Deeper Action

ICCV 2021 Workshop

ICCV DeeperAction Challenge - MultiSports Track on Spatio-temporal Action Detection

Yixuan Li Lei Chen Runyu He Zhenzhi Wang Gangshan Wu Limin Wang

State Key Laboratory for Novel Software Technology, Nanjing University, China

Track 2, DeeperAction, ICCV 2021

南京大學 NANJING UNIVERSITY

MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

Task: Spatio-Temporal Action Detection

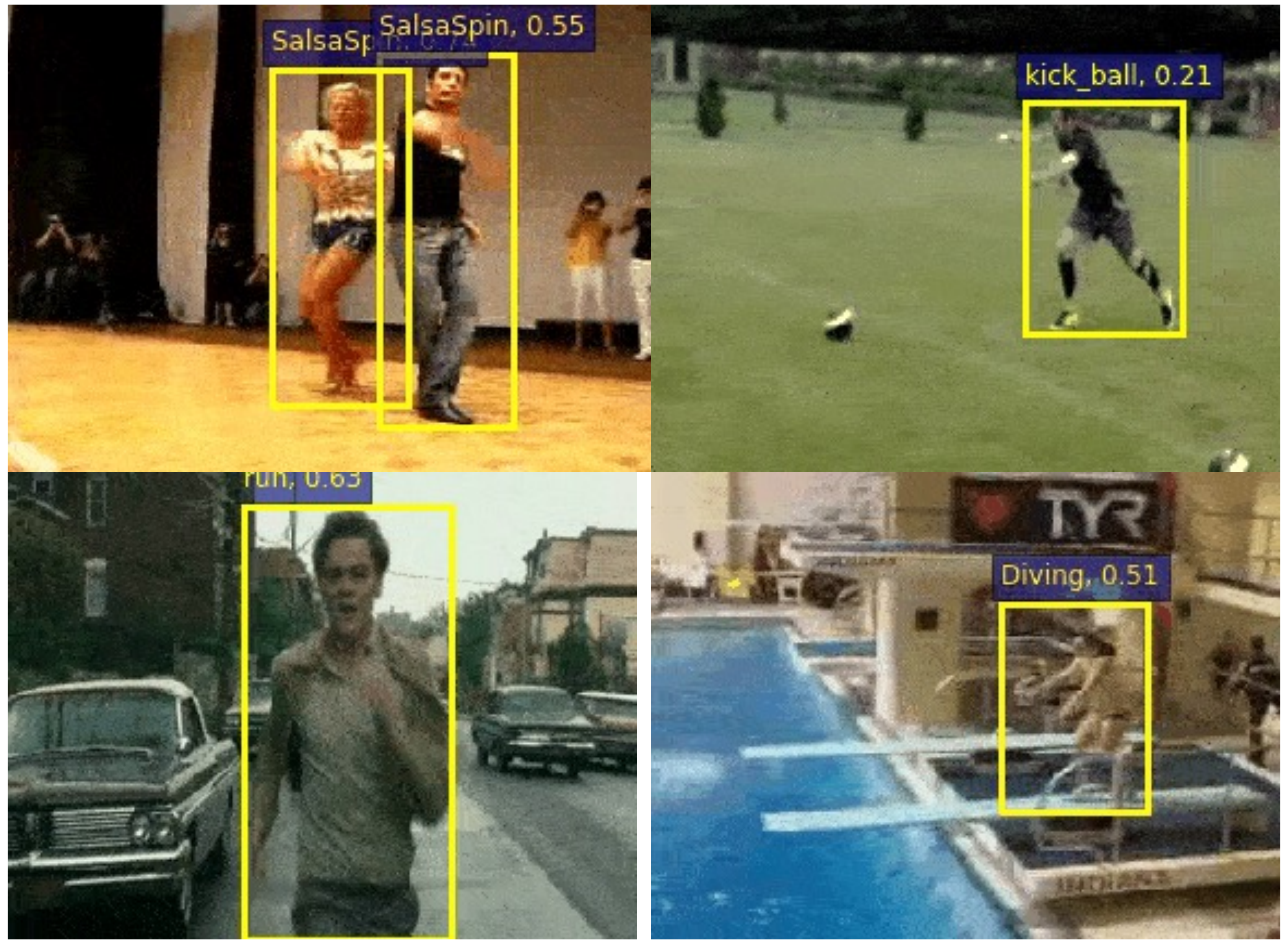
Input

- untrimmed video

Video Input → Tube Output

Output

- action labels
- temporal boundaries
- actor trackings



Deeper Action

Outline

ICCV 2021 Workshop

DataSet Introduction

Competition Introduction

Part 1

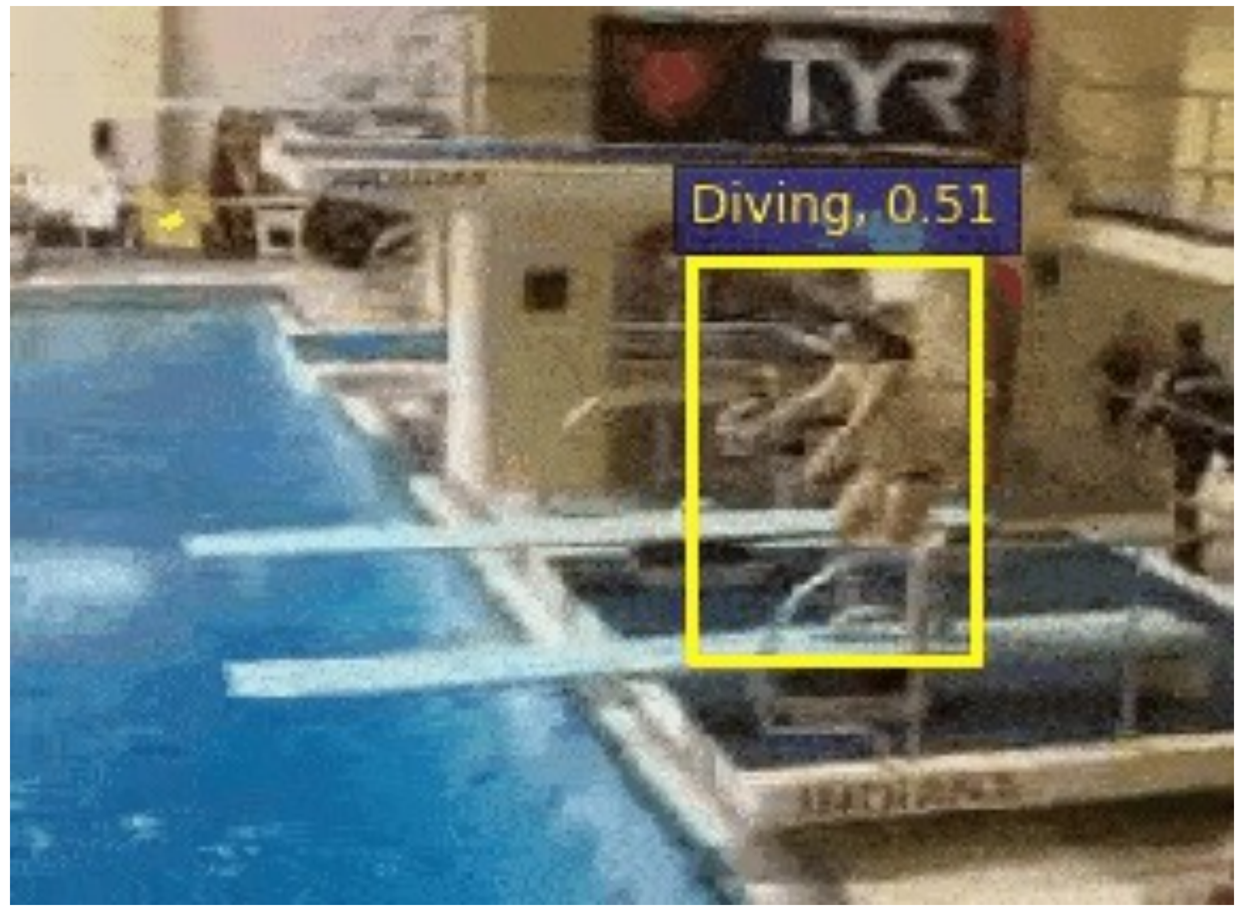
