FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Action Understanding

Jinglin Xu[1] Guohao Zhao[2] Sibo Yin[2] Wenhao Zhou[1] Yuxin Peng[2*]
[1] School of Intelligence Science and Technology, University of Science and Technology Beijing
[2] Wangxuan Institute of Computer Technology, Peking University
xujinglinlove@gmail.com; ssee7235@gmail.com; 2000012982@stu.pku.edu.cn;
m202320876@xs.ustb.edu.cn; pengyuxin@pku.edu.cn

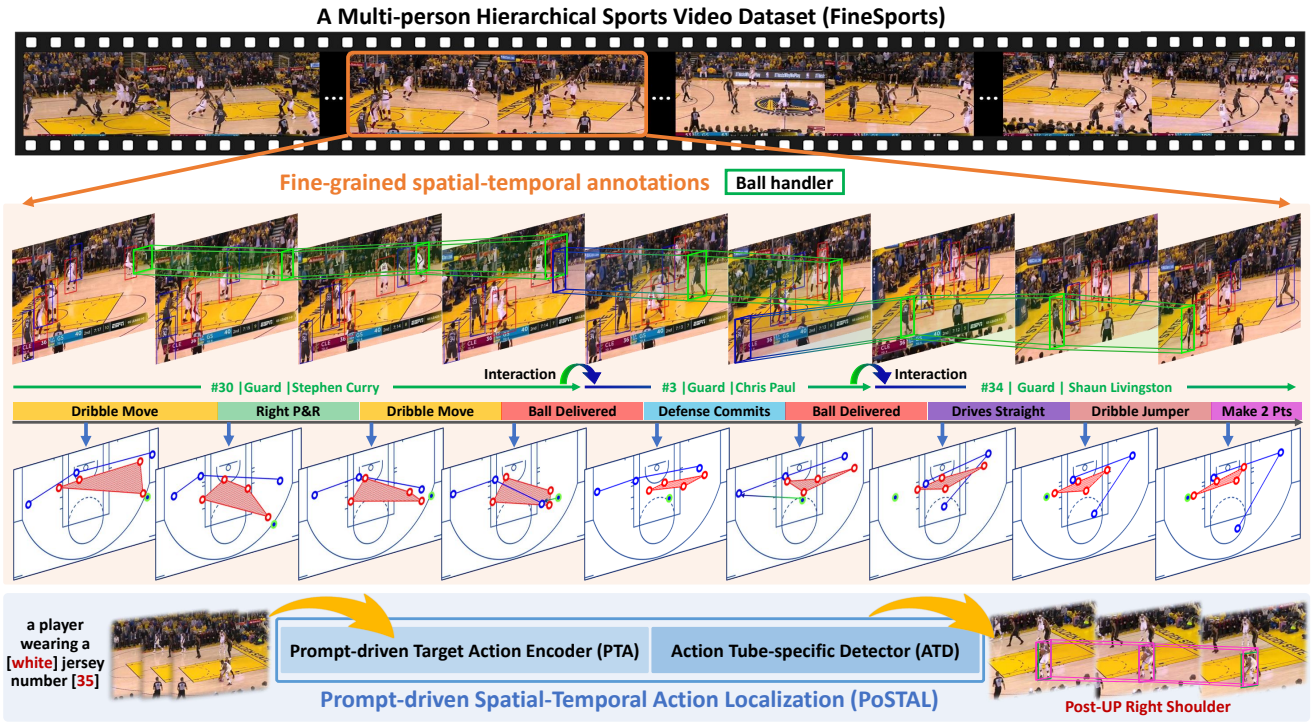

Figure 1. An overview of the ***FineSports*** dataset and new prompt-driven spatial-temporal action location approach, ***PoSTAL***. FineSports is a multi-person basketball sports video dataset with high-quality fine-grained annotations on action procedures covering target players' movements and multiple players' interactions. It provides the potential for capturing target player movement in 2D and proposing PoSTAL with better sports analysis ability via constructing a prompt-driven target action encoder and an action tube-specific detector.

Abstract

Fine-grained action analysis in multi-person sports is complex due to athletes' quick movements and intense physical confrontations, which result in severe visual obstructions in most scenes. In addition, accessible multi-person sports video datasets lack fine-grained action annotations in both space and time, adding to the difficulty in fine-grained action analysis. To this end, we construct a new multi-person basketball sports video dataset named ***FineSports****, which contains fine-grained semantic and spatial-temporal annotations on 10,000 NBA game videos, covering 52 fine-grained action types, 16,000 action instances, and 123,000 spatial-temporal bounding boxes. We also propose a new prompt-driven spatial-temporal action location approach called* ***PoSTAL****, composed of a prompt-driven target action encoder (****PTA****) and an action tube-specific detector (****ATD****) to directly generate target action tubes with fine-grained action types without any off-line proposal generation. Extensive experiments on the FineSports dataset demonstrate that PoSTAL outperforms state-of-the-art methods. Data and code are available at* https://github.com/PKU-ICST-MIPL/FineSports_CVPR2024.

*Corresponding author.

1. Introduction

In recent years, human action understanding has emerged as a hot research topic that involves interdisciplinary collaborations in computer science, human kinetics, and behavioral science. It has various applications in practical scenarios, such as autonomous driving [7, 39], abnormal monitoring [34, 40], and sports analysis [14, 41]. With advances in computer vision techniques, deep learning-based video understanding approaches [8, 13, 21, 27, 35] have achieved remarkable performance on various human action understanding tasks. While these approaches are data-driven, most publicly accessible action video datasets (e.g., ActivityNet [6], Kinetics [3], and AVA [12]) generally lack high-quality fine-grained annotations, which leaves fine-grained action analysis of spatial, temporal, and semantic relationships difficult and severely hinders the development of spatial-temporal models for fine-grained action understanding.

