FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding

Dian Shao Yue Zhao Bo Dai Dahua Lin

CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
{sd017, zy317, bdai, dhlin}@ie.cuhk.edu.hk

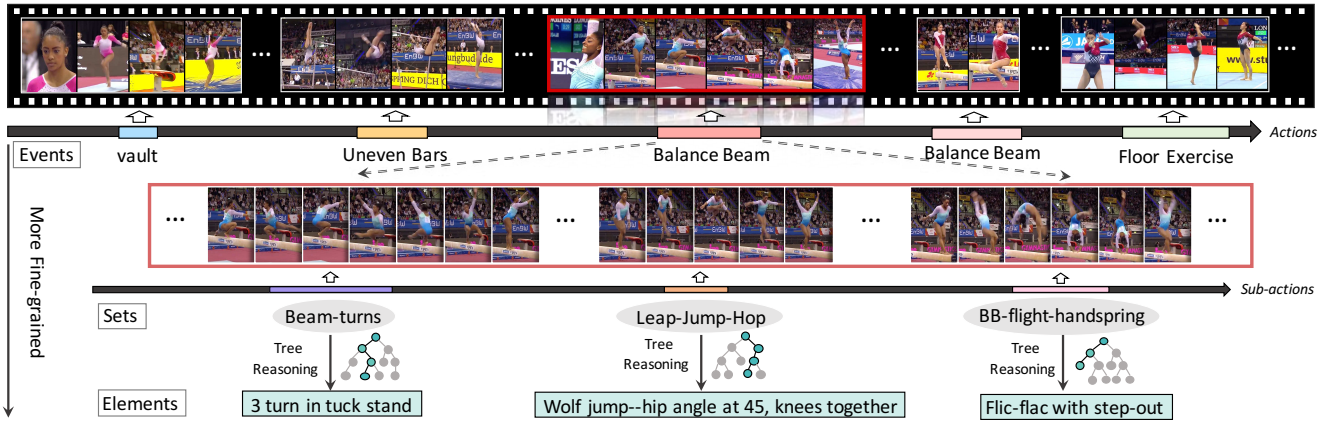


Figure 1: An overview of the *FineGym* dataset. We provide coarse-to-fine annotations both temporally and semantically. There are three levels of categorical labels. The temporal dimension (represented by the two bars) is also divided into two levels, *i.e.*, actions and sub-actions. Sub-actions could be described generally using set categories or precisely using element categories. Ground-truth element categories of sub-action instances are obtained via manually constructed decision-trees.

Abstract

On public benchmarks, current action recognition techniques have achieved great success. However, when used in real-world applications, e.g. sport analysis, which requires the capability of parsing an activity into phases and differentiating between subtly different actions, their performances remain far from being satisfactory. To take action recognition to a new level, we develop FineGym[1], a new dataset built on top of gymnastic videos. Compared to existing action recognition datasets, FineGym is distinguished in richness, quality, and diversity. In particular, it provides temporal annotations at both action and sub-action levels with a three-level semantic hierarchy. For example, a "balance beam" event will be annotated as a sequence of elementary sub-actions derived from five sets: "leap-jump-hop", "beam-turns", "flight-salto", "flight-handspring", and "dismount", where the sub-action in each set will be further annotated with finely defined class labels. This new level of granularity presents significant challenges for action recognition, e.g. *how to parse the temporal structures from a coherent action, and how to distinguish between subtly different action classes. We systematically investigate representative methods on this dataset and obtain a number of interesting findings. We hope this dataset could advance research towards action understanding.*

[1]Dataset and codes can be found at https://sdolivia.github.io/FineGym/

1. Introduction

The remarkable progress in action recognition [39, 42, 40, 25, 37, 49], particularly the development of many new recognition models, *e.g.* TSN [44], TRN [55], and I3D [3], have been largely driven by large-scale benchmarks, such as ActivityNet [11] and Kinetics [23]. On these benchmarks, latest techniques have obtained very high accuracies.

Even so, we found that existing techniques and the datasets that underpin their development are subject to an important limitation, namely, they focus on *coarse-grained* action categories, *e.g. "hockey"* vs. *"gymnastics"*. To dif-

ferentiate between these categories, the background context often plays an important role, sometimes even more significant than the action itself. However, in certain areas, coarse-grained classification is not enough. Take sport analytics for example, it usually requires a detailed comparison between fine-grained classes, *e.g*. different moves during a vault. For such applications, the capability of fine-grained analysis is needed. It is worth noting that the *fine-grained* capability here involves two aspects: 1) *temporal:* being able to decompose an action into smaller elements along the time axis; 2) *semantical:* being able to differentiate between sub-classes at the next level of the taxonomic hierarchy.

To facilitate the study of fine-grained action understanding, we develop *FineGym*, short for *Fine-grained Gymnastics*, which is a large-scale high-quality action dataset that provides fine-grained annotations. Specifically, *FineGym* has several distinguished features: 1) *Multi-level semantic hierarchy.* All actions are annotated with semantic labels at three levels, namely *event*, *set*, and *element*. Such a semantic hierarchy provides a solid foundation for both coarse- and fine-grained action understanding. 2) *Temporal structure.* All action instances of interest in each video are identified, and they are manually decomposed into sub-actions. These annotated temporal structures also provide important support to fine-grained understanding, from another aspect. 3) *High quality.* All videos in the dataset are high-resolution records of high-level professional competitions. Also, careful quality control is enforced to ensure the accuracy, reliability, and consistency of the annotations. These aspects together make it a rich dataset for research and a reliable benchmark for assessment. Moreover, we have summarized a systematic framework for collecting data and annotations, *e.g*. labeling via decision trees, which can also be applied to the construction of other datasets with similar requirements.

