

MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions

Yixuan Li Lei Chen Runyu He Zhenzhi Wang Gangshan Wu Limin Wang*
 State Key Laboratory for Novel Software Technology, Nanjing University, China

Abstract

Spatio-temporal action detection is an important and challenging problem in video understanding. The existing action detection benchmarks are limited in aspects of small numbers of instances in a trimmed video or relatively low-level atomic actions. This paper aims to present a new multi-person dataset of spatio-temporal localized sports actions, coined as MultiSports. We first analyze the important ingredients of constructing a realistic and challenging dataset for spatio-temporal action detection by proposing three criteria: (1) motion dependent identification, (2) with well-defined boundaries, (3) relatively high-level classes. Based on these guidelines, we build the dataset of MultiSports v1.0 by selecting 4 sports classes, collecting around 3200 video clips, and annotating around 37790 action instances with 907k bounding boxes. Our datasets are characterized with important properties of strong diversity, detailed annotation, and high quality. Our MultiSports, with its realistic setting and dense annotations, exposes the intrinsic challenge of action localization. To benchmark this, we adapt several representative methods to our dataset and give an in-depth analysis on the difficulty of action localization in our dataset. We hope our MultiSports can serve as a standard benchmark for spatio-temporal action detection in the future. Our dataset website is at [this https URL](#).

1. Introduction

Spatio-temporal human action detection in untrimmed videos is of great importance for many applications, such as surveillance and sports analysis. Recently, recognizing actions from short trimmed videos has achieved considerable progress [51, 3, 47, 41, 48, 49], but the existing models can not be applied for grounded video analysis. Meanwhile, although temporal action detection methods [65, 29, 27, 59, 63] for untrimmed videos can distinguish intervals of human actions out of trivial backgrounds, they are still unable to spatially differentiate multiple con-

current human actions, which is vital for real-world video analysis applications.

Current spatio-temporal action detection benchmarks can be mainly classified into two categories: 1) Densely annotated high-level actions such as J-HMDB [19] and UCF101-24 [45], but their video clips typically only have a single person doing some semantically simple and temporally repeated actions. Moreover, chances are that characteristic visual scenes have provided enough cues for predicting actions due to coarse-grained action categories and therefore it weakens the real complexity of this task and fails to highlight the importance of fine-grained motion of human actions; 2) Sparsely annotated actions such as AVA [14] and DALY [53], that fail to provide clear temporal action boundaries, and thus might be unsuitable for modeling actions with fast movement, and temporal evolution. Meanwhile, their atomic actions rarely require reasoning.

Based on the analysis above, we argue that a new benchmark is necessary to advance the research of spatio-temporal action localization. The benchmark should satisfy several important requirements to cover the realistic challenges of this task. *First*, it should disentangle the influence of background scenes and human motions for predicting action categories and focus on motion information itself of human actions. This could be cases of multiple people performing different and even fine-grained actions concurrently in the same scene. *Second*, to address the inherently confusing human action boundaries in time, actions should be both semantically and temporally well-defined with a consensus among humans. *Finally*, considering the complexity of real-world applications, actions should be high-level which requires modeling subtle human pose motions, long-term semantics, possible interactions between humans, objects and scenes, and reasoning.

To facilitate the study of spatio-temporal action localization, we develop *MultiSports*, short for *Multi-person Sports Actions*. The dataset is large-scale, high-quality and contains 25fps frame-by-frame annotations of multi-person action labels with precise temporal boundaries. An example clip has been visualized in Figure 1. We choose sports actions for the following reasons. First of all, actions in pro-

*Corresponding Author (lmwang@nju.edu.cn).

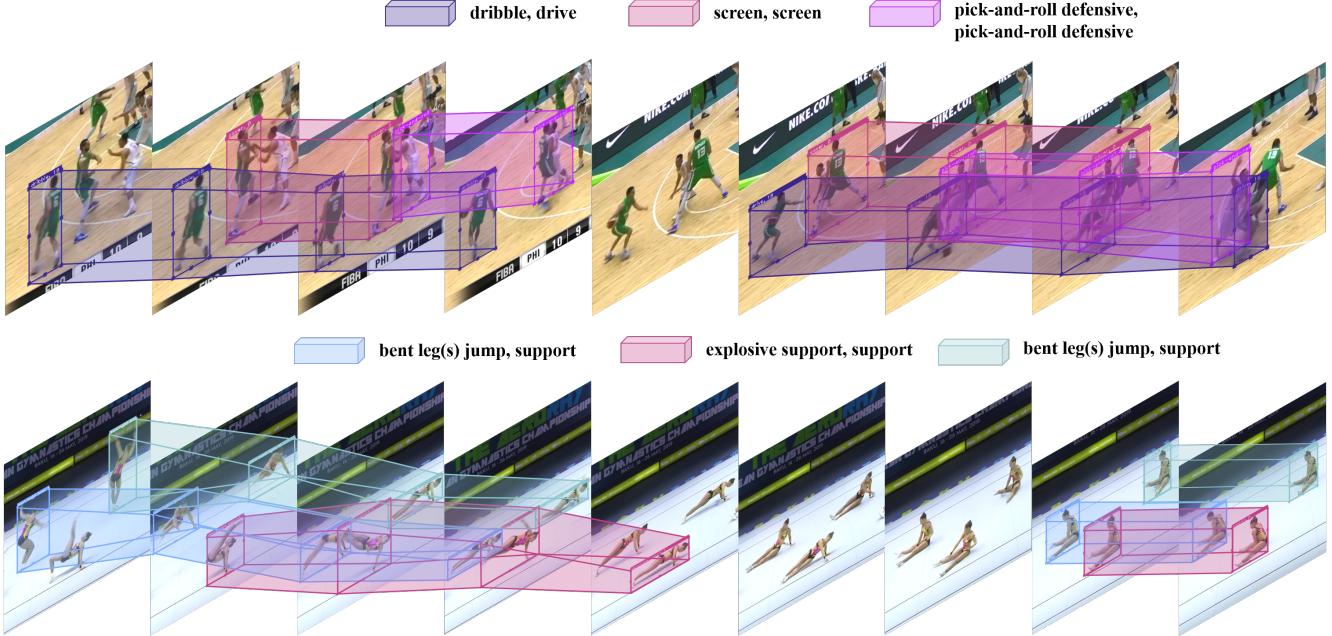


Figure 1. The 25fps tubelets of bounding boxes and fine-grained action category annotations in the sample frames of *MultiSports* dataset. Multiple concurrent action situations frequently appear in *MultiSports* with many starting and ending points in the long untrimmed video clips. The frames are cropped and sampled by stride 5 or 7 for visualization propose. Tubes with the same color represent the same person.

fessional sports competitions have well-defined categories and boundaries. The action vocabulary consists of 66 action labels collected from 4 sports (basketball, volleyball, football and aerobic gymnastics), all of which are defined by either professional athletes or official documentations [8]. Besides, there are plenty of multiple concurrent action samples in sports competitions. Also, the backgrounds are far less characteristic and action recognition cannot get much help from the background information. Meanwhile, the advantage of sports actions also lies in the complex human-object-scene interactions compared with atomic actions. In football, for example, although the athlete may take only 0.5s to kick the ball, we may need up to 5s' context to understand whether it is pass, long ball, through ball or cross. Likewise, when faced with similar motions of shooting in basketball, the position of the athlete becomes critical for determining whether it is a 2-point shot or a 3-point shot.

