F^{3}Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos

Anonymous authors
Paper under double-blind review

Abstract

Analyzing Fast, Frequent, and Fine-grained (F^3) events presents a significant challenge in video analytics and multi-modal LLMs. Current methods struggle to identify events that satisfy all the F^3 criteria with high accuracy due to challenges such as motion blur and subtle visual discrepancies. To advance research in video understanding, we introduce F^{3}Set, a benchmark that consists of video datasets for precise F^3 event detection. Datasets in F^{3}Set are characterized by their extensive scale and comprehensive detail, usually encompassing over 1,000 event types with precise timestamps and supporting multi-level granularity. Currently F^{3}Set contains several sports datasets, and this framework may be extended to other applications as well. We evaluated popular temporal action understanding methods on F^{3}Set, revealing substantial challenges for existing techniques. Additionally, we propose a new method, F^{3}ED, for F^3 event detections, achieving superior performance. The dataset, model, and benchmark code are available at `https://github.com/F3Set/F3Set`.

1 Introduction

Recognizing sequences of fast (fast-paced), frequent (many actions in a short period), and fine-grained (diverse types) events with precise timestamps (with a tolerance of 1-2 frames) is a challenging problem for both current video analytics methods and multi-modal large language models (LLMs). Despite advancements in fine-grained action recognition [31; 53; 46], temporal action localization [55; 7; 39; 54], segmentation [62; 33; 66; 2], and video captioning [58; 51; 43; 36], limited focus has been directed towards this problem. This task is critical for various real-world applications, such as sports analytics, where action forecasting [20; 60], strategic and tactical analysis [10; 42], and player performance evaluation [11; 50] depend on a *detailed* understanding of event sequences. Other examples include industrial inspection [41], crucial for detecting subtle irregularities in high-speed production lines to ensure quality and safety; computer vision in autonomous driving [26], essential for accurate and instantaneous vehicle control and obstacle detection; and surveillance [48], important for the precise identification of abnormal or sudden events to enhance security. However, existing methods and datasets foundational to their development only *partially* address the F^3 scenario.

To facilitate the study of F^3 events understanding, we propose a new benchmark, F^{3}Set, for precise temporal events detection and recognition. F^{3}Set datasets usually have a large number of event types (on the order of 1,000), annotated with exact timestamps, and offer multi-level granularity to capture comprehensive event details. Although F^3 is a general problem, creating such a dataset requires domain-specific knowledge for labeling and processing, thus, in this paper, we use tennis as a case study. We also introduce a general annotation pipeline and toolchain to support domain experts in creating new F^3 datasets. Using this pipeline, we have also been building datasets for table tennis and badminton, and a community of users are actively expanding these with other applications.

Unlike other video analysis tasks, tennis actions are characterized by their rapid succession and diversity as illustrated in Figure 1. Understanding detailed event attributes like shot direction, technique, and outcome is essential. For instance, analyzing patterns in serve directions (e.g., “T”, “body”, “wide”, defined in Appendix B) or success rates can reveal a player’s habits and skill levels, offering strategic insights for competitive advantage. This detailed analysis supports coaches and players in developing tailored strategies against different opponents. However, detecting F^3 events from videos poses significant challenges, such as subtle visual differences, motion-induced blurring,

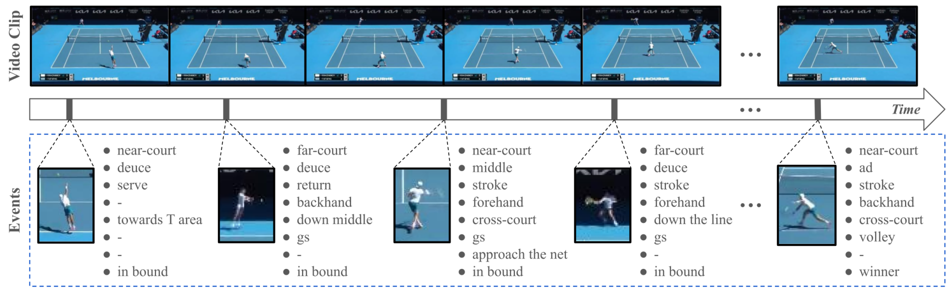

Figure 1: Example of detecting fast, frequent, and fine-grained events with precise moments.

and the need for precise event localization. Current video understanding methods are inadequately equipped to address these challenges. For instance, traditional fine-grained action recognition [53; 9] typically assigns a single label to an entire video rather than identifying a sequence of events. Temporal action localization (TAL) and temporal action segmentation (TAS) often depend on pre-trained or modestly fine-tuned input features [38; 15], which lack the specificity required to capture the subtle and domain-specific visual details necessary for recognizing diverse events with temporal precision. Some studies [25; 36] attempt to address these issues through *dense* frame sampling and end-to-end training. However, this makes targeted events temporally sparse (e.g., only a few events over hundreds of consecutive frames). As a result, long-term temporal correlation modules on dense visual features struggle to capture event-wise causal correlations effectively.

Moreover, Large Language Models (LLMs) [49; 56; 37] have expanded their capabilities to include multi-modal inference, encompassing text, visuals, and audio. Recognizing the potential, we conducted preliminary experiments on F^{3}Set using GPT-4 and observed that it understood basic video contexts, such as sports types, contextual information (e.g., court type and scoreboard), and simple actions. However, it struggles with understanding F^3 events and temporal relations between frames (e.g., shot directions). See Appendix A for details. Consequently, GPT-4 yields poor results compared to the other methods for F^3 problems, and we do not use it in the experiment. By introducing F^{3}Set, we hope it can help advance multi-modal LLM capabilities in F^3 video understanding in the future.

Leveraging F^{3}Set, we extensively evaluate existing temporal action understanding methods, aiming to reveal the challenges of F^3 event understanding. To provide guidelines for future research, we conduct a number of ablation studies on modeling choices. Addressing the shortcomings of existing methods, we also propose a simple yet efficient model, F^{3}ED, that is designed for F^3 event detection tasks. It outperforms existing models and can serve as a baseline for further development.