Taking advantage of the new exploration space offered by *FineGym*, we conducted a series of empirical studies, with the aim of revealing the challenges of fine-grained action understanding. Specifically, we tested various action recognition techniques and found that their performance on fine-grained recognition is still far from being satisfactory. In order to provide guidelines for future research, we also revisited a number of modeling choices, *e.g*. the sampling scheme and the input data modalities. We found that for fine-grained action recognition, 1) sparsely sampled frames are not sufficient to represent action instances. 2) Motion information plays a significantly important role, rather than visual appearance. 3) Correct modeling of temporal dynamics is crucial. 4) And pre-training on datasets which target for coarse-grained action recognition is not always beneficial. These observations clearly show the gaps between coarse- and fine-grained action recognition.

Overall, our work contributes to the research of action understanding in two different ways: 1) We develop a new dataset *FineGym* for fine-grained action understanding, which provides high-quality and fine-grained annotations. In particular, the annotations are in three semantic levels, namely *event*, *set*, and *element*, and two temporal levels, namely *action* and *sub-action*. 2) We conduct in-depth studies on top of *FineGym*, which reveal the key challenges that arise in the fine-grained setting, which may point to new directions of future research.

2. Related Work

Coarse-grained Datasets for Action Recognition. Being the foundation of more sophisticated techniques, the pursuit of better datasets never stops in the area of action understanding. Early attempts could be traced back to KTH [35] and Weizmann [1]. More challenging datasets are proposed subsequently, including UCF101 [40], Kinetics [3], ActivityNet [11], Moments in Time [32], and others [25, 18, 50, 2, 52, 38, 33, 22, 46, 31, 20]. Some of them also provide annotations beyond category labels, ranging from temporal locations [18, 50, 2, 11, 52, 38] to spatial-temporal bounding boxes [33, 22, 46, 31, 20]. However, all of these datasets target for coarse-grained action understanding (*e.g*. *hockey*, *skateboarding*, etc.), in which the background context often provides distinguishing signals, rather than the actions themselves. Moreover, as reported in [44, 29], sometimes a few frames are sufficient for action recognition on these datasets.

Fine-grained Datasets for Action Recognition. There are also attempts towards building datasets for fine-grained action recognition [6, 34, 19, 15, 24, 29]. Specifically, both Breakfast [24] and MPII-Cooking 2 [34] provides annotations for individual steps of various cooking activities. In [24] the coarse actions (*e.g*. *Juice*) are decomposed into action units (*e.g*. cut orange), and in [34] the verb parts are defined to be fine-grained classes (*e.g*. cut in *cutting onion*). Something-Something [19] collects 147 classes of daily human-object interactions, such as *moving something down* and *taking something from somewhere*. Diving48 [29] is built on 48 fine-grained diving actions, where the labels are combinations of 4 attributes, *e.g*. *back+15som+15twis+free*. Compared to these datasets, our proposed *FineGym* has the following characteristics: 1) the structure hierarchy is more sophisticated (2 temporal levels and 3 semantic levels), and the number of finest classes is significantly larger (e.g. 530 in FineGym vs. 48 in Breakfast); 2) the actions in FineGym involve rapid movements and dramatic body deformations, raising new challenges for recognition models; 3) the annotations are obtained with reference to expert knowledge, where a unified standard is enforced across all classes to avoid ambiguities and inconsistencies.

Methods for Action Recognition. Upon *FineGym* we have empirically studied various state-of-the-art action recog-

nition methods. These methods could be summarized in three pipelines. The first pipeline adopts a 2D CNN [39, 44, 13, 10] to model per-frame semantics, followed by a 1D module to account for temporal aggregation. Specifically, TSN [44] divides an action instance into multiple segments, representing the instance via a sparse sampling scheme. An average pooling operation is used to fuse per-frame predictions. TRN [55] and TSM [30] respectively replace the pooling operation with a temporal reasoning module and a temporal shifting module. Alternatively, the second pipeline directly utilizes a 3D CNN [42, 3, 43, 45, 8] to jointly capture spatial-temporal semantics, such as Non-local [45], C3D [42], and I3D [3]. Recently, an intermediate representation (*e.g*. human skeleton in [48, 4, 5]) is used by several methods, which could be described as the third pipeline. Besides action recognition, other tasks of action understanding, including action detection and localization [14, 47, 54, 21, 16, 36], action segmentation [26, 9], and action generation [28, 41], also attract many researchers.

3. The FineGym Dataset

The goal of our *FineGym* dataset is to introduce a new challenging benchmark with high-quality annotations to the community of action understanding. While more types of annotations will be included in succeeding versions, current version of *FineGym* mainly provides annotations for *fine-grained* human action recognition on gymnastics.