Compared to understanding other activities, fine-grained action analysis in videos of team sports, such as basketball, volleyball, and football, is more challenging due to their *chaotic* nature. For example, in an NBA game, players often move quickly and unpredictably in offensive and defensive situations, and frequently gather in a tiny area or even "pile up"— over the top of each other. These typical complicated behaviors exacerbate the issues of *motion blur* and *occlusion* in action understanding. More precisely, multi-person sports are characterized by (1) subjects with dynamic relationships, e.g., ball handler and other players, (2) rapid changes, e.g., in offensive and defensive situations, and (3) extreme body postures, e.g., in some "Backboard" or "P&R" scenarios. All these make fine-grained action analysis more challenging. In addition, many non-players are in the videos, including multiple referees on the court and hundreds of audiences in the background, leading to a lot of noise during fine-grained action analysis. Considering the above factors, we build a new multi-person sports video dataset with fine-grained annotations, named ***FineSports***, to support strongly and benchmark fine-grained action understanding research.

We collect 10,000 video sequences from the official NBA archive and employ three basketball association athletes to assist in the construction of the lexicon that guides the annotation work. Specifically, FineSports annotations comprehensively provide fine-grained action categories and spatial-temporal structures of videos. The former introduces two-level action types, covering 12 coarse-grained and 52 fine-grained action categories. The latter densely labels target players' spatial bounding boxes and temporal boundaries, as shown in Fig. 1. FineSports provides an exacting benchmark for human action understanding, which supports various tasks, such as fine-grained action recognition and spatial-temporal action localization.

Among downstream tasks of human action understanding, spatial-temporal action localization (STAL) is particularly required to perceive actions' semantic and spatial-temporal structures. Given a video containing the target player description, STAL aims to detect the target action tube by a sequence of bounding boxes in space and time, as well as the corresponding action class. In this work, we propose a new prompt-driven spatial-temporal action localization approach, named ***PoSTAL***, which consists of two core components: (1) a prompt-driven target action encoder (***PTA***) and (2) an action tube-specific detector (***ATD***). PoSTAL applies the PTA module to extract target action features guided by descriptive words and then designs the ATD module for obtaining a set of target action tubes and associated action classes simultaneously. Conducting extensive experiments on the proposed FineSports dataset, our PoSTAL outperforms the state-of-the-art methods, demonstrating the usefulness of FineSports and the effectiveness of PoSTAL.

The contributions of this paper can be summarized as (1) We build a new multi-person sports video dataset with fine-grained annotations, named ***FineSports***. It contains 10,000 basketball game videos, covering 12 action types and 52 sub-action types, providing 123,014 spatial bounding boxes and 32,096 temporal boundaries of associated fine-grained sub-actions. (2) We propose a new prompt-driven spatial-temporal action location (STAL) approach, named ***PoSTAL***, for the task of spatial-temporal action localization in the multi-person scenario. (3) Extensive experiments demonstrate the usefulness of FineSports and the effectiveness of PoSTAL on the STAL task.

2. Related Work

Fine-grained Sports Video Datasets. As shown in Tab. 1, existing sports video datasets providing fine-grained annotations can be roughly divided into four categories based on different action understanding tasks: recognition, localization, detection, and assessment. In action recognition, Diving48 [1] built a diving dataset annotated by combining four components (i.e., back, somersault, twist, and free). FineGym [32] was a hierarchical dataset with temporally and semantically coarse-to-fine annotations. FSD-10 [24] constructed a figure skating dataset for fine-grained sports content analysis. In temporal action localization, TAPOS [33] constructed an Olympics sports video dataset with sub-action annotations for studying temporal action parsing. MCFS [25] introduced a fine-grained dataset annotated by three semantic levels for the temporal action segmentation task. In spatial-temporal action localization (detection), MultiSports [23] developed a large-scale fine-grained dataset with dense annotations in both space and time for spatial-temporal action detection. In action quality assessment, FP-Basket [2] was a first-person basketball dataset for estimating the performance assessment of basketball players. MTL-AQA [29] assessed action quality via constructing multi-task networks. FineDiving [41] was a

Dataset	# Sam	# Act	# Sub	# Bb	# Ins	Task
Volleyball [16] (CVPR'16)	15	6	7	-	-	R
FP-Basket.[2] (ICCV'17)	500	3	-	-	-	A
Diving48 [1] (ECCV'18)	18404	4	48	-	-	R
MLB-YT [31] (CVPRW'18)	6418	-	9	-	-	R
GolfDB [28] (CVPRW'19)	1400	-	8	-	-	L
Fis-V [38] (CSVT'19)	500	13	-	-	-	A
MTL-AQA [29] (CVPR'19)	1412	16	-	-	-	R/A
NBA [42] (ECCV'20)	9172	-	9	-	-	R
FineGym [32] (CVPR'20)	4883	15	530	-	32k	R
TAPOS [33] (CVPR'20)	16294	-	21	-	-	L
FSD-10 [24] (arXiv'20)	1484	3	10	-	-	R
MultiSports^{A} [23] (ICCV'21)	800	-	21	325k	8.7k	D/L/R
MultiSports^{V} [23] (ICCV'21)	800	-	12	193k	7.6k	D/L/R
MultiSports^{F} [23] (ICCV'21)	800	-	15	225k	12k	D/L/R
MultiSports^{B} [23] (ICCV'21)	800	-	18	213k	9k	D/L/R
MCFS [25] (AAAI'21)	271	22	130	-	-	L
FineDiving [41] (CVPR'22)	3000	52	29	-	10k	L/A
LOGO [44] (CVPR'23)	200	-	12	-	-	L/A
SportsMOT [4] (ICCV'23)	240	-	-	1.6M	-	T
FineFS [17] (MM'22)	1167	3	16	-	-	L/R/A
RFSJ [26] (MM'22)	1304	-	23	-	-	R/A
FineSports (Ours)	**10000**	**12**	**52**	**123k**	**16k**	**D/L/R**