Contributions. The key contributions of this paper are as follows:

- We create F^{3}Set, a new benchmark with datasets that feature over 1,000 precisely timestamped event types with multi-level granularity, designed to challenge and advance the state-of-the-art in temporal action understanding.
- We introduce a general annotation toolchain that enables domain experts to create new F^3 datasets.
- We propose an end-to-end model named F^{3}ED, which can accurately detect F^3 event sequences from videos through visual features and contextual sequence refinement.
- We assess the performance of leading temporal action understanding methods on F^{3}Set through comprehensive evaluations and ablation studies and provide an analysis of the results.

2 Related Work

Temporal event understanding. F^{3}Set targets the detection of F^3 events with precise temporal annotations. This task is highly related to temporal action understanding tasks, which aims to recognize and locate the timestamp of actions that happen within an untrimmed video stream. It can be further categorized into sub-tasks such as temporal action localization (TAL) [55; 7; 39; 54], temporal action segmentation (TAS) [62; 33; 66; 2], and temporal action spotting (TASpot) [44; 21; 25] determined by the temporal duration and density of targeted events. Many TAL and TAS methods focus on coarse-grained and daily activities (e.g., standing up, pouring water) [28; 30], which may span from seconds to minutes. Typically, these methods extract features using video encoders pre-trained on action recognition datasets like Kinetics-400 [29], followed by training

Table 1: Comparison of existing F^3 related datasets and F^3Set. "Evt. Len." is the average duration of each event, and "# Evt. / sec" is the average number of events per second.

Datasets	# Vid.	# Clips.	Avg. Clip Len.	# Classes	Evt. Len.	# Evt. / sec
(a) Fine-grained						
FineAction [40]	-	16,732	149.5s	101	6.9s	0.3
ActivityNet [5]	-	19,994	116.7s	200	49.2s	0.01
FineGym [53]	303	32,697	50.3s	530	1.7s	0.3
(b) Fast						
CCTV-Pipe [41]	575	575	549.3s	16	< 0.1s	0.02
SoccerNetV2 [12]	9	9	99.6min	12	< 0.1s	0.3
(c) Fast & Frequent						
FineDiving [64]	135	3,000	4.2s	29	1.1s	$\sim$1
ShuttleSet [61]	44	3,685	10.9s	18	< 0.1s	$\sim$1
P^2ANet [3]	200	2,721	360.0s	14	< 0.1s	$\sim$2
(d) Fast & Frequent & Fine-grained						
F^3Set	114	11,584	8.4s	1,108	< 0.1s	$\sim$1

and evaluating a detection head with these features. However, this paradigm often struggles with fine-grained, domain-specific events. Even with end-to-end training, video is encoded into non-overlapping segments [4] or downsampled [39], yielding temporally coarse features inadequate for precise event spotting. TASpot addresses the issue of detecting the precise moment in time events occur through *dense* frame sampling and prediction [25; 22]. However, it struggles with challenges such as events are temporally sparse. As a result, long-term temporal correlation modules on dense visual features fail to capture event-wise causal correlations effectively. To address these limitations and advance video understanding in F^3 events, we introduce F^3Set, which tackles the intersectional challenges of existing temporal action understanding methods.

Existing F^3 related datasets. Although datasets have been developed for temporal action understanding, few focus on the F^3 events. Table 1 compares existing datasets with F^3Set by scale ("# Vid", "# Clips") and characteristics like action speed ("Evt. Len."), frequency ("Evt. / sec"), and granularity ("# Classes"), which correspond to "fast", "frequent", and "fine-grained" respectively. Datasets such as THUMOS14 [28] and Breakfast [30] focus on coarse-grained actions, where background context provides clear cues, and actions span seconds to minutes. In contrast, FineAction [40] and ActivityNet [5] cover a wide range of daily activities with diverse action categories, while FineGym [53] delves into detailed action types within gymnastics. Like FineGym, F^3Set emphasizes domain-specific granularity with subtle visual differences but encounters additional challenges due to faster and more frequent actions. Besides, unlike FineGym's typical single-player focus, F^3Set (e.g., tennis) features two players and a fast-moving ball, with both players rapidly moving across the court, occupying only small portions of the scene, thus increasing task difficulty. CCTV-Pipe [41] targets temporal defect detection in urban pipe systems, providing single-frame annotations for rapid event detection, though it is limited in frequency and event types. Research in the sports domain has explored the detection of fast and frequent actions. FineDiving [64] segments diverse diving events, while ShuttleSet [61] and P^2ANet [3] focus on identifying strokes in fast-paced racket sports. Volleyball [27] and NSVA (basketball) [63] focus on team sports understanding and video captioning, while SoccerNetV2 [12] ball action spotting task focus on identifying the timing and type of ball-related actions. However, these datasets typically cover coarser event types and are limited to specific F^3 aspects.

In contrast, our proposed F^3Set is characterized by 1) *rapid* events that occur instantaneously, 2) *high frequency* of approximately one event per second, and 3) *extensive granularity* with a larger number of detailed event classes. These attributes introduce novel challenges.

3 F^3SET

Recognizing the limitations in existing video datasets for F^3 event understanding, we introduce F^3Set, a new benchmark for precise temporal F^3 events detection and recognition. Given the need for domain-specific knowledge in creating F^3 datasets, this section uses tennis as a case study to

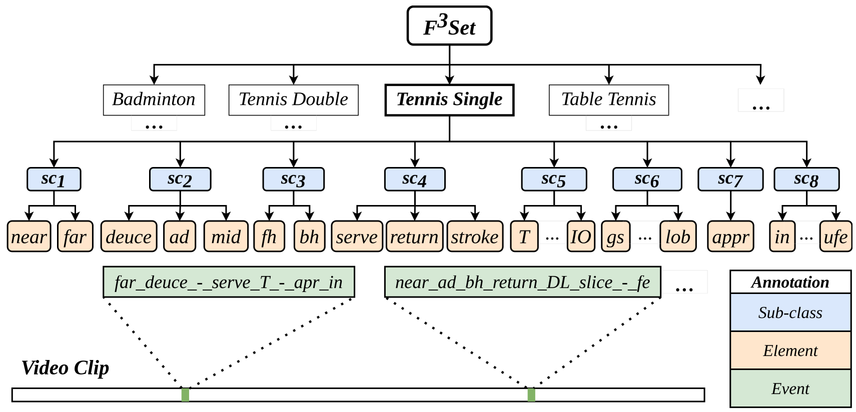

Figure 2: Breakdown of F^{3}Set event class annotation.

demonstrate F^{3}Set's event description, construction, and properties. We have proposed a general annotation pipeline and toolchain that enables domain experts to create new F^3 datasets across various applications. Using our toolchain, we have built F^3 datasets for badminton, table tennis, and tennis doubles. A community of users are actively expanding these with other applications (see link).