DataSet

Introduction

Current Benchmarks

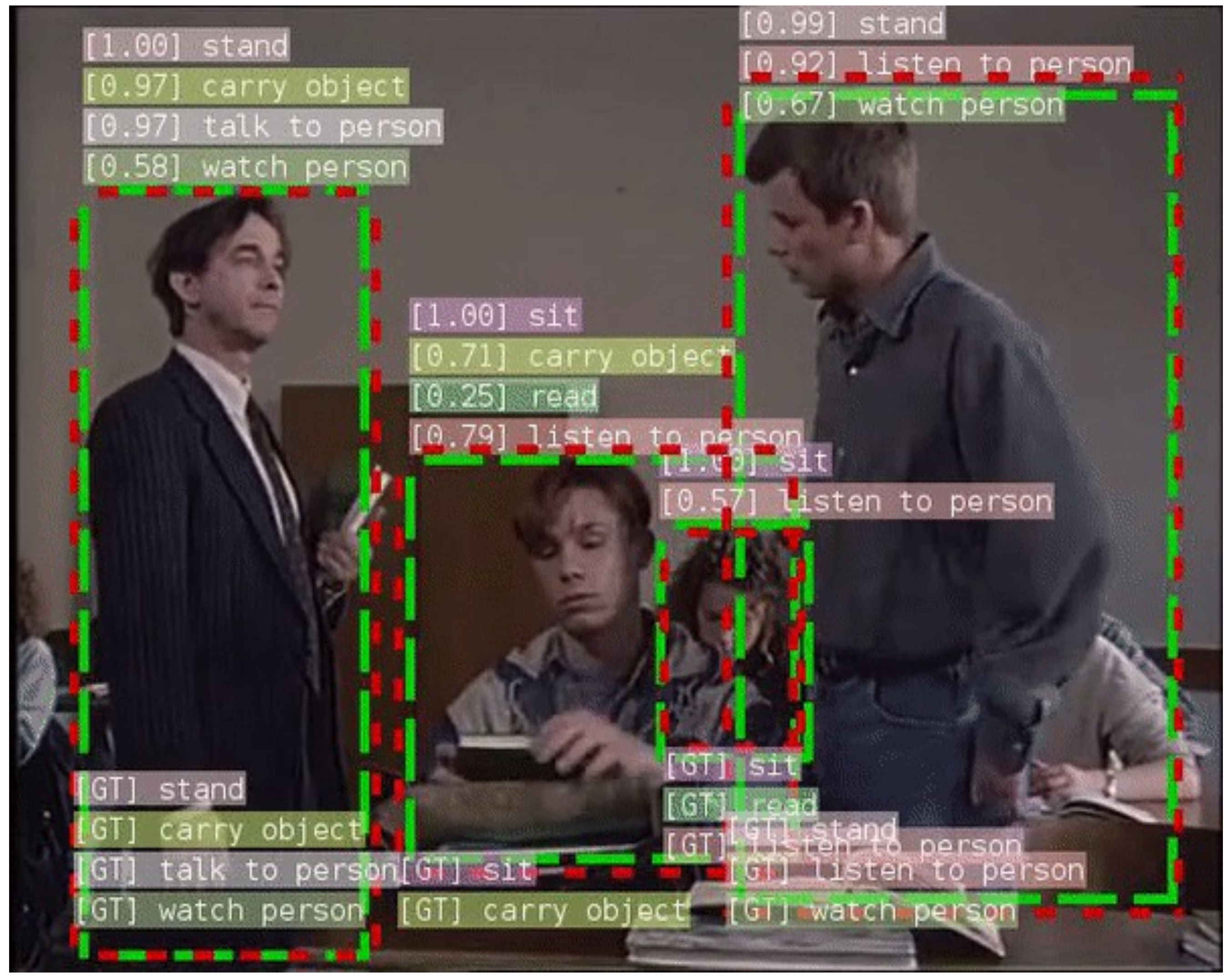
UCF101-24 / JHMDB

- Dense annotations (25 FPS).
- Single-person scenes (most videos).
- Coarse-grained actions.



AVA

- Sparse annotations (1 FPS).
- Atomic actions.
- Without clear temporal boundaries.



Motivation

Expected Features

- Multi-person scenes.
- Dense annotations (25 FPS).
- Well-defined temporal boundaries.
- Fine-grained and complex actions.

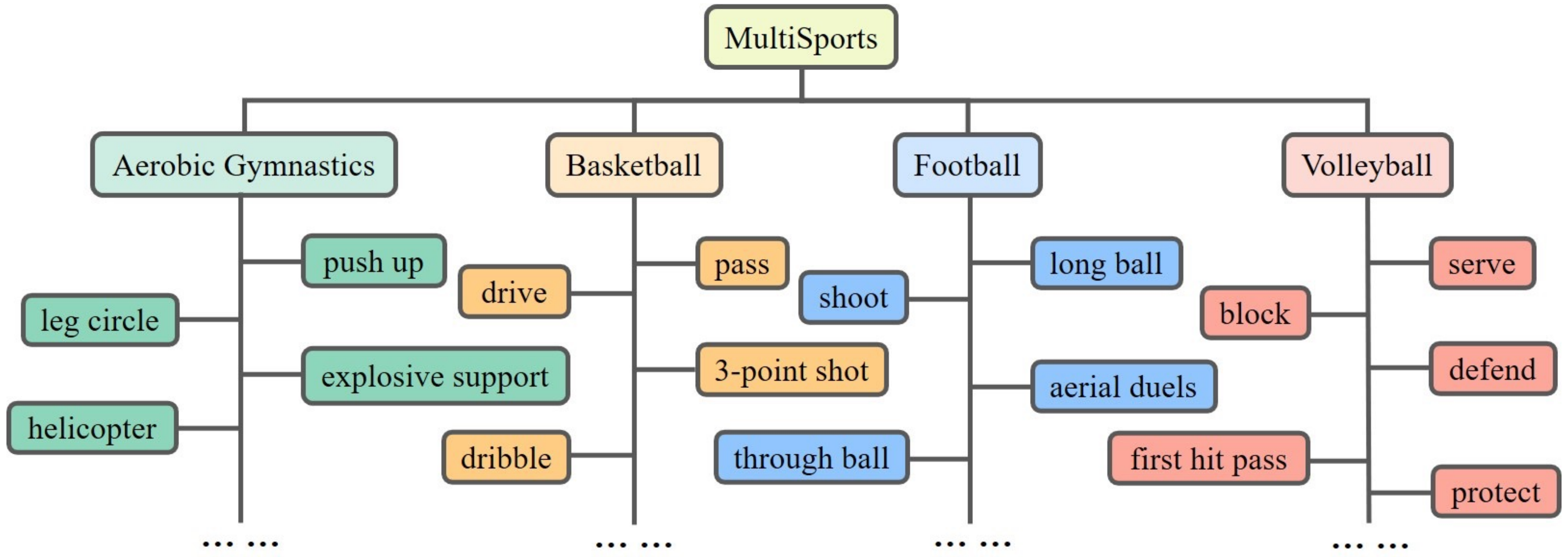
Deeper Action

Annotation Process

ICCV 2021 Workshop

Action vocabulary generation

- Official documentations for aerobic gymnastics.
- Athletes set the rules in an iterative way for ball sports.



