ETH zürich

Deeper Action

ECCV TEL AVIV 2022

Computer Vision Lab

Spatiotemporal Action Detection Under Large Motion

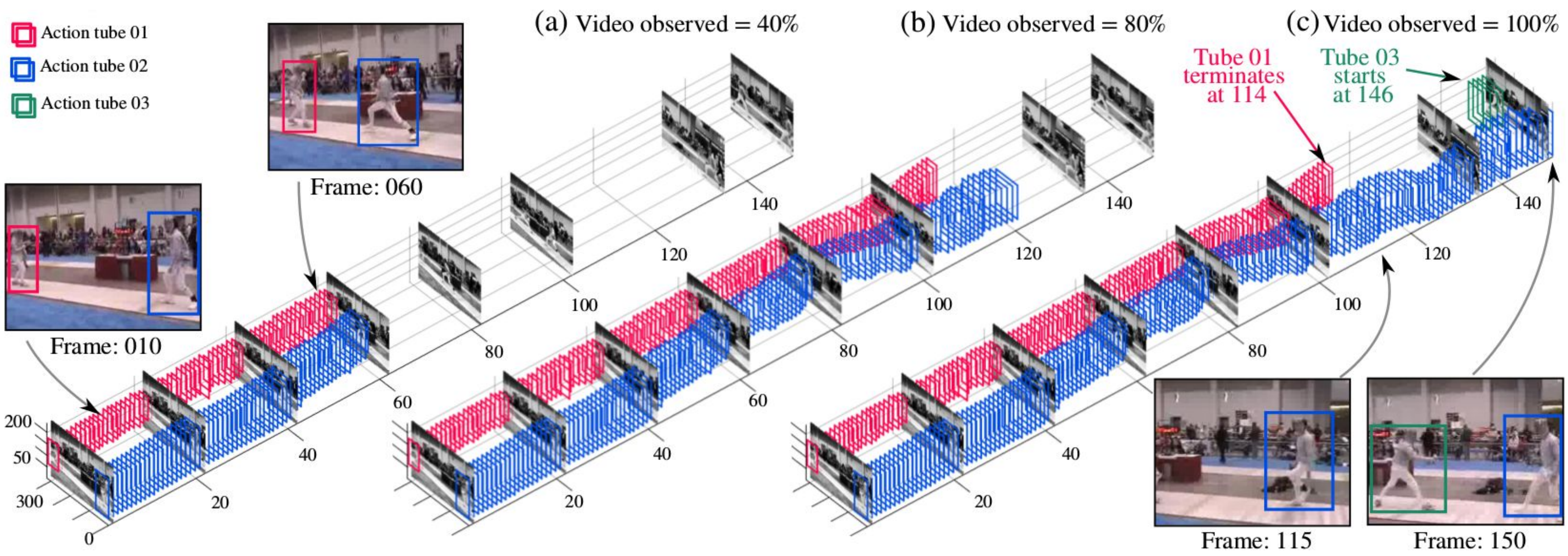
Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu and Prof. Luc Van Gool

Accepted at WACV 2023

Outline

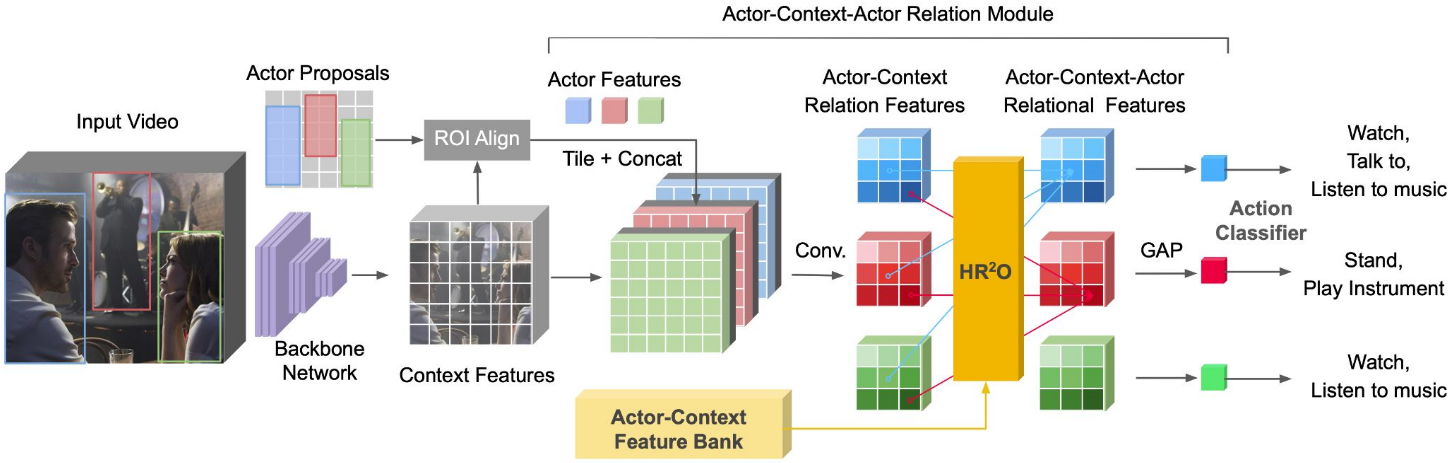
- Problem statement
- Key Insight
- Method
- Results & Analysis
- Q&A

Spatiotemporal Action Tube Detection



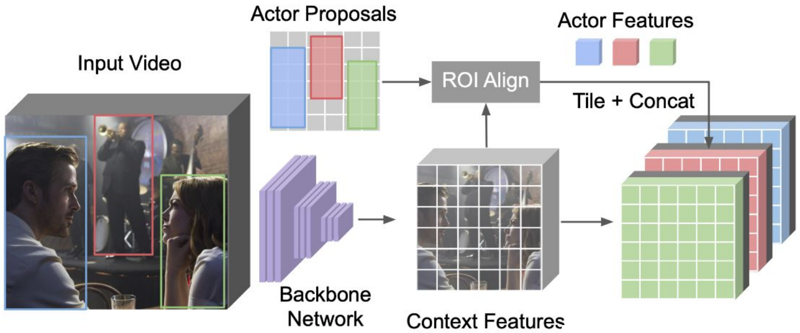
Singh et al. ICCV 2017

Key-frame based methods

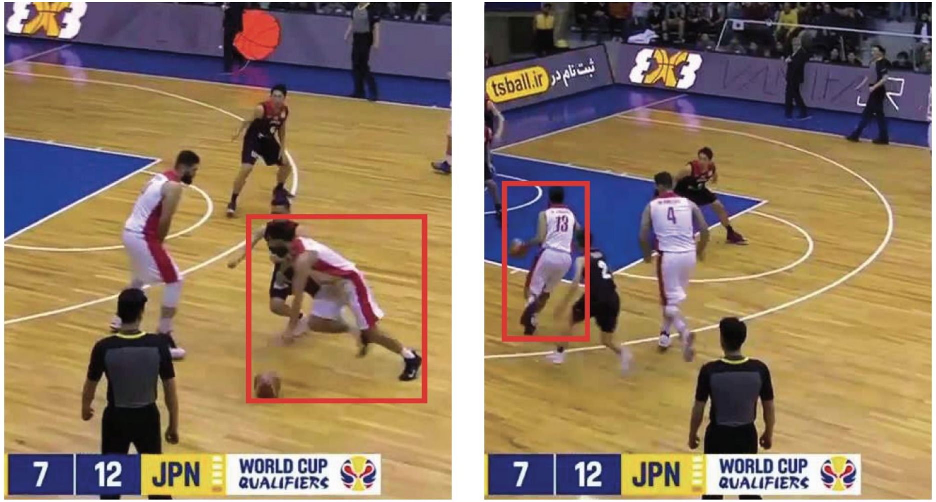


Pan et al. CVPR 2021
Feichtenhofer et al. 2019
Gu et al CVPR 2018 (AVA dataset)