3.1 Event description

Lexicon. In tennis, a standard court can be divided into three regions on each side: the deuce court, the middle court, and the ad court. The initial shot of a point is called a "serve". The server can aim to hit the ball in the T, Body (B), or Wide (W) area of the service box. The shot taken by the receiver after a serve is called a "return" if it lands in bounds after crossing the net. Subsequent shots are referred to as "strokes". Players can hit the ball "cross-court" (CC), "down the line" (DL), "down the middle" (DM), "inside-in" (II), or "inside-out" (IO) at either their "forehand" (fh) or "backhand" (bh) side. Generally, a player can "approach" (apr) the net on a ball that lands around the service line or shorter. There are different shot techniques, such as "ground stroke/top spin" (gs), "slice", "volley", and "lob", under certain conditions. Each shot has four possible outcomes: "in-bound", "winner", "forced error", and "unforced error". More detailed definitions can be found in Appendix B.

F^3 events. Formally, each event (tennis) consists of 8 *sub-classes*, denoted as $sc_1, sc_2, ..., sc_8$:

- sc_1 – *hit by which player*: (1) near- or (2) far-end player;
- sc_2 – *hit from which court location*: (3) deuce, (4) middle, or (5) ad court;
- sc_3 – *hit at which side of the body*: (6) forehand or (7) backhand;
- sc_4 – *shot type*: (8) serve, (9) return, or (10) stroke;
- sc_5 – *shot direction*: (11) T, (12) B, (13) W, (14) CC, (15) DL, (16) DM, (17) II, or (18) IO;
- sc_6 – *shot technique*: (19) gs, (20) slice, (21) volley, (22) lob, (23) drop, or (24) smash;
- sc_7 – *player movement*: (25) approach;
- sc_8 – *shot outcome*: (26) in, (27) winner, (28) forced error, or (29) unforced error.

Altogether, there are 29 *elements* and 1,108 *event types* based on various combinations (Figure 2).

Similarly, for other domains, badminton contains 6 *sub-classes*, 28 *elements* and 1008 *event types*; table tennis contains 7 *sub-classes*, 23 *elements* and 1296 *event types*; and tennis doubles contain 26 *elements* and 744 *event types*. Compared to existing racket sports video datasets [3; 61], F^{3}Set offers additional dimensions, such as shot direction and outcomes, which are crucial for identifying playing patterns and success rates. Please refer to Appendix C for more details.

3.2 Dataset construction

Video collection. We collected high-resolution videos of singles tennis matches spanning from 2012 to 2023, sourced from YouTube. These matches encompass Grand Slams, Olympic games, and major ATP and WTA tournaments. The videos feature a variety of court surfaces—including hard, clay, and grass—and showcase both male and female players, as well as competitors with

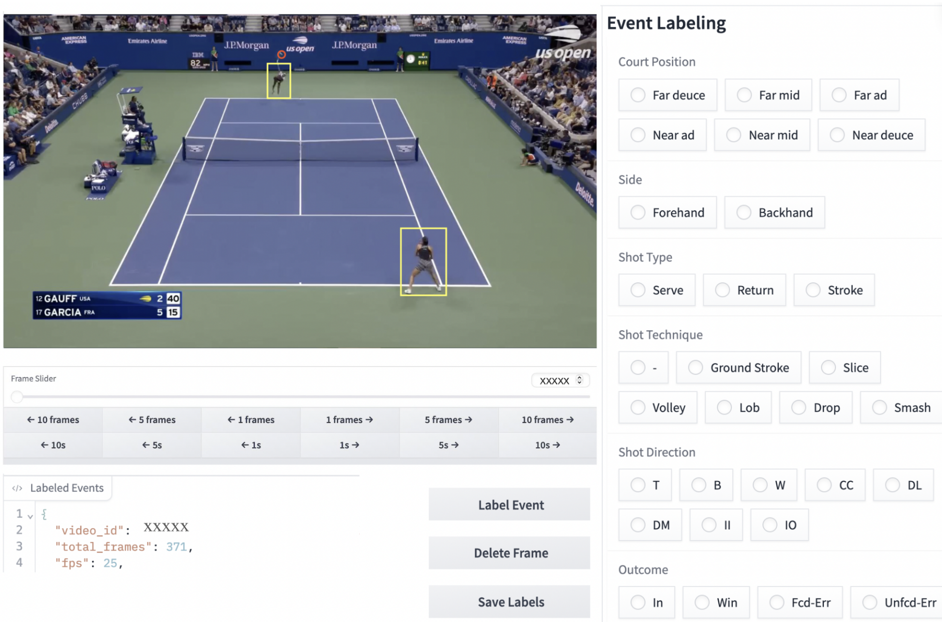

Figure 3: An interface of the labeling tool. The panel on the right is application-customizable.

different handedness (right-handed and left-handed). They provide comprehensive content that includes complete tennis rallies, extensive match footage, and detailed player information.

Annotation. After data collection, we conduct a three-stage process for data annotation, in which we try to maximize automation to reduce human labor. Our pipeline can be easily applied to different sports broadcast videos and broader domains. The annotation process comprises:

- *Video segmentation*: The first stage is to segment a full broadcast video into shorter clips using a context-aware scene detector [1] that automatically identifies jump cuts within the video.
- *Clip selection*: The second stage is to select targeted clips (e.g., clips contain tennis rallies) using a Siamese network to compare each clip with a "base image" indicative of the scene of interest.
- F^3 *event annotation*: The final stage is to identify the precise event moments (e.g., frames when a player hits the ball) and record the corresponding event types through an annotation tool.