Practically, categories of actions and sub-actions in *FineGym* are organized according to a three-level hierarchy, namely *events*, *sets*, and *elements*. Events, at the coarsest level of the hierarchy, refer to actions belonging to different gymnastic routines, such as *vault* (VT), *floor exercise* (FX), *uneven-bars* (UB), and *balance-beam* (BB). Sets are mid-level categories, describing sub-actions. A set holds several technically and visually similar elements. At the finest granularity are element categories, which equips sub-actions with more detailed descriptions than the set categories. *e.g*. a sub-action instance of the set *beam-dismounts* could be more precisely described as *double salto backward tucked* or other element categories in the set. Meanwhile, *FineGym* also provides two levels of temporal annotations, namely locations of all events in a video and locations of sub-actions in an action instance (*i.e*. event instance). Figure 2 reveals the annotation organization of *FineGym*.

Below we at first review the key challenges when building *FineGym*, followed by a brief introduction on the construction process, which covers both data preparation, annotation collection and quality control. Finally, statistics and properties of *FineGym* are elaborated.

3.1. Key Challenges

Building such a complex and fine-grained dataset brings a series of unprecedented challenges, including: (1) *How to collect data?* Generally, data for large-scale action datasets are mainly collected in two ways, namely crawling from the Internet and self-recording from invited workers. However, while fine-grained labels of *FineGym* contain rich details, *e.g*. *double salto backward tucked with 2 twist*, videos collected in these ways can hardly match the details precisely. Instead, we collect data from video records of high-level professional competitions. (2) *How to define and organize the categories?* With the rich granularities of *FineGym* categories and the subtle differences between instances of the finest categories, manually defining and organizing *FineGym* categories as in [19, 34] is impractical. Fortunately, we could resort to official documentation provided by experts [7], which naturally define and organize *FineGym* categories in a consistent way. This results in 530 well defined categories. (3) *How to collect annotations?* As mentioned, the professional requirements and subtle differences of *FineGym* categories prevent us from utilizing crowdsourcing services such as the Amazon Mechanical Turk. Instead, we hire a team trained specifically for this job. (4) *How to control the quality?* Even with a trained team, the richness and diversity of possible annotations inevitably require an effective and efficient mechanism for quality control, without which we may face serious troubles as errors would propagate along the hierarchies of *FineGym*. We thus enforce a series of measures for quality control, as described in 3.2.

3.2. Dataset Construction

Data Preparation. Our procedure for data collection takes the following steps. We start by surveying the top-level gymnastics competitions held in recent years. Then, we collect official video records of them from the Internet, ensuring these video records are complete, distinctive and of high-resolutions, *e.g*. 720P and 1080P. Finally, we cut them evenly into chunks of 10-minutes for further processing. Through these steps, the quality of data is ensured by the choice of official video records. The temporal structures of actions and sub-actions are also guaranteed as official competitions are consistent and rich in content. Moreover, data redundancy is avoided through manual checking.

Annotation Collection. We adopt a multi-stage strategy to collect the annotations for both the three-level semantic category hierarchy (*i.e*. event, set and element labels) and the two-level temporal structures of action instances. The whole annotation process is illustrated in Figure 1, and described as follows: 1) Firstly, annotators are asked to accurately locate the start and end time of each complete gymnastics routine (*i.e*. *an complete action instance containing several sub-actions*) in a video record, and then select the correct event label for it. In this step, we discard all incomplete routines, such as the ones that have an interruption. 2) Secondly, 15 sets from 4 events are selected from the latest official codebook [7], for they provide more

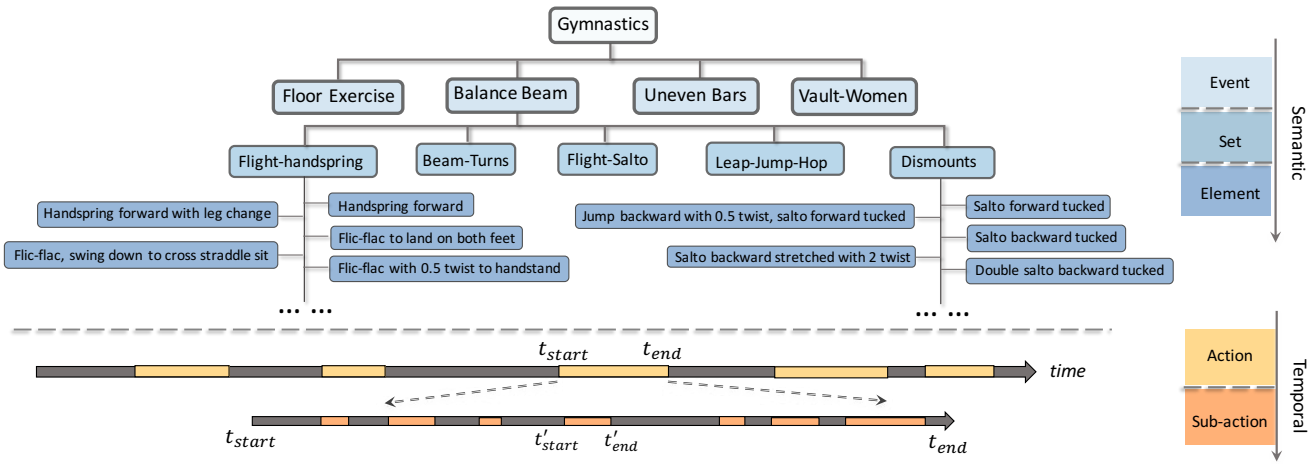


Figure 2: *FineGym* organizes both the semantic and temporal annotations hierarchically. The upper part shows three levels of categorical labels, namely *events* (*e.g. balance beam*), *sets* (*e.g. dismounts*) and *elements* (*e.g. salto forward tucked*). The lower part depicts the two-level temporal annotations, *i.e.* the temporal boundaries of actions (in the top bar) and sub-action instances (in the bottom bar).