We conduct exhaustive labeling of frame-wise bounding boxes at 25 FPS and identify fine-grained action categories by a two-stage annotation procedure: a team of professional athletes of corresponding sport to annotate the temporal and categorical labels and a team of crowd-sourced annotators to finish the bounding boxes with the help of tracking method FCOT [7]. The two-stage annotation procedure as well as careful quality control together guarantee consistent and clean annotations. Although we select the competitions at the clip level and do not introduce bias to action categories, the action instances naturally result in a long-tailed distribution [16] as illustrated in Figure 3. Typical actions

(pass in football, pass in basketball, etc.) occur far more often than rare ones (balance turn in aerobic gymnastics, dribble steal in basketball, etc.). That is what happens in competitions so action localization models need to face the realistic long-tailed distribution instead of stopping at artificially balanced datasets. Therefore we do not manually adjust the dataset by approaches like thresholding the number of instances. Meanwhile, we do attach importance to the quality of the dataset. All videos in our dataset are high-resolution records of professional competitions covering a diversity of countries and performance levels.

Given the well-defined and dense-annotated actions in *MultiSports*, we benchmark spatio-temporal action detection on this challenging dataset. We perform a series of empirical studies with several recent state-of-the-art action detector methods. Compared with previous action detection benchmarks such as J-HMDB [19], UCF101-24 [45], our *MultiSports* is quite challenging with a much lower frame mAP and video mAP. We also introduce a detailed error analysis. These results demonstrate that our *MultiSports* benchmark reveals the realistic challenges of spatio-temporal action detection, such as fast movement and large deformation of actors, action occlusion, small sizes of both humans and objects, subtle differences between fine-grained action categories, the interactions between human and context and accurate temporal localization. We hope our *MultiSports* could serve as a standard benchmark to advance the area of spatio-temporal action localization in the future.

In summary, our contribution mainly lies in the following two aspects. 1) We develop a new benchmark *MultiSports* for the research of spatio-temporal action localization with well-defined but realistically difficult human actions as well as multi-person, high-quality and 25fps frame-wise annotations from four popular sports, which is representative for many real-life applications. 2) We conduct comprehensive studies and systematic error analysis on *MultiSports*, which reveals the key challenges of this area and may hopefully forward future research in this area.

2. Related Works

Action recognition datasets. Early datasets of action recognition mainly focus on action classification. Those datasets, including KTH [38], Weizmann [2], UCF-101 [45] and HMDB [23], contains manually trimmed short clips to capture semantics of a single action. Their human action cues, however, are overwhelmed by signals of background scenes. Recently, large-scale video classification datasets such as Sports-1M [21], YouTube-8M [1] and Kinetics [3] have been created for feature representation learning and serves as pre-training in downstream tasks, but appearance cues still play a important role here. Something-something [13] and FineGym [39], with plenty of fine-grained action categories, effectively reduce the influences of background scenes and reveal some key challenges of modeling a single action. They share the similar property of capturing motion cues with MultiSports, but only have one concurrent action therefore we address a different need with them.

Temporal localization datasets such as ActivityNet [15], THUMOS14 [18], MultiTHUMOS [61] and Charades [40] provides temporal localization annotations for each action of interest in untrimmed videos, but unlike MultiSports, they do not provide spatial annotations and could not identify multiple concurrent actions for multiple people.

Previous spatio-temporal action localization datasets such as UCF Sports [36], UCF101-24 [45], J-HMDB [19] typically evaluate spatio-temporal action localization for short videos with only a single person and coarse-grained action categories. Our *MultiSports* significantly differs from them in several aspects: multiple concurrent actions by multiple people; less characteristic background scenes; the larger number of action and fine-grained categories; more fast movement and large deformation; and significantly more instances per clip. Recently, a new type of extensions such as DALY [53], AVA [14] and AVA-Kinetics [24] adopt sparse annotations of daily life actions, either in composite or atomic forms, to reduce human labors of annotating and increase the scale of datasets. It may be a good way for evaluating daily life actions without fast movement and large deformation, but unsuitable for areas like sports analysis, since it often requires continuous anno-

tations of all human actions of interest. MEVA [6] is a security dataset, which provides spatial-temporal annotations and some other modality annotations. While our sports actions are more complex and fast-changing than MEVA. Different from previous datasets, our *MultiSports* proposes a more difficult benchmark with multi-person, fine-grained setting and frame-by-frame annotations, which focuses on the sports domain.

Methods for spatio-temporal action localization. Most recent approaches for UCF101-24 and JHMDB can be classified into two categories: frame-level detectors and clip-level detectors. Many efforts have been made to extend an image object detector to the task of spatio-temporal action localization at the frame level [12, 50, 33, 37, 43, 52], where the resulting per-frame detections are then linked to generate final tubes. Although flows could be used to capture motion cues, frame-level detector fails to fully utilize temporal information. To model temporal structures for action detection, some clip-level approaches or action tubelet detectors [17, 25, 20, 60, 26, 64, 44] have been proposed. ACT [20] took several frames as input and detected tubelets regressed from anchor cuboids. STEP [60] progressively refined the proposals by a few steps to solve the large displacement problem and utilized longer temporal information. MOC-detector [26] proposed an anchor-free tubelet detector by treating action instances as trajectories of moving points. For AVA, many methods [10, 11, 46, 55, 56] have been proposed to better make use of spatio-temporal information for atomic action classification.

3. The MultiSports Dataset

Our *MultiSports* dataset aims to introduce a new challenging benchmark with high-quality annotations to the area of spatio-temporal action localization, which differs from previous ones in large-scale, multi-person, well-defined temporal boundary and fine-grained high-level action categories.

Below we introduce the details of our dataset. Sec. 3.1 explains our annotation procedure, which consists of four stages: action vocabulary generation, data preparation, action annotation and person bounding-box tracking. Statistics and characteristics of *MultiSports* are elaborated in Sec. 3.2 and Sec. 3.3.

3.1. Dataset Construction

Action Vocabulary Generation. We select basketball, volleyball, football and aerobic gymnastics as our domain because of their multi-person setting, less ambiguous actions and well-defined temporal boundary. For aerobic gymnastics, we use the official documentations [8]. In addition, we only select *difficulty elements* and discard *movement patterns*. For three ball sports, we use an iterative way to generate our action vocabulary in each sport: we initialize an

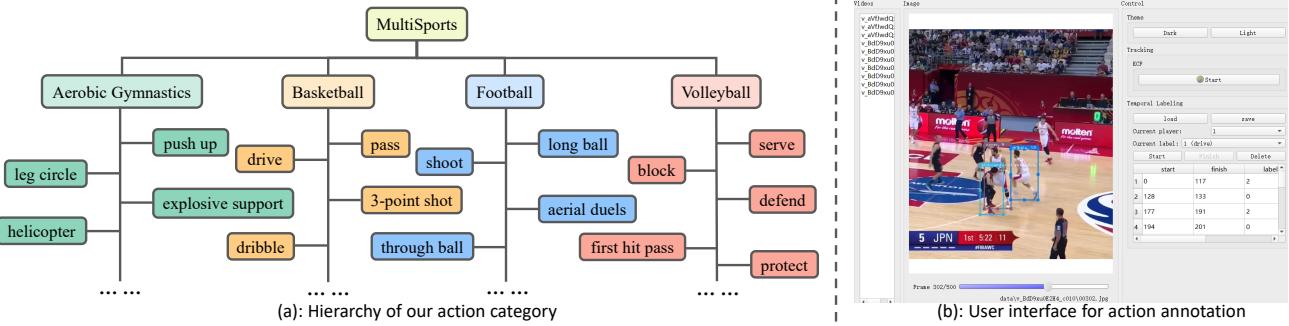


Figure 2. The action vocabulary hierarchy and annotator interface of the *MultiSports* dataset. (a) Our *MultiSports* has a two-level hierarchy of action vocabularies, where the actions of each sport are fine-grained. (b) Details of annotations can be found in Sec 3.1.

action list by the suggestions of athletes and write a handbook to regulate the definition of action boundaries. Then we let several annotators try to annotate the data, where inaccurate definitions of action boundaries, ambiguities between action categories and missed action categories will be reported. We iteratively adjust our action list and handbook by the feedback of annotators several times before we start massive annotating, which results in the final action hierarchy shown in Figure 2(a). Note that the annotators of action categories and temporal boundaries are professional athletes of the corresponding sport, so their feedback is essential for a well-defined action vocabulary in practice. To keep action boundaries accurate and make our dataset suitable for spatio-temporal action localization, especially in the sports domain, we do not count common actions such as run or stand in our action vocabulary. We also exclude foul in ball sports. Because in the 2D video records, we recognize fouls most from the referee’s reaction but not from the actor’s motion. What’s worse, it’s hard to identify who fouls due to occlusion.