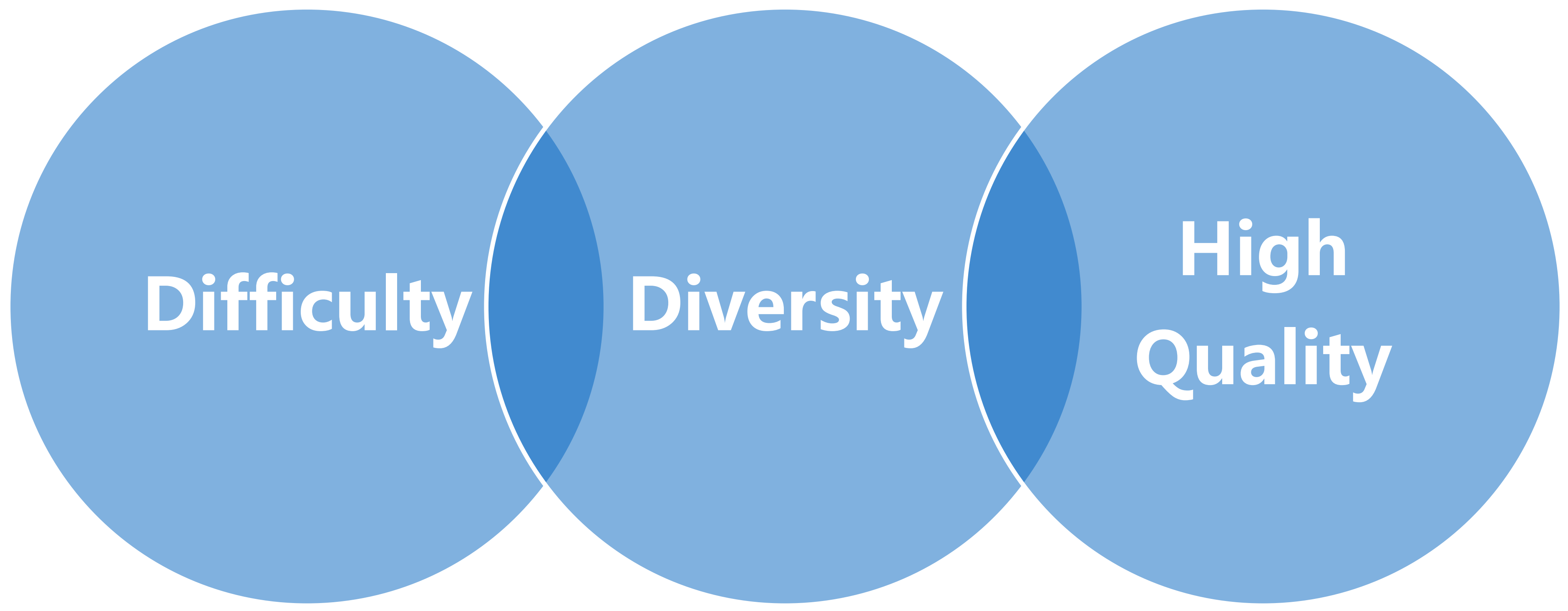
Data Preparation

- 720P or 1080P professional competitions.
- Different levels, countries and genders.

Annotation Process

Two Stage Action Annotation

- Athletes annotate action label, boundary and the first frame box.
- FCOT tracker [1] + Crowd-sourced annotators adjust boxes of tracking results at each frame.



Quality Control

- Double check actions and boundaries for each clip.
- Double check boxes in 5 FPS for each instance.

[1] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Fully convolutional online tracking. *CoRR*, abs/2004.07109, 2020.

Statistics

Compare with other datasets

- More fine-grained actions categories.
- More instances and instances per clip.
- The largest number of bounding boxes.

Long-tailed distribution.

Large variations of action instance duration.

	anno type	# act.	# inst.	avg act./vid. dur.	# bbox
J-HMDB [20]	Tube	21	928	1.2s / 1.2s	32k
UCF101-24 [46]	Tube	24	4458	5.1s / 6.9s	574k
AVA V2.1 [15]*	Frame	80	~56000†	Sparse‡ / 15m	426k
AVA-Kinetics [25]*	Frame	80	~186000†	-	590k
HACS [64]	Segment	200	140k	33.2s / 148.7s	-
FineGym V1.0 [40]	Segment	530	32697	1.7s / 10m	-
Aerobic gym.	Tube	21	8703	1.5s / 30.7s	325k
Volleyball	Tube	12	7645	0.7s / 10.5s	139k
Football	Tube	15	12254	0.7s / 22.6s	225k
Basketball	Tube	18	9099	0.9s / 19.7s	213k
Ours in total	Tube	66	37701	1.0s / 20.9s	902k

Table 2. Comparison of statistics between existing action detection datasets and our MultiSports v1.0. (* only train and val sets' ground-truths are available; *Tube* with class, temporal boundary and spatial localization; *Frame* with class and spatial localization; *Segment* with class and temporal boundary; † number of person tracklets, each of which has one or more action labels; ‡ 1fps action annotations)

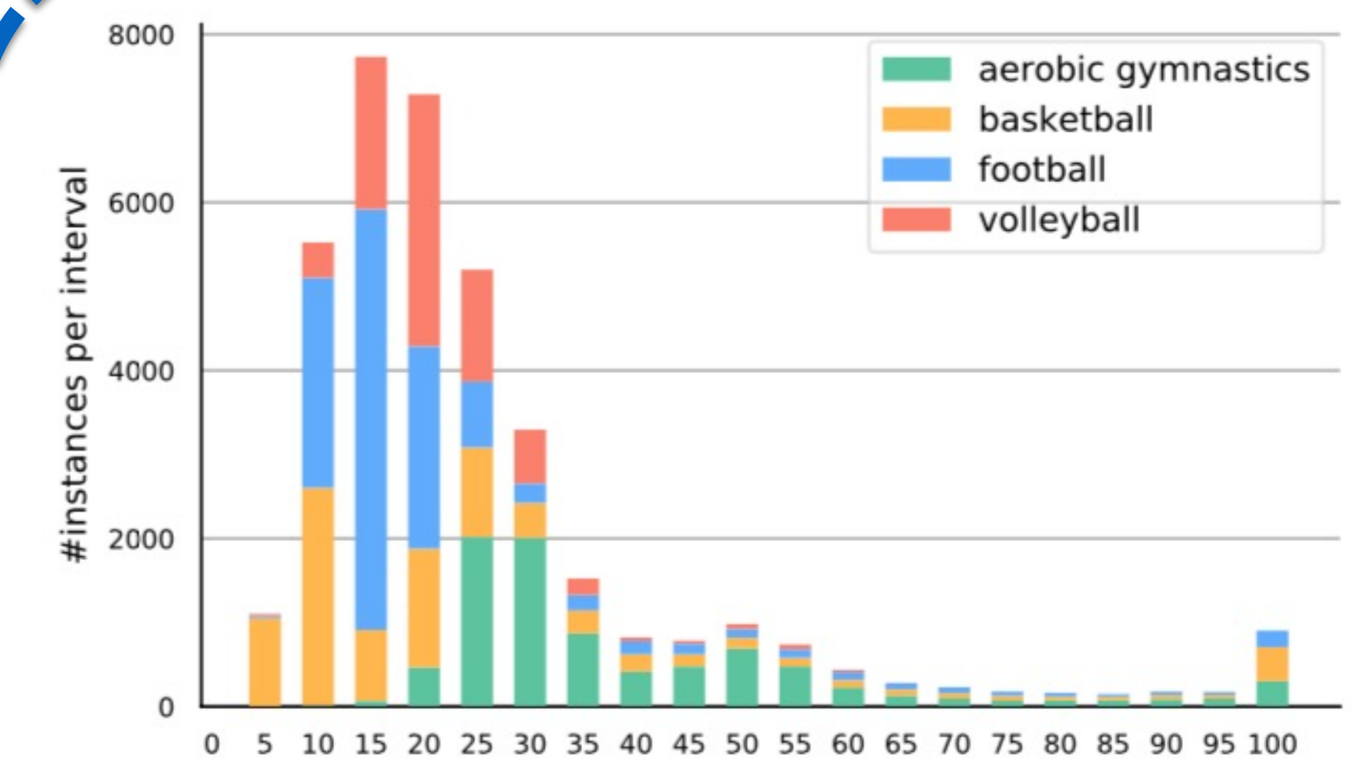


Figure 4. Statistics of action instance duration in MultiSports, where the x-axis is the number of frames and we count all instances longer than 95 frames in the last bar.