Table 1. Comparison of representative fine-grained sports video datasets. # Sam, # Act, # Sub, and # Bb indicate the numbers of samples, action types, sub-action types, and bounding boxes, respectively. # Ins denotes the number of fine-grained action instances. R, L, T, D, and A indicate action recognition, temporal action localization, tracking, spatial-temporal action localization, and action quality assessment tasks. The superscripts A, V, F, and B indicate four sports: Aerobic gymnastics, Volleyball, Football, and Basketball.

diving dataset with fine-grained annotations of action procedures. LOGO [44] was a multi-person long-form video dataset based on artistic swimming competitions with detailed annotations on action and formation. FineFS [17] was a large-scale fine-grained figure skating dataset involving RGB videos and estimated skeleton sequences. RFSJ [26] provided a figure skating jumping dataset with replay information and fine-grained annotations. Compared to previous efforts, our FineSports shows the advantages in the (1) scale of samples, (2) variety and hierarchy of action types, and (3) number of fine-grained instances, primarily supporting the spatial-temporal action localization task.

Spatio-temporal Action Localization. Emerging as a pivotal area of research in action understanding, spatial-temporal localization [17, 22, 26, 43, 45] focuses on identifying both where (spatial) and when (temporal) the target actions occur within video sequences at the frame or video level. Previous efforts often processed videos per frame and predicted bounding boxes for each frame, which were then concatenated as the final result. Despite the intuition of this paradigm, frame-level localization [10, 30] primarily focused on the spatial information intra-frame through utilizing 2D CNN-based detection networks and region proposal networks to obtain detection results. With recent advancements in 3D CNN-based video understanding, video-level localization employing a backbone, such as I3D [3], CSN [36], and Video-SwinT [27], has become the mainstream paradigm [11, 17, 20, 22, 45]. Approaches under this paradigm not only capture spatial information within individual frames but also model the temporal dynamics between frames. For example, YOWO [20] presented a two-branch framework with 3D-ResNext [13] followed by channel fusion and attention mechanism to enhance spatial-temporal feature aggregation. LUSD-NET [17] fused spatial-temporal features with original sequence features to enhance perceiving long sequences and further applied it to more tasks of fine-grained action understanding. MOC [22] proposed an action tubelet detection framework, which considers action as a track of moving points and detects the action's center, movement, and bounding box, respectively, via a tri-branch structure. TubeR [45] utilized CSN [36] to encode the spatial-temporal features of videos and employed a set of tubelet queries to get spatio-temporal localization results. Unlike the above methods, our PoSTAL follows the video-level paradigm and utilizes textual prompts to explicitly guide learning the target action's spatial-temporal features, significantly distinguished from previous methods.

3. The FineSports Dataset

This section introduces a new multi-person basketball sports video dataset, ***FineSports***, from its construction, statistics, and characteristics.

3.1. Dataset Construction

Collection. Considering the popularity and scene complexity, we collect real professional game videos from the NBA official replay archive. FineSports contains 10,000 videos while retaining only overhead shots from game courts and filtering out other extreme situations, e.g., overexposure, slow playbacks, and unrelated camera switches. Although each action type in the NBA game has typical movement patterns and player formations, we refine the procedural steps of players' actions via a series of sub-action types. Every game drive, shoot, and defense is meticulously labeled.

Lexicon. According to the FIBA rule, we define a rigorous lexicon for fine-grained annotations. Three basketball association athletes with expert-level knowledge are employed to guarantee its precision. After iterative revisions, the resulting lexicon ensures FineSports annotations maintain accurate spatial-temporal boundaries of actions in practice and are suitable for the spatial-temporal action localization task. Specifically, it is organized by fine-grained semantic and spatial-temporal structures, each containing two-level annotations, as shown in Fig. 2 and Fig. 3.

For semantic structure in Fig. 2, the action-level labels describe the coarse-grained action types of players in a valid action procedure, covering twelve categories: *Drive*, *Drib-*

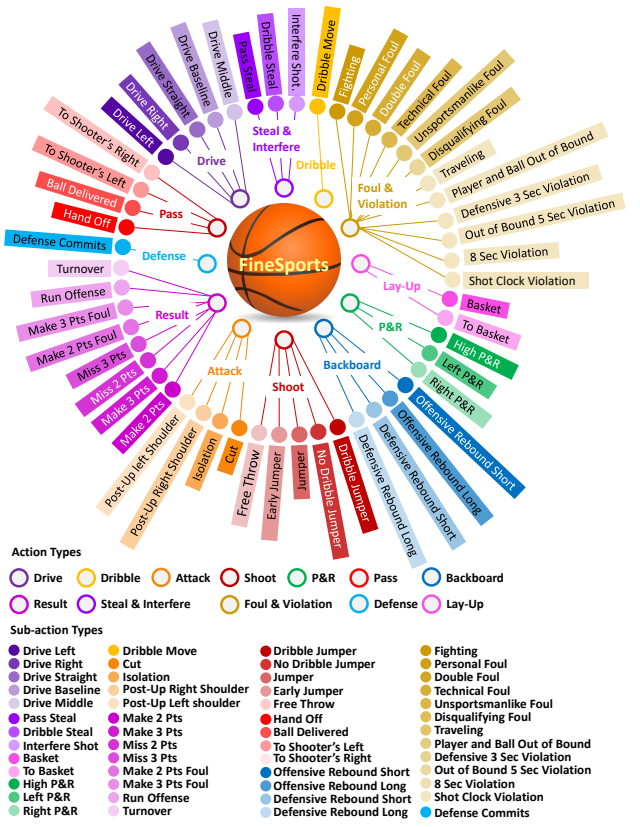

Figure 2. The two-level semantic structure of FineSports. An action type denotes the coarse-grained action categories the player performs while being the ball handler, covering 12 categories. A sub-action type is the component of an action type, with 52 categories explicitly describing the procedural steps involved in each target player's (ball handler) activity. Different sub-action types refine the corresponding action type within each color branch.