Cuboid Feature Aggregation & Large Motion



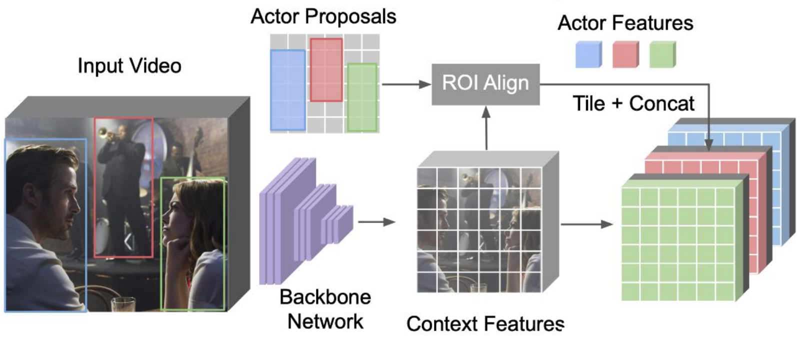
Pan et al. CVPR 2021



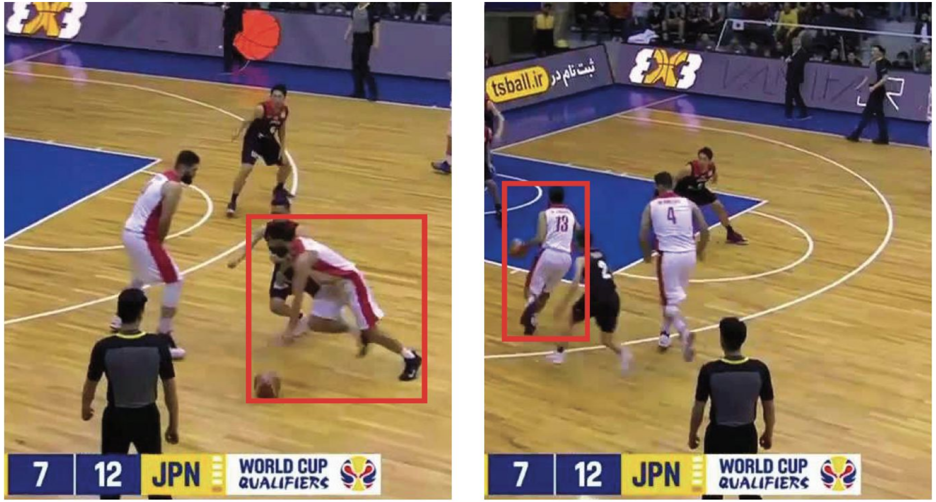
(b) Basketball drive

Will it generate reasonable features for all keyframes?

Cuboid Feature Aggregation & Large Motion



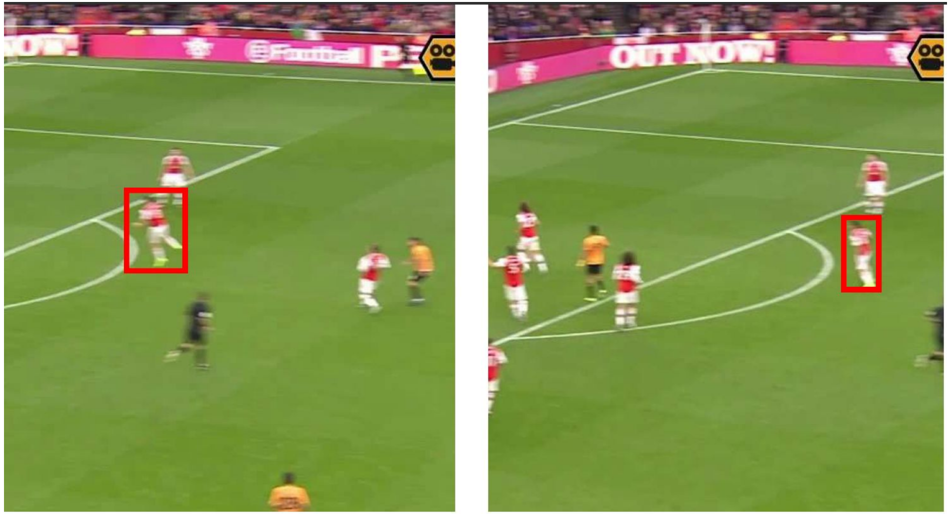
Pan et al. CVPR 2021



(b) Basketball drive

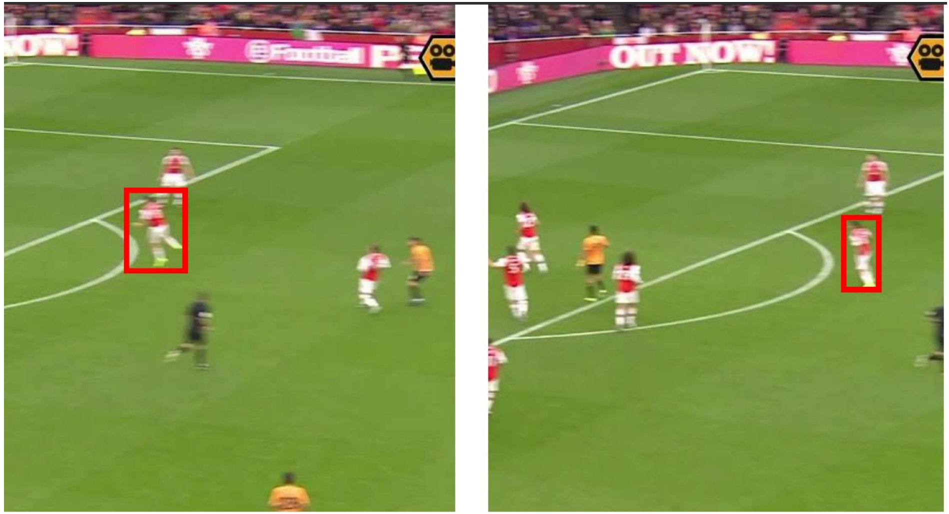
Will it generate reasonable features for all keyframes? NO

Large Motion how & why?