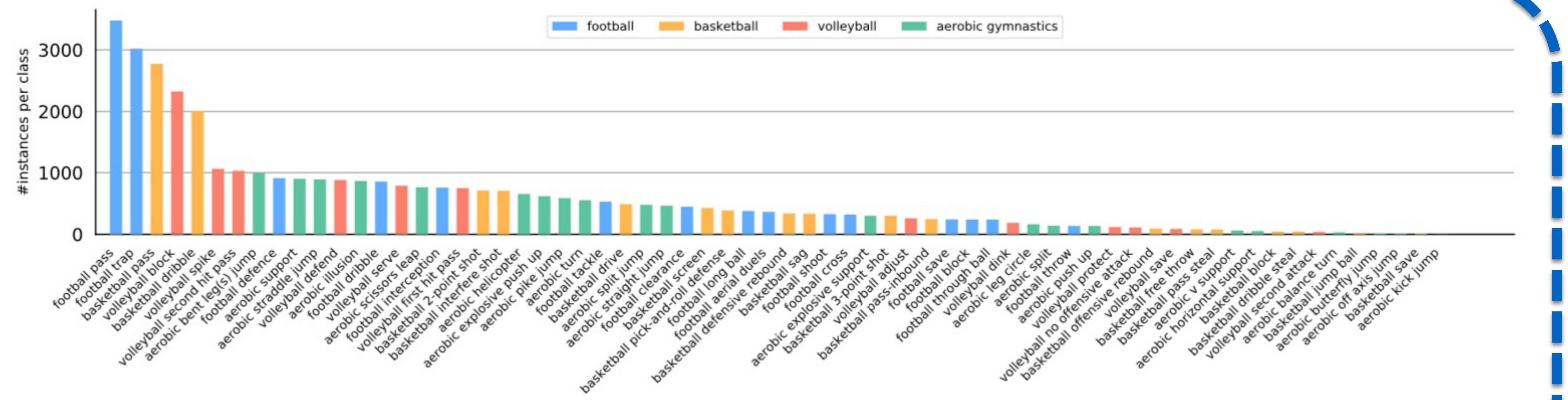


Figure 3. Statistics of each action class's data size in MultiSports, which is sorted by descending order with 4 colors indicating 4 different sports. For actions in the different sports sharing the same name, we add the name of sports before them.

Spatio-Temporal Action Detection Results

Deeper Action

ICCV 2021 Workshop

UCF101-24 / JHMDB methods

- Low performance on MultiSports.
- Largest performance drop occurs on frame-level detector ROAD.

AVA methods

- More evident performance gap between two methods on MultiSports.
- Actions with intense motion gain large improvement.

Method	Res	MultiSports			UCF101-24			JHMDB			AVA
		F@0.5	V@0.2	V@0.5	F@0.5	V@0.2	V@0.5	F@0.5	V@0.2	V@0.5	F-mAP@0.5
ROAD [44]	300 × 300	3.90	0.00	0.00	70.7	69.8	40.9	-	60.8	59.7	-
YOWO [23]	224 × 224	9.28	10.78	0.87	71.10	72.97	46.42	74.51	88.05	82.57	-
MOC [27] (K=7)	288 × 288	22.51	12.13	0.77	78.0	82.8	53.8	70.8	77.3	77.2	-
MOC [27] (K=11)	288 × 288	25.22	12.88	0.62	-	-	-	-	-	-	-
SlowOnly Det., 4 × 16 [11]	short side 256	16.70	15.71	5.50	-	-	-	-	-	-	20.02
SlowFast Det., 4 × 16 [11]	short side 256	27.72	24.18	9.65	-	-	-	-	-	-	24.56

Table 3. Comparison of the state-of-the-art methods on MultiSports, UCF101-24, JHMDB and AVA.