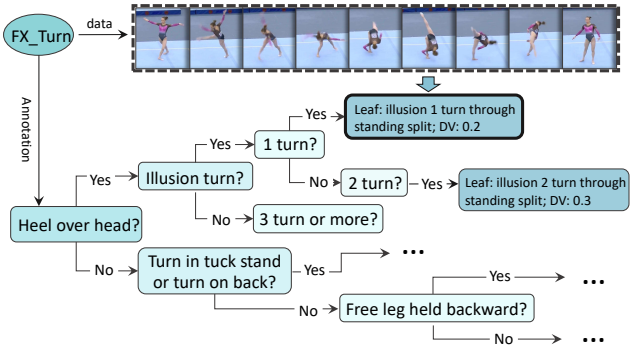


Figure 3: Illustration of the decision-tree based reasoning process for annotating element labels within a given set (*e.g. FX-turns*).

distinctive element-level classes. We further discard the element-level classes that have visually imperceptible differences and unregulated moves. Consequently, when given an event, an annotator will locate all the sub-actions from the defined sets and provide their set-level labels. 3) Each sub-action further requires an element label, which is hard to decide directly. We thus utilize a decision-tree[2] consisting of attribute-based queries to guide the decision. Starting from the root, which has a set label, an annotator travels on the tree until he meets a leaf node, which has an element label. See Figure 3 for a demonstration.

Quality Control. To build a high-quality dataset which offers clean annotations at all hierarchies, we adopt a series of mechanisms including: training annotators with domain-specific knowledge, pretesting the annotators rigorously before formal annotation, preparing referential slides as well as demos, and cross-validating across annotators.

[2]Details of the decision-trees are included in the supplemental material.

event	# set cls	# element cls	# inst	# sub_inst
VT	1	67	2034	2034
FX	5	64+20+23+4	912	8929
BB	5	58+25+26+26	976	11586
UB	4	52+22+57+2	961	10148
10 in total	15	530	4883	32697

Table 1: The statistics of *FineGym* v1.0.

3.3. Dataset Statistics

Table 1 shows the statistics of *FineGym* v1.0, which is used for empirical studies in this paper.[3] Specifically, *FineGym* contains 10 event categories, including 6 male events and 4 female events. Particularly, we selected 4 female events therefrom to provide more fine-grained annotations. The number of instances in each element category ranges from 1 to 1,648, reflecting the natural heavy-tail distribution of them. 354 out of the defined 530 element categories have at least one instance.[4] To meet different demands, besides the naturally imbalanced setting, we also provide a more balanced setting by thresholding the number of instances. Details are included in Sec.4.2. In terms of other statistics, there are currently 303 competition records, amounted to $\sim$ 708 hours. For the 4 events with finer

[3]More data is provided in the v1.1 release, see the webpage for details.

[4]The overall distribution of element categories is presented in the supplementary material.

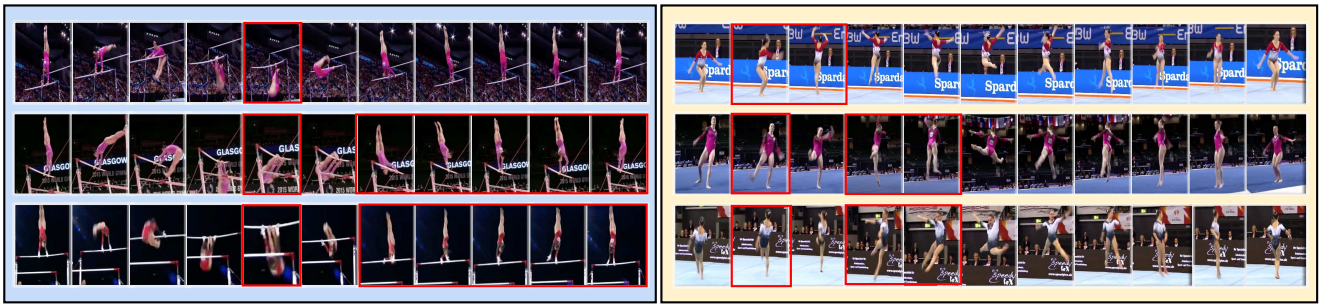


Figure 4: Examples of fine-grained sub-action instances in *FineGym*. The left part shows instances belonging to three element categories within the set *UB-circles*, from top to bottom: *"clear pike circle backward with* 1 *turn"*, *"pike sole circle backward with* 1 *turn"*, and *"pike sole circle backward with* 0.5 *turn*. On the right, there are instances from three element categories of the set *FX-leap-jump-hop*, from top to bottom: *"split jump with* 1 *turn"*, *"split leap with* 1 *turn"*, and *"switch leap with* 1 *turn"*. It can be seen such fine-grained instances contain subtle and challenging differences. Best viewed in high resolution.

annotations, *Vault* has a relatively shorter duration (8s in average) and intenser motions, while others have a relatively longer duration (55s in average). Being temporally more fine-grained, the annotated sub-action instances usually cover less than 2 seconds, satisfying the prerequisite to being temporally short for learning more fine-grained information [19].