ble, *Attack*, *Shoot*, *Pick&Roll* (i.e., *P&R*), *Pass*, *Backboard*, *Defense*, *Lay-Up*, *Steal&Interfere*, and *Foul&Violation*. The step-level labels depict the fine-grained sub-action types of procedural steps during the action procedure, where adjacent steps within the action procedure belong to different sub-action types. New group-level plays can be generated through the combination of sub-action types. For instance, the action type *Drive* is divided into five sub-action types: *Drive Left*, *Drive Right*, *Drive Straight*, *Drive Baseline*, and *Drive Middle*. During the *Drive* situation, the ball-handle player executing the step belonging to the sub-action type *Drive Straight* (or *Drive Right*) will directly affect the sub-action type of the subsequent step *Dribble Jumper* (or *To Basket*). *Dribble Jumper* is the sub-action type belonging to the action type *Shoot*, while *To Basket* is the sub-action type belonging to the action type *Lay-Up*. It can be seen that executed steps belonging to different sub-action types may lead to different plays.

For spatial-temporal structure in Fig. 3, we provide the spatial location and temporal boundary of sub-action types within the action procedure performed by each ball handler. In the spatial dimension, the step-level labels are the bounding boxes of all players within each frame, including ball handlers, teammates, and opponents, which can be used to identify critical sub-actions and develop winning strategies. In the temporal dimension, the step-level labels are the beginning and end frames of sub-actions in the action procedure, which parses its internal structure and aligns fine-grained action semantics. As shown in Fig. 3, the switched frames of sub-actions (i.e., *High P&R*, *Defense Commits*, *Ball Delivered*, *No Dribble Jumper*, and *Miss 2Pts*) are the 47th, 49th, 56th, and 64th frames, respectively.

Annotation. The annotation process contains two stages:

(1) **Spatial annotations.** We adopt MixSort-OC [4] to track all players in each video to obtain their bounding boxes on the court, where each player has a unique ID throughout the video. If all tracking results are correct, we select the last frame and manually mark the ball handler. For example, *#23 LeBron James* is the ball handler, denoted as *#23 James*. Our annotation system can automatically identify the bounding boxes of #23 James among all frames and store them as his spatial annotations. If the tracking results of *#23 James* are incorrect after a certain frame, we need to find this frame (e.g., the 37th frame) and repeat the above operations to store the correct spatial annotations of *#23 James* before the 37th frame. Given the correct ID of *#23 James* is 0 but was incorrectly tracked as ID 3 from the 38th frame to the end, we select the last frame and manually mark *#23 James*, whose ID is 3, as the ball handler. Our annotation system can automatically identify the bounding boxes of *#23 James* from the 38th frame to the end and store these as his spatial annotations. We divide all tracking results into the ball handler (the target player) and other players. The spatial annotations of other players are processed and stored in the same way. With the help of MixSort-OC, crowdsourced annotators can adjust bounding boxes of tracking results of all players at each frame.

(2) **Temporal annotations.** Based on our lexicon, we annotate the temporal boundaries of valid action segments of ball handlers in each video. For instance, in Fig. 3, we first annotate the temporal boundary of the valid action segment from the beginning of the *High P&R* step to the end of the *No Dribble Jumper* step. In this action segment, we annotate the switched frames of sub-actions *High P&R*, *Defense Commits*, *Ball Delivered*, and *No Dribble Jumper* in sequence. The sub-action *Miss 2Pts* belongs to the action type *Result*, i.e., the result of performing the action procedure.

3.2. Dataset Statistics

The FineSports dataset consists of 10,000 video samples, covering 12 action types and 52 sub-action types. The average video duration is 11.74 seconds. The distribution of fine-grained sub-action types can be found in Fig. 4 and more

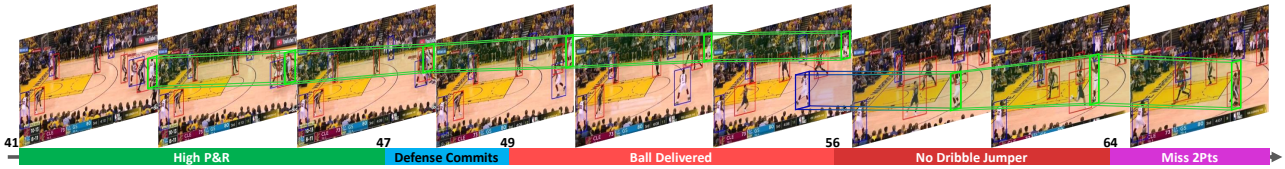

Figure 3. The spatial-temporal structure of fine-grained action types of the target players (green bounding boxes).

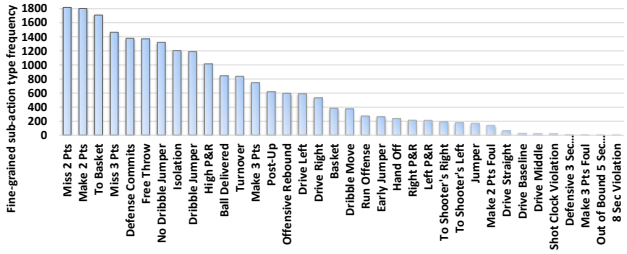

Figure 4. Statistic of FineSports. The distribution of fine-grained sub-action types.

detailed information on FineSports is reported in Tab. 1.

3.3. Dataset Characteristics

Tab. 1 compares FineSports with recent fine-grained sports video datasets. Specifically, compared to representative fine-grained sports video datasets for action recognition (e.g., FineGym [32] and NBA [42]), temporal action localization (e.g., GolfDB [28] and TAPOS [33]), and action quality assessment (e.g., Diving48 [1] and FineDiving [41]), our FineSports contains frame-level spatial annotations of 52 fine-grained sub-action types and lay the groundwork for achieving more challenging spatial-temporal action detection task. Compared to well-known MultiSports [23], our FineSports has a larger data scale and more refined sub-action types than each sport (i.e., Aerobic gym., Volleyball, Football, and Basketball) in MultiSports. For example, the sport basketball in MultiSports has 800 video samples covering 18 fine-grained categories, smaller than 10000 video samples spanning 52 fine-grained sub-action types in FineSports. We also compare FineSports and SportsMOT [4], where the former contains more fine-grained action annotations while the latter contains tracking results of multiple objects. Thanks to MixSort-OC proposed in SportsMOT, it helps us design a practical toolbox to improve efficiency during fine-grained annotation. FineSports is a larger and finer sports video dataset, enhancing the development of fine-grained analysis techniques in team sports and group activities.