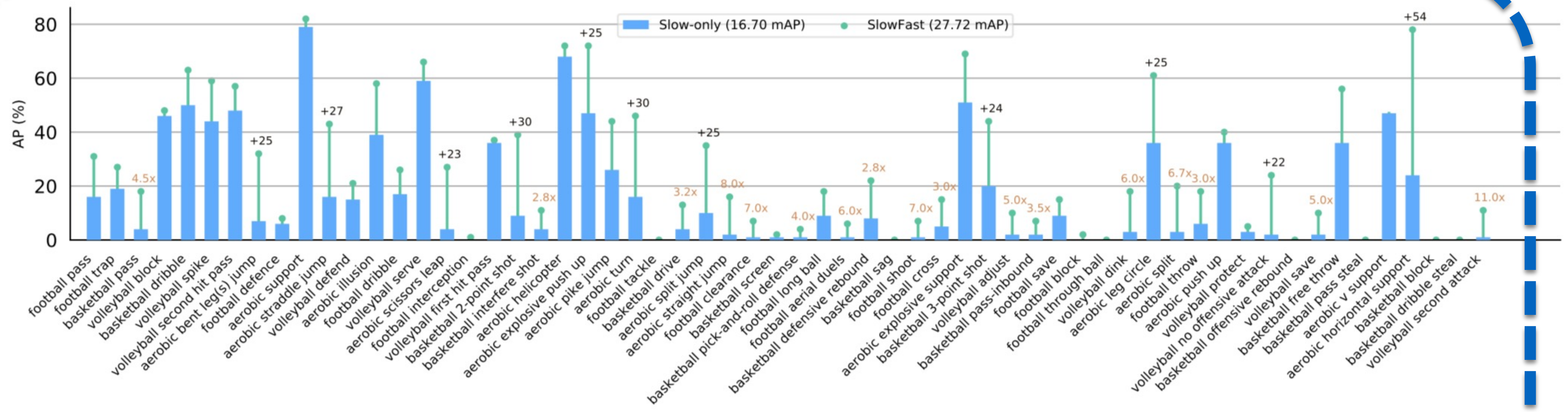


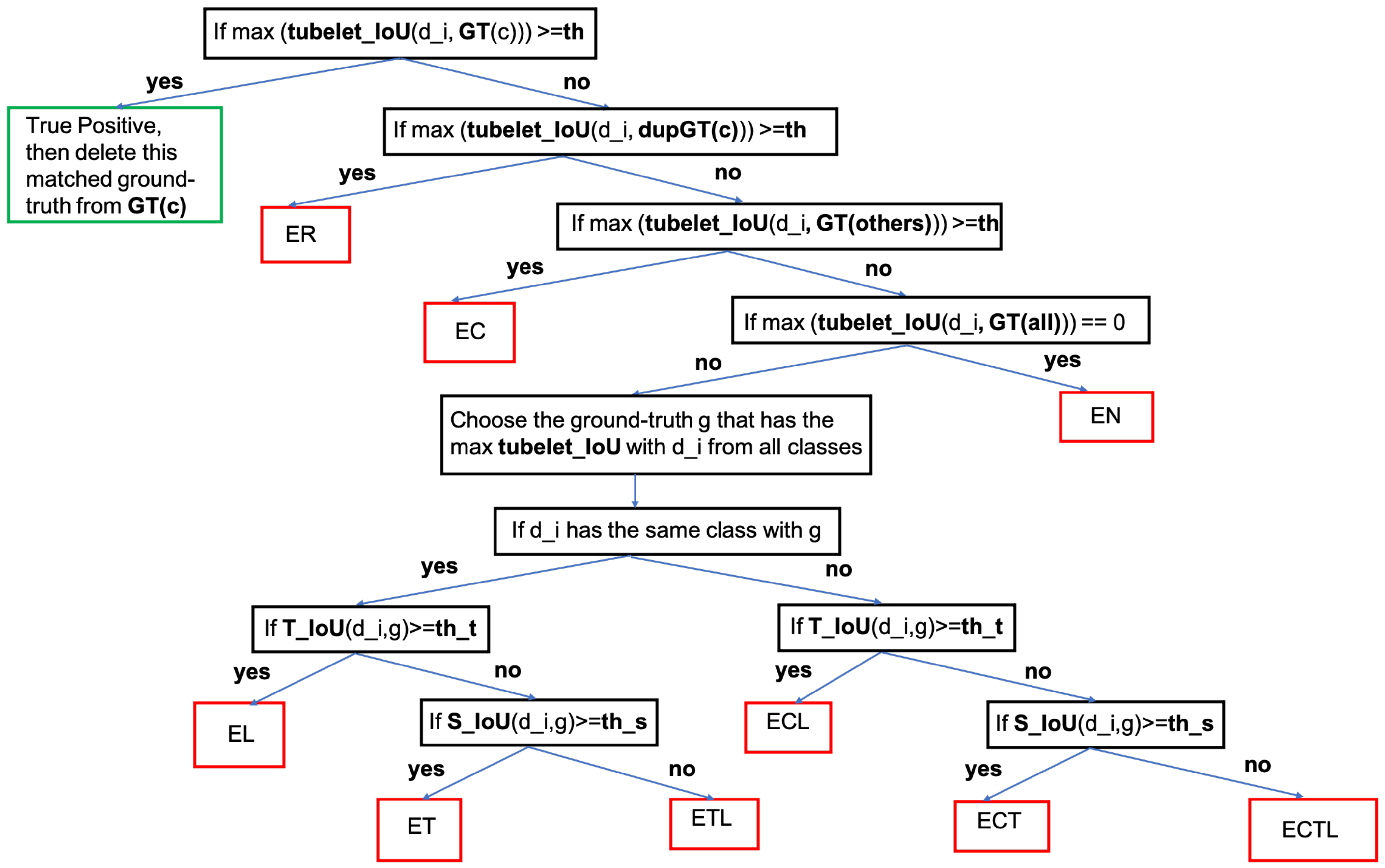
Figure 5. SlowOnly vs. SlowFast frame-mAP. Categories are sorted by descending order on the number of instances.

Challenges

Error Analysis (Video mAP)

- E_R : Repeat Error.
- E_N : No spatio-temporal interaction with any GT.
- E_M : Ground-truth missing.
- E_T: Only temporal localization error.
- E_C : Only classification error.
- E_L : Only spatial localization error.
- E_{CT}, E_{CL}, E_{TL} , E_{CTL}: Contain many kinds of error.

For each detected tubelet d_i from a sorted list by descending order of confidence score of class c.
Notation: **th:** threshold; **th_t :** the square root of **th**; **th_s:** the square root of **th; GT(c):** set of ground-truths of class c; **dupGT(c):** copy of **GT(c)** ; **GT(others) :** set of all ground-truths that not in class c; **GT(all):** set of all ground-truths; **T_IoU :** the temporal domain IoU; **S_IoU:** the average of the IoU between the overlapping frames; **tubelet_IoU: T_IoU* S_IoU.**



Challenges

SlowFast

- Make fewer false positive predictions than MOC but still miss many hard examples.
- Classification is hard for SlowFast.

MOC

- Classification is the biggest problem for MOC.
- Temporal localization is more difficult than spatial localization.

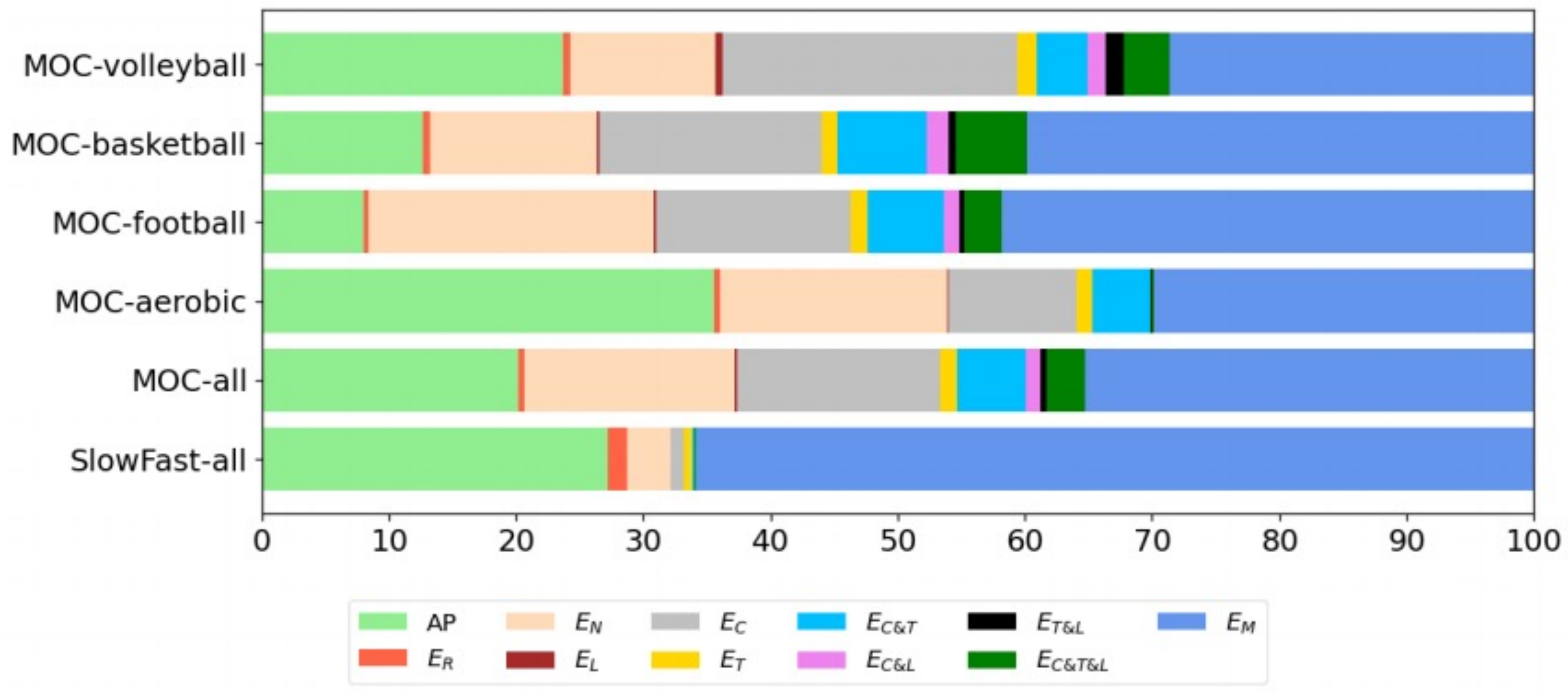


Figure 6. Error Analysis. AP is the correct detection. The threshold for a ground-truth matched by a detection is 0.1

Classification > Temporal localization > Spatial localization

Analysis

The importance of temporal information.