Data Preparation. After choosing the four sports, we survey their competitions held in recent years by querying the name of sports like volleyball and the name of competition levels like Olympics and World Cup on YouTube and collect videos from top search results. For each searched video, we only select high-resolution, *e.g.* 720P or 1080P, competition records and then manually cut them into clips of minutes where we explicitly reduce the number of shot changes in clips to make our data suitable for this task. Through these official records, which are consistent and rich in content, the quality of data is guaranteed.

Action Annotation. Since our action annotations are both difficult in fine-grained categories and exhaustive in 25fps frame-wise bounding boxes, we naturally decompose our annotation procedure into two stages: firstly a team of professional athletes to annotate action categories, temporal action boundaries and the bounding box of the first frame of each action for efficiency, accuracy and consistency; and then a team of crowd-sourced annotators to finish the rest bounding boxes because the ambiguity of spatial human

boundaries is much less than that of temporal action boundaries. They both use the interface shown in Figure 2(b), where annotators in the first stage add records of the starting and ending frame and the action label to the file and annotators in the second stage adjust bounding boxes for each record to finish the annotations of actions. To ensure the consistency of action boundaries, which tends to be ambiguous and remains a big challenge of most datasets concerning temporal action annotations, we write a handbook to regulate the definition of action boundaries as mentioned above. For example, our handbook unifies the annotations of *passing* as starting from ball-controlling-leg leaving the ground and ending with this leg touching the ground again. **Person Bounding-box Tracking.** As mentioned above, a team of crowd-sourced annotators without domain knowledge finishes the bounding boxes for each instance got in the first stage with the help of FCOT [7] tracking method. First, FCOT [7] tracks the bounding boxes frame-by-frame starting from the first frame. Then annotators manually refine each frame precisely.

Quality Control. For the first stage of annotation, at least one annotator with domain knowledge per sport are responsible to double-check the annotations by correcting wrong or inaccurate ones and also *adding missing annotations* for a higher recall, *e.g.*, adding missed defence action in football and modifying inconsistent action boundaries. For the second stage, we double-check each instance by playing it in 5fps and correct the inaccurate bounding boxes.

3.2. Dataset Statistics

Our *MultiSports* contains 66 fine-grained action categories from four different sports, selected from 247 competition records. The records are manually cut into 800 clips per sport to keep the balance of data size between sports, where we discard intervals with only background scenes, such as award, and select the highlights of competitions as video clips for action localization. Table 1 compares the statistics of *MultiSports* v1.0 and existing datasets. Since AVA [14], AVA-Kinetics [24] and DALY [53] only have as sparse as 1fps annotations of bounding boxes and they

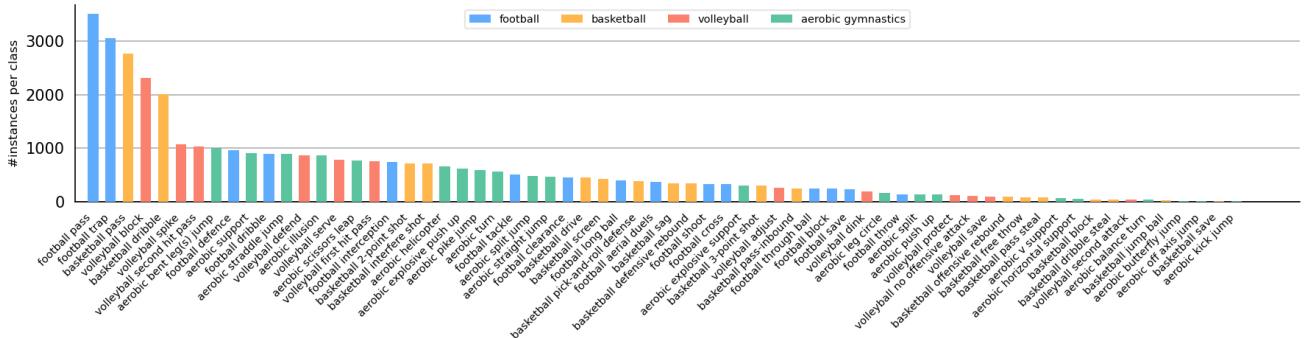


Figure 3. Statistics of each action class’s data size in *MultiSports*, which is sorted by descending order with 4 colors indicating 4 different sports. For actions in the different sports sharing the same name, we add the name of sports after them. The natural long-tailed distribution of action categories raises new challenges for action localization models.

	# act.	# inst.	avg act./vid. dur.	# bbox
J-HMDB [19]	21	928	1.2s / 1.2s	32k
UCF101-24 [45]	24	4458	5.1s / 6.9s	574k
DALY [53]	10	3637	Sparse [‡] / 3m45s	~29k
AVA V2.1 [14]*	80	~56000 [†]	Sparse [‡] / 15m	426k
AVA-Kinetics [14]*	80	~186000 [†]	-	590k
FineGym V1.0 [39]	530	32697	1.7s / 10m	-
Aerobic gym.	21	8703	1.5s / 30.9s	325k
Volleyball	12	7637	0.7s / 9.6s	139k
Football	15	12394	0.7s / 22.8s	229k
Basketball	18	9056	0.9s / 20.5s	214k
Ours in total	66	37790	1.0s / 21.0s	907k

Table 1. Comparison of statistics between existing action localization datasets and our *MultiSports* v1.0. (* only train and val sets’ ground-truths are available, [†] number of person tracklets, each of which has one or more action labels, [‡] 1fps action annotations have no clear action boundaries)

mainly focus on action localization in daily life rather than sports with fast movement where continuous tubelets are required, we do not directly compare to them. Specially, AVA-Kinetics adds Kinetics trimmed clip with single class and annotates one key frame per clip, which focuses on action recognition. *MultiSports* distinguishes with existing datasets such as J-HMDB [19] and UCF101-24 [45] in longer untrimmed video clips (21.0s vs. 1.2s or 6.9s), more fine-grained action categories (66 vs. 21 or 24), much more instances (37790 vs. 928 or 4458) and instances per video clip (11.8 vs. 1 or 1.4), which raises new challenges for modeling fast changing actions of multiple people in a longer video. Our *MultiSports* also has the biggest number of bounding boxes among all existing datasets, providing large-scale supervision signals for action localization in sports, where the fast changing spatial information plays an important role here. In order to get well-defined action boundaries, we only count the common part of human motions in ball sports, e.g., passing a basketball starts from the player straightening his arms but does not include holding the ball and doing fake actions. Therefore the average action duration is smaller than UCF101-24 [45], which often contains temporally repeated actions such as riding horses

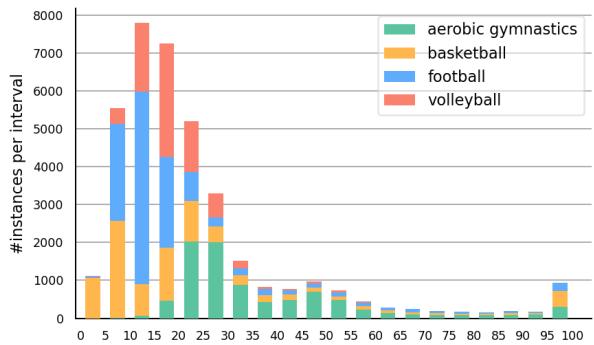


Figure 4. Statistics of action instance duration in *MultiSports*, where the x-axis is the number of frames and we count all instances longer than 95 frames in the last bar. Our action instances have a large variance in duration, resulting in challenges in modeling varying temporal structures.

or fencing.