4. The Proposed Approach: PoSTAL

This section systematically presents our approach for spatio-temporal action localization in the multi-person scenario. The main idea is to construct a new prompt-driven spatio-temporal action localization approach, named ***PoSTAL***, which formulates each target player's action procedure as a series of tubes with consistent semantics and correspondences in space and time. The overall architecture of our approach is illustrated in Fig. 5.

4.1. Problem Formulation

Our PoSTAL formulates the spatial-temporal action localization task as a multi-task learning problem that inputs a video snippet and outputs the action tube of each target player and associated fine-grained action class. The action tube consists of a sequence of bounding boxes of the target player, including the spatial locations and temporal boundaries of the target action being executed. Concretely, given a video snippet $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times 3}$ and a pair of descriptive words (i.e., Color and Number) based on the appearance characteristics of the target player, PoSTAL first applies a prompt-driven target action encoder (PTA, denoted as $\mathcal{P}$) to extract target action features guided by descriptive words and then designs an action tube-specific detector (ATD, denoted as $\mathcal{D}$) to simultaneously obtain a set of target action tube $\hat{\mathbf{Y}}$ and associated fine-grained action class $\hat{y}$, which can be presented as:

$$\hat{\mathbf{Y}}, \hat{y} = \mathcal{D}(\mathcal{P}(\mathbf{X}, \textit{text})) \tag{1}$$

where *text* is such a description "a player wearing a [Color] jersey number [Number]". $\hat{\mathbf{Y}} \in \mathbb{R}^{NT' \times 4}$ is the coordinates of N tubes and each tube is across T' frames. $\hat{y} \in \mathbb{R}^{N \times K}$ denotes predicted action classes for N tubes, belonging to K fine-grained action categories. $\mathcal{P}$ can be seen as a spatial-temporal vision-language module meticulously for learning prompt-driven target action features in videos and serving for subsequent action tube detection. $\mathcal{D}$ is designed as a multi-task learning module for simultaneously achieving target action tube regression and fine-grained classification.

4.2. Prompt-driven Spatial-Temporal Action Localization (PoSTAL)

PoSTAL comprises two core components: a prompt-driven target action encoder (PTA) and an action tube-specific detector (ATD).

Prompt-driven Target Action Encoder (PTA). We design the PTA module $\mathcal{P}$ by a spatial-temporal vision-language cross-attention to learn the target action representations precisely guided by the appearance characteristics of the target player and the associated fine-grained sub-action type.

We first encode the *text* and a fine-grained sub-action type of the target athlete in the prompt embedding space. After-

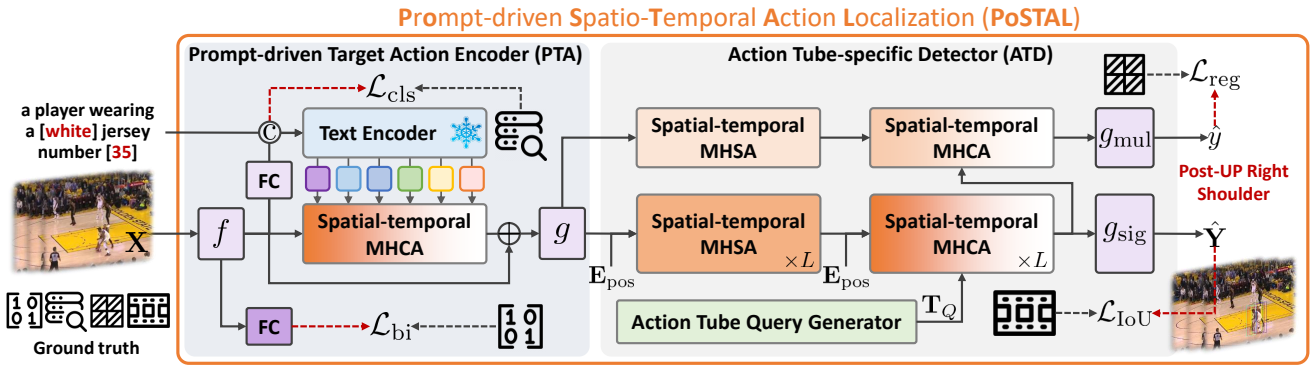

Figure 5. The architecture of the proposed PoSTAL method. It inputs a video sequence with descriptive words of the target action and outputs the target action tubes with fine-grained action types.

ward, we integrate the prompt embedding with video features to learn prompt-driven target action representations via a spatial-temporal vision-language cross-attention. Notable, the fine-grained sub-action type in the prompt embedding space is encoded by video features, not the ground-truth label. It can be presented as $\mathbf{B}=\mathcal{B}(text,\tilde{y})$ where $\mathbf{B}\in\mathbb{R}^{L_{\text{tx}}\times C'}$ indicates the prompt embedding, $\mathcal{B}$ stands for the BERT [5] model followed by a projection function, and $\tilde{y}=\text{FC}(f(\mathbf{X}))$ is the predicted fine-grained sub-action type in the prompt embedding space by using a channel-separated convolutional network f followed by a fully connected layer (FC), and $f(\mathbf{X})\in\mathbb{R}^{T'\times H'\times W'\times C'}$. Besides, we apply another FC to determine whether the current video snippet contains the bounding box of the target action, that is, $\tilde{y}_{\text{bi}}=\text{FC}_{\text{bi}}(f(\mathbf{X}))$.