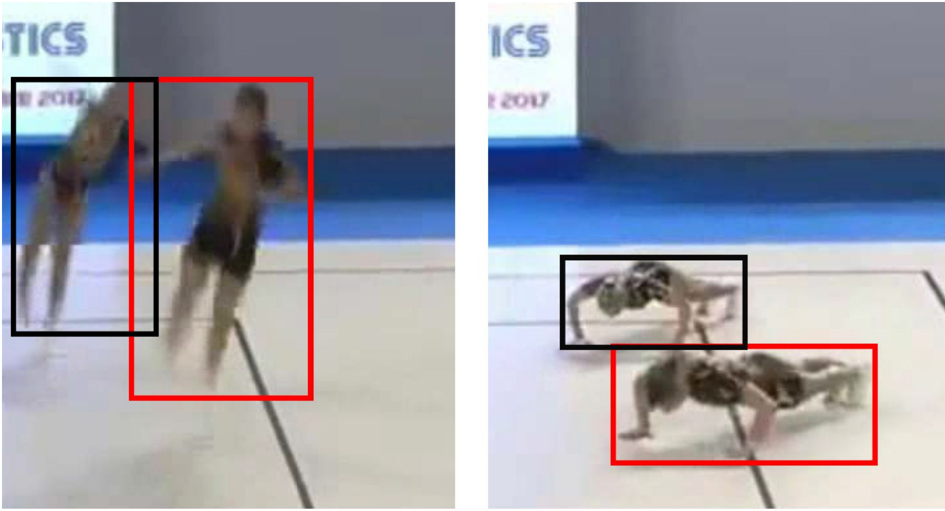


(a) Football block
(Large camera motion)

Large Motion how & why?

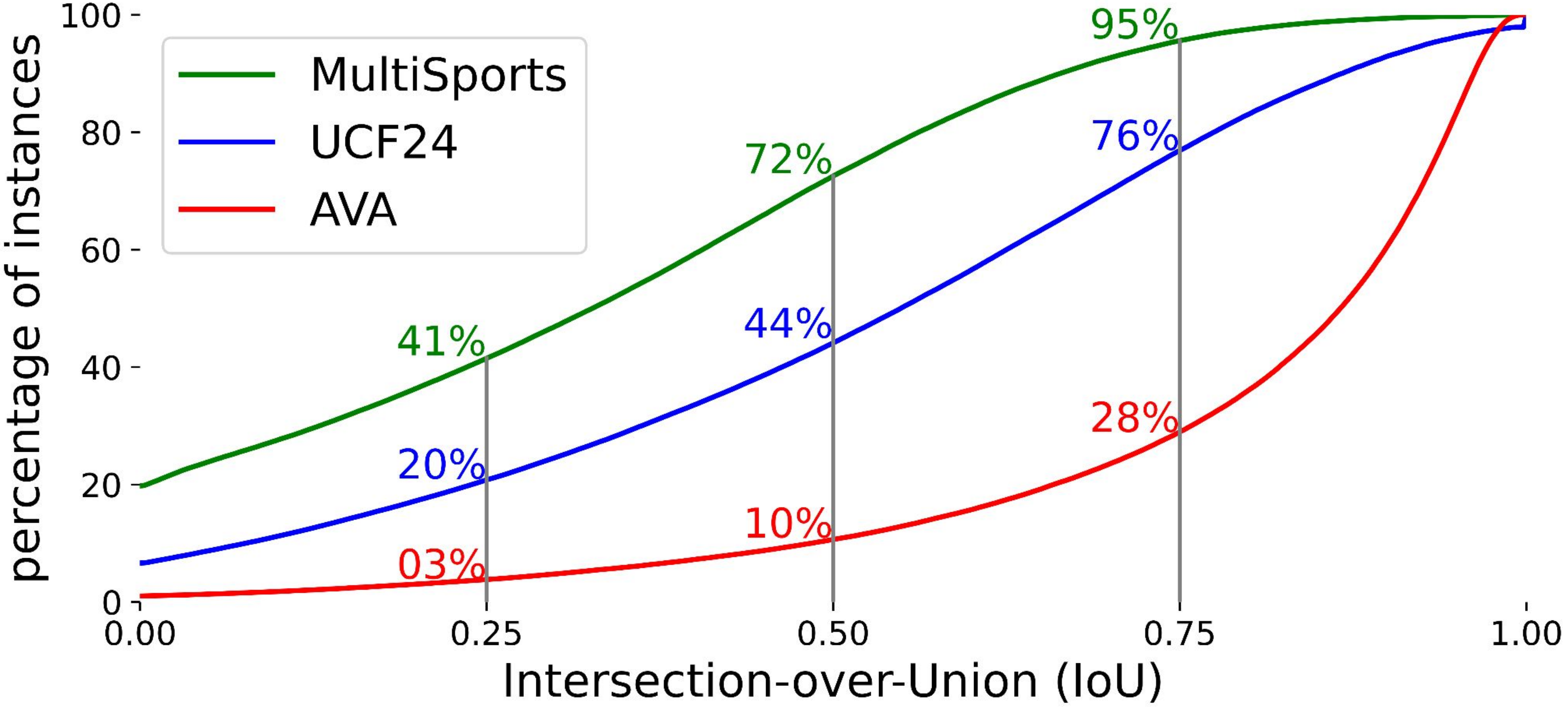


(a) Football block
(Large camera motion)



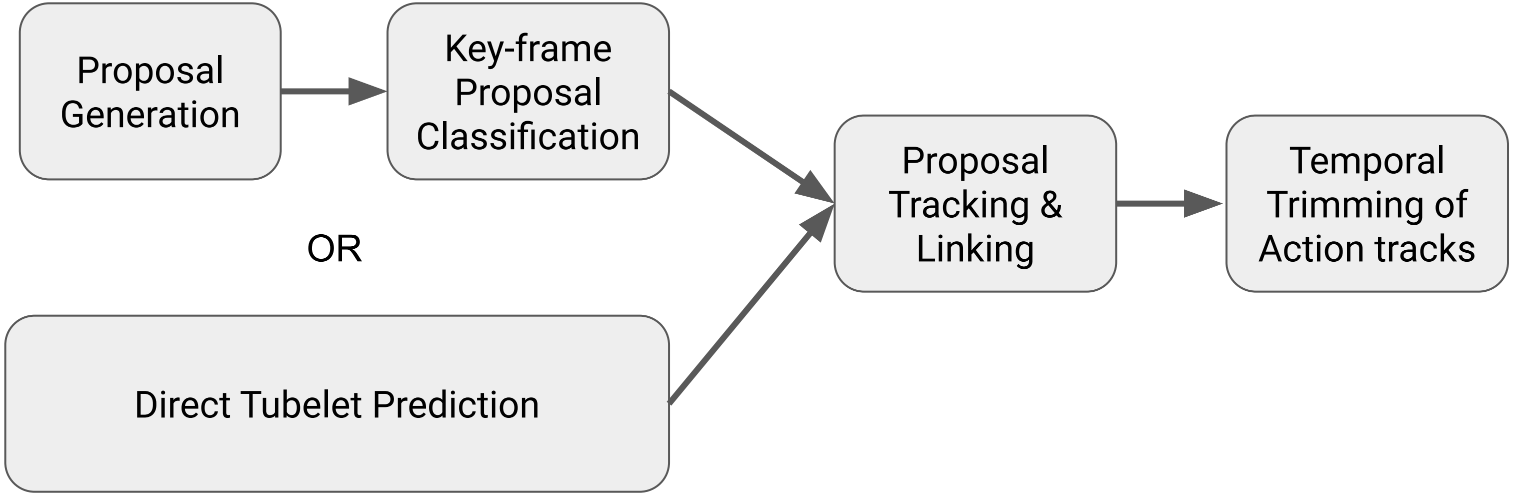
(c) Aerobic pike jump
(Fast action)

How much large motion is there in datasets?