3.4. Dataset Properties

FineGym has several attracting properties that distinguish it from existing datasets.

High Quality. The videos in *FineGym* are all official recordings of top-level competitions, action instances in which are thus professional and standard. Besides, over 95% of these videos are of high resolutions (720P and 1080P), so that subtle differences between action instances are well preserved, leaving a room for future annotations and models. Also, due to the utilization of a well-trained annotation team and official documents of category definitions and organizations, annotations in *FineGym* are consistent and clean across different aspects.

Richness and Diversity. As discussed, *FineGym* contains multiple granularities both semantically and temporally. While the number of categories increases significantly when we move downwards along the semantic hierarchy, the varying dynamics captured in temporal granularities lay a foundation for more comprehensive temporal analysis. Moreover, *FineGym* is also rich and diverse in terms of viewpoints and poses. For example, many rare poses are covered in *FineGym* due to actions like twist and salto.

Action-centric Instances. Unlike several existing datasets where the background is also a major factor for distinguishing different categories, all instances in *FineGym* have relatively consistent backgrounds. Moreover, being the same at first glance, instances from two different categories may only have subtle differences, especially at the finest semantic granularity. *e.g.* the bottom two samples in the right of Figure 4 differ in whether the directions of legs and the turn are consistent at the beginning. We thus believe *FineGym* is one of the challenging datasets that requires more focus on the actions themselves.

Model Info		*Event*		*Set*	
Method	**Modality**	**Mean**	**Top-1**	**Mean**	**Top-1**
TSN [44]	RGB	98.42	98.18	89.85	95.25
	Flow	93.40	93.25	91.64	96.42
	2Stream	98.47	99.86	91.97	97.69

Table 2: Results of Temporal Segment Network (TSN) in terms of coarse-grained (*i.e.* event and set level) action recognition.

Decision Trees of Element Categories. As we annotate element categories using manually built decision trees consisting of attribute-based queries, the path from a tree's root to one of its leaf node naturally offers more information than just an element label, such as the attribute sets and the difficulty score of an element. Potentially one could use these decision trees for prediction interpretation and reasoning.

4. Empirical Studies

On top of *FineGym*, we systematically evaluate representative action recognition methods across multiple granularities, and also include a demonstrative study on a typical action localization method using MMAction [53]. All training protocols follow the original papers unless stated otherwise. Our main focus is on understanding fine-grained actions (*i.e.* element-level), whose challenging characteristics could lead to new inspirations. Finally, we provide some heuristic observations for future research on this direction.

Model Info		*Gym288*		*Gym99*	
Method	**Modality**	**Mean**	**Top-1**	**Mean**	**Top-1**
Random	-	0.3	-	1.0	–
ActionVLAD [17]	RGB	16.5	60.5	50.1	69.5
TSN [44]	RGB	26.5	68.3	61.4	74.8
	Flow	38.7	78.3	75.6	84.7
	2Stream	37.6	79.9	76.4	86.0
TRN [55]	RGB	33.1	73.7	68.7	79.9
	Flow	42.6	79.5	77.2	85.0
	2Stream	42.9	81.6	79.8	87.4
TRNms [55]	RGB	32.0	73.1	68.8	79.5
	Flow	43.4	79.7	77.6	85.5
	2Stream	43.3	82.0	80.2	87.8
TSM [30]	RGB	34.8	73.5	70.6	80.4
	Flow	46.0	81.6	80.3	87.1
	2Stream	46.5	83.1	81.2	88.4
I3D [3]	RGB	27.9	66.7	63.2	74.8
I3D* [3]	RGB	28.2	66.1	64.4	75.6
NL I3D [45]	RGB	27.1	64.0	62.1	73.0
NL I3D* [45]	RGB	28.0	67.0	64.3	75.3
ST-GCN [48]	Pose	11.0	34.0	25.2	36.4

(a) Results of *elements across all events*.

Model Info		**VT**, *6cls*		**FX**, *35cls*	
Method	**Modality**	**Mean**	**Top-1**	**Mean**	**Top-1**
Random	-	16.7	-	2.9	–
ActionVLAD [17]	RGB	32.7	44.6	56.4	65.0
TSN [44]	RGB	27.8	46.6	58.6	67.5
	Flow	23.1	42.6	70.7	78.5
	2Stream	27.0	47.5	73.1	81.6
TRN [55]	RGB	32.1	48.0	65.8	72.0
	Flow	28.9	44.2	74.9	81.2
	2Stream	31.4	47.1	77.5	84.6
TRNms [55]	RGB	31.5	46.6	66.6	73.4
	Flow	29.1	43.9	74.8	81.1
	2Stream	30.1	47.3	78.2	84.9
TSM [30]	RGB	29.2	42.2	62.2	68.8
	Flow	26.2	42.4	76.2	81.9
	2Stream	28.8	44.8	76.9	83.6
I3D [3]	RGB	31.5	42.1	53.7	59.5
I3D* [3]	RGB	33.4	47.8	52.2	60.2
NL I3D [45]	RGB	30.6	46.0	53.4	59.8
NL I3D* [7]	RGB	30.8	47.3	50.9	57.6
ST-GCN [48]	Pose	19.5	38.8	35.3	40.1

(b) Results of *elements within a event*.