The first two steps are automated and applicable to a range of sports videos, facilitating the efficient breakdown of lengthy videos into relevant clips. For the final phase, we have developed an interactive interface to support manual annotation, as demonstrated in Figure 3. This tool enables users to navigate through each clip with options for rapid advancement (e.g., in 1-second increments) or more detailed review (e.g., frame by frame), facilitating quicker identification of key event moments (e.g., tennis hitting moments). Additionally, it features functionality for selecting shot types and identifying location data (e.g., court position) by directly clicking on the video. The location of each click is concurrently displayed on the video for immediate verification. Object-level detection can be applied to facilitate annotation. Furthermore, to reduce labeling errors, our labeling tool also incorporates a fool-proof design to avoid unintended manual clicking errors or misjudgment. This tool is adaptable to other sports by integrating domain-specific knowledge and can be applied to broader applications.

Our annotation team consists of 8 members. We provided them with specialized training and rigorous pre-tests before beginning the official annotation work, along with supporting materials such as slides and demonstrations. Each annotator was assigned an equal portion of the dataset, totaling 1,450 clips (rallies) each. The manual labeling takes roughly 30 hours to finish all 1,450 clips. Following the initial annotation phase, we conducted multiple rounds of cross-validation involving random sampling of rallies and quality checks among annotators to ensure the accuracy of the event-based labels. In cases where conflicting annotations arose, annotators were asked to input the labels they believed to be correct. The final label was determined based on a majority vote among the annotators.

3.3 Data statistics and properties

The F^{3}Set tennis dataset consists of 114 high-resolution broadcast tennis singles matches featuring 75 professional players (30 men and 45 women) with frame rates ranging from 25 to 30 FPS. Among these players, 68 are right-handed, and 7 are left-handed. This dataset encompasses 11,584 video clips, each capturing a tennis rally with an average duration of approximately 8.4 seconds. These clips collectively contain 42,846 tennis shots, with each rally comprising between 1 to 34 shots. We employ a training, validation, and testing split of 3:1:1, with the training and validation sets drawn from the same video sources, while the test set features clips from distinct videos. Additional statistical insights can be found in Appendix D.

Event Timestamp. Unlike typical TAL and TAS tasks, where an action spans several frames or seconds, the duration of actions in racket sports is often ambiguous. Thus, stroke actions are defined as instantaneous events, recording only the moment of ball-racket contact [57] as shown in Figure 1.

Multi-level granularity. Depending on the requirement of the analytics task, the dataset can focus on a subset of the entire sub-classes. Therefore, we define a parameter $G \in \mathcal{P}(\{sc_1, \dots, sc_8\})$, where $\mathcal{P}(\{sc_1, \dots, sc_8\})$ denotes the power set of $\{sc_1, \dots, sc_8\}$, to determine what sub-classes to choose and form different levels of granularity. The multi-level granularity provides high flexibility for various real-world tasks (e.g., tennis strategy analytics). Here we define three different levels of granularity: $G_{low} = \{sc_1, sc_3, sc_4, sc_8\}$, representing a coarse granularity with 4 sub-classes, 11 elements, and 38 possible event types; $G_{mid} = \{sc_1, \dots, sc_6\}$, offering a finer granularity with 6 sub-classes, 24 elements, and 365 possible event types; and $G_{high} = \{sc_1, \dots, sc_8\}$, which is the most fine-grained level, encompassing all 29 elements and 1,108 event types. This multi-level granularity enhances the flexibility and applicability of our dataset for diverse real-world tasks.

4 Our Proposed Approach: F^{3}ED

Acknowledging the challenges and limitations of existing approaches, we propose a simple yet effective method named **F**ast **F**requent **F**ine-grained **E**vent **D**etection network (F^{3}ED), illustrated in Figure 4. It is designed for F^3 event detection and can serve as a baseline for further development.

Problem formulation. Let $X \in \mathbb{R}^{H\times W\times 3\times N}$ denote the input, consisting of N RGB frames of size $H \times W$. The output is a sequence of M event-timestamp pairs $((E_1, t_1), \dots, (E_M, t_M))$, where E_i is the event type with C classes and t_i is the corresponding timestamp for $i \in \{1, \dots, M\}$. Additionally, each event E_i can also be expressed as a vector $[e_{i,1}, \dots, e_{i,K}]$, with each element $e_{i,j} \in \{0, 1\}$ indicating the presence or absence of the j^{th} element in event E_i, where j is an integer $j \in \{1, \dots, K\}$. The parameter K, which defines the *number of elements* in each event vector.

Video Encoder (VE). The first stage of both baselines and our model will extract spatial-temporal frame-wise features. The video encoder (VE) consists of a visual backbone, followed by a bidirectional GRU to capture long-term visual dependencies: $\mathbf{F}_{emb} = \text{VE}(X)$, with $\mathbf{F}_{emb} \in R^{N\times d'}$.

Event Localizer (LCL). Utilizing the frame-wise features $\mathbf{F}_{emb}$, the event localizer (LCL) employs a fully connected network with a Sigmoid activation function to perform dense binary classification, aiming to accurately identify specific event instances. For an N-frame clip, the output is represented as $(\hat{p}_1, \dots, \hat{p}_N)$, where each $\hat{p}_i$ denotes the probability that an event occurs at the corresponding timestamp: $(\hat{p}_1, \dots, \hat{p}_N) = \textit{Sigmoid}(\text{LCL}(\mathbf{F}_{emb}))$. Ground truth labels $(p_1, \dots, p_N)$ with $p_i \in \{0, 1\}$ are used to compute the discrepancy between the predicted probabilities and the actual values using binary cross-entropy loss as: $L_{LCL} = \frac{1}{N}\sum_{i=1}^{N} p_i \cdot log(\hat{p}_i) + (1 - p_i) \cdot log(1 - \hat{p}_i)$.

