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