Model Info		**FX-S1**, *11cls* **UB-S1**, *15cls*			
Method	**modality**	**Mean**	**Top-1**	**Mean**	**Top-1**
Random	-	9.1	-	6.7	–
ActionVLAD [17]	RGB	45.0	52.3	51.9	64.6
TSN [44]	RGB	31.2	49.9	44.8	65.6
	Flow	69.6	78.0	65.3	78.9
	2Stream	68.2	78.5	65.0	80.0
TRN [55]	RGB	58.2	55.0	53.6	70.9
	Flow	73.3	79.9	71.5	82.5
	2Stream	74.4	81.9	83.0	71.0
TRNms [55]	RGB	58.5	64.4	55.8	71.4
	Flow	75.8	82.6	70.8	82.2
	2Stream	72.9	80.8	70.8	83.2
TSM [30]	RGB	45.6	53.3	50.9	66.4
	Flow	75.8	81.7	73.1	82.5
	2Stream	72.9	79.4	70.1	80.8
I3D [3]	RGB	33.3	38.9	32.2	49.1
I3D* [3]	RGB	36.1	42.9	31.0	48.1
NL I3D [45]	RGB	31.4	39.0	29.3	48.5
NL I3D* [45]	RGB	35.8	40.1	26.9	48.5
ST-GCN [48]	Pose	21.6	30.8	13.7	28.1

(c) Results of *elements within a set*.

Table 3: Element-level action recognition results of representative methods. Specifically, results of recognizing element categories across all events, within an event, and within a set, are respectively included in (a), (b), and (c).

4.1. Event-/Set-level Action Recognition

We present a brief demonstrative study for the event and set level action recognition, as their characteristics resemble the coarse-grained action recognition that is well studied in multiple benchmarks. Specifically, we choose the widely adopted Temporal Segment Networks (TSN) [44] as the representative. It divides an instance into 3 segments and samples one frame from each segment to form the input. Visual appearance (RGB) and motion (Optical Flow) features of the input frames are separately processed in TSN, making it a good choice for comparing the contribution of each feature source. The results of event and set level action recognition are listed in Table 2, from which we observe: 1) 3 frames, accounting for less than 5% of all frames, are sufficient for recognizing event and set categories, suggesting categories at these two levels could be well classified using isolated frames. 2) Compared to motion features, appearance features contribute more at the event-level, and vice versa at the set level. This means the reliance on static visual cues such as background context is decreased as we step into a finer granularity. Such trend continues and becomes clearer at the finest granularity, as shown in the element-level action recognition.

4.2. Element-level Action Recognition

We mainly focus on the element-level action recognition, which raises significant challenges for existing methods. Specifically, representative methods belonging to various pipelines are selected, including 2D-CNN (*i.e*. TSN [44], TRN [55], TSM [30], and ActionVLAD [17]), 3D-CNN methods (*i.e*. I3D [3], Non-local [45]), as well as a skeleton-based method ST-GCN [48].

These methods are thoroughly studied in three sub-tasks, namely recognition of *elements across all events*, *elements within an event*, and *elements within a set*, as shown in Figure 3 (a), (b), and (c) respectively. For *elements across all events*, we adopt both a natural long-tailed setting and a more balanced setting, respectively referred to as *Gym288* and *Gym99*. Details of these settings are included in the supplemental material. For *elements within an event*, we separately select from *Gym99* all elements of two specific events, namely *Vault (VT)* and *Floor Exercise (FX)*. The elements of *FX* come from 4 different sets, while the elements of *VT* come from a single set (*VT* is a special event with only one set). Finally for *elements within a set*, we select the set *FX-G1* covering leaps, jumps and hops of *FX*, and the set *UB-G1* covering circles in *Uneven Bars (UB)*.

From the results of these tasks in Table 3, we have summarized several observations. (1) Given the long-tail nature of the instance distribution, all methods are shown to overfit to the elements having the most number of instances, especially on the setting *Gym288*. (2) Due to the subtle differences between elements, visual appearances in the form of RGB values contribute significantly less than that in coarse-grained action recognition. And motion features contribute a lot in most cases except for the *Vault* in *elements within an event*, for motion dynamics of elements in *Vault* are very intense. (3) Capturing temporal dynamics is important as TRN and TSM outperform TSN by large margins. (4) I3D and Non-local network pre-trained on ImageNet and Kinetics obtain similar results with 2D-CNN methods, which may be due to the large gap between temporal patterns of element categories and those from Kinetics. (5) Skeleton-

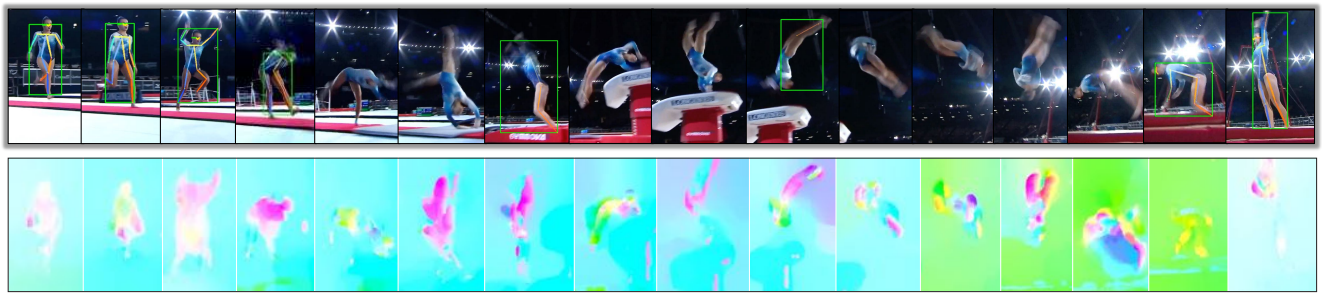


Figure 5: **Top-row** represents the results of person detection and pose estimation [12, 27] for a *Vault* routine, and the **bottom-row** visualizes the optical flow [51] features. It can be seen that detections and pose estimations of the gymnast are missed in multiple frames, especially in frames with intense motion. Best viewed in high resolution.

