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 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) multi-person scenes and motion dependent identification, (2) with well-defined boundaries, (3) relatively fine-grained classes of high complexity. Based on these guidelines, we build the dataset of MultiSports v1.0 by selecting 4 sports classes, collecting 3200 video clips, and annotating 37701 action instances with 902k bounding boxes. Our datasets are characterized with important properties of high diversity, dense annotation, and high quality. Our MultiSports, with its realistic setting and detailed annotations, exposes the intrinsic challenges of spatio-temporal action detection. To benchmark this, we adapt several baseline methods to our dataset and give an in-depth analysis on the action detection results 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 https://deeperaction.github.io/multisports/.*

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 [44, 3, 40, 35, 41, 42], but these classification models can not be directly applied for video analysis in a multi-person scene. Meanwhile, although temporal action detection methods [56, 26, 25, 50, 53] for untrimmed videos can distinguish intervals of human actions from background, they are still unable to spatially detect multiple concurrent human actions, which is important in real-world applications of video analysis.

Current spatio-temporal action detection benchmarks can be mainly classified into two categories: 1) Densely annotated high-level actions such as J-HMDB [17] and UCF101-24 [38]. Their clips only have a single person doing some semantically simple and temporally repeated actions. Typically, the scene context can provide enough cues for recognizing these coarse-grained action categories. Thus, these benchmarks might be impractical for real-world applications such as surveillance, where it is required to deal with more fine-grained actions in a multi-person scene; 2) Sparsely annotated atomic actions such as AVA [12]. They fail to provide clear temporal action boundaries, and simply focus on frame-level spatial localization of atomic actions. This setting removes the requirements of temporal localization for action detection algorithms. Meanwhile, their atomic actions rarely require the complex reasoning over the actors and their surrounding environment.

Based on the analysis above, we argue that a new benchmark is necessary to advance the research of spatio-temporal action detection. The benchmark should satisfy several important requirements to cover the realistic challenges of this task. 1) There should be multiple persons performing different actions concurrently in the same scene, where the background information is not sufficient for action recognition and motion itself of the actor plays a significant role. 2) To address the inherently confusing human action boundaries in time, actions should be both semantically and temporally well-defined with a consensus among humans. 3) Considering the complexity of real-world applications, actions should be fine-grained which requires accurate human pose and motion information, long-term temporal structure, possible interactions between humans, objects and scenes, and reasoning over their relations.

Following the above guidelines, we develop the *MultiSports* dataset, short for *Multi-person Sports Actions*. The dataset is large-scale, high-quality, multi-person, and contains fine-grained action categories with precise and dense annotations in both spatial and temporal domains. The ac-

✉: Corresponding author (lmwang@nju.edu.cn).