Multi-label Event Classifier (MLC). Upon detecting events, we proceed to categorize them into specific types using a multi-label classification module (MLC). This module, a fully connected network, takes the identified event features f_i from $\mathbf{F}_{emb}$ as inputs to predict the event types: $\hat{E}_i = \textit{Sigmoid}(\text{MLC}(f_i)) = [\hat{e}_{i,1}, \dots, \hat{e}_{i,K}]$, where K denotes the number of elements, f_i represents the features for the event at the i^{th} frame, $\hat{E}_i$ is the predicted event type, and $\hat{e}_{i,j} \in [0, 1]$ is the probability of $\hat{E}_i$ containing the j^{th} element. For a video clip with M events, the ground truths are given as $(E_1, \dots, E_M)$ with each E_i represented as a vector of K elements $[e_{i,1}, \dots, e_{i,K}]$. The loss can be represented by $L_{MLC} = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{K}\sum_{j=1}^{K} e_{i,j} \cdot log(\hat{e}_{i,j}) + (1 - e_{i,j}) \cdot log(1 - \hat{e}_{i,j}))$.

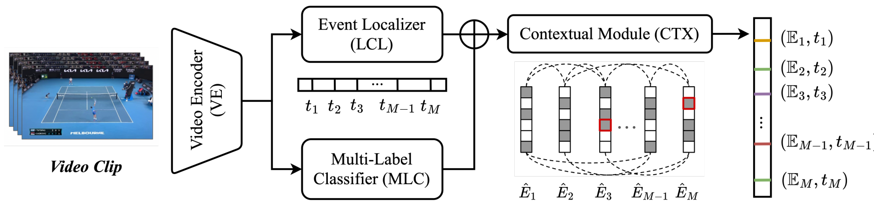

Figure 4: Overview of F[3]ED. RGB images are processed by VE to capture frame-wise spatial-temporal features, which are passed to LCL to identify event timestamps and MLC to predict labels. Outputs from LCL and MLC are combined ('plus' symbol) to form an event representation sequence and refined by CTX module. 'Red squares' represent errors from purely visual predictions.

Contextual module (CTX) Video encoders often struggle to extract insightful visual features from fast-paced videos due to motion blur, and objects of interest, such as players, may only occupy a small portion of the frame. This can result in the loss of crucial visual details for fine-grained action classification, particularly when resizing images to 224×224. Selecting the best-predicted event types naively might, therefore, produce invalid event sequences. To address this, we introduce a contextual module (CTX), designed to concurrently learn contextual knowledge from event sequences during end-to-end training: $(\mathbb{E}_1, \ldots, \mathbb{E}_M) = \text{CTX}(\hat{E}_1, \ldots, \hat{E}_M)$. CTX employs a bidirectional GRU to process the predicted event sequence $\hat{E}$ and outputs a refined sequence $\mathbb{E}_i = [\mathbb{e}_1, \ldots, \mathbb{e}_k]$, integrating both visual-based predictions and contextual correlations across events. The loss is calculated for each refined event: $L_{CTX} = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{K}\sum_{j=1}^{K} e_{i,j} \cdot log(\mathbb{e}_{i,j}) + (1 - e_{i,j}) \cdot log(1 - \mathbb{e}_{i,j}))$.

5 Experiments

In this section, we benchmark existing temporal action understanding methods, including TAL, TAS, and TASpot, on the F[3]Set dataset and conduct a series of ablation studies.

Evaluation metrics. We evaluate using two types of metrics: (1) the sequence-wise edit score [32], which assesses the similarity between event sequences using the Levenshtein or edit distance—measuring the minimum number of insertions, deletions, and substitutions needed to convert one sequence into another, and (2) the mean F1 score within a tight temporal tolerance of ± 1 frame. The mean F1 score considers a prediction accurate only if its timestamp aligns within ± 1 frame of the ground truth and the class label is correctly identified. Given the long-tail distribution of event types in the dataset, where some events are extremely rare, we report two variants of the mean F1 score to ensure a balanced evaluation: $F1_{evt}$, the average F1 score across all event types, and $F1_{elm}$, the average F1 score across all elements, which typically presents a more balanced distribution. These metrics align with evaluation standards in similar tasks [25; 24].

Baselines. Existing temporal action understanding frameworks typically incorporate two key components: a *video encoder* for visual feature extraction and a *head module* for specific tasks such as detection or segmentation. Applying these models directly to our study presents challenges, as they generally utilize a two-stage training process—employing a static, pre-trained video encoder for feature extraction and training only the head module. This approach often fails to capture fine-grained, domain-specific events due to its reliance on temporally coarse, non-overlapping, or downsampled video segments. To address these limitations, we have adapted these temporal action understanding methods to develop new baselines better suited for detecting F[3] events. Given the rapid pace and short duration of tennis shots, it is crucial to utilize frame-wise feature extraction [8] (discussed in Section 5.2). Besides, end-to-end training with video encoder fine-tuning is required to capture the subtle event differences. Moreover, the classification of some sub-classes (e.g., shot direction, outcome) demands long-term temporal reasoning to integrate information from subsequent frames

Consequently, we focus on three established feature extractors: TSN [59], TSM [35], and SlowFast [19], which are known for their efficiency in frame-wise feature extraction and end-to-end training. We pair each encoder with five representative head module architectures from existing methods: MS-TCN [18] and ASFormer [66] from TAS, G-TAD [65] and ActionFormer [67] from TAL and E2E-Spot [25] from TASpot, to establish a set of new baseline models for our study. To identify hitting moments and their respective event types, frame-wise dense *multi-class* classification is applied to identify each frame as either background or one of the event types.

Table 2: Experimental results on F^{3}Set (tennis) with 3 levels of granularity.