***GymFine*, mAP@α**						
Temporal level	0.50	0.60	0.70	0.80	0.90	Avg
Action	60.0	57.9	57.1	54.6	35.0	49.4
Sub-action	22.2	15.4	9.2	3.9	0.6	9.6

Table 4: Temporal action localization results of SSN [54] at coarse- (*i.e.* actions) and fine-grained (*i.e.* sub-actions) levels. The metric is mAP@tIoU. The average (Avg) values are obtained by ranging tIoU thresholds from 0.5 to 0.95 with an interval of 0.05.

# Frame	*Gym99*	UCF101 [40]	ActivityNet v1.2 [11]
1	35.46	85.0	82.0
3	61.4	86.5	83.6
5	70.8	**86.7**	**84.6**
7	74.4	86.4	84.0
12	**78.82**	-	-

Table 5: Performances of TSN when varying the number of sampled frames during training.

based ST-GCN struggles due to the challenges in skeleton estimation on gymnastics instances, as shown in Figure 5.

4.3. Temporal Action Localization

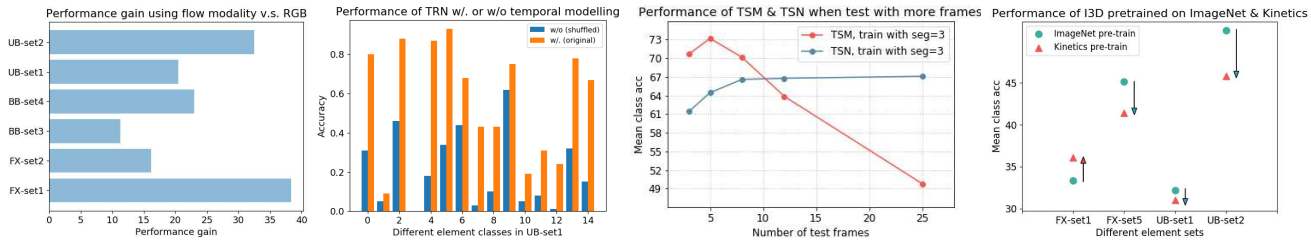
We also include an illustrative study for temporal action localization, as *FineGym* could support a wide range of tasks. Practically, temporal action localization could be conducted for event actions within video records or sub-actions within action instances, resulting in two sub-tasks. We select Structured Segment Network (SSN) [54] as the representative, relying on its open-sourced implementation. The results of SSN on these two tasks are listed in Table 4, where localizing sub-actions is shown to be much more challenging than localizing actions. While the boundaries of actions in a video record are more distinctive, identifying the boundaries of sub-actions may require a comprehensive understanding of the whole action.

4.4. Analysis

In this section, we enumerate the key messages we have observed in the conducted empirical studies.

Is sparse sampling sufficient for action recognition? The sparse sampling scheme has been widely adopted in action recognition, due to its high efficiency and promising accuracy demonstrated in various datasets [40, 11]. However, this trend does not hold for element-level action recognition in *FineGym*. Table 5 lists the results of TSN [44] on the subset *Gym99* as well as existing datasets, where we adjust the number of input frames. Compared to saturated results on existing datasets using only few frames, the result of TSN on *Gym99* steadily increases as the number of frames increases, and saturates at 12 frames which account for 30% of all frames. These results indicate that every frame counts in fine-grained action recognition on *FineGym*.

How important is temporal information? As shown in Figure 6a, motion features such as optical flows could capture frame-wise temporal dynamics, leading to better performance of TSN [44]. Many methods have also designed innovative modules for longer-term temporal modeling, such as TRN [55] and TSM [30]. To study them, for the temporal reasoning module in TRN, we shuffle the input frames during testing, and observe significant performance drops in Figure 6b, indicating temporal dynamics indeed play an important role in *FineGym*, and TRN could capture it. Moreover, for the temporal shifting module in TSM, we conduct a scheme where we start by training a TSM with 3 input frames, then gradually increase the number of frames during testing. Taking TSN for a comparison, Figure 6c includes the resulting curves, where the performance of TSM drops sharply when the number of testing frames is very different from that in training, and TSN maintains its performance as only temporal average pooling is applied in it. These results again verify that temporal dynamics is essential on *FineGym*, so that a very different number of frames leads to significantly different temporal dynamics. To summarize, optical flows could capture some extent of temporal



(a) (b) (c) (d)

Figure 6: (a) Per-class performances of TSN with motion and appearance features in 6 element categories. (b) Performances of TRN on the set *UB-circles* using ordered or shuffled testing frames. N (c) Mean-class accuracies of TSM and TSN on *Gym99* when trained with 3 frames and tested with more frames. (d) Per-class performances of I3D pre-trained on Kinetics and ImageNet in various element categories. Best viewed in high resolution.

dynamics, but not all. Fine-grained action recognition of motion-intense actions heavily relies on temporal dynamics modeling.