The number of instances in each action category ranges from 3 to 3,514, reflecting the natural long-tailed distribution [16] as illustrated in Figure 3. The long-tailed action categories raise new challenges for action localization models. Figure 4 shows the distribution of action instance duration. The variability of action instance duration and even more differences in action duration distributions between sports raise new challenges for action localization models to accurately localize temporal boundary. Moreover, action instances in *MultiSports* are often related with longer temporal context and interactions with information of other spatial areas. The large variance and inherent difficulty require a more dynamic way of modeling temporal structures. We also evaluated output of FCOT [7] and results are shown in Table 2. We adopt success and precision metrics proposed in OTB100 [57]. Aerobic turned out harder in both success and precision aspects.

	Aerobic gym.	Volleyball	Football	Basketball
Success	0.66	0.72	0.77	0.66
Precision	0.67	0.93	0.92	0.72

Table 2. Tracking results on different sports

Our training/validation/test sets are split at the clip level, where the numbers of instances in each sport between sets are manually controlled as roughly 3:1:2 split.

3.3. Dataset Characteristics

Our *MultiSports* has several distinguishing characteristics that different from existing datasets.

Difficulty and Diversity. As discussed, *MultiSports* is difficult in several aspects comparing to existing datasets: 1) fine-grained, fast changing actions and a larger number of categories with long-tailed distributions; 2) frequently appeared multi-person situations, which prevents the model from distinguishing action categories from backgrounds and require models to capture subtly different motion cues; 3) longer video clips and more instances per video clip; 4) the fast movement, large deformation and occlusion of actions in sports; 5) the large variance of action instance duration and distributions between sports, which makes it difficult to localize temporal boundary. Besides, our video clips are selected from diverse performance levels with diverse countries, making the dataset a good representative of real-life scenes for sports analysis.

High Quality. The videos of *MultiSports* are high-resolution (720P or 1080P) competition records, which well preserve details of small humans and objects, leaving the room for modeling actions from distance such as in football. Besides, with the help of our annotation team composed by professional athletes from university team, our action categories and their corresponding action boundaries are well-defined for more effective evaluation of action localization. The professional annotators and careful quality control together provide consistent and clean annotations.

4. Experiments and Analysis

4.1. Datasets and Metrics

MultiSports benchmark. To guarantee enough instances for each class despite the severely unbalanced distribution (Figure. 3), we artificially split the instances into the training set and the validation set. Following AVA [14], we evaluate on 60 classes that have at least 25 instances in validation and test splits to benchmark performance. We resize the whole dataset into 720P. In total, our released version contains 18422 training examples from 1574 clips and 6577 validation instances from 555 clips. We will provide the detailed ratio of training and validation instances in the Appendix A. All those instances are selected from 3200 clips covering 247 matches. Unless otherwise mentioned, we report the results trained on the training set and evaluated on the validation set. The testing set includes approximately 13000 instances and will be released later.

Metrics. Following the standard practice [52, 20], we utilize frame-mAP and video-mAP to evaluate action local-

ization performance. For video-mAP, we use the 3D IoU, which is defined as the temporal domain IoU of two tracks, multiplied by the average of the IoU between the overlapping frames. The threshold is 0.5 for frame-mAP, 0.2 and 0.5 for video-mAP.

4.2. Spatio-temporal Action Detection Results

On top of *MultiSports*, we evaluate several representative action detection methods of UCF101-24 [45] and AVA [14] in Table 3. For SlowOnly Det. and SlowFast Det., we use the code in MMAAction2 [5]. We use the official released code for ROAD, YOWO and MOC. More details about the methods are provided in the Appendix B.

For UCF101-24 [45] and JHMDB [19], which have dense temporal annotations and high-level actions as *MultiSports*, we find that these methods achieve satisfactory performance on them but obtain low performance on *MultiSports* (frame-mAP of **25.22%**, video-mAP@0.2 of **12.88%** and video-mAP@0.5 of **0.62%** for MOC [26]). The largest performance drop occurs on ROAD [43], a frame-level action detector that performs action detection at each frame independently without exploiting temporal information. UCF101-24 [45] and JHMDB [19] have only one category per video. Characteristic visual scenes provide enough cues for predicting their coarse-grained actions. However, *MultiSports* has a similar background in the same sport, where the background does not provide decisive information for the fine-grained actions.

For AVA [14], which has only sparse temporal annotation and atomic actions, we observe that the performance gap between SlowFast Det. [10] and SlowOnly Det. [10] on *MultiSports* is far more than on AVA (frame-mAP gap of **11.02%** vs. **4.54%**). This indicates that the sports actions need a higher frame rate to capture motion at fine temporal resolution than the general atomic actions of AVA. As shown in Figure 5, many aerobic actions gain large absolute improvement, such as aerobic turn (+30 AP) and aerobic horizontal support (+54 AP), because aerobic actions' deformation and displacement is the largest among the four sports and benefit more from this enhancement. We also observe a large relative increase in other sports, such as basketball pass, football clearance, volleyball second attack, which have short temporal duration and intense motion.

4.3. Error Analysis

In this section, we examine the cause of errors to better understand *MultiSports*' challenges. Based on ACT [20]'s frame-mAP error analysis, which is for the dataset with one action category per video, we propose a new detailed error analysis in video-mAP. We classify the detection errors into 10 mutually exclusive categories to analyze which percentage of the mAP is lost: E_R : a detection result targets at a ground-truth tube that has already be matched. E_N :

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 [43]	300 × 300	3.90	0.00	0.00	70.7	69.8	40.9	-	60.8	59.7	-
YOWO [22]	224 × 224	9.28	10.78	0.87	71.10	72.97	46.42	74.51	88.05	82.57	-
MOC [26] (K=7)	288 × 288	22.51	12.13	0.77	78.0	82.8	53.8	70.8	77.3	77.2	-
MOC [26] (K=11)	288 × 288	25.22	12.88	0.62	-	-	-	-	-	-	-
SlowOnly Det., 4 × 16 [10]	short side 256	16.70	15.71	5.50	-	-	-	-	-	-	20.02
SlowFast Det., 4 × 16 [10]	short side 256	27.72	24.18	9.65	-	-	-	-	-	-	24.56

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

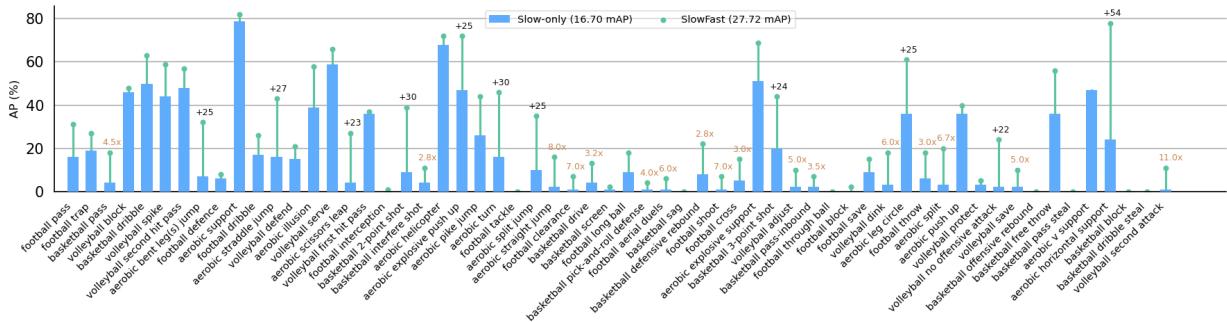


Figure 5. SlowOnly vs. SlowFast frame-mAP. Categories are sorted by descending order on the number of instances.