K	MultiSports F@0.5	MultiSports V@0.2	MultiSports V@0.5	UCF101-24 F@0.5	UCF101-24 V@0.2	UCF101-24 V@0.5
1	14.61	12.53	1.06	68.33	65.47	31.50
3	17.22	11.88	0.76	69.94	75.83	45.94
5	19.29	11.81	**0.98**	71.63	77.74	49.55
7	22.51	12.13	0.77	**73.14**	**78.81**	**51.02**
9	24.22	11.72	0.57	72.17	77.94	50.16
11	**25.22**	**12.88**	0.62	-	-	-
13	24.28	11.23	0.57	-	-	-

Table 4. Exploration study of MOC on the *MultiSports* and UCF101-24 with different tubelet length K.

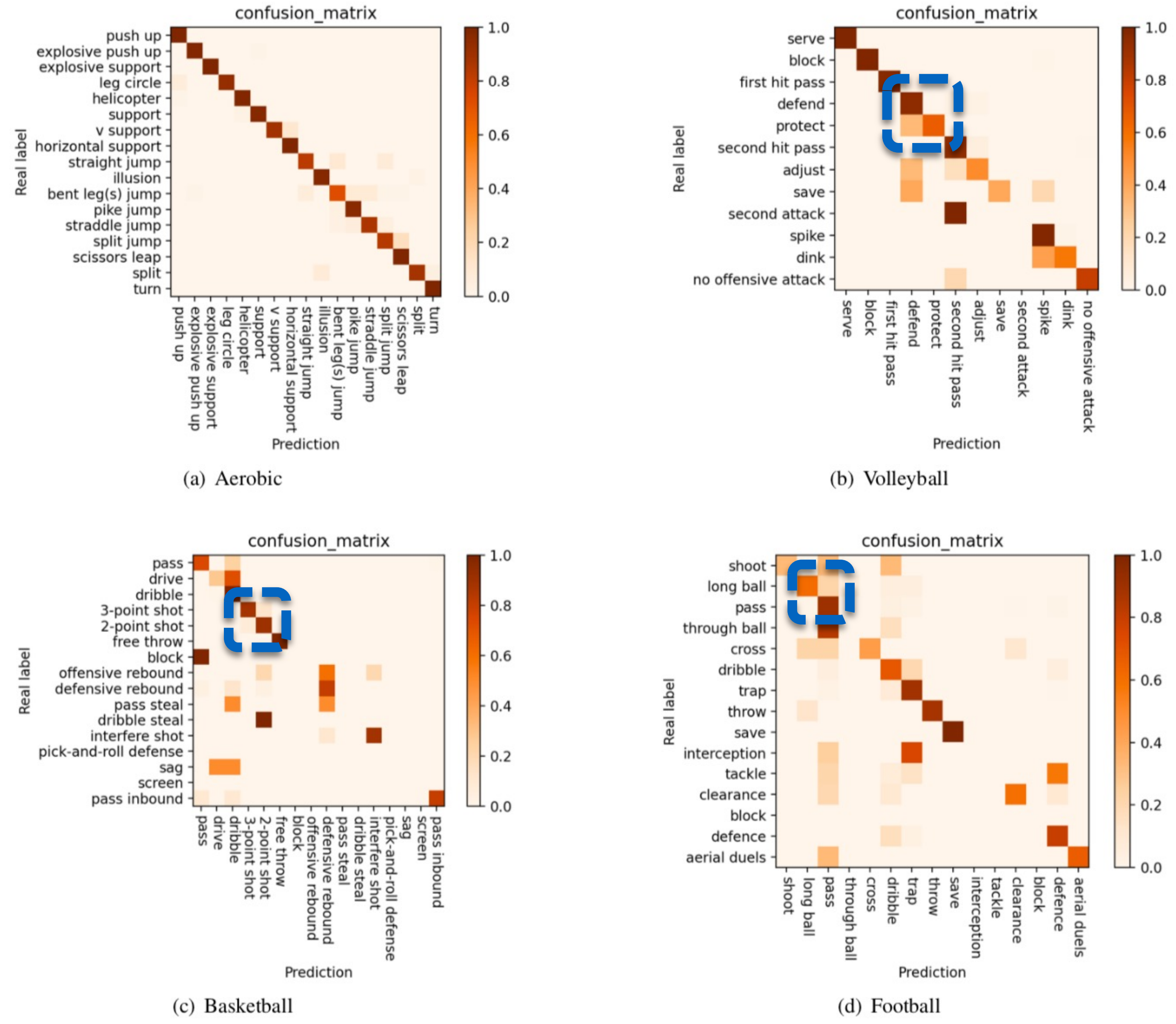
Trimmed vs. untrimmed settings.

Estimation	MultiSports F@0.5	MultiSports V@0.2	MultiSports V@0.5	AVA F-mAP@0.5
Untrimmed	27.72	24.18	9.65	22.57
Trimmed	38.71	24.95	18.34	24.56

Table 5. Test SlowFast Det. on AVA and *MultiSports* with trimmed way and untrimmed way.

Which action categories are challenging?

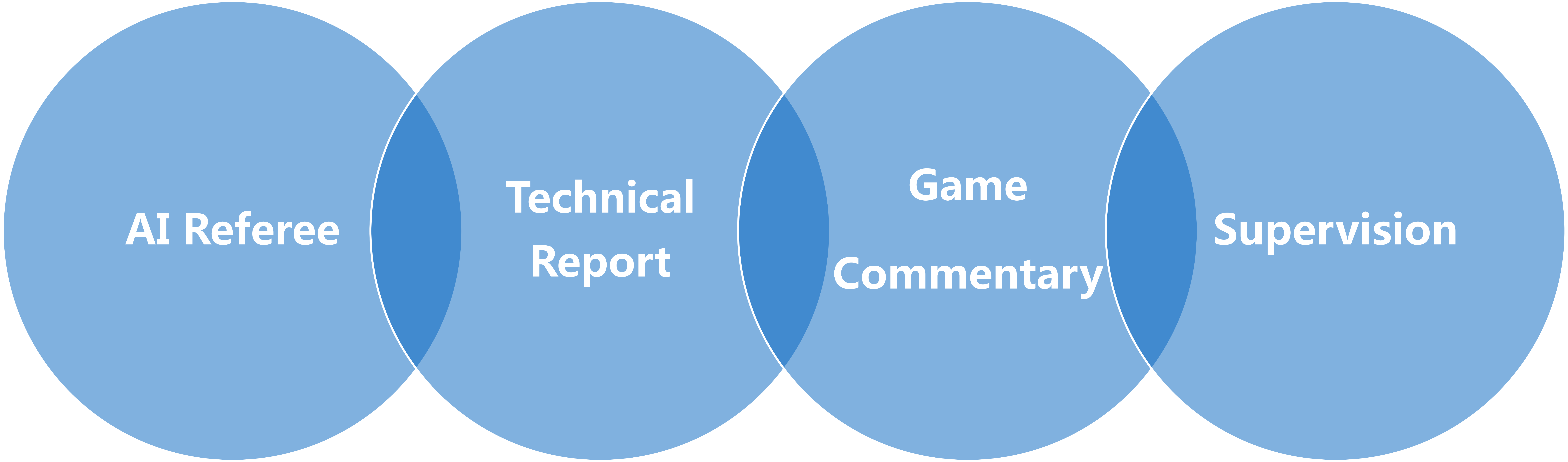
- Context modeling, e.g. basketball 2-point shot vs. 3-point shot.
- Reasoning, e.g. volleyball protect vs. defend.
- Long temporal modeling, e.g. football long ball vs. pass.