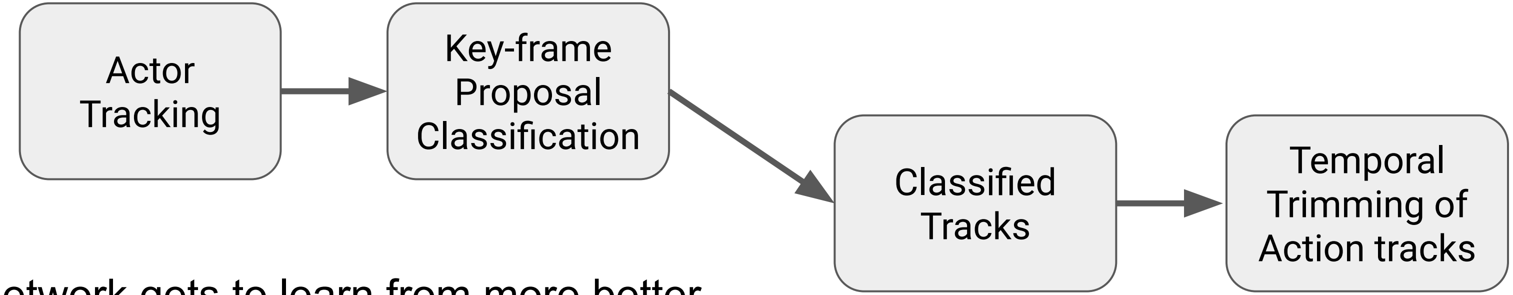
What Can we do?

What Tube Detection Requires?


Proposal Generation
Key-frame Proposal Classification
OR
Direct Tubelet Prediction
Proposal Tracking & Linking
Temporal Trimming of Action tracks

MOC, Li et al. ECCV 22
TUBER, Zhao et al. CVPR 22
ACT, Kalogeiton et al. ICCV 17
AMTNET, Saha et al. ICCV 17

Key question: would early linking help?



Actor Tracking

Key-frame Proposal Classification

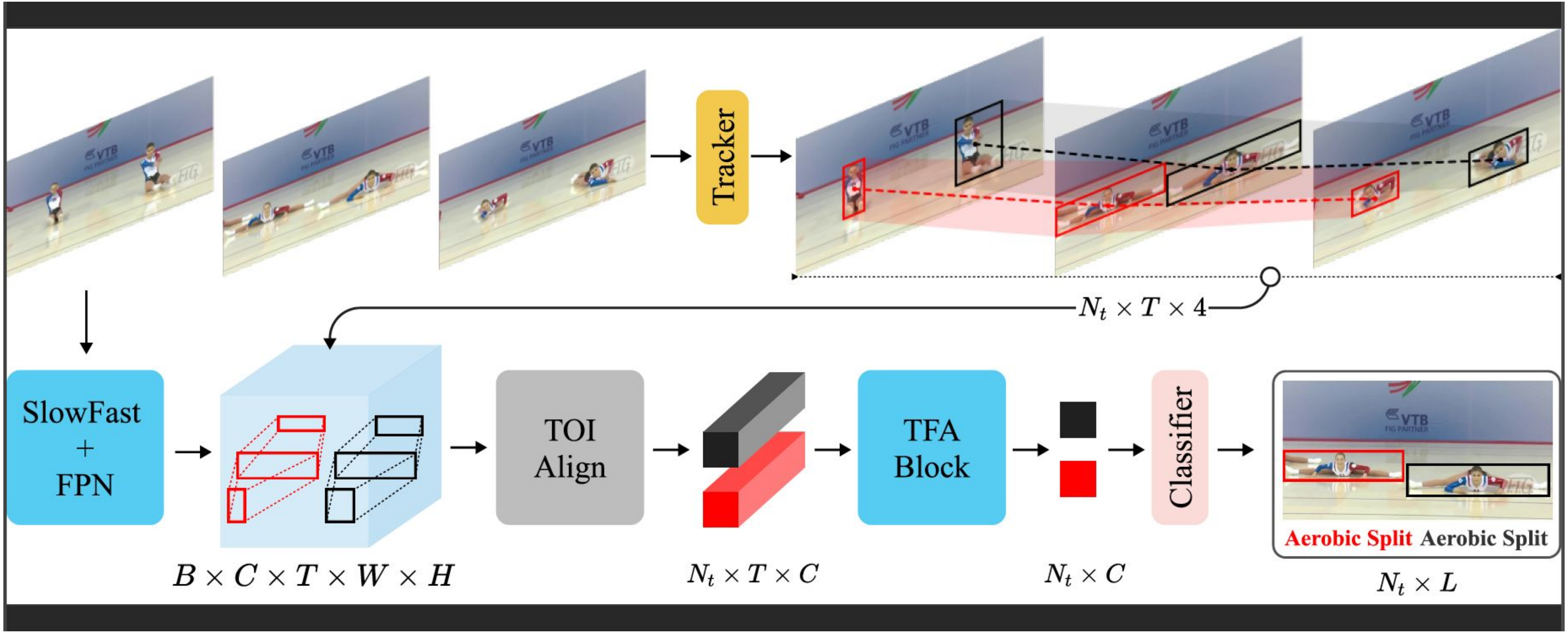
Classified Tracks

Temporal Trimming of Action tracks

⬆ Network gets to learn from more better feature accumulation

⬇ Tracking has to be Good

Track Aware Action Detector (TAAD)



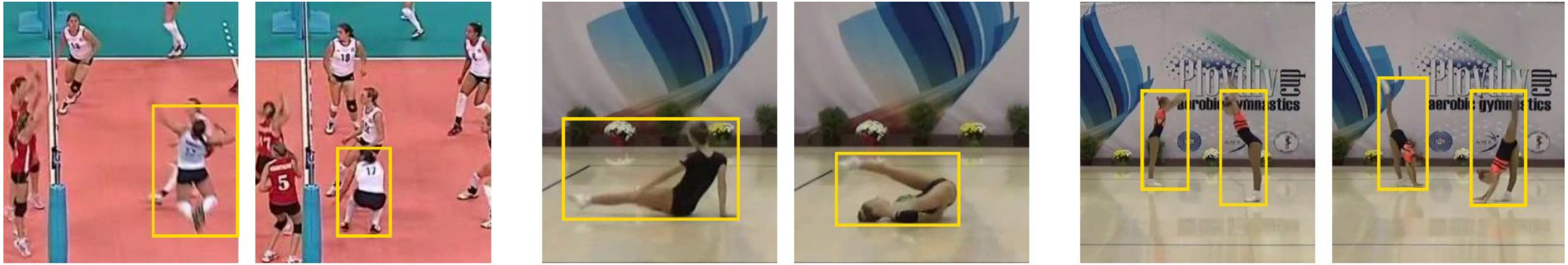
Temporal Feature Aggregation

- Maxpool
- ASPP
- TCN

- CovNxt block
- Swin Block
- MHA

How Do We Analyse Results?