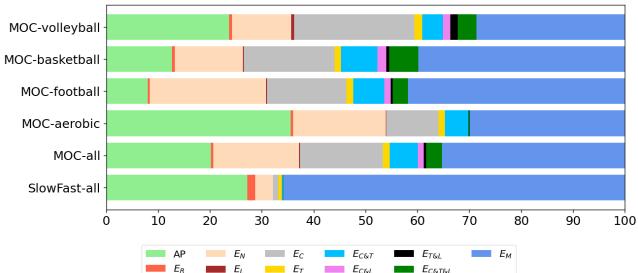


Figure 6. Error Analysis. AP is the correct detection. The threshold for a ground-truth matched by a detection is 0.1

a detection result that has no spatial-temporal intersection with any ground-truth tubes and appears out of thin air. $\mathbf{E_L}$: a detection result that has the correct action class, accurate temporal localization and inaccurate spatial localization. $\mathbf{E_C}$: a detection result that has the wrong action class, accurate temporal localization and accurate spatial localization. $\mathbf{E_T}$: a detection result that has the correct action class, accurate spatial localization and inaccurate temporal localization. $\mathbf{E_{C\&T}}$, $\mathbf{E_{C\&L}}$, $\mathbf{E_{T\&L}}$, $\mathbf{E_{C\&T\&L}}$: a detection that is inaccurate in multiple aspects while acceptable in other (if any) aspect. For example, $\mathbf{E_{C\&T}}$ refers to results with wrong action class, inaccurate temporal prediction yet accurate spatial localization. $\mathbf{E_M}$: ground-truth tubes that have not been matched by any detection results. The first nine categories cover the false positive predictions. The partition can be explained with a decision tree. The decision tree is attached in the Appendix C.

As shown in Figure 6, despite the relatively low recall, SlowFast Det. achieves higher video-mAP than MOC be-

cause it makes fewer false positive predictions. This can be explained by the fact that SlowFast Det. uses faster rcnn [35] finetuned on *MultiSports* as person detector to greatly decline the number of person boxes without actions. However, there are still many hard examples missed by SlowFast Det. For MOC, E_C and E_N are the most common errors among false positive detections, indicating the difficulty of our fine-grained action classification. Detection results with E_N errors means the model indeed detects the person spatio-temporally but unable to identify the action class correctly as the background class. E_N errors are also a result of the training strategy of MOC where only the frames temporally inside action instances are sampled for training so that although there are negative samples in other spatial location of these frames, the detector does not have enough amount of negative samples with people doing not one of sports actions. What's more, $E_{C\&T}$, $E_{C\&T\&L}$ and E_T are also a large portion of the rest errors (where $E_{C\&T} > E_{C\&T\&L} > E_T$), indicating more temporal errors with inaccurate action boundaries than spatial errors among the spatio-temporal localization errors for current method. Therefore a more effective way of modeling temporal structure (potentially with long-term modeling) is needed. Typical error visualization is shown in Figure 7.

4.4. Ablation Study

How important is temporal information? The tubelet length K is important in MOC [26] and we report results on UCF101-24 [45] and *MultiSports* with different K in Table 4. For frame-mAP, we can find that *MultiSports* can benefit more from longer temporal context than UCF101-

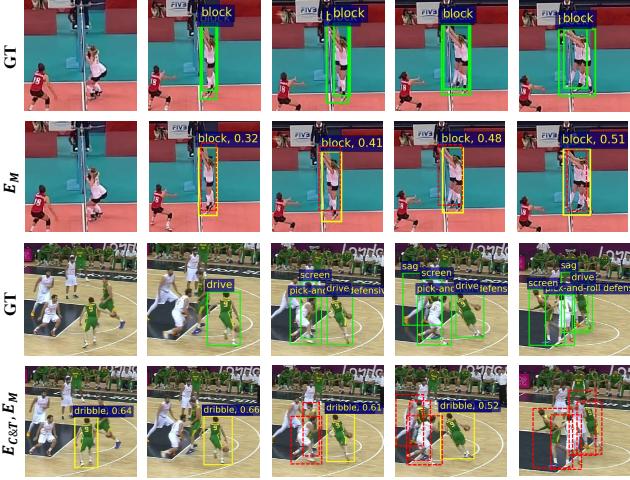


Figure 7. Visualization of typical errors in our *MultiSports*. Green boxes are the ground-truths. Yellow boxes are the detections. Red boxes are the missed ground-truths. 1st and 2nd row: missed detection due to occlusion. 3rd and 4th row: $E_{C\&T}$: drive is misclassified as dribble and also has inaccurate action boundary; E_M : missed detections of screen, pick-and-roll defensive and sag.

24, although Table 1 shows that the average action duration of *MultiSports* is smaller than UCF101-24. For video-mAP, the result doesn't increase as frame-mAP. We analyze there are two reasons. First, MOC points out that predicting movement is harder with input length being longer. What's worse, the category in our *MultiSports* have large deformation and displacement and MOC's Movement Branch can not predict accurately, which will harm the video level detection seriously. Second, Figure 4 shows the variability of action instance duration. The ratio is 9.2% for instances duration minus 7 and 23.1% for minus 11. The fixed clip length K (e.g., 11) will damage temporal localization ability. In summary, longer temporal context, more accurate movement estimation and more flexible temporal localization are needed for our *MultiSports*.

Which action categories are challenging? Figure 5 shows that not all categories yield better performance with more data. Categories highly correlated with scenes (such as basketball free throw) or aerobics basic categories (such as aerobic horizontal support and V support) achieve high performance with fewer data. Note that aerobics contains basic and combined categories, where combined action contains the motion of basic action and its own core motion, thus longer temporal context is needed for the combined action. In contrast, categories with short temporal duration and intense motion (such as football pass, basketball pass and football interception) achieve low performance with lots of data. By observing the confusion matrix, we summarize some other challenges, (1) Context modeling, such as basketball 2-point shot vs. 3-point shot (2) Reasoning, such as for volleyball protect vs. defend, we need to focus on

K	MultiSports			UCF101-24		
	F@0.5	V@0.2	V@0.5	F@0.5	V@0.2	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.

Estimation	MultiSports			AVA F-mAP@0.5
	F@0.5	V@0.2	V@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.

whether the ball was blocked back or was spiked by an opponent several frames earlier. (3) Long temporal modeling, such as football long ball vs. pass, they have the similar motion but need to identify how long the ball will be passed. The confusion matrix and detailed analysis will be provided in the Appendix C.