		F^{3}Set (G_{high})			F^{3}Set (G_{mid})			F^{3}Set (G_{low})		
Video encoder	Head arch.	F1$_{evt}$	F1$_{elm}$	Edit	F1$_{evt}$	F1$_{elm}$	Edit	F1$_{evt}$	F1$_{elm}$	Edit
TSN [59]	MS-TCN [18]	15.9	59.8	53.5	23.2	60.9	65.8	45.7	70.4	72.8
	ASformer [66]	11.9	54.3	49.8	17.3	56.1	62.5	40.3	67.3	70.3
	G-TAD [65]	6.0	47.5	24.7	14.1	52.1	48.6	19.9	57.4	44.7
	ActionFormer [67]	18.4	60.6	55.2	24.8	61.9	67.3	48.7	70.6	72.2
	E2E-Spot [25]	24.7	65.3	60.1	31.5	66.2	71.0	53.5	73.6	75.0
SlowFast [19]	MS-TCN [18]	17.2	63.1	56.2	24.3	65.5	70.3	47.4	73.1	73.5
	ASformer [66]	14.1	60.8	55.3	20.3	62.8	69.4	44.8	72.9	71.9
	G-TAD [65]	23.0	66.1	64.0	29.6	66.5	74.2	53.3	76.0	77.9
	ActionFormer [67]	28.7	70.0	67.6	35.5	70.9	76.4	59.3	77.1	81.5
	E2E-Spot [25]	25.9	69.4	65.7	33.8	70.4	75.4	55.5	76.5	79.5
TSM [35]	MS-TCN [18]	21.7	67.3	58.6	30.4	69.5	73.0	50.2	74.0	75.3
	ASformer [66]	17.6	61.9	57.5	25.5	64.0	74.2	46.0	72.9	74.0
	G-TAD [65]	16.9	62.5	55.2	29.8	66.9	74.8	39.8	70.1	67.2
	ActionFormer [67]	22.4	65.7	60.3	31.0	68.2	74.7	52.4	73.8	74.9
	E2E-Spot [25]	31.4	71.4	68.7	39.5	72.3	77.9	60.6	78.4	82.1
TSM[35]	F^{3}ED	**40.3**	**75.2**	**74.0**	**48.0**	**76.5**	**82.4**	**68.4**	**80.0**	**87.2**

Implementation details. We implement and train models on F^{3}Set in an end-to-end manner. The video encoder takes video clip X down-scaled and cropped to 224 $\times$ 224 to extract frame-wise visual features. Subsequently, each head module processes per-frame features to identify a sequence of F^3 events and their timestamps. For more implementation details, please refer to Appendix E.

5.1 Results and analysis.

The evaluation results presented in Table 2 provide several critical insights into the performance of various methods across different levels of granularity (G_{low}, G_{mid}, and G_{high}). A general trend emerges where performance decreases as granularity increases, underscoring the growing challenges associated with finer granularity. While certain methods demonstrate some robustness, the overall efficacy across all approaches remains suboptimal, particularly at higher levels of granularity, indicating the challenge of precise F^3 event detection task.

Simple 2D CNNs such as TSN, which process frames independently, are inadequate for F^3 event detection due to their inability to capture critical spatial-temporal correlations between frames, which are essential for distinguishing visually similar events. Without modeling temporal dynamics, these approaches struggle to differentiate events that may appear identical when viewed frame-by-frame, leading to significantly lower performance, particularly at higher granularity levels.

Head modules such as transformer-based ActionFormer, and GRU-based E2E-Spot, generally outperform other methods. This advantage highlights their effectiveness in capturing long-term temporal dependencies through end-to-end training. Notably, E2E-Spot consistently outperforms ActionFormer across most settings, suggesting that GRU-based architectures may offer an advantageous trade-off between efficiency and representational power for certain types of temporal correlations.

Interestingly, the combination of TSM with E2E-Spot outperforms the more complex SlowFast model, indicating that increasing the complexity of the video encoder does not necessarily translate to better performance. Instead, it is more important for a video encoder to capture subtle visual differences over short temporal durations, which are crucial for F^3 event detection. This result suggests that the capability of capturing subtle temporal cues and representation is more impactful than the model complexity.

Our proposed F^{3}ED model, leveraging the TSM video encoder, achieves the best performance among all granularity levels. This is attributable to two key design choices: the multi-label classifier and the contextual module. Detailed discussions of these design elements are presented in the next section.

5.2 Ablation Study

We selected the highest-performing baseline model (TSM + E2E-Spot) as our default configuration for the subsequent ablation studies. More ablation studies can be found in Appendix F.

Table 3: Ablation and analysis experiments. The default model takes stride size 2 and clip length 96.

	F^3Set (G_{high})			F^3Set (G_{mid})			F^3Set (G_{low})		
Experiment	$F1_{evt}$	$F1_{elm}$	Edit	$F1_{evt}$	$F1_{elm}$	Edit	$F1_{evt}$	$F1_{elm}$	Edit
Default (stride 2, clip len 96)									
TSM + E2E-Spot	31.4	71.4	68.7	39.5	72.3	77.9	60.6	78.4	82.1
(a) Feature extractor									
I3D [6] (clip-wise)	22.7	59.7	68.7	27.1	60.7	74.2	51.9	67.7	78.3
VTN [47] (video transformer)	14.8	58.3	56.7	20.0	59.4	68.2	39.7	63.1	73.1
ST-GCN++ [16] (skeleton-based)	25.4	62.1	56.1	32.4	63.9	63.5	55.1	69.4	73.2
PoseConv3D [17] ((skeleton-based))	20.1	54.5	53.2	26.0	55.4	61.9	48.8	63.0	69.7
(b) Stride size									
Stride = 4	25.9	69.2	62.7	33.4	69.9	73.0	60.0	77.9	78.8
Stride = 8	14.0	56.7	44.3	18.5	57.4	54.8	40.4	67.0	59.2
(c) without GRU	27.6	69.0	60.6	38.0	71.3	75.3	54.7	74.1	73.4
(d) Clip length									
Length = 32	26.3	67.4	54.5	35.5	69.4	71.8	53.2	75.1	68.9
Length = 64	30.7	71.2	67.4	38.6	72.4	77.5	58.4	77.9	81.1
Length = 192	29.3	70.3	65.7	37.3	71.4	77.0	58.8	77.1	80.4
(e) Multi-label	37.9	74.3	71.7	45.9	75.6	80.1	66.6	80.1	85.1
(f) Multi-label + CTX (Transformer)	39.0	74.3	72.8	50.5	75.5	81.8	63.4	79.6	86.8
Multi-label + CTX (BiGRU)	40.3	75.2	74.0	48.0	76.5	82.4	68.4	80.0	87.2

Feature extractor. An effective feature extractor is crucial for accurate F^3 event detection. Below, we summarize some key findings (details in Appendix F). *(1) Frame-wise feature extraction outperforms clip-wise methods*, which divide inputs into non-overlapping segments. Experiments show clip-wise methods produce temporally coarse features and hinder precise event detection. *(2) Transformer-based video encoders* such as VTN [47] struggle on F^3Set due to high computational costs and limited ability to effectively capture short-term temporal correlations. *(3) In addition to RGB inputs, we also experimented with skeleton-based pose estimation methods*, including ST-GCN++ [16] and PoseConv3D [17] with human key points as input. While they excel in efficiency and interpretability, they lack critical details like shot direction, limiting performance on F^3Set.