Definition of motion type



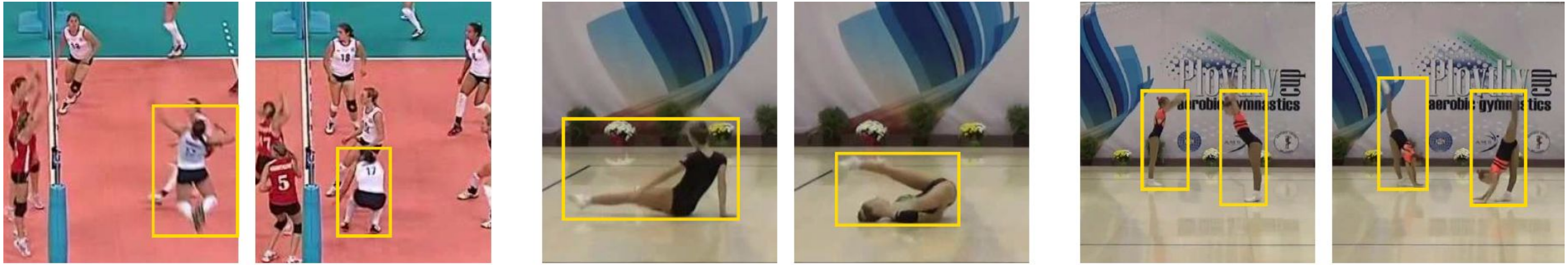
(a) Large movement
IoU: 0.00

(b) Medium movement
IoU: 0.44

(c) Small movement
IoU: 0.85

How Do We Analyse Results?

Definition of motion type



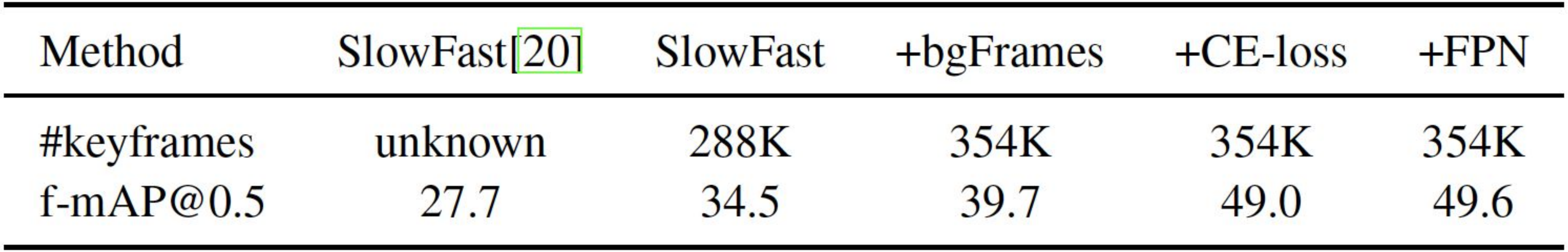
(a) Large movement
IoU: 0.00

(b) Medium movement
IoU: 0.44

(c) Small movement
IoU: 0.85

$$\text{MultiSports} = \begin{cases} \text{Large}, & \text{IoU} \in [0.00, 0.21] \\ \text{Medium}, & \text{IoU} \in [0.21, 0.51] \\ \text{Small}, & \text{IoU} \in [0.51, 1.00] \end{cases}$$

Baseline Improvements



Method	SlowFast[20]	SlowFast	+bgFrames	+CE-loss	+FPN
#keyframes	unknown	288K	354K	354K	354K
f-mAP@0.5	27.7	34.5	39.7	49.0	49.6

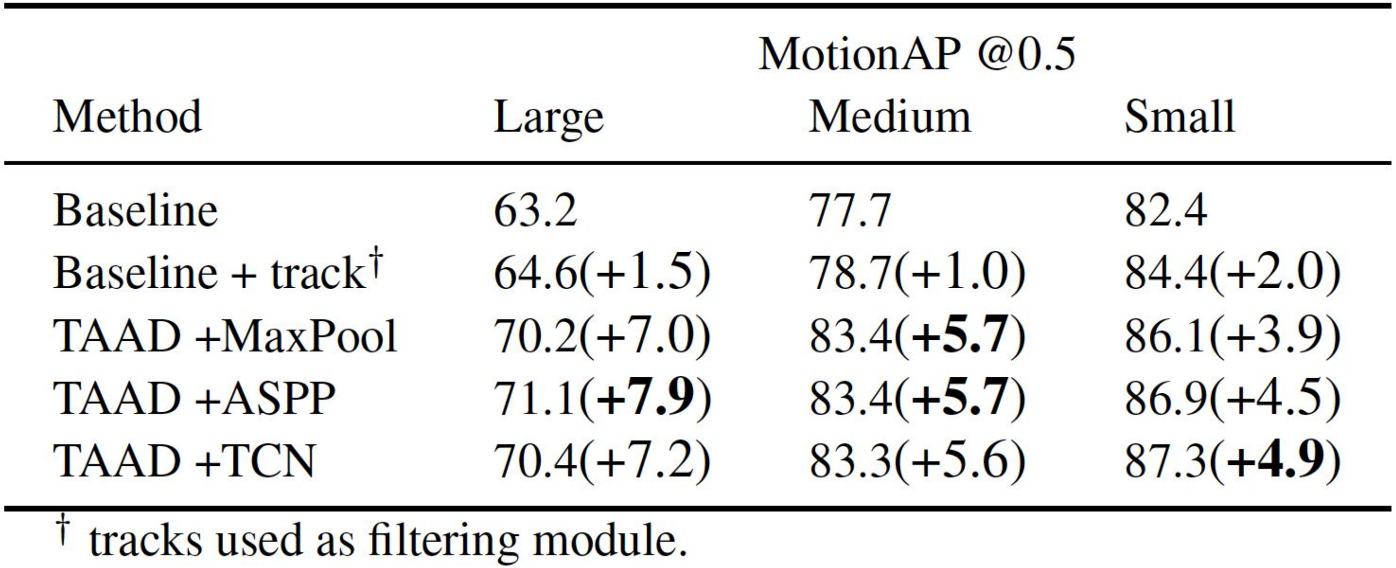
- SlowFastR50-8x8
- Input : 32 frames
- Batch Size: 32
- Optimiser SGD with 0.05 LR

Impact of adding Background Frame Training in Baseline

Boxtype	#keyframes	Trimmed	Untrimmed
GTframes-GTboxes	288K	47.0	32.2
+GTframes-proposals	288K	48.3	34.5
+every8thBGframe	355K	48.8	39.7 (+5.2)
+every6thBGframe	376K	49.3	40.5
+every4thBGframe	421K	49.3	41.5
+every2ndBGframe	553K	49.2	42.3