How important is the temporal localization? *MultiSports* has well-defined temporal boundaries and dense temporal annotations. We estimate SlowFast Det. with both the untrimmed and trimmed way on *MultiSports* and AVA in the Table 5. The trimmed way only estimates performance on the frames having annotations and untrimmed way estimates performance on all frames. We find that AVA only drops 2% while our dataset drops 11%, which indicates that temporal localization is really important in our dataset. In addition, video-mAP@0.5 drops far more than video-mAP@0.2. This demonstrates that temporal localization is important for high-quality action tube detection. We believe that this characteristic will inspire new advances in the field of spatial-temporal action localization.

5. Conclusion

In this paper, we introduce the *MultiSports* dataset with 25fps spatio-temporal annotations of sports actions over four sports. *MultiSports* distinguishes from existing action detection datasets in many aspects: 1) fine-grained and temporally well-defined categories; 2) multi-person situation; 3) the fast movement, large deformation and occlusion of sports actions; 4) the large variance of action instance duration. We have empirically investigated representative methods on the *MultiSports* dataset. Our error analysis on the detector's performance leads to many attractive findings that are beneficial for future research of action detection models.

Appendix A: More Dataset Details

A.1 Train split vs. Validation split

In order to guarantee enough instances for each class despite the severely unbalanced distribution, we artificially split the instances into the training set and the validation set in Table 6. To avoid data leakage from the training set to the validation/testing set, we ensure that data from the same match should be used for only one purpose. In other words, clips in the validation set cannot come from the matches covered in the training set. Unless otherwise mentioned, we report the results trained on the training set and evaluated on the validation set.

A.2 Comparison with other type of Dataset

MEVA [6] is a new security dataset, whose data is from RGB and thermal IR cameras, UAV footage and GPS locations for the actors. It defines 37 activities (66 for *MultiSports*) with 17055 instances (37790 for *MultiSports*), where 29 activities are about person and 8 activities are about vehicle. The categories in this dataset are atomic, such as *person_close_trunk* and *person_stand_up*, which are different from our high-level and complicated sports categories. What's more, most of the categories in MEVA are daily actions, whose deformation and displacement are not large. Although it is a multi-person dataset, we believe our *MultiSports* can bring new challenges different from MEVA.

Appendix B: Method Details

ROAD [43] is a deep-learning framework for real-time action localisation and classification. It adopts SSD [32] to regress and classify action detection boxes in each frame independently, which does not utilize temporal information. Then, the frame detections are linked into action tubes by an online algorithm. Here we use the python linking code provided by MOC [26] instead of the original MATLAB code. Following the settings of ROAD on UCF101-24 [45], we use an ImageNet pretrained VGG16 [42] network. We first try an initial learning rate of 1e-4 as their setting on UCF101-24, but the loss diverges into infinity after 20 iterations. The reported experiment on our *MultiSports* adopts an initial learning rate of 1e-5. We use SGD optimizer and the learning rate is reduced to its $\frac{1}{10}$ after 30000, 60000 iterations, which is the same as their practice on UCF101-24. The maximum iteration number is 120000.

YOWO [22] is a frame-level action detector with two branches. A 2D-CNN branch extracts the spatial features of the key frame while a 3D-CNN branch extracts spatio-temporal features of the key frame and the previous n ($n=16$) frames. Then, the features of two branches are fused by a channel fusion and attention mechanism(CFAM) module and finally passed to a convolution layer to predict the

action class and bounding box in Yolov2 [34] manner. Finally, the frame detections are linked into action tubes by a dynamic programming algorithm. Note that the linking algorithm in YOWO is trimmed, thus we use the same linking algorithm as MOC on *MultiSports*. We use 2D Darknet-19 backbone pretrained on PASCAL VOC [28] and 3D ResNeXt-101 backbone pretrained on Kinetics [3]. To utilize multiple GPUs, we modified the batch size to 80 and the initial learning rate to 8e-4. Following the training strategy of YOWO on UCF101-24 [45], we adopt SGD optimizer and the learning rate is reduced to its $\frac{1}{2}$ after 30000, 40000, 50000, 60000 iterations. The epoch maximum is set to 5. Note that YOWO only estimates performance on the frames having annotations, thus frame-mAP we report on UCF101-24 is much lower than in the original paper.

MOC [26] is an anchor-free tubelet-level action detector with three branches, which firstly takes K frames as input, then outputs K frame tubelet results and finally links these tubelets into tubes with a common matching strategy. We use DLA34 [62] as the backbone network, which is pre-trained on COCO [31]. Following the training strategy of MOC on UCF101-24 [45], we use the Adam optimizer with the learning rate 5e-4. The learning rate is reduced to its $\frac{1}{10}$ after epoch 6 and 8. The epoch maximum is set to 12.

SlowFast Det. [10] firstly uses a person detector on the key frame to localize for region proposal. Then, each 2D ROI at the key frame is extended into a 3D ROI by replicating it along the temporal dimension. Finally, it extracts ROI features from the backbone features for predicting category. The person detector is a Faster R-CNN with a ResNeXt-101-FPN [58, 30] backbone, which is pre-trained on ImageNet [19] and the COCO human keypoint images [31]. The backbone is the variant of SlowFast or SlowOnly, which sets the spatial stride of res_5 to 1 and uses a dilation of 2 for its filters. Note that we use the code in MMAAction2 [5]. The results on AVA [14] and our *MultiSports* in the paper are all produced by it. We use the pre-computed proposals for AVA from previous work [10, 54]. Following previous work [10, 54], we fine-tune the person detector on our *MultiSports* with MMDetection [4]. We use the SGD optimizer with the learning rate 0.0025 and finetune 2 epochs on our *MultiSports*. The person detector produces 96.16 AR@100 on our *MultiSports* validation set. The detected boxes with confidence of > 0.9 are selected for action detection on both datasets. Our backbones are based on ResNet50, which are pre-trained on Kinetics-400 [3]. The $T \times \tau$ is set to 4×16 . The α is set to 8 for SlowFast. . We use a step-wise learning rate, reducing the learning rate $10 \times$ after epoch 10 and 15. We train for 20 epochs with a linear warm-up for the first 5 epochs. The initial learning rate is set to 0.1125 for SlowFast and 0.2 for SlowOnly. SlowFast and Slowonly Det. use the same link algorithm as MOC.

	Volleyball	Football	Basketball	Aerobic	All
instance ratio	3549:1294	6144:2153	4532:1715	4197:1415	18422:6577
clip ratio	402:130	402:132	379:147	391:146	1574:555
competition ratio	32:11	36:12	34:14	23:8	125:45

Table 6. Train split vs Validation split

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 ; $\text{GT}(c)$: set of ground-truths of class c ; $\text{dupGT}(c)$: copy of $\text{GT}(c)$; $\text{GT}(\text{others})$: set of all ground-truths that not in class c ; $\text{GT}(\text{all})$: set of all ground-truths; T_{IoU} : the temporal domain IoU; S_{IoU} : the average of the IoU between the overlapping frames; $\text{tubelet}_{\text{IoU}}$: $T_{\text{IoU}} * S_{\text{IoU}}$.