Does pre-training on large-scale video datasets help? Considering the number of parameters in 3D-CNN methods, *e.g*. I3D [3], usually they are pre-trained first on large-scale datasets, *e.g*. Kinetics, which indeed leads to a performance boost [3, 52]. For example, the Kinetics pre-trained I3D could promote the recognition accuracy from 84.5% to 97.9% on UCF101 [40]. However, on *FineGym*, such a pre-training scheme is not always helpful, as shown in Figure 6d. One potential reason is the large gaps in terms of temporal patterns between coarse- and fine-grained actions.

What can not be handled by current methods/modules? By carefully observing the confusion matrices [5], we summarize some points that are challenging for existing methods. (1) *Intense motion*, especially in different kinds of saltos (often finished within 1 second), as shown in the last several frames of Figure 5. (2) *Subtle spatial semantics*, which involves differences in body parts such as legs are whether bent or straight, and human-object relationships. (3) *Complex temporal dynamics*, such as the direction of motion, and the degree of rotation. (4) *Reasoning*, such as counting the times of saltos. We hope the high-resolution and professional data of *FineGym* could help future researches aiming for these points. In addition, *FineGym* poses higher requirements for methods that have an intermediate representation, *e.g*. human skeleton, which is hard to be estimated on *FineGym* due to the large diversity in human poses. See Figure 5 for a demonstration.

5. Potential Applications and Discussion

While more types of annotations will be added subsequently, the high-quality data of *FineGym* has offered a foundation for various applications, besides coarse- and fine-grained action recognition and localization, including but not limited to (1) *auto-scoring*, where difficult scores are given for each element category in the official documents, and we could also estimate the quality scores based on visual information, resulting in a gymnastic auto-scoring framework. (2) *Action generation*, where the consistent background context of fine-grained sub-actions could help generative models focus more on the action themselves, and the standard and diverse instances in *FineGym* could facilitate exploration. (3) *Multi-attribute prediction*, for which the attribute ground-truths of the element categories are immediately ready due to the use of decision trees. (4) *Model interpretation and reasoning*, which could benefit from the manually built decision trees, as shown in Figure 3.

FineGym may be used to conduct more empirical studies on model designs, such as *how to strike a balance between accuracy and efficiency when dealing with highly informative yet subtly different actions?* And *how to model the complex temporal dynamics efficiently, effectively and robustly?*

6. Conclusion

In this paper, we propose *FineGym*, a dataset focusing on gymnastic videos. *FineGym* differs from existing action recognition datasets in multiple aspects, including the high-quality and action-centric data, the consistent annotations across multiple granularities both semantically and temporally, as well as the diverse and informative action instances. On top of *FineGym*, we have empirically investigated representative methods at various levels. These studies not only lead to a number of attractive findings that are beneficial for future research, but also clearly show new challenges posed by *FineGym*. We hope these efforts could facilitate new advances in the field of action understanding.

Acknowledgements. We sincerely thank the outstanding annotation team for their excellent work. This work is partially supported by SenseTime Collaborative Grant on Large-scale Multi-modality Analysis and the General Research Funds (GRF) of Hong Kong (No. 14203518 and No. 14205719).

[5] See supplementary material for examples.

References

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

[2] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *ECCV*, 2018. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 2, 3, 6, 8

[4] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, pages 3218–3226, 2015. 3

[5] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033, 2018. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2

[7] Fdration Internationale de Gymnastique (FIG). *2017 2020 CODE OF POINTS for Womens Artistic Gymnastics*, 2017 (accessed October 28, 2019). 3, 6

[8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 2017. 3

[9] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *CVPR*, pages 6508–6516, 2018. 3

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 3

[11] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 2, 7

[12] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 7

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 3

[14] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795, 2013. 3

[15] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. 2014. 2

[16] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, June 2019. 3

[17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, pages 971–980, 2017. 6

[18] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015. 2

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, and et al. The something something video database for learning and evaluating visual common sense. *ICCV*, Oct 2017. 2, 3, 5

[20] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 2

[21] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *ICCV*, pages 5822–5831, 2017. 3

[22] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, Dec. 2013. 2

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, pages 780–787, 2014. 2

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 1, 2

[26] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pages 156–165, 2017. 3

[27] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 7

[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, pages 600–615, 2018. 3

[29] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018. 2

[30] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 3, 6, 7

[31] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, pages 437–453. Springer, 2016. 2

[32] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. 2

[33] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2

[34] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016. 2, 3

[35] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36. IEEE, 2004. 2

[36] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, pages 200–216, 2018. 3

[37] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *CVPR*, 2020. 1

[38] 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. Springer, 2016. 2

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 1, 3

[40] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 1, 2, 7, 8

[41] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *CVPR*, 2019. 3

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

[43] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 3

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 2, 3, 5, 6, 7

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3, 6

[46] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 2

[47] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5783–5792, 2017. 3

[48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3, 6

[49] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *the CVPR*, 2020. 1

[50] 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. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 2

[51] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 7

[52] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*, 2019. 2, 8

[53] Yue Zhao, Haodong Duan, Yuanjun Xiong, and Dahua Lin. MMAction. https://github.com/open-mmlab/mmaction, 2019. 5

[54] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 3, 7

[55] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. *ECCV*, 2018. 1, 3, 6, 7