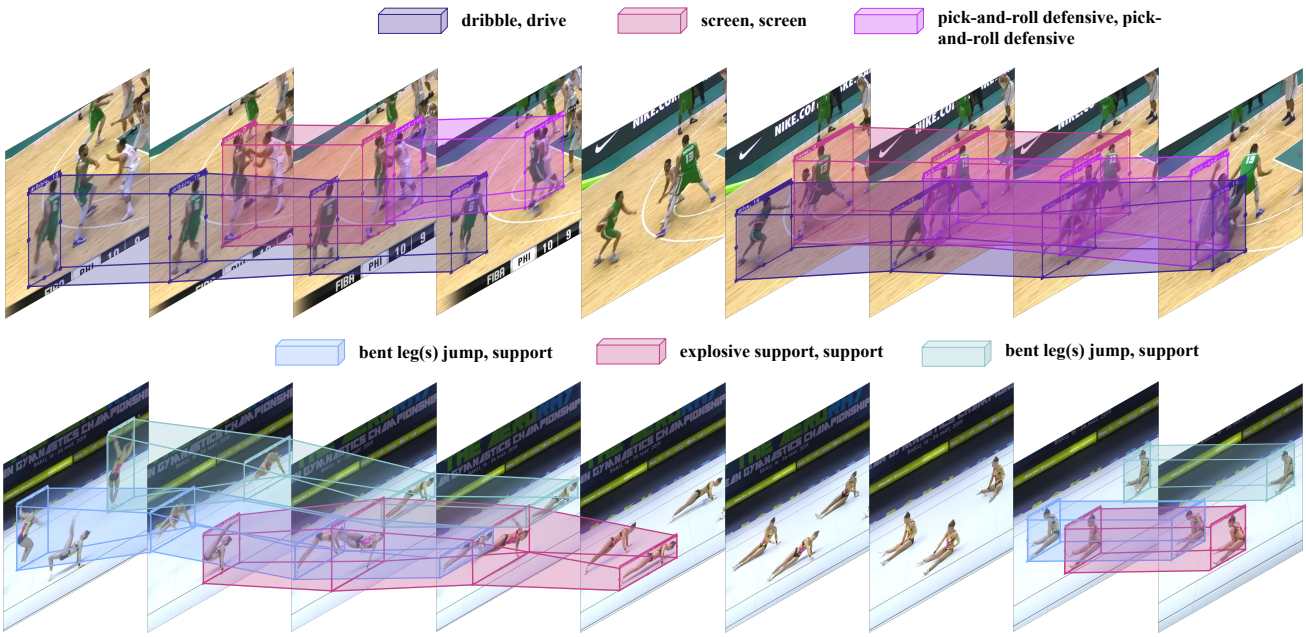

Figure 1. The 25fps tubelets of bounding boxes and fine-grained action category annotations in MultiSports dataset. Multiple concurrent action situations frequently appear in MultiSports with many starting and ending points in the long untrimmed video clips. The frames are cropped and sampled by stride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

tion vocabulary consists of 66 action classes collected from 4 sports (basketball, volleyball, football and aerobic gymnastics). An example clip has been visualized in Figure 1. We choose these four sports for the following reasons. 1) There are plenty of multiple concurrent action instances in sports competitions. Also, the background is far less characteristic and cannot provide sufficient information for fine-grained action recognition. 2) Sports actions have well-defined categories and boundaries. These boundaries are defined by either professional athletes or official documentations [7]. 3) Due to the complex competition rules, recognizing sports action generally requires to model the long-term structure and the human-object-scene interactions. For example, in football, although the athlete may take only 0.5s to kick the ball, we may need up to 5s context to recognize whether it is pass, long ball, through ball, or cross.

In practice, we conduct exhaustive annotations of 25 fps frame-wise bounding boxes and fine-grained action categories in a two-stage procedure: 1) a team of professional athletes of corresponding sport to annotate the temporal and category labels, and 2) a team of crowd-sourced annotators to finish the bounding boxes with the help of tracking method FCOT [6]. This two-stage annotation procedure as well as careful quality control together can guarantee consistent and clean annotations. To ensure the visual quality, all videos in our dataset are high-resolution records of professional competitions from a diversity of countries and different performance levels.

Given the well-defined and dense-annotated action instances in *MultiSports v1.0*, we benchmark spatio-temporal action detection on this challenging dataset. We perform empirical studies with several recent state-of-the-art action detector methods. Compared with previous action detection benchmarks such as J-HMDB [17] and UCF101-24 [38], our MultiSports is quite challenging with a much lower frame mAP and video mAP. We also introduce a detailed error analysis on detection results and try to provide more insights on spatio-temporal action detection. According to our analysis on MultiSports benchmark, we figure out several challenges of spatio-temporal action detection that needs to be addressed, such as capturing subtle differences between fine-grained action categories, performing accurate temporal localization, dealing with action occlusion and modeling long-range context. We hope MultiSports could serve as a standard benchmark to advance the area of spatio-temporal action detection in the future. MultiSports sptatio-temporal action detection is currently a track of DeeperAction challenge at ICCV 2021 https://deeperaction.github.io/.

In summary, our main contribution is twofold. 1) We develop a new benchmark MultiSports of spatio-temporal action detection for well-defined and realistically difficult human actions in a multi-person scene, providing high-quality and 25fps frame-wise annotations from four sports. 2) We conduct extensive studies and systematic error analysis on MultiSports, which reveals the key challenges of spatio-temporal action detection and hopefully can facilitate future research in this area.