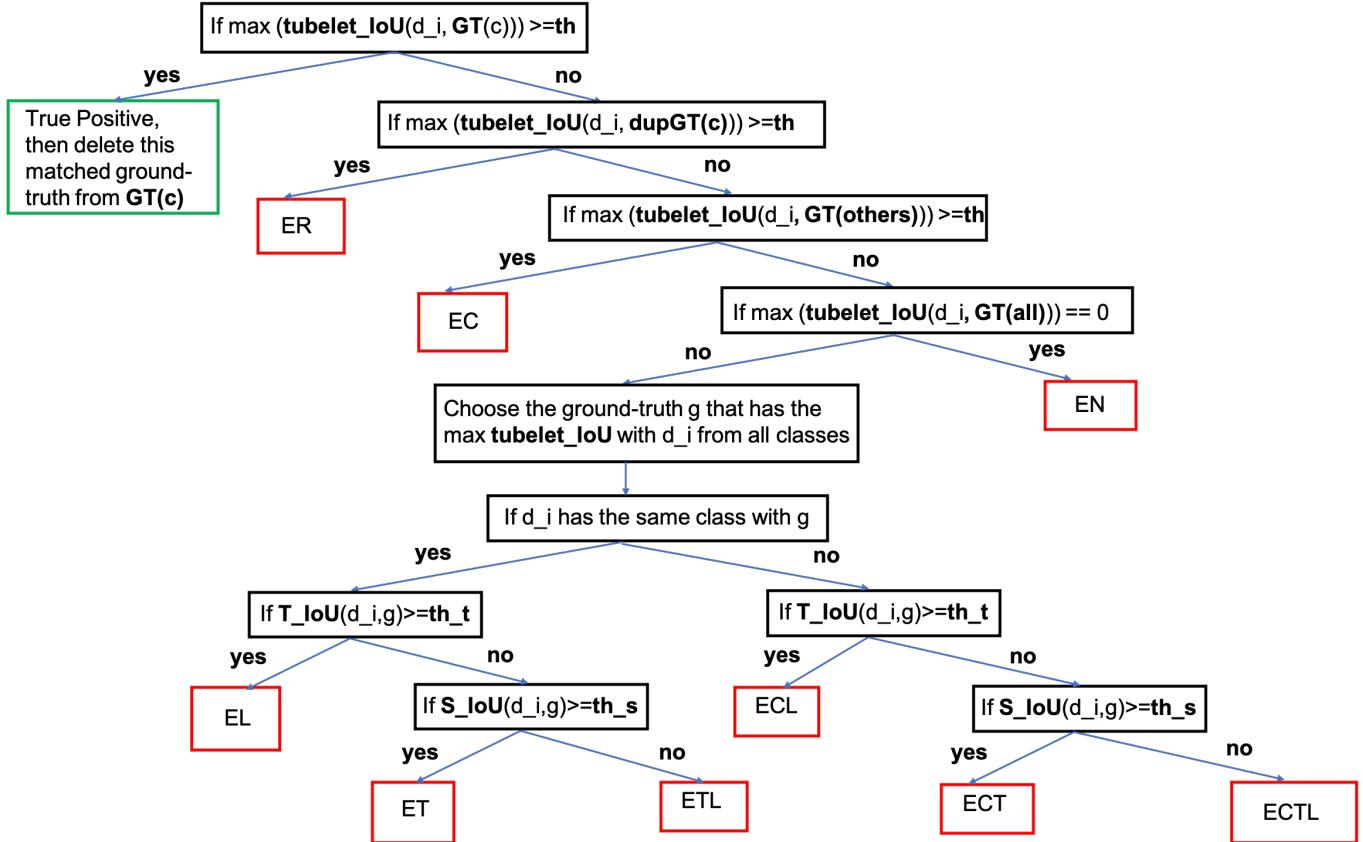


Figure 8. Error Tree

Appendix C: Error Analysis

C.1 Error Tree

To further understand the difficulty in our *MultiSports* dataset, we classify the detection errors into 10 different categories in a tree structure as shown in Figure 8 (code in *VideomAP_error.py*), which are:

- E_R (Errors of repeated detections): a detection result that has tubelet IoU larger than a threshold and the right action class with some ground-truth tubelets, but

the ground-truths have been matched by other detection results before with a confidence score larger than it.

- E_N (Errors of not matched): a detection result that has no intersection with any ground-truth tubelets of any class, indicating there should be no detection results but it appears out of thin air.
- E_L (Errors of spatial localization): a detection result that has the same action class and temporal IoU larger than a threshold with some ground-truth, but it has a



Figure 9. More detailed visualizations on our *MultiSports* dataset with our novel error categories of video-mAP. Green boxes are the ground-truths. Yellow boxes are the detections. 1st and 2nd row: E_M : missed detection of defence; E_C : tackle is misclassified as defence; E_T : dribble has inaccurate action ending boundary. 3rd and 4th row: E_C : turn is misclassified as illusion in the last picture in 4th row; E_T : turn has inaccurate action boundary. 5th and 6th row: E_N : detection results contain that athletes actually doing none of sports actions but the model identifies first hit pass and second hit pass for them. 7rd and 8th row: $E_C \& T$: drive is misclassified as dribble and also has inaccurate action boundary; E_M : missed detections of interfere shot and 2-point shot.

low average spatial bounding box IoU in the area of the temporal intersection of ground-truth tubelets and it so that a lower tubelet IoU than the required threshold.

- E_C (Errors of classification): a detection result that has the tubelet IoU larger than a threshold with a ground-truth, but its action class is not the same with the

ground-truth's class.

- E_T (Errors of temporal localization): a detection result that has the same action class and average spatial bounding box IoU larger than a threshold with some ground-truth in the area of the temporal intersection of ground-truth tubelets and it, but low temporal IoU with ground-truths so that a lower tubelet IoU than the required threshold.
- $E_{C\&T}$ (Errors of both classification and temporal localization): a detection result that has average spatial bounding box IoU larger than a threshold with some ground-truth tubelets in the area of the temporal intersection of ground-truth tubelets and it, but both low temporal IoU and wrong action class.
- $E_{C\&L}$ (Errors of both classification and spatial localization): a detection result that has temporal IoU larger than a threshold with some ground-truth tubelets, but both wrong action class and low average spatial bounding box IoU with some ground-truth in the area of the temporal intersection of ground-truth tubelets and it.
- $E_{T\&L}$ (Errors of both temporal and spatial localization): a detection result in which we first select the ground-truth tubelet from all action classes that has the maximum tubelet IoU with the detection result, then we find they share the same action classs, but both temporal IoU and average IoU of spatial bounding boxes lower than a threshold.
- $E_{C\&T\&L}$ (Errors of classification, temporal and spatial localization): a detection result that has some intersection with some ground-truth tubelets, which is different with EN, but wrong action class and both the temporal and average bounding box IoU lower than a threshold.
- E_M (Errors of missed detections): ground truth tubelets that have not been matched by any detection results.

C.2 More Visualization of Error Analysis

As shown in Figure 9, we collect more visualizations of MOC(K=11) as a supplementary of Figure 7.

C.3 Confusion Matrix

We draw the confusion matrices of the predictions which are classified into AP and E_C in Figure 10. We observe that the aerobic performs best because its categories relate only to individual actions. Actions having similar motions but different spatio-temporal contexts tend to confuse. For example, 1) drive vs. dribble in basketball, drive emphasizes on breaking through defender and being closer to the basket, which needs to model person-person interaction and spatial

localization; 2) through ball vs. pass in football, through ball will break through the opponent's line of defense and be passed in front of the teammate, which needs long-term temporal modeling and reasoning. 3) offensive rebound vs. defensive rebound, the difference is whether the offensive player or defensive player gains control of the ball; 4) defend vs. protect in volleyball, we need to focus on whether the ball was blocked back or was spiked by an opponent several frames earlier.