Sparse sampling. Increasing the stride size allows for a broader temporal coverage within a fixed sequence length. This sparse sampling technique is prevalent in many video understanding tasks [39; 34], offering high efficiency and reasonable accuracy. However, this approach proves inadequate for our task, where events are characterized by their rapid occurrence, frequency, and fine granularity. As illustrated in Table 3(b), increasing the stride size to 4 and 8 leads to a marked decline in performance, underscoring the importance of dense sampling for detecting F^3.

Long-term temporal reasoning. The default model employs a spatio-temporal video encoder (TSM), complemented by a bidirectional Gated Recurrent Unit [14] (GRU) head for enhanced long-term temporal integration. To assess the necessity of long-term temporal reasoning, we replaced the GRU module with a fully connected layer. The results, presented in Table 3(c), indicate a significant performance decline relative to the original configuration. This finding highlights the essential role of long-term temporal reasoning in analyzing sub-classes such as shot direction, outcomes, and player movements that require information from subsequent frames.

Clip length. The sensitivity of sequence models to varying input clip lengths, which encapsulate different temporal contexts, is notable. In F^3Set, the incidence of F^3 events correlates directly with clip length. Table 3(d) shows that shorter clips result in fewer events per sequence, hindering the model's ability to leverage long-term dependencies among consecutive events effectively. Conversely, while longer clip lengths yield improved results, the marginal gains diminish with increasing length.

Multi-class versus multi-label classification. The challenge of modeling over 1,000 possible event type combinations as a multi-class classification problem is formidable. For example, consider two events, E_1 (far_ad_bh_stroke_DL_*slice*_apr_in) and E_2 (far_ad_bh_stroke_DL_*drop*_apr_in), which differ only in shot technique (*slice* vs. *drop*). Although similar, multi-class classification treats these as distinct classes, thus reducing training efficiency and exacerbating the long-tail distribution bias towards more frequent classes. A more natural approach is multi-label classification, where each event can belong to multiple sub-class elements (e.g., ['far', 'ad', 'serve', 'W', 'in']). Thus, $E1$ and

Table 4: Experimental results on other "semi-F^3" datasets.

	ShuttleSet [61]		FineDiving [64]		FineGym [53]		SoccerNetV2 [12]		CCTV-Pipe [41]	
Head arch.	F1$_{evt}$	Edit	F1$_{evt}$	Edit	F1$_{evt}$	Edit	F1$_{evt}$	Edit	F1$_{evt}$	Edit
MS-TCN [18]	70.3	74.4	65.7	92.2	57.6	65.3	43.4	74.5	25.8	31.3
ASformer [66]	55.9	70.6	49.9	87.6	53.6	66.3	46.3	76.1	15.4	33.4
G-TAD [65]	48.2	61.1	52.1	82.6	45.8	51.4	42.3	72.3	31.3	33.6
ActionFormer [67]	62.1	67.5	68.3	92.4	54.0	59.7	43.0	64.6	18.8	29.5
E2E-Spot [25]	70.2	75.0	75.8	93.7	62.1	65.4	46.2	72.9	27.2	35.2
F^{3}ED	70.7	77.1	77.6	95.1	70.9	70.7	48.1	76.6	37.0	39.5

$E2$ only differ in shot technique but are identical in other aspects. This adjustment facilitates more effective training and shows an increase in performance, as shown in Table 3(e).

Contextual knowledge. Beyond the statistical results in Table 2, analysis of predicted event sequences reveals that current baselines may produce invalid sequences due to logical errors or uncommon practices. For instance, a right-handed player cannot logically direct a forehand shot from the deuce court as "II" or "IO". Similarly, an event ending in a winner or error should logically conclude the sequence. Additionally, it is uncommon for a player to hit with backhand when the ball is played to their forehand side. Further examples are detailed in Appendix G. These observations indicate that existing baselines fail to effectively capture event-wise contextual correlations. By adding the CTX module, the performance further increases as shown in Table 3(f). We also compared BiGRU and Transformer Encoder for the CTX module. BiGRU performed slightly better, likely due to its efficiency in modeling short *event sequences* (usually < 20 per clip) with fewer parameters.

5.3 Generalizability to "semi-F^3" data

F^3 task possesses broad applicability across numerous real-world domains, such as sports, autonomous driving, surveillance, and production line inspection. Nevertheless, creating such a F^3 dataset necessitates substantial expertise and extensive labeling efforts. We have found that existing video datasets often fail to fully address all three dimensions of the F^3 task—"fast", "frequent", and "fine-grained". In this section, we conducted experiments on several "semi-F^3" datasets that partially meet these criteria, including Shuttleset [61] for badminton (racket sport), FineDiving [64] for diving (individual sports), FineGym [53] for gymnastics (individual sports), SoccerNetV2 [45] (team sports), and CCTV-Pipe [41] for pipe defect detection (industrial application). We report only the F1$_{evt}$ and Edit score, as not all datasets necessitate multi-label classification given their limited event types. For the video encoder, we chose TSM, which consistently outperforms the others on average.

Performance across different domains can vary significantly depending on the difficulty of tasks and the scale of datasets. For instance, the CCTV-Pipe dataset, targeting temporal defect localization in urban pipe systems, shows suboptimal performance due to factors such as ambiguous single-frame annotations for each defect, multiple defects at the same time, long-tailed distribution of defect types, and limited dataset size. Our performance is better than the results reported in [41]. Generally, methods that effectively handle F^{3}Set tend to perform well across other applications, as indicated in Table 4. Our F^{3}ED outperforms existing baselines in all datasets, demonstrating its robust generalizability for detecting "semi-F^3" events across various domains. While F^3 event detection benefits from accurate event localization, a high-performing LCL module is not a hard prerequisite (see Appendix H). Therefore, our method can be generalized and benefit broader applications.