2. Related Work

Action recognition datasets. Early datasets of action recognition mainly focus on action classification. Those datasets, including KTH [32], Weizmann [2], UCF-101 [38] and HMDB [21], contains manually trimmed short clips to capture semantics of a single action. Their human action cues, however, are overwhelmed by signals of background scenes. Multi-MiT [27] is a multi-label action recognition dataset, which may have several concurrent actions but do not provide temporal duration and spatial annotations. Recently, large-scale video classification datasets such as Sports-1M [19], YouTube-8M [1] and Kinetics [3] have been created for feature representation learning and serve as pre-training in downstream tasks, but appearance cues still play a important role here. Something-something [11] and FineGym [33], 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 action detection datasets such as ActivityNet [13], HACS [54], THUMOS14 [16], MultiTHUMOS [52] and Charades [34] provide temporal action detection 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 detection datasets, such as UCF Sports [30], UCF101-24 [38] and J-HMDB [17], typically evaluate spatio-temporal action detection 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 [46], AVA [12] and AVA-Kinetics [22] 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 annotations of all human actions of interest. MEVA [5] is a security dataset, which provides spatial-temporal annotations and some other modality annotations. But 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, well-defined boundaries, fine-grained setting and frame-by-frame annotations, which focuses on the sports domain.

Spatio-temporal action detection. 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 detection at the frame level [10, 43, 28, 31, 36, 45], 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 [15, 23, 18, 51, 24, 55, 37] have been proposed. ACT [18] took several frames as input and detected tubelets regressed from anchor cuboids. STEP [51] progressively refined the proposals by a few steps to solve the large displacement problem and utilized longer temporal information. MOC-detector [24] proposed an anchor-free tubelet detector by treating action instances as trajectories of moving points. For AVA, many methods [8, 9, 39, 47, 48] 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 detection, which differs from previous ones in multi-person scene, well-defined temporal boundaries, and fine-grained action categories. Sec. 3.1 introduces our annotation procedure. Statistics and characteristics of MultiSports are elaborated in Sec. 3.2 and Sec. 3.3.

3.1. Dataset Construction

Action vocabulary generation. We select sports of basketball, volleyball, football and aerobic gymnastics, because of their multi-person setting, less ambiguous actions and well-defined temporal boundary. For aerobic gymnastics, we use the official documentations [7]. In practice, we only select *difficulty elements* and discard *movement patterns*. For the remaining ball sports, we use an iterative way to generate our action vocabulary in each sport: we initialize an action list by the suggestions of athletes and write a handbook to clarify 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 collected from their feedback. We iteratively adjust our action list and handbook according to the feedback 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 sports, so their feedback is important for building a well-defined action vocabulary in practice. To keep action boundaries accurate

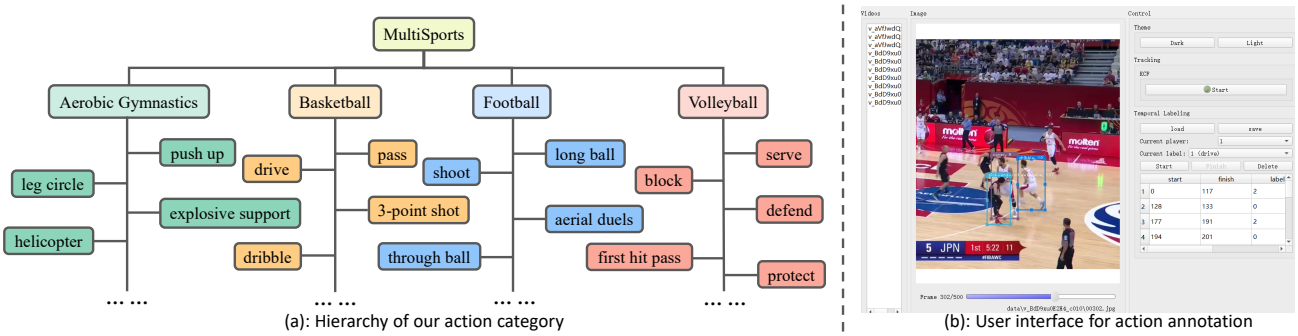

Figure 2. The action vocabulary hierarchy and annotator interface of the MultiSports dataset. (a) Our MultiSports has a two-level hierarchy of action vocabularies, where the actions of each sport are fine-grained. (b) Details of annotations can be found in Sec 3.1.

and make our dataset suitable for spatio-temporal action detection, we do not count common and atomic 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 instead of the actor's motion. What is worse, it is hard to identify who fouls due to occlusion.