Backbone : Slowfast8x8-R50

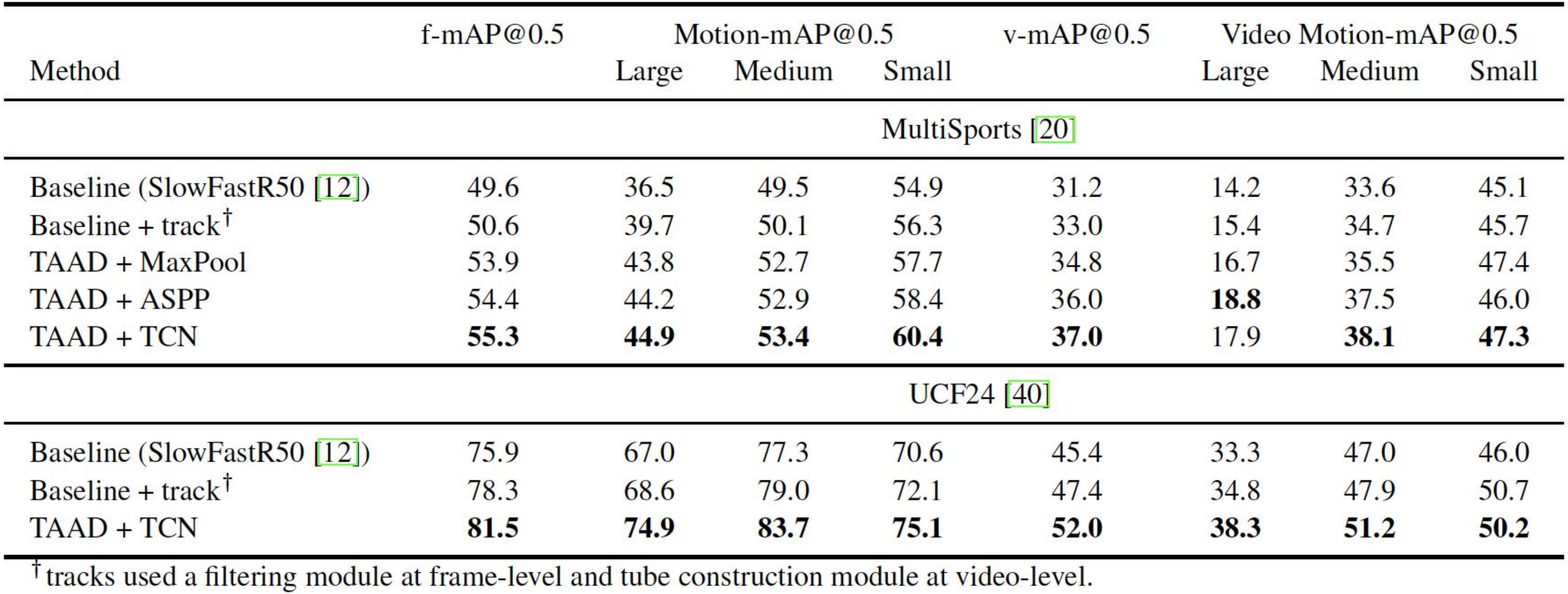
Motion-wise results (MotionAP)



	MotionAP @0.5		
Method	Large	Medium	Small
Baseline	63.2	77.7	82.4
Baseline + track[†]	64.6(+1.5)	78.7(+1.0)	84.4(+2.0)
TAAD +MaxPool	70.2(+7.0)	83.4(**+5.7**)	86.1(+3.9)
TAAD +ASPP	71.1(**+7.9**)	83.4(**+5.7**)	86.9(+4.5)
TAAD +TCN	70.4(+7.2)	83.3(+5.6)	87.3(**+4.9**)

[†] tracks used as filtering module.

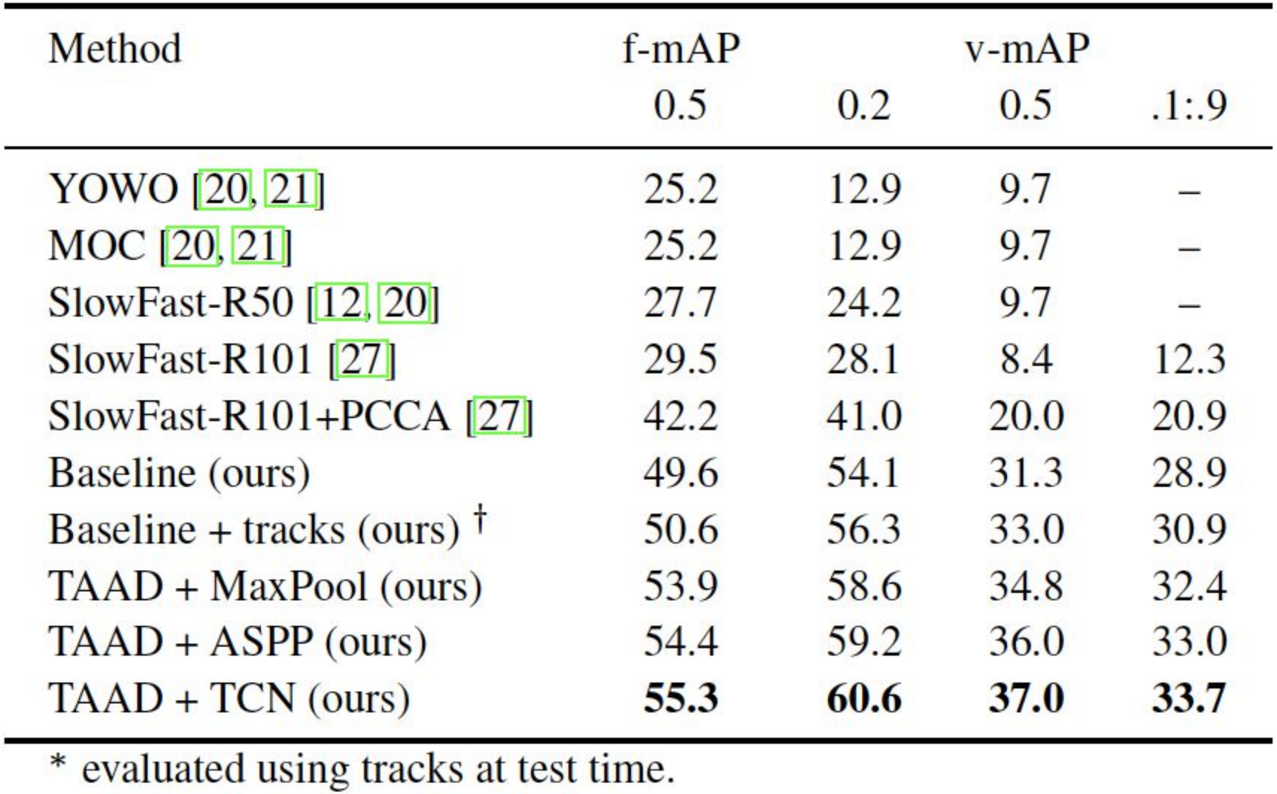
Motion-wise results (Motion-mAP)



	f-mAP@0.5	Motion-mAP@0.5			v-mAP@0.5	Video Motion-mAP@0.5		
Method		Large	Medium	Small		Large	Medium	Small
				MultiSports [20]				
Baseline (SlowFastR50 [12])	49.6	36.5	49.5	54.9	31.2	14.2	33.6	45.1
Baseline + track†	50.6	39.7	50.1	56.3	33.0	15.4	34.7	45.7
TAAD + MaxPool	53.9	43.8	52.7	57.7	34.8	16.7	35.5	47.4
TAAD + ASPP	54.4	44.2	52.9	58.4	36.0	**18.8**	37.5	46.0
TAAD + TCN	**55.3**	**44.9**	**53.4**	**60.4**	**37.0**	17.9	**38.1**	**47.3**
				UCF24 [40]				
Baseline (SlowFastR50 [12])	75.9	67.0	77.3	70.6	45.4	33.3	47.0	46.0
Baseline + track†	78.3	68.6	79.0	72.1	47.4	34.8	47.9	50.7
TAAD + TCN	**81.5**	**74.9**	**83.7**	**75.1**	**52.0**	**38.3**	**51.2**	**50.2**

† tracks used a filtering module at frame-level and tube construction module at video-level.