Given the prompt embedding $\mathbf{B}$ and video features $f(\mathbf{X})$, we implement a multi-head cross-attention (MHCA) model along the spatial and temporal dimensions, called spatial-temporal MHCA, where the *query*, *key*, and *value* are as $\mathbf{Q}=\mathbf{W}_Q f(\mathbf{X})$, $\mathbf{K}=\mathbf{W}_K\mathbf{B}$, and $\mathbf{V}=\mathbf{W}_V\mathbf{B}$. The *value* is aggregated with spatial-temporal cross-attention $\mathbf{A}_P^S$ to generate prompt-driven target action representation $\mathbf{X}_P$, achieving the prompt-driven target action encoding. This process can be formulated as

$$\mathbf{A}_P^S=\text{softmax}(\mathbf{Q}\otimes\mathbf{K}^\top/\sqrt{C'/H}), \tag{2a}$$
$$\mathbf{X}_P=g(\mathbf{X}'_P+f(\mathbf{X})),\ \mathbf{X}'_P=\mathbf{A}_P^S\otimes\mathbf{V} \tag{2b}$$

where $\mathbf{X}_P\in\mathbb{R}^{T'\times H'\times W'\times C_\mathcal{T}}$, g is a projection function changed the channel dimension from C' to $C_\mathcal{T}$, and H is the number of attention heads.

Action Tube-specific Detector (ATD). After obtaining prompt-driven target action representation $\mathbf{X}_P$, we design the ATD module for predicting the target action's spatial locations, temporal boundaries, and fine-grained sub-action type belonging. ATD is composed of two core blocks: a single-level action tube-specific transformer $\mathcal{T}_{\text{sig}}=\{\mathcal{E}_{\text{sig}},\mathcal{D}_{\text{sig}}\}$ and a multi-level action tube-specific transformer $\mathcal{T}_{\text{mul}}=\{\mathcal{E}_{\text{mul}},\mathcal{D}_{\text{mul}}\}$, where $\mathcal{T}_{\text{sig}}$ is to localize each target action in space and time, and $\mathcal{T}_{\text{mul}}$ is to recognize its fine-grained category, and $\mathcal{T}_{\text{sig}}$ is the basis of $\mathcal{T}_{\text{mul}}$. We feed $\mathbf{X}_P$ into $\mathcal{T}_{\text{sig}}$ and $\mathcal{T}_{\text{mul}}$ in parallel.

In the single-level branch, $\mathbf{X}_P$ is enhanced by the spatial-temporal 3D-aware position embedding $\mathbf{E}_{\text{pos}}$ before inputting it into the action tube-specific encoder $\mathcal{E}_{\text{sig}}$. The result of $\mathcal{E}_{\text{sig}}$ is further enhanced by the position embedding $\mathbf{E}_{\text{pos}}$ and then serves as the input to the decoder $\mathcal{D}_{\text{sig}}$, where $\mathcal{E}_{\text{sig}}$ contains L spatial-temporal MHSA layers and $\mathcal{D}_{\text{sig}}$ contains L spatial-temporal MHCA layers. This process can be written as

$$\mathbf{X}_P^E=\mathcal{E}_{\text{sig}}(\mathbf{X}_P+\mathbf{E}_{\text{pos}}), \tag{3a}$$
$$\mathbf{X}_P^D=\mathcal{D}_{\text{sig}}(\mathbf{T}_Q,\mathbf{X}_P^E+\mathbf{E}_{\text{pos}}) \tag{3b}$$

where $\mathbf{X}_P^E\in\mathbb{R}^{T'H'W'\times C_\mathcal{T}}$, $\mathbf{X}_P^D\in\mathbb{R}^{L\times NT\times C_\mathcal{T}}$ stores L outputs of L MHCA layers for feeding into the multi-level branch, and $\mathbf{T}_Q\in\mathbb{R}^{N\times T\times C_\mathcal{T}}$ is a set of learnable action tube queries approximated by 3D cuboids [15, 43, 45]. To localize the target action tube in space and time, we input $\mathbf{X}_P^D$ into an MLP block g_{sig} that contains three layers with two RELU non-linearity and outputs the bounding boxes of target action among T frames, i.e., $\hat{\mathbf{Y}}=g_{\text{sig}}(\mathbf{X}_P^D[-1])$, where $\mathbf{X}_P^D[-1]$ indicates the last output of $\mathbf{X}_P^D$ and $\hat{\mathbf{Y}}\in\mathbb{R}^{NT\times 4}$.

In the multi-level branch, $\mathbf{X}_P$ is processed by a `repeat` operation to ensure that the dimensions of the query and key are matched in the attention model, denoted as $\tilde{\mathbf{X}}_P$. $\mathcal{T}_{\text{mul}}$ consists of a spatial-temporal MHSA layer ($\mathcal{E}_{\text{mul}}$) and a spatial-temporal MHCA layer ($\mathcal{D}_{\text{mul}}$). Therefore, the fine-grained action recognition task can be formulated as

$$\tilde{\mathbf{X}}_P^D=\mathcal{D}_{\text{mul}}(\tilde{\mathbf{X}}_P^E,\mathbf{X}_P^D),\ \tilde{\mathbf{X}}_P^E=\mathcal{E}_{\text{mul}}(\tilde{\mathbf{X}}_P) \tag{4a}$$
$$\hat{y}=g_{\text{mul}}(\tilde{\mathbf{X}}_P^D) \tag{4b}$$

where $\tilde{\mathbf{X}}_P^E\in\mathbb{R}^{L\times T'H'W'\times C_\mathcal{T}}$, $\tilde{\mathbf{X}}_P^D\in\mathbb{R}^{L\times NT\times C_\mathcal{T}}$, g_{mul} is a fully connected layer for fine-grained action recognition. Predicting $\hat{y}$ is necessary since metrics can only be calculated for the action tube with correct classification.