Data preparation. After choosing the four sports, we search for their competition videos by querying the name of sports like volleyball and the name of competition levels like Olympics and World Cup on YouTube, and then download videos from top search results. For each video, we only select high-resolution, *e.g.* 720P or 1080P, competition records and then manually cut them into clips of minutes, with less shot changes in each clip and to be more suitable for action detection. These official records share consistent and rich content, and can guarantee a high-quality dataset.

Action annotation. Since our annotations are difficult in labeling fine-grained categories and exhaustive in determining 25fps frame-wise bounding boxes, we naturally decompose our annotation procedure into two stages: 1) A team of professional athletes generate records of the action label, the starting and ending frame, and the person box in the starting frame, which can ensure the efficiency, accuracy and consistency of our annotation results; 2) With the help of FCOT [6] tracking algorithms, a team of crowd-sourced annotators adjust bounding boxes of tracking results at each frame for each record. The ambiguity of spatial human boundaries is much less than that of fine-grained action categories and temporal action boundaries. They use the interface shown in Figure 2(b).

To ensure the consistency of action temporal boundaries, which tends to be ambiguous and remains as a big challenge for most temporal action detection datasets, we write a handbook to clarify the definition of action boundaries as mentioned above. For example, our handbook unifies the annotations of *football pass* as starting from the ball-controlling-leg leaving the ground and ending with this leg touching the ground again. The annotation handbook is provided in the supplementary material.

Person bounding-box tracking. As mentioned above, we first tack each record generated by professional athletes and then employ crowd-sourced annotators to refine the bounding boxes at each frame. Specifically, we use FCOT [6] to track the bounding boxes frame-by-frame. We find this tracking-to-refinement labeling process can not only speed up the annotation process, but also increase the annotation quality by enforcing workers to focus on determining precise boundary of each box.

We also evaluate the output of FCOT [6] and results are shown in Table 1. We adopt success and precision metrics proposed in OTB100 [49]. Aerobic turned out the hardest 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 1. Tracking results on different sports

Quality control. For the first stage of annotation, every clip has at least one annotator with domain knowledge double-checking the annotations. We correct wrong or inaccurate ones and also add 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 manually correct the inaccurate bounding boxes.

3.2. Dataset Statistics

Our *MultiSports* v1.0 contains 66 fine-grained action categories from four sports, and has videos selected from 247 competitions. The videos are manually cut into 800 clips per sport to keep data balance between sports. We discard intervals with only background scenes, such as award, and select the highlights of competitions as clips for action detection. Table 2 compares the annotation types and statistics of MultiSports v1.0 with the existing datasets. AVA [12] only has sparse and 1fps annotations of bounding boxes, which fails to provide clear temporal action boundaries and focuses on atomic action recognition. AVA-Kinetics [22] uses part of 10s clips of the Kinetics [3] and annotates one

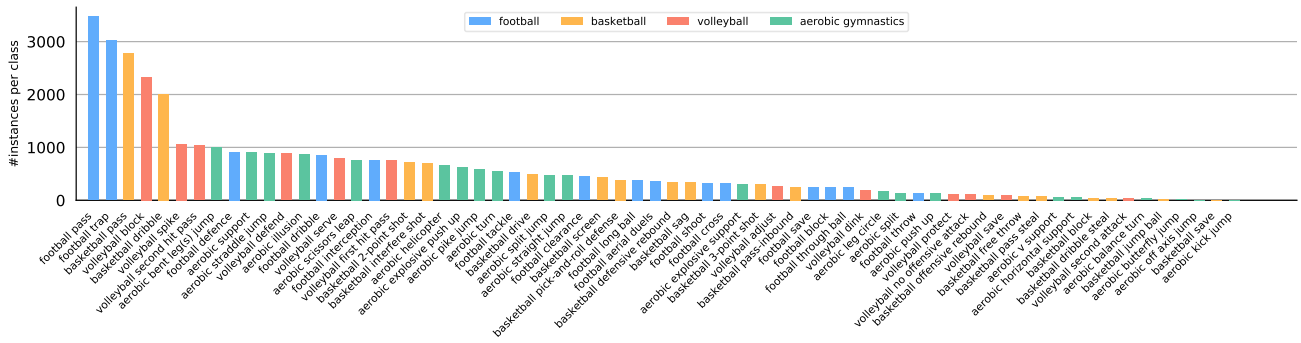

Figure 3. Statistics of each action class's data size in MultiSports, which is sorted by descending order with 4 colors indicating 4 different sports. For actions in the different sports sharing the same name, we add the name of sports before them.