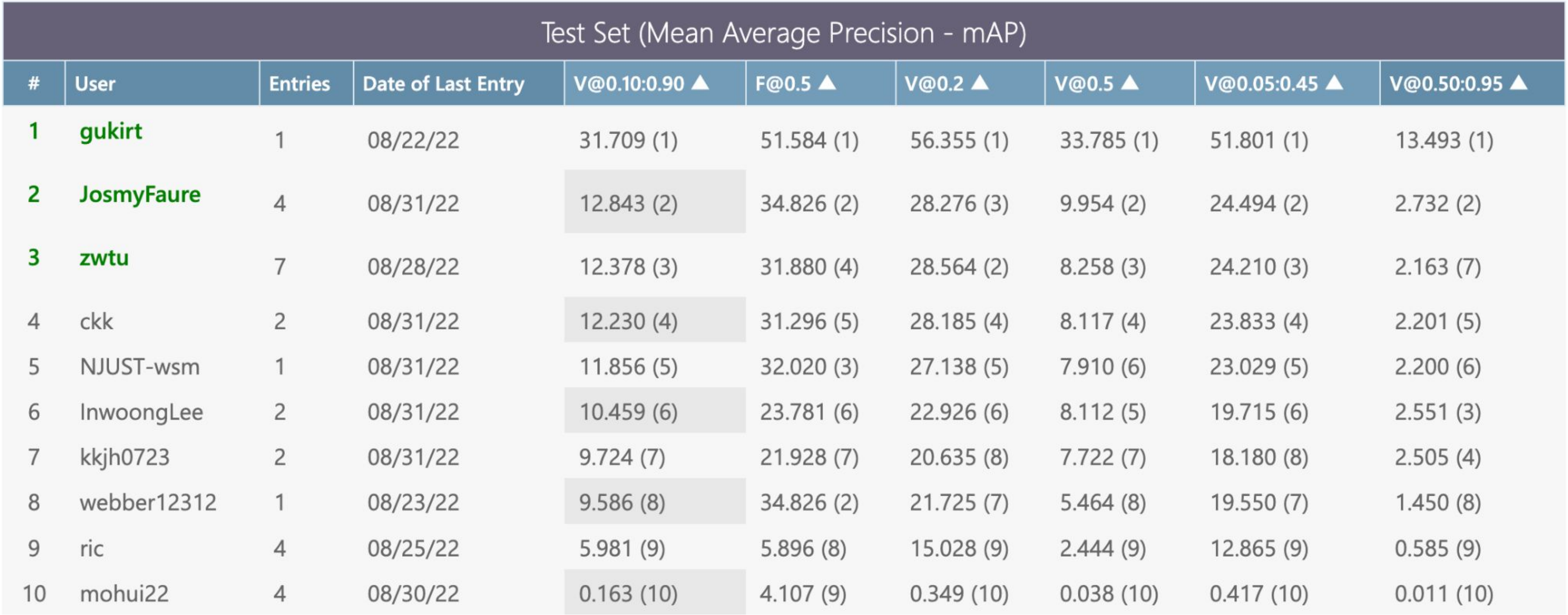
State-of-the-art Comparison



Method	f-mAP	v-mAP		
	0.5	0.2	0.5	.1:.9
YOWO [20, 21]	25.2	12.9	9.7	–
MOC [20, 21]	25.2	12.9	9.7	–
SlowFast-R50 [12, 20]	27.7	24.2	9.7	–
SlowFast-R101 [27]	29.5	28.1	8.4	12.3
SlowFast-R101+PCCA [27]	42.2	41.0	20.0	20.9
Baseline (ours)	49.6	54.1	31.3	28.9
Baseline + tracks (ours) †	50.6	56.3	33.0	30.9
TAAD + MaxPool (ours)	53.9	58.6	34.8	32.4
TAAD + ASPP (ours)	54.4	59.2	36.0	33.0
TAAD + TCN (ours)	**55.3**	**60.6**	**37.0**	**33.7**

* evaluated using tracks at test time.

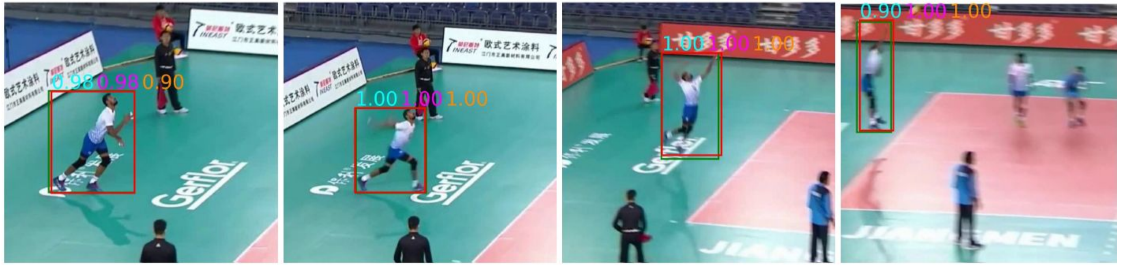
Test-set Results



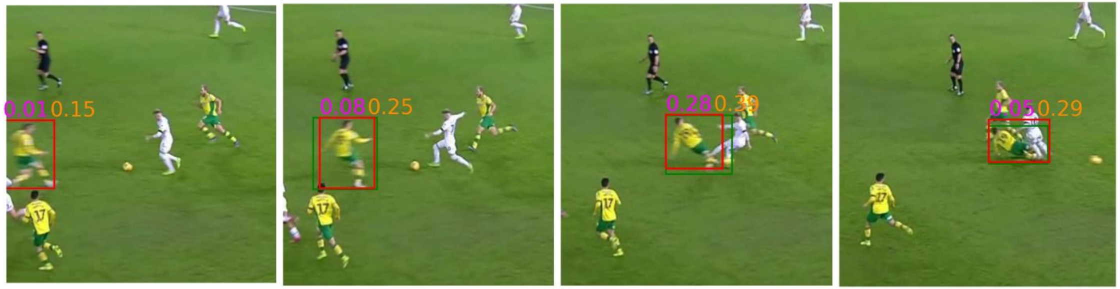
Test Set (Mean Average Precision - mAP)