(a) Aerobic (b) Volleyball (c) Basketball (d) Football

Figure 10. Confusion Matrix of SlowFast Det. on different sports.

Potential Applications



Conclusion

Introduce the MultiSports dataset.

- ⇥ Raise new challenges for recognizing fine-grained action classes.
- ⇥ Require accurate localization of action boundaries in multiple-person situations.
- ⇥ High quality video data and dense annotations.
- ⇥ High diversity in competition levels, countries and genders.

Investigate several action detection baseline methods on MultiSports.

Provide detailed error analysis and ablation studies.

Part 2

Competition

Introduction

MulitSports Track

- Validation Phase: 2021.06.01-2021.08.31
- Testing Phase: 2021.09.01-2021.09.12

Deeper Action

ICCV DeeperAction Challenge - MultiSports Track on Spatiotemporal Action Detection

Organized by yixuanli

The challenge is Track 2 at ICCV DeeperAction Challenge. This track is for multi-person spatiotemporal action localization in sports videos.

Jun 01, 2021-Sep 12, 2021

187 participants

Edit	Unpublish	Participants	Submissions	Dumps

Evaluation

Video mAP

- 3D IoU: temporal IoU of two tracks × average of IoU between the overlapping frames.
- Threshold: 0.2, 0.5, 0.05:0.45, 0.5:0.95, 0.1:0.9
- Rank according to the **V@0.1:0.9**

Frame mAP

- Threshold: 0.5

Statistics

Valid Participants: 187

Valid Teams: 7 (Val Phase) + 10 (Test Phase)

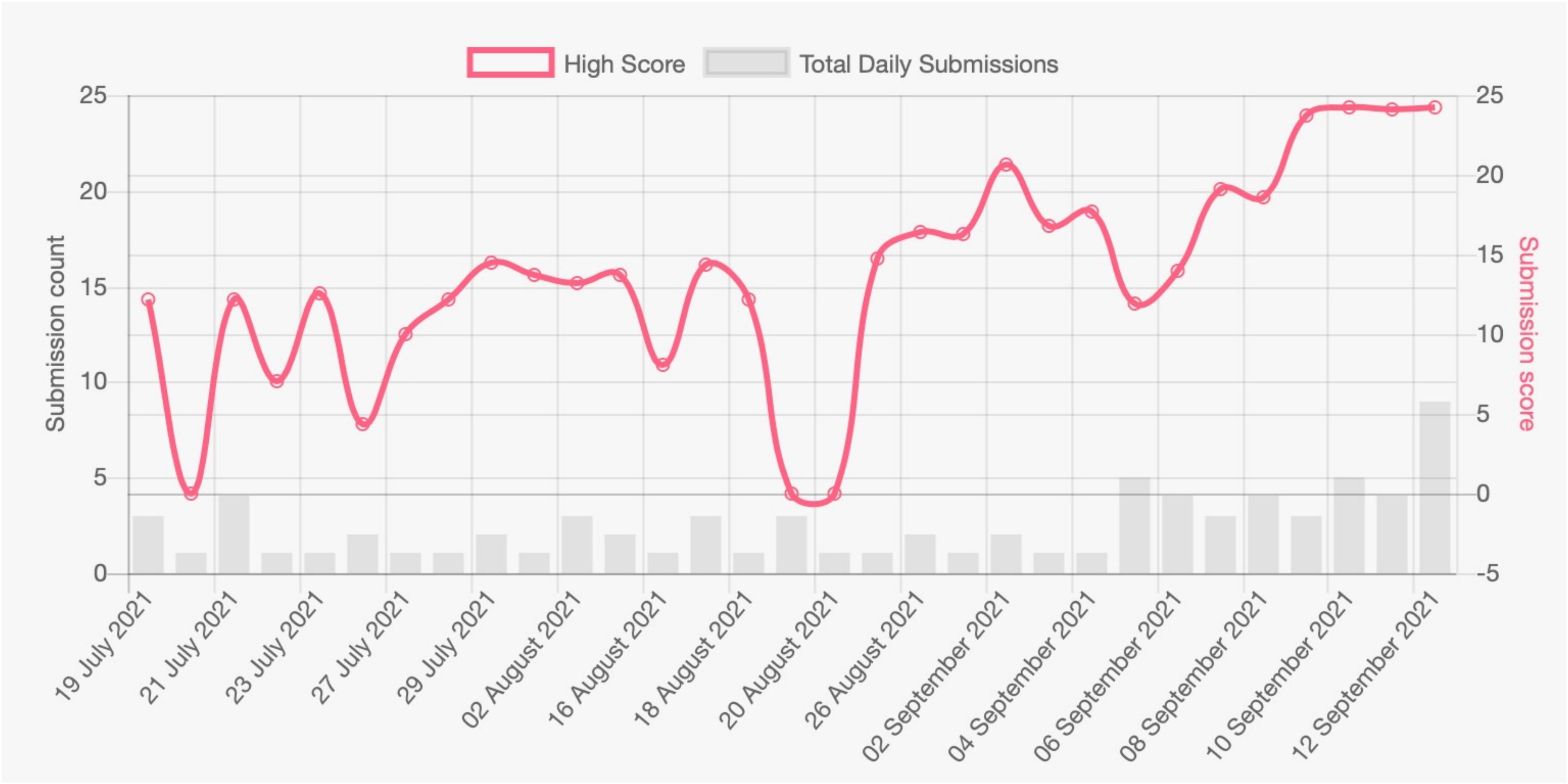
Deeper Action

ICCV 2021 Workshop

Results

Valid Submission: 34 (Val Phase) + 42 (Test Phase)

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	ningzhiqing	4	09/12/21	24.235 (1)	48.675 (1)	48.596 (1)	22.823 (1)	43.564 (1)	7.166 (1)
2	wings8643	8	09/07/21	19.132 (2)	29.872 (2)	35.045 (2)	20.826 (2)	32.477 (2)	7.112 (2)
3	yixuanli	2	09/05/21	11.923 (3)	28.485 (3)	25.780 (3)	9.888 (3)	22.506 (3)	2.651 (3)
4	ckk	5	09/05/21	7.092 (4)	1.188 (8)	14.516 (4)	6.240 (4)	13.055 (4)	1.810 (4)