	anno type	# act.	# inst.	avg act./vid. dur.	# bbox
J-HMDB [17]	Tube	21	928	1.2s / 1.2s	32k
UCF101-24 [38]	Tube	24	4458	5.1s / 6.9s	574k
AVA V2.1 [12]*	Frame	80	~56000†	Sparse‡ /15m	426k
AVA-Kinetics [22]*	Frame	80	~186000†	-	590k
HACS [54]	Segment	200	140k	33.2s / 148.7s	-
FineGym V1.0 [33]	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)

key frame per clip without any temporal boundary annotations either. Our annotation type is different from theirs. MultiSports distinguishes with existing datasets such as J-HMDB [17] and UCF101-24 [38] in longer untrimmed video clips (20.9s vs. 1.2s or 6.9s), more fine-grained action categories (66 vs. 21 or 24), much more instances (37701 vs. 928 or 4458), and more instances per video clip (11.8 vs. 1 or 1.4), which raises new challenges of modeling fast movement and fine-grained actions of multiple people in a longer video. Our MultiSports also has the largest number of bounding boxes among all existing datasets. We find that fine-grained category and well-defined boundary usually greatly shorten the action duration, which agrees with FineGym [33]. Also, we only keep the common part of actions in ball sports for well-defined boundaries. For instance, *basketball pass* starts from the player pushing the ball outwards with his arms, but does not include holding the ball and doing fake actions. Therefore our average action duration is smaller than UCF101-24 and HACS [54], which contains coarse-grained and temporally repeated actions such as *volleyball* in HACS and *riding horses* in UCF101-24.

As shown in Figure 3, the instance number of each action category ranges from 3 to 3,477, showing the natural

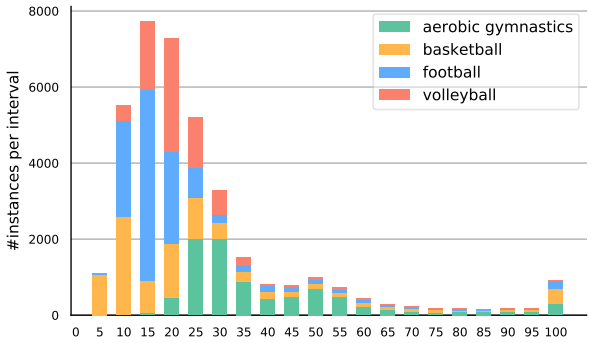

Figure 4. Statistics of action instance duration in MultiSports, where the x-axis is the number of frames and we count all instances longer than 95 frames in the last bar.

long-tailed distribution [14]. The long-tailed action categories also raise new challenges for action detection models. Figure 4 shows the distribution of action instance duration. The large variations of action instance duration add more difficulty for action detection models to accurately localize temporal boundary. Moreover, action instances in MultiSports are often related with longer temporal context and interactions with context. These inherent challenges of MultiSports require a more powerful and flexible temporal modeling scheme for action detection.

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

3.3. Dataset Characteristics

Our *MultiSports* has several distinguishing characteristics compared with existing datasets.

Difficulty. As discussed above, MultiSports is difficult in several aspects comparing to existing datasets: 1) multi-person situations of different concurrent actions, which prevents the model from distinguishing action categories only with backgrounds and requires models to capture subtly different motion cues; 2) a larger number of fine-grained categories with a long-tailed distribution; 3) the large variance of action instance duration, which makes it difficult to lo-

calize the temporal boundary; 4) the fast movement, large deformation and occlusion of actions in sports.

High Quality. The videos of MultiSports are with high-resolution (720P or 1080P) competition records, which can preserve details of small humans and objects. Besides, with the help of our annotation team composed of professional athletes, our action categories and their corresponding action boundaries are precisely annotated. The professional annotators and careful quality control is able to provide consistent and clean annotations.