6 Conclusion and Future Work

In this study, we addressed the challenge of analyzing fast, frequent, and fine-grained (F^3) events from videos by introducing F^{3}Set, a benchmark for precise temporal F^3 event detection. F^{3}Set datasets usually feature detailed event types (approximately 1,000), annotated with precise timestamps, and provide multi-level granularity. We have also developed a general annotation toolchain that enables domain experts to create F^3 datasets, thereby facilitating further research in this field. Moreover, we proposed F^{3}ED, an end-to-end model that effectively detects complex event sequences from videos, using a combination of visual features and contextual sequence refinement. Our comprehensive evaluations and ablation studies of leading methods in temporal action understanding on F^{3}Set highlighted their performance and provided critical insights into their capabilities and limitations. Moving forward, we aim to extend the scope of F^3 task to more real-world scenarios and advance the development of F^3 video understanding.

REFERENCES

[1] Pyscenedetect. `https://www.scenedetect.com/`.

[2] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision*, pp. 52–68. Springer, 2022.

[3] Jiang Bian, Xuhong Li, Tao Wang, Qingzhong Wang, Jun Huang, Chen Liu, Jun Zhao, Feixiang Lu, Dejing Dou, and Haoyi Xiong. P2anet: A large-scale benchmark for dense action detection from table tennis match broadcasting videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–23, 2024.

[4] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. 2019.

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[7] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1130–1139, 2018.

[8] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13801–13810, 2022.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

[10] Tom Decroos, Jan Van Haaren, and Jesse Davis. Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 223–232, 2018.

[11] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1851–1861, 2019.

[12] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE/CVF conference on computer vision and pattern recognition*, pp. 4508–4519, 2021.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[14] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597–1600. IEEE, 2017.

[15] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[16] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7351–7354, 2022.

[17] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2969–2978, 2022.

[18] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *The IEEE/CVF conference on computer vision and pattern recognition*, pp. 3575–3584, 2019.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *The IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

[20] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 3342–3351, 2017.

[21] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2021.

[22] Silvio Giancola, Anthony Cioppa, Julia Georgieva, Johsan Billingham, Andreas Serner, Kerry Peek, Bernard Ghanem, and Marc Van Droogenbroeck. Towards active learning for action spotting in association football videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5097–5107, 2023.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[24] Yuchen He, Zeqing Yuan, Yihong Wu, Liqi Cheng, Dazhen Deng, and Yingcai Wu. Vistec: Video modeling for sports technique recognition and tactical analysis. *arXiv preprint arXiv:2402.15952*, 2024.

[25] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*, pp. 33–51. Springer, 2022.

[26] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 954–960, 2018.

[27] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *The IEEE conference on computer vision and pattern recognition*, pp. 1971–1980, 2016.

[28] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[29] 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.

[30] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*, Lake Placid, Mar 2016.

[31] Colin Lea, René Vidal, and Gregory D Hager. Learning convolutional action primitives for fine-grained action recognition. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 1642–1649. IEEE, 2016.

[32] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–165, 2017.

[33] Zhe Li, Yazan Abu Farha, and Jurgen Gall. Temporal action segmentation from timestamp supervision. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8365–8374, 2021.

[34] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3320–3329, 2021.

[35] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *The IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

[36] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17949–17958, 2022.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[38] Meng Liu, Liqiang Nie, Yunxiao Wang, Meng Wang, and Yong Rui. A survey on video moment localization. *ACM Computing Surveys*, 55(9):1–37, 2023.

[39] Xiaolong Liu, Song Bai, and Xiang Bai. An empirical study of end-to-end temporal action detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20010–20019, 2022.

[40] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950, 2022.

[41] Yi Liu, Xuan Zhang, Ying Li, Guixin Liang, Yabing Jiang, Lixia Qiu, Haiping Tang, Fei Xie, Wei Yao, Yi Dai, et al. Videopipe 2022 challenge: Real-world video understanding for urban pipe inspection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 4967–4973. IEEE, 2022.

[42] Zhaoyu Liu, Kan Jiang, Zhe Hou, Yun Lin, and Jin Song Dong. Insight analysis for tennis strategy and tactics. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1175–1180. IEEE, 2023.

[43] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[44] Hiroaki Minoura, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, Mitsuru Nakazawa, Yeongnam Chae, and Björn Stenger. Action spotting and temporal attention analysis in soccer videos. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pp. 1–6. IEEE, 2021.

[45] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5073–5084, 2023.

[46] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 122–132, 2020.

[47] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3163–3172, 2021.

[48] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pp. 3153–3160. IEEE, 2011.

[49] OpenAI. Gpt-4 technical report, 2023.

[50] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–27, 2019.

[51] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8347–8356, 2019.

[52] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

[53] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *The IEEE/CVF conference on computer vision and pattern recognition*, pp. 2616–2625, 2020.

[54] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18857–18866, 2023.

[55] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1049–1058, 2016.

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[57] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. Ttnet: Real-time temporal and spatial video analysis of table tennis. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 884–885, 2020.

[58] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *The IEEE conference on computer vision and pattern recognition*, pp. 7622–7631, 2018.

[59] 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*, 41(11):2740–2755, 2018.

[60] Wei-Yao Wang, Hong-Han Shuai, Kai-Shiang Chang, and Wen-Chih Peng. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4219–4227, 2022.

[61] Wei-Yao Wang, Yung-Chang Huang, Tsi-Ui Ik, and Wen-Chih Peng. Shuttleset: A human-annotated stroke-level singles dataset for badminton tactical analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5126–5136, 2023.

[62] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 34–51. Springer, 2020.

[63] Dekun Wu, He Zhao, Xingce Bao, and Richard P Wildes. Sports video analysis on large-scale data. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2022.

[64] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2949–2958, 2022.

[65] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10156–10165, 2020.

[66] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

[67] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pp. 492–510. Springer, 2022.