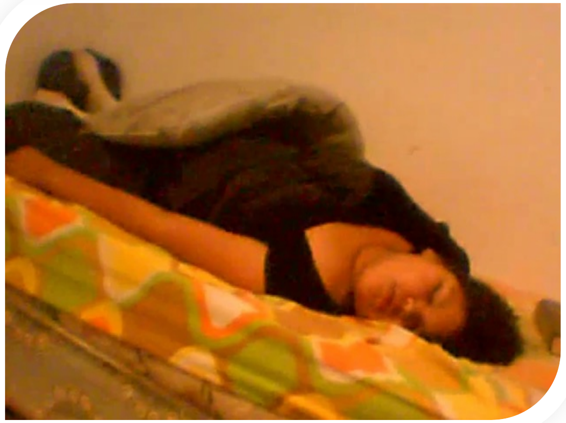
semi-supervised domain adaptation





Actor-shift: success case

Source domain

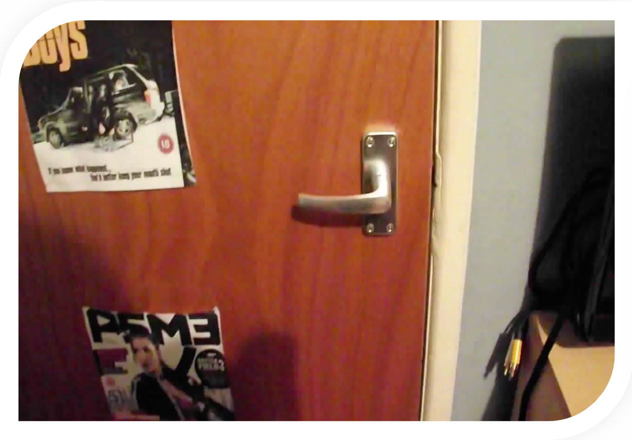


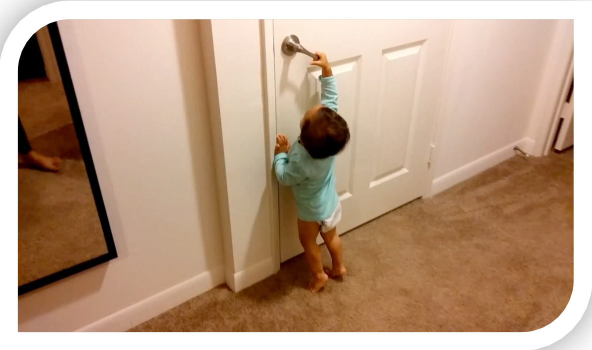
Target domain

Encoder + recognizer
Groundtruth: *sleeping*
Prediction: *sleeping*
Confidence: 0.76

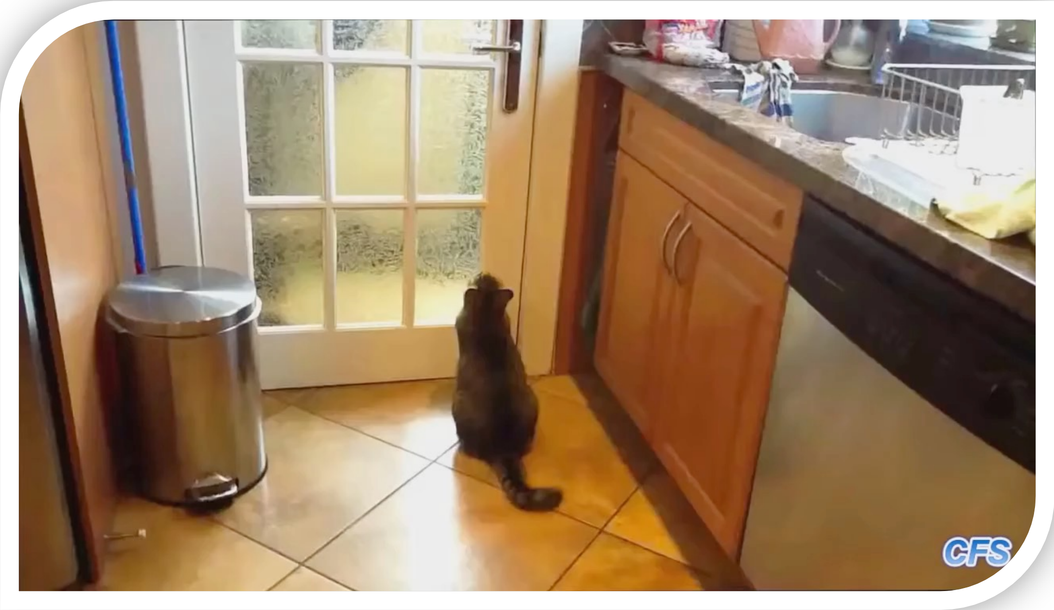
Actor-shift: success case

Source domain





Target domain



Encoder + recognizer
Groundtruth: *opening door*
Prediction: *opening door*
Confidence: 0.85

Actor-shift: failure case

Source domain

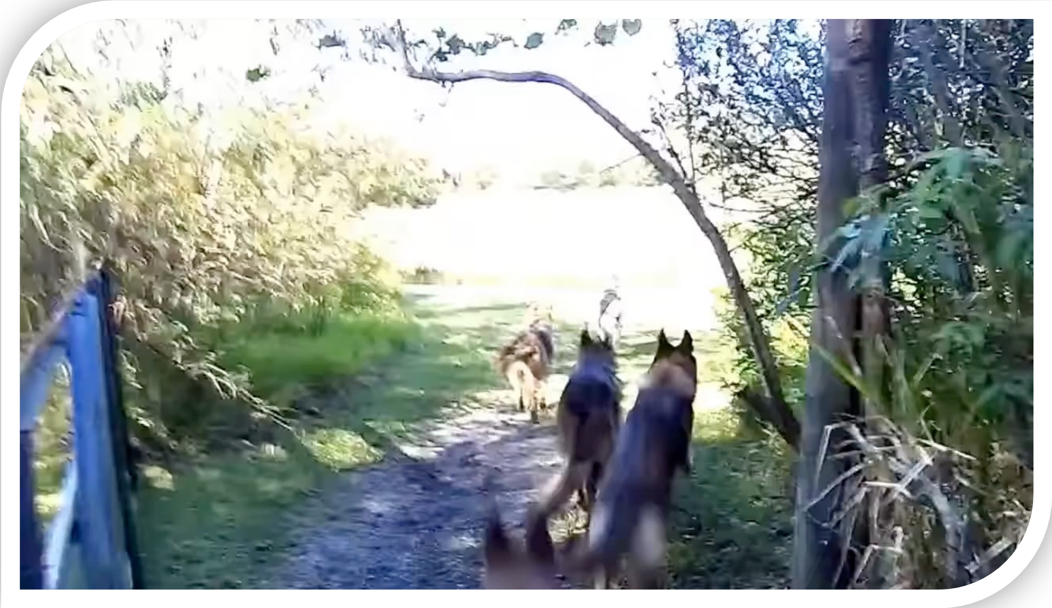
Target domain

Encoder + recognizer
Groundtruth: *drinking*
Prediction: *eating*
Confidence: 0.35

Actor-shift: failure case

Source domain

Target domain



Encoder + recognizer
Groundtruth: *running*
Prediction: *swimming*
Confidence: 0.48

Conclusions

Video understanding treated by many as **glorified image** recognition problem.

We presented **holistic video** perspective based on **spatiotemporal tubelets**.

Showed invariant properties of **sound for hard activity recognition** conditions.

Thank you

www.ceessnoek.info