Diversity. Our video clips are selected from competitions of different performance levels with diverse countries and genders, making the dataset less biased and good balanced for realistic sports analysis.

Application. This task has many application scenarios for sports analysis. Combined with Re-ID techniques, we can automatically perform game commentary, AI referee and technical statistics. It can also assess the player abilities and provide information for developing the training plan and game strategy, and trading players between clubs.

4. Experiments and Analysis

4.1. Datasets and Metrics

MultiSports benchmark. To build a solid action detection benchmark, we manually split the instances into the training set, validation set, and testing set. Due to the long-tailed distribution of action instance numbers, following AVA [12], we only 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, the current version contains 18,422 training instances from 1,574 clips and 6,577 validation instances from 555 clips. We provide the detailed ratio of training and validation instances for each sport in the supplementary material. All those instances are selected from 3200 clips covering 247 competition records. Unless otherwise mentioned, we report the results trained on the training set and evaluated on the validation set. The testing set includes 1071 clips and we withhold the annotations in the public release.

Metrics. Following the standard practice [45, 18], we utilize frame-mAP and video-mAP to evaluate action detection 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 overlapped frames. The threshold is 0.5 for frame-mAP, 0.2 and 0.5 for video-mAP.

4.2. Spatio-temporal Action Detection Results

We evaluate several representative action detection methods on *MultiSports* and compare their performance on the UCF101-24 [38], JHMDB [17], and AVA [12] in Table 3. For SlowOnly Det. and SlowFast Det., we use the code in MMAction2 [4]. We use the official released code for ROAD, YOWO and MOC. More details about the methods are provided in the supplementary material.

For UCF101-24 [38] and JHMDB [17], which have dense annotations of high-level actions as MultiSports, we find that these methods achieve good 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 [24]). In our dataset, the largest performance drop occurs on ROAD [36], which is a frame-level action detector that performs action detection at each frame independently without exploiting temporal information. UCF101-24 [38] and JHMDB [17] 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 fails to provide sufficient information for fine-grained action recognition. Meanwhile, our temporal boundary annotation is more precise and requires more accurate localization in temporal domain.

For AVA [12], which has only sparse annotations of atomic actions, we observe that the performance gap between SlowFast Det. [8] and SlowOnly Det. [8] on MultiSports is more evident 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 fast motion at a finer temporal granularity. As shown in Figure 5, many aerobic actions gain large absolute improvement, such as aerobic turn (+30 AP) and aerobic horizontal support (+54 AP). We analyze that aerobic actions' deformation and displacement is the largest among the four sports and benefit more from this finer temporal analysis. We also observe a large performance increase in other sports, such as basketball pass, football clearance and volleyball second attack, which have short temporal duration and intense motion.

4.3. Error Analysis

In this section, we analyze the cause of errors to better understand *MultiSports*' challenges. Based on ACT [18] frame-mAP error analysis, which is designed 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. $\mathbf{E_R}$: a detection result targets at a ground-truth tube that has already been matched. $\mathbf{E_N}$: 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

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 [36]	300 × 300	3.90	0.00	0.00	70.7	69.8	40.9	-	60.8	59.7	-
YOWO [20]	224 × 224	9.28	10.78	0.87	71.10	72.97	46.42	74.51	88.05	82.57	-
MOC [24] (K=7)	288 × 288	22.51	12.13	0.77	78.0	82.8	53.8	70.8	77.3	77.2	-
MOC [24] (K=11)	288 × 288	25.22	12.88	0.62	-	-	-	-	-	-	-
SlowOnly Det., 4 × 16 [8]	short side 256	16.70	15.71	5.50	-	-	-	-	-	-	20.02
SlowFast Det., 4 × 16 [8]	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.