#	User	Entries	Date of Last Entry	V@0.10:0.90 ▲	F@0.5 ▲	V@0.2 ▲	V@0.5 ▲	V@0.05:0.45 ▲	V@0.50:0.95 ▲
1	gukirt	1	08/22/22	31.709 (1)	51.584 (1)	56.355 (1)	33.785 (1)	51.801 (1)	13.493 (1)
2	JosmyFaure	4	08/31/22	12.843 (2)	34.826 (2)	28.276 (3)	9.954 (2)	24.494 (2)	2.732 (2)
3	zwtu	7	08/28/22	12.378 (3)	31.880 (4)	28.564 (2)	8.258 (3)	24.210 (3)	2.163 (7)
4	ckk	2	08/31/22	12.230 (4)	31.296 (5)	28.185 (4)	8.117 (4)	23.833 (4)	2.201 (5)
5	NJUST-wsm	1	08/31/22	11.856 (5)	32.020 (3)	27.138 (5)	7.910 (6)	23.029 (5)	2.200 (6)
6	InwoongLee	2	08/31/22	10.459 (6)	23.781 (6)	22.926 (6)	8.112 (5)	19.715 (6)	2.551 (3)
7	kkjh0723	2	08/31/22	9.724 (7)	21.928 (7)	20.635 (8)	7.722 (7)	18.180 (8)	2.505 (4)
8	webber12312	1	08/23/22	9.586 (8)	34.826 (2)	21.725 (7)	5.464 (8)	19.550 (7)	1.450 (8)
9	ric	4	08/25/22	5.981 (9)	5.896 (8)	15.028 (9)	2.444 (9)	12.865 (9)	0.585 (9)
10	mohui22	4	08/30/22	0.163 (10)	4.107 (9)	0.349 (10)	0.038 (10)	0.417 (10)	0.011 (10)

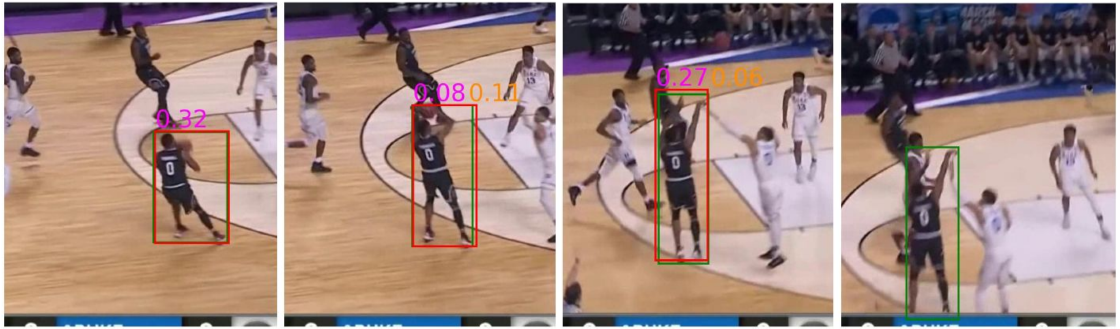
Visuals



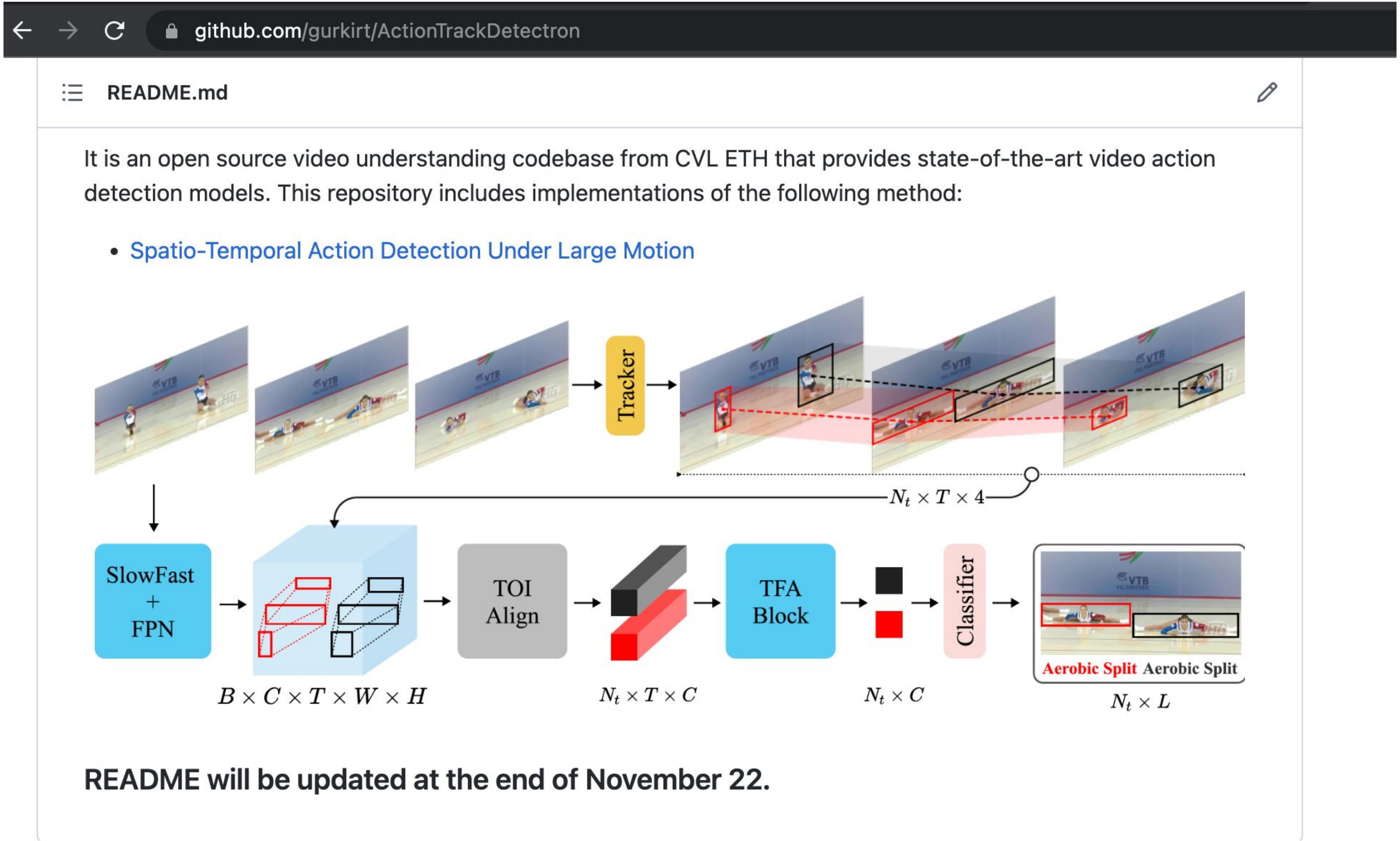
(a) Volleyball-serve: Large-motion: Speed 0.17 IoU; Overlap: Baseline 79%, ASPP 79%, TCN 79 %



(b) Football-steal: Large-motion: Speed 0.03 IoU; Overlap: ASPP 77%, TCN 77%



(c) Basketball-3-point-shot: Large-motion: Speed 0.07 IoU; Overlap: ASPP 68%, TCN 57 %



github.com/gurkirt/ActionTrackDetectron

README.md

It is an open source video understanding codebase from CVL ETH that provides state-of-the-art video action detection models. This repository includes implementations of the following method:

- Spatio-Temporal Action Detection Under Large Motion

README will be updated at the end of November 22.

Discussion / Q&A