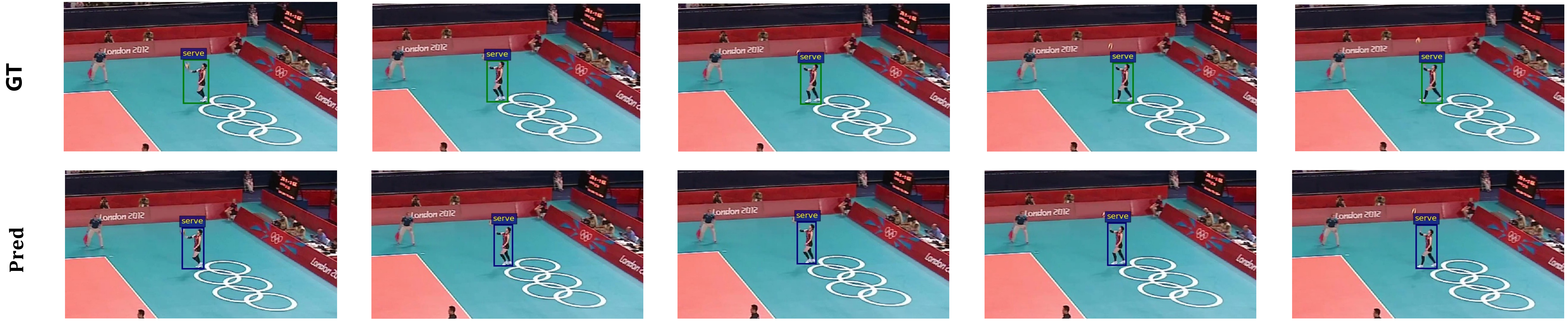


Deeper Action

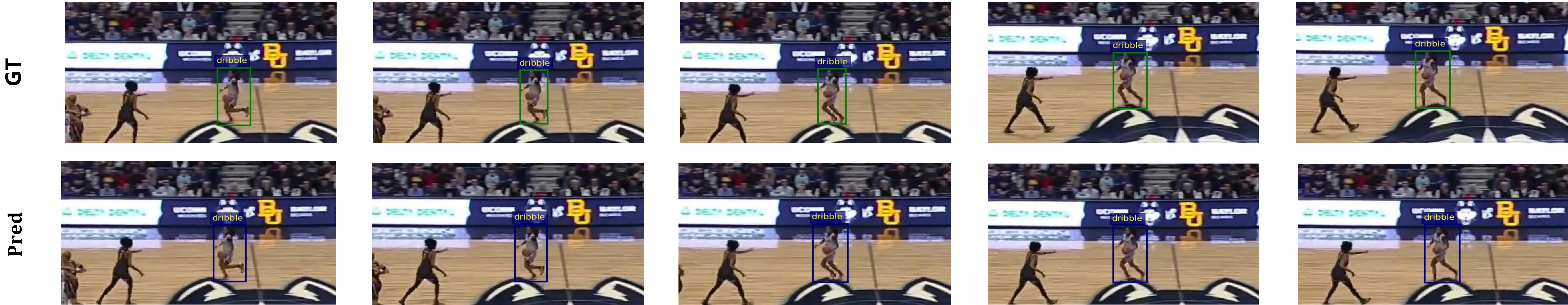
Simple Examples

ICCV 2021 Workshop

- Background provides much information. Motion pattern is simple.



- No need for modeling interactions between person, objects and scenes. Motion pattern is simple.



Deeper Action

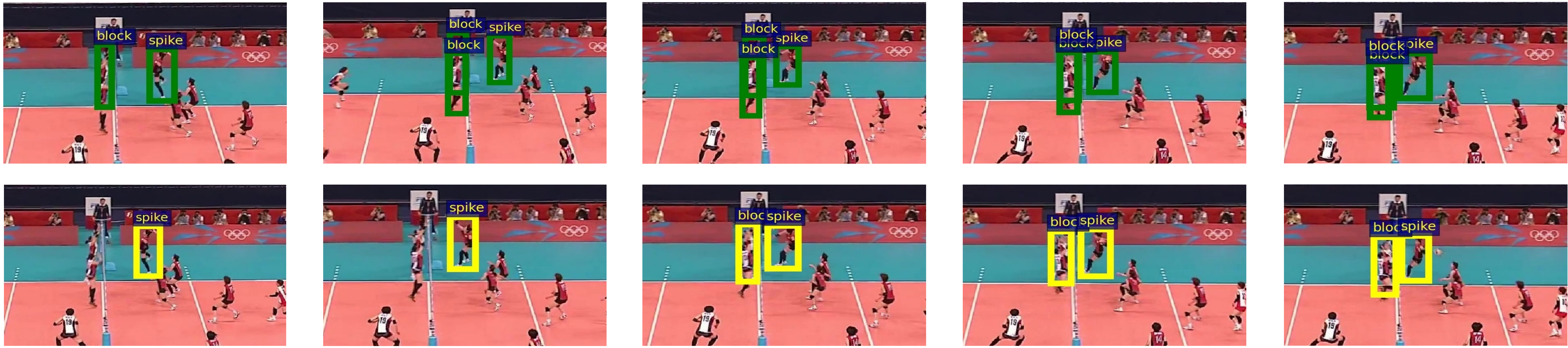
Hard Examples

ICCV 2021 Workshop

- Missed detection due to occlusion. Inaccurate action boundaries.

GT

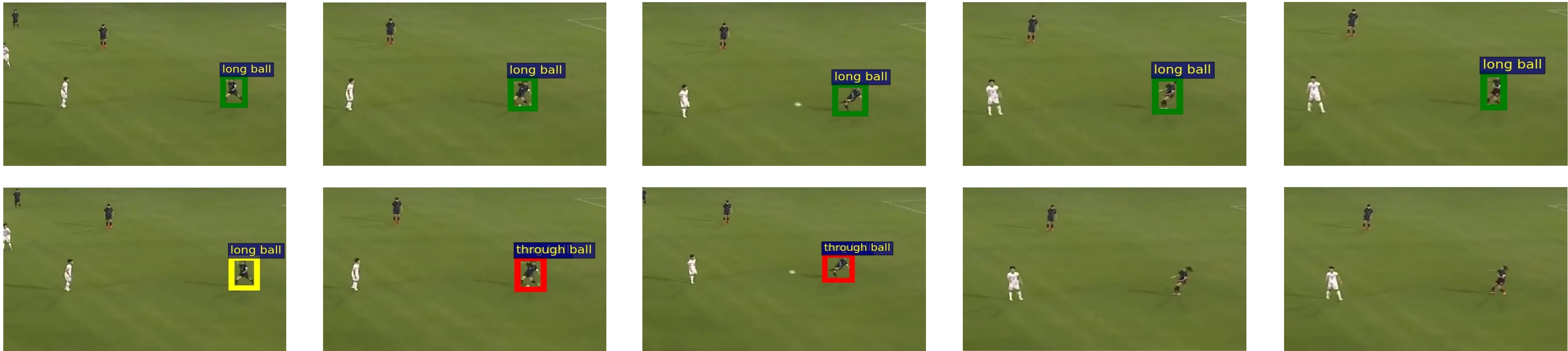
E_M, E_T



- Failing to model the interactions between person, objects and scenes.

GT

E_C, E_T



Deeper Action

Hard Examples

ICCV 2021 Workshop

- Fine-grained human motion pattern.

- Failing to model the interactions between person, objects and scenes. Inaccurate temporal boundary.

Deeper Action

1st Place Winners

ICCV 2021 Workshop

Person-Context Cross Attention for Spatio-Temporal Action Detection

Zhiqing Ning[1*] Qiaokang Xie[2†*] Wengang Zhou[2] Liangwei Wang[1] Houqiang Li[2]
[1]Huawei Noah's Ark Lab [2]University of Science and Technology of China
{ningzhiqing3, wangliangwei}@huawei.com
xieqiaok@mail.ustc.edu.cn {zhwg, lihq}@ustc.edu.cn

Deeper Action

2nd Place Winner

ICCV 2021 Workshop

DeeperAction workshop at ICCV 2021: MultiSports Challenge on Spatio-Temporal Action Detection Track Technical Report: A Solution to Detect Key Actions in Complicated Multi-person Scene

Yanbin Chen, Jiangyuan Mei, Zhicai Ou, Feifei Feng and Jian Tang
AIIC vision group, Midea Group
2388 Houhai avenue, Shenzhen, Guangdong, China
{chenyb60, meijy3, zhicai.ou, feifei.feng and tangjian22}@midea.com

MideaGroup
humanizing technology

Deeper Action

3^{rd} Place Winner

ICCV 2021 Workshop

LCTS: Longest Continuous Temporal Sequences for Action Detection

Shaomeng Wang, Yan Song, Keke Chen, Zeyu Zhou, Rui Yan, Xiangbo Shu, Jinhui Tang

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
Nanjing, China

wangshaomen@gmail.com

Thanks !

Homepage: https://deeperaction.github.io/multisports/

Github: https://github.com/MCG-NJU/MultiSports/