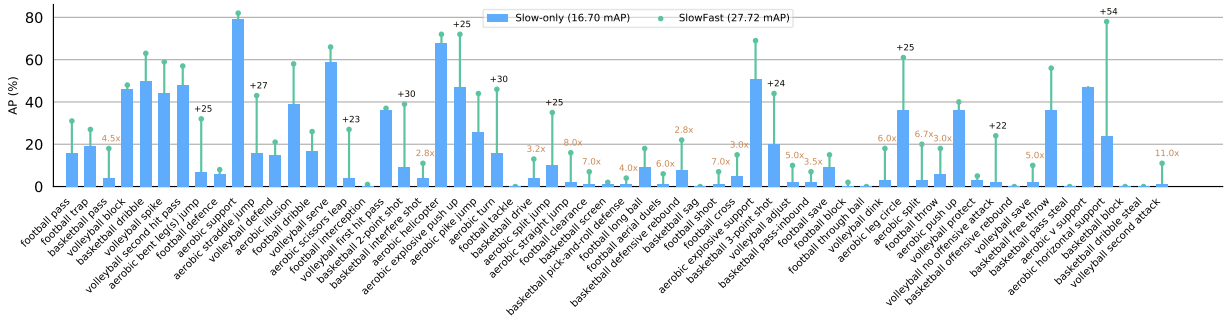

Figure 5. SlowOnly vs. SlowFast frame-mAP. Categories are sorted by descending order on the number of instances.

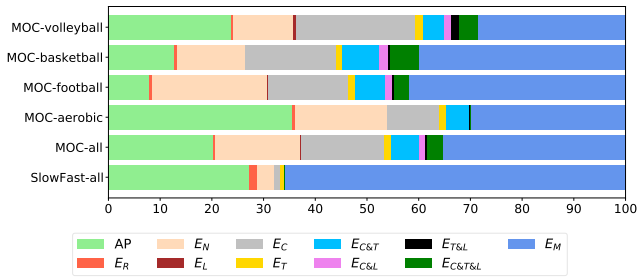

Figure 6. Error Analysis. AP is the correct detection. The threshold for a ground-truth matched by a detection is 0.1. Recall is $1 - E_M$

localization. $\mathbf{E_{C\&T}}$, $\mathbf{E_{C\&L}}$, $\mathbf{E_{T\&L}}$, $\mathbf{E_{C\&T\&L}}$: a detection that is inaccurate in corresponding aspects while acceptable in other aspect (if any). For example, $\mathbf{E_{C\&T}}$ refers to results with wrong action class, inaccurate temporal localization 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 which is attached to our supplementary material. The code is provided at https://github.com/MCG-NJU/MultiSports.

As shown in Figure 6, despite the relatively low recall, SlowFast Det. achieves higher video-mAP than MOC because it makes much fewer false positive predictions. This can be explained by the fact that SlowFast Det. uses Faster RCNN [29] finetuned on MultiSports as person detector to greatly avoid the 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 detection results, 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 error is 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 for people without doing any sports action. What is 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 localization errors for current methods. Therefore we need a more effective way of modeling temporal boundary. 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 [24] and we report results on UCF101-24 [38] and *MultiSports* with different K in Table 4. For frame-mAP, we can find that MultiSports can benefit more from longer temporal context than UCF101-24, in spite of the shorter action duration of MultiSports than UCF101-24 as shown in Table 2. For video-mAP, the result does not increase as frame-mAP. We analyze there are two reasons. First, predicting movement in MOC is harder with longer input length. What is worse, the categories in MultiSports have large deformation and displacement, and MOC Movement Branch can not predict them accurately, which harms the video level detection seriously. Second,

Figure 7. Visualization of typical errors in 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.

Figure 4 shows the variability of action duration. The ratio is 9% for instances duration less than 7 and 23% for less than 11. The fixed clip length K (e.g. 11) will damage temporal detection ability. So, we need to consider longer temporal context, more accurate movement estimation and flexible temporal detection for MultiSports.

Which action categories are challenging? Figure 5 shows that not all categories yield better performance with more training samples. Categories highly correlated with scenes (such as basketball free throw) or aerobics basic categories (such as aerobic horizontal support and V support) can still achieve high performance with fewer samples. Note that aerobics contains basic and complex categories, where complex action combines the motion of basic action and its own core motion, thus longer temporal context is required for these complex actions. In contrast, categories with short temporal duration and intense motion (such as football pass, basketball pass and football interception) achieve low performance even though with lots of training samples. By observing the confusion matrix in supplementary materials, we summarize other common 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 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.

Trimmed vs. untrimmed settings. *MultiSports* has well-defined and high-quality temporal boundaries. We evaluate the performance of SlowFast Det. under both the untrimmed and trimmed setting on MultiSports and AVA datasets. The results are reported in Table 5. The trimmed setting only evaluates the performance on the frames having annotations and the untrimmed setting reports the performance on all frames. We find that it only drops 2% on AVA while 11% on our dataset, 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.