4.3. Training and Inference

Training. Given a video snippet $\mathbf{X}$ and a pair of descriptive words *text*, the entire framework of PoSTAL is optimized by minimizing the loss function

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cls}}(y, \tilde{y}) + \lambda_2 \mathcal{L}_{\text{bi}}(y_{\text{bi}}, \tilde{y}_{\text{bi}}) + \lambda_3 \mathcal{L}_{\text{reg}}(y, \hat{y}) + \lambda_4 \mathcal{L}_{\text{iou}}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_5 \mathcal{L}_{\text{bbox}}(\mathbf{Y}, \hat{\mathbf{Y}}) \quad (5)$$

where $\mathbf{Y}$ and y denote the ground truth coordinate and fine-grained sub-action type, and $\hat{\mathbf{Y}}$ and $\hat{y}$ are corresponding predictions. $\tilde{y}$ represents the predicted embedding of fine-grained sub-action type in prompt embedding space. $\tilde{y}_{\text{bi}}$ is supervised by an indicator that outputs 1 if the current video snippet contains the target action bounding box, and otherwise, it outputs 0. $\mathcal{L}_{\text{bi}}$ is a binary cross entropy loss. $\mathcal{L}_{\text{cls}}$ computes a cross-entropy loss for fine-grained sub-action classification in the prompt embedding space. $\mathcal{L}_{\text{reg}}$ indicates the cross-entropy loss for the action tube with correct fine-grained sub-action. $\mathcal{L}_{\text{iou}}$ denotes the per-frame bounding box matching error. $\mathcal{L}_{\text{bbox}}$ calculates the ℓ_1 loss of the bounding box coordinate regression. λ_1, λ_2, λ_3, λ_4, and λ_5 are hyper-parameters to balance various losses. During training, we utilize *text* as input to participate in encoding prompt-driven target action representations in the prompt embedding space.

Inference. Given a test video snippet $\mathbf{X}$, we can get the predicted target action tubes and the corresponding fine-grained sub-action type by $\hat{\mathbf{Y}} = g_{\text{sig}}(\mathbf{X}_P^D[-1])$ and $\hat{y} = g_{\text{mul}}(\tilde{\mathbf{X}}_P^D)$. Concretely, we choose the center frame of $\hat{\mathbf{Y}}$ as the prediction of each frame, i.e., $\hat{\mathbf{Y}}_{\text{pred}} \in \mathbb{R}^{N \times 4}$. For a test video with T_{all} frames, we can obtain N action tubes among T_{all} frames, where each action tube is across K fine-grained action categories with a certain confidence. During inference, we utilize the visual features of each testing video to fill in the *text*'s descriptive words (i.e., Color and Number) to obtain prompt-driven target action representations.

5. Experiments

5.1. Evaluation Metrics

To evaluate state-of-the-art spatial-temporal action localization methods on the FineSports dataset, we followed recent STAL methods (e.g., MOC [22] and TubeR [45]) and adopted frame-level mean average precision (frame-mAP) [18] and video-level mean average precision (video-mAP) [37] at different thresholds of intersections over union (IoU). Frame-mAP@θ means that if the IoU between the detection and ground truth is greater than θ, then this detection is correct. For video-mAP, the IoU between videos is an average time across the per-frame IoU.

5.2. Implementation Details

In PoSTAL, we utilized the CSN-152 network pre-trained on the Kinetics-400 [19] dataset as the backbone of the video extractor. During training, the learning rate of the backbone was 1e-5, and that of other components was 1e-4, and the weight decay was 1e-4. AdamW was the optimizer. The cosine annealing strategy was adopted to adjust the learning rate, with the first two epochs as warmups. The batch size was 16. We adopted the frozen BLIP as the text encoder in the PTA module. In the ATD module, $\mathcal{E}$ and $\mathcal{D}$ contained eight spatial-temporal MHSA and MHCA layers (i.e., $L=8$), respectively. For $\mathbf{T}_Q$, we set N as 6 and utilized bipartite matching [9] to match 6 tubes with the ground truth to compute the matching loss. The hyper-parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ were set as $\{1, 1, 1, 2, 5\}$.

Method	Metrics			Year
	F@0.5	V@0.2	V@0.5	
MOC [22]	19.21	/	/	ECCV'20
TubeR [45]	19.48	28.91	17.76	CVPR'22
PoSTAL (Ours)	**21.54**	**31.18**	**24.31**	

Table 2. Quantitative comparison with the state-of-the-art methods on FineSports. F@0.5: frame-mAP with $\theta=0.5$. V@0.2: video-mAP with $\theta=0.2$. V@0.5: video-mAP with $\theta=0.5$.

5.3. Results and Analysis

Spatial-temporal Action Localization on FineSports. In Tab. 2, we compare our PoSTAL with the state-of-the-art (SOTA) methods under the metrics frame-mAP and video-mAP. The overall results are relatively lower, primarily due to the multi-person sports challenges in FineSports, which involve occlusion and more interactive action details, more akin to real sports scenarios. Even under such challenging conditions, our PoSTAL achieves state-of-the-art performance in spatial-temporal action localization. Compared with MOC [22], our PoSTAL exhibits 2.33% improvement on frame-mAP since PoSTAL introduces action tube queries to represent target actions and produce their bounding boxes, achieving greater accuracy and effectiveness. Compared with TubeR [45], our PoSTAL outperforms TubeR by 2.06% on frame-mAP since incorporating learnable prompts in PTA can guide the generation of bounding boxes more accurately and enhance the frame-level result.

Ablations. We conducted a series of ablation experiments to demonstrate the effectiveness of PTA and ATD.

(1) **Effects of different PTA settings.** PoSTAL utilizes target players' descriptive words (i.e., Color and Number) and employs a classification network to predict the fine-grained action categories in PTA, enabling the model to better focus on the action regions of target players within each video frame. Besides, PoSTAL sets word embeddings used in PTA as a set of learnable parameters during training. We report the results of using target players' descriptive words and learnable word embeddings in Tab. 3. We observe that without descriptive words, the performance of PoSTAL

Figure 6. Visualization of PoSTAL. The left indicates the fine-grained sub-action types for each target player. The magenta target action tubes are the predictions, while the green bounding boxes are the ground truth. The right denotes the descriptive words of each target player.