References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.
- [2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402. IEEE Computer Society, 2005.
- [3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE Computer Society, 2017.
- [4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tian-heng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [5] MMACTION2 Contributors. Openmmlab’s next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmaction2>, 2020.
- [6] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1060–1068, 2021.
- [7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Fully convolutional online tracking, 2021.
- [8] Federation Internationale de Gymnastique. Aerobic gymnastics-code of points. *FIG Aerobic Gymnastics FIG Executive Committee*, 2017.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210. IEEE, 2019.
- [11] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.
- [12] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, pages 759–768. IEEE Computer Society, 2015.

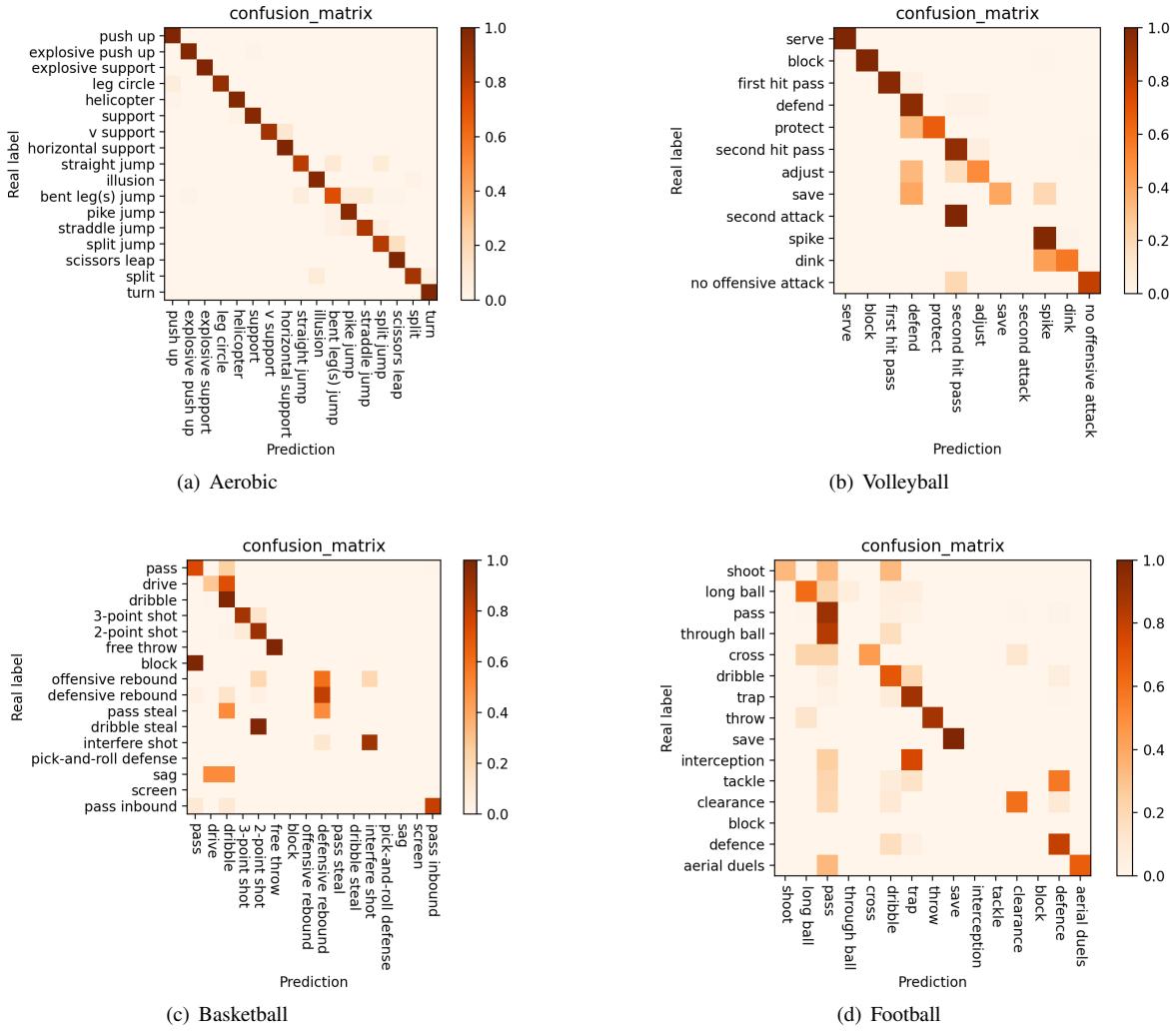


Figure 10. Confusion Matrix of SlowFast Det. on different sports.

- [13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5843–5851. IEEE Computer Society, 2017.
- [14] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056. IEEE Computer Society, 2018.
- [15] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970. IEEE Computer Society, 2015.
- [16] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *CoRR*, abs/1709.01450, 2017.
- [17] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (T-CNN) for action detection in videos. In *ICCV*, pages 5823–5832. IEEE Computer Society, 2017.
- [18] Haroon Idrees, Amir Roshan Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.*, 155:1–23, 2017.
- [19] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199. IEEE Computer Society, 2013.
- [20] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4415–4423. IEEE Computer Society, 2017.
- [21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732. IEEE Computer Society, 2014.

- [22] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.
- [23] Hildegarde Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE Computer Society, 2011.
- [24] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.
- [25] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, pages 306–322. Springer, 2018.
- [26] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *ECCV (16)*, volume 12361 of *Lecture Notes in Computer Science*, pages 68–84. Springer, 2020.
- [27] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3888–3897. IEEE, 2019.
- [28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.
- [29] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV (4)*, volume 11208 of *Lecture Notes in Computer Science*, pages 3–21. Springer, 2018.
- [30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [33] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 744–759. Springer, 2016.
- [34] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6517–6525. IEEE Computer Society, 2017.
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [36] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*. IEEE Computer Society, 2008.
- [37] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *BMVC*. BMVA Press, 2016.
- [38] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *ICPR (3)*, pages 32–36. IEEE Computer Society, 2004.
- [39] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2613–2622. IEEE, 2020.
- [40] Gunnar A. Sigurdsson, Gülcin Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 510–526. Springer, 2016.
- [41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [43] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, pages 3657–3666. IEEE Computer Society, 2017.
- [44] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *CVPR*, pages 11987–11995. Computer Vision Foundation / IEEE, 2019.
- [45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.
- [46] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020.
- [47] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE Computer Society, 2015.
- [48] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439. IEEE Computer Society, 2018.
- [49] Limin Wang, Yu Qiao, and Xiaou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314. IEEE Computer Society, 2015.

- [50] Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, pages 2708–2717. IEEE Computer Society, 2016.
- [51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV* (8), volume 9912 of *Lecture Notes in Computer Science*, pages 20–36. Springer, 2016.
- [52] Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, pages 3164–3172. IEEE Computer Society, 2015.
- [53] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *CoRR*, abs/1605.05197, 2016.
- [54] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. *CoRR*, abs/1812.05038, 2018.
- [55] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.
- [56] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020.
- [57] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.
- [58] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [59] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5794–5803. IEEE Computer Society, 2017.
- [60] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S. Davis, and Jan Kautz. STEP: spatio-temporal progressive learning for video action detection. In *CVPR*, pages 264–272. Computer Vision Foundation / IEEE, 2019.
- [61] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *Int. J. Comput. Vis.*, 126(2-4):375–389, 2018.
- [62] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.
- [63] Runhao Zeng, Wenbing Huang, Chuang Gan, Mingkui Tan, Yu Rong, Peilin Zhao, and Junzhou Huang. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7093–7102. IEEE, 2019.
- [64] Jiaojiao Zhao and Cees G. M. Snoek. Dance with flow: Two-in-one stream action detection. In *CVPR*, pages 9935–9944. Computer Vision Foundation / IEEE, 2019.
- [65] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2933–2942. IEEE Computer Society, 2017.