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@0.5	V@0.2	V@0.5	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.

5. Conclusion

In this paper, we have introduced the *MultiSports* dataset with dense spatio-temporal annotations of actions from four sports. MultiSports distinguishes from the existing action detection datasets in many aspects: 1) raising new challenges for recognizing fine-grained action classes; 2) requirement of accurate localization of well-defined boundaries in multiple-person situations; 3) high quality video data and dense annotations; 4) potential applications in sports analysis; 5) less biased dataset with high diversity in competition levels, countries and genders. We have empirically investigated several action detection baseline methods on the MultiSports dataset. Our error analysis and ablation studies on the detection results uncover several insightful findings that are beneficial for the future research of spatio-temporal action detection.

Acknowledgements. This work is supported by National Natural Science Foundation of China (No. 62076119, No. 61921006), Program for Innovative Talents and Entrepreneur in Jiangsu Province, and Collaborative Innovation Center of Novel Software Technology and Industrialization. Thanks to professional athletes of Nanjing University varsities and MCG students for annotating this dataset.

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. 3

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005. 3

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1, 3, 4

[4] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 6

[5] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. MEVA: A large-scale multiview, multimodal video dataset for activity detection. In *WACV*, pages 1059–1067, 2021. 3

[6] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Fully convolutional online tracking. *CoRR*, abs/2004.07109, 2020. 2, 4

[7] Federation Internationale de Gymnastique. Aerobic gymnastics-code of points. *FIG Aerobic Gymnastics FIG Executive Committee*, 2017. 2, 3

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 3, 6, 7

[9] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, pages 244–253, 2019. 3

[10] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, pages 759–768, 2015. 3

[11] 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, 2017. 3

[12] 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, 2018. 1, 3, 4, 5, 6

[13] 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, 2015. 3

[14] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *CoRR*, abs/1709.01450, 2017. 5

[15] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (T-CNN) for action detection in videos. In *ICCV*, pages 5823–5832, 2017. 3

[16] 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.*, pages 1–23, 2017. 3

[17] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 1, 2, 3, 5, 6

[18] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4415–4423, 2017. 3, 6

[19] 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, 2014. 3

[20] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *CoRR*, abs/1911.06644, 2019. 7

[21] Hildegard 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, 2011. 3

[22] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *CoRR*, abs/2005.00214, 2020. 3, 4, 5

[23] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV*, pages 306–322, 2018. 3

[24] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *ECCV*, pages 68–84, 2020. 3, 6, 7

[25] 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, 2019. 1

[26] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–21, 2018. 1

[27] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A. McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *CoRR*, abs/1911.00232, 2019. 3

[28] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV*, pages 744–759, 2016. 3

[29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 7

[30] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 3

[31] 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*, 2016. 3

[32] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, pages 32–36, 2004. 3

[33] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2613–2622, 2020. 3, 5

[34] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 3

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 1

[36] 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, 2017. 3, 6, 7

[37] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *CVPR*, pages 11987–11995, 2019. 3

[38] 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. 1, 2, 3, 5, 6, 7

[39] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *ECCV*, pages 71–87, 2020. 3

[40] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1

[41] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439, 2018. 1

[42] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 1

[43] Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, pages 2708–2717, 2016. 3

[44] 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*, pages 20–36, 2016. 1

[45] Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, pages 3164–3172, 2015. 3, 6

[46] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *CoRR*, abs/1605.05197, 2016. 3

[47] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 3

[48] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware RCNN: A baseline for action detection in videos. In *ECCV*, pages 440–456, 2020. 3

[49] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1834–1848, 2015. 4

[50] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5794–5803, 2017. 1

[51] 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, 2019. 3

[52] 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.*, pages 375–389, 2018. 3

[53] 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, 2019. 1

[54] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: human action clips and segments dataset for recognition and temporal localization. In *ICCV*, pages 8667–8677, 2019. 3, 5

[55] Jiaojiao Zhao and Cees G. M. Snoek. Dance with flow: Two-in-one stream action detection. In *CVPR*, pages 9935–9944, 2019. 3

[56] 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, 2017. 1