PTA Settings	Metrics		
	F@0.5	V@0.2	V@0.5
w/o Descriptive Words	20.67	24.69	12.44
w/o Learnable Embeddings	19.45	31.37	18.19
PTA (Ours)	**21.54**	**31.18**	**24.31**

Table 3. Effects of different PTA settings.

# Tube Query (N)	Metrics		
	F@0.5	V@0.2	V@0.5
2	20.16	32.72	21.34
6	**21.54**	**31.18**	**24.31**
10	21.79	31.41	23.46

Table 4. Effects of different numbers of tube queries.

drops 0.87% on FineSports, while without learnable word embeddings, the performance drops 2.09%. The descriptive words of each target player can help the model precisely focus on the target action region while setting the word embeddings as learnable, which allows the model to learn instructive prompts for target actions dynamically.

(2) **Effects of different numbers of tube queries.** To study the benefit of learnable tube queries, we set different numbers of action tube queries (N) as 2, 6, and 10 in the ATD module. The results are summarized in Tab. 4. The performance of PoSTAL can achieve a better balance among three metrics when $N=6$, while $N=2$ leads to 1.38% and 2.97% performance degradation on F@0.5 and V@0.5, respectively. $N=10$ leads to 0.85% performance degradation on V@0.5. In fact, V@0.5 is a more rigorous metric for evaluating STAL performance.

Visualization. Fig. 6 presents the visualization results of PoSTAL. The magenta tube represents the spatial-temporal action localization results of PoSTAL, and the green bounding box refers to the ground-truth bounding boxes. The NAB game scenarios are complicated, with multiple players wearing the same color jerseys, rapid movement, and occlusion, which present challenges to spatial-temporal action localization for the target player. PoSTAL can still obtain target action tubes across different fine-grained sub-action types.

6. Conclusion and Discussion

In this paper, we constructed a new multi-person sports video dataset, FineSports, which consists of 10,000 NBA game videos, covering 52 fine-grained sub-action types while providing fine-grained semantic and spatial-temporal annotations of target players. We also proposed an end-to-end spatial-temporal action localization approach, PoSTAL, which employs a prompt-driven target action encoder (PTA) and an action tube-specific detector (ATD) to obtain target action tubes with fine-grained semantics. We conducted extensive experiments on FineSports and observed that PoSTAL achieves state-of-the-art performance, demonstrating its usefulness and effectiveness.

Limitations. FineSports only contains NBA games, which needs to be generalized to more multi-person sports like baseball, football, and volleyball, meeting the need for fine-grained action understanding across various sports.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (62373043, 61925201, 62132001) and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Resound: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018. 2, 3, 5

[2] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *ICCV*, pages 2177–2185, 2017. 2, 3

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

[4] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023. 3, 4, 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[6] 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. 2

[7] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. *arXiv preprint arXiv:2403.00436*, 2024. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2

[9] Dobrik Georgiev and Pietro Lió. Neural bipartite matching. *arXiv preprint arXiv:2005.11304*, 2020. 7

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

[11] Alexey Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lučić, Cordelia Schmid, and Anurag Arnab. End-to-end spatio-temporal action localisation with video transformers. *arXiv preprint arXiv:2304.12160*, 2023. 3

[12] 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

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 2, 3

[14] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8(6), 2022. 2

[15] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *ICCV*, pages 5822–5831, 2017. 6

[16] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016. 3

[17] Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *ACM MM*, pages 8590–8597, 2023. 3

[18] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4405–4413, 2017. 7

[19] 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. 7

[20] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019. 3

[21] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *ICCV*, pages 1632–1643, 2023. 2

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

[23] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *ICCV*, pages 13536–13545, 2021. 2, 3, 5

[24] Shenlan Liu, Xiang Liu, Gao Huang, Lin Feng, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Hong Qiao. Fsd-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312*, 2020. 2, 3

[25] Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset. In *AAAI*, pages 2163–2171, 2021. 2, 3

[26] Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. A figure skating jumping dataset for replay-guided action quality assessment. In *ACM MM*, pages 2437–2445, 2023. 3

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2, 3

[28] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *CVPRW*, pages 2553–2562, 2019. 3, 5

[29] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *CVPR*, pages 304–313, 2019. 2, 3

[30] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, pages 744–759. Springer, 2016. 3

[31] AJ Piergiovanni and Michael S Ryoo. Fine-grained activity recognition in baseball videos. In *CVPRW*, pages 1740–1748, 2018. 3

[32] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 2, 3, 5

[33] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In *CVPR*, pages 730–739, 2020. 2, 3, 5

[34] Han Sun and Yu Chen. Real-time elderly monitoring for senior safety by lightweight human action recognition. In *ISMICT*, pages 1–6, 2022. 2

[35] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, pages 10078–10093, 2022. 2

[36] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561, 2019. 3

[37] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, pages 3164–3172, 2015. 7

[38] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *TCSVT*, 30(12):4578–4590, 2019. 3

[39] Feiyi Xu, Feng Xu, Jiucheng Xie, Chi-Man Pun, Huimin Lu, and Hao Gao. Action recognition framework in traffic scene for autonomous driving system. *TITS*, 23(11):22301–22311, 2022. 2

[40] Jinglin Xu, Guangyi Chen, Jiwen Lu, and Jie Zhou. Unintentional action localization via counterfactual examples. *TIP*, 31:3281–3294, 2022. 2

[41] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 2, 3, 5

[42] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *ECCV*, pages 208–224, 2020. 3, 5

[43] 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, 6

[44] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, pages 2405–2414, 2023. 3

[45] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *CVPR*, pages 13598–13607, 2022. 3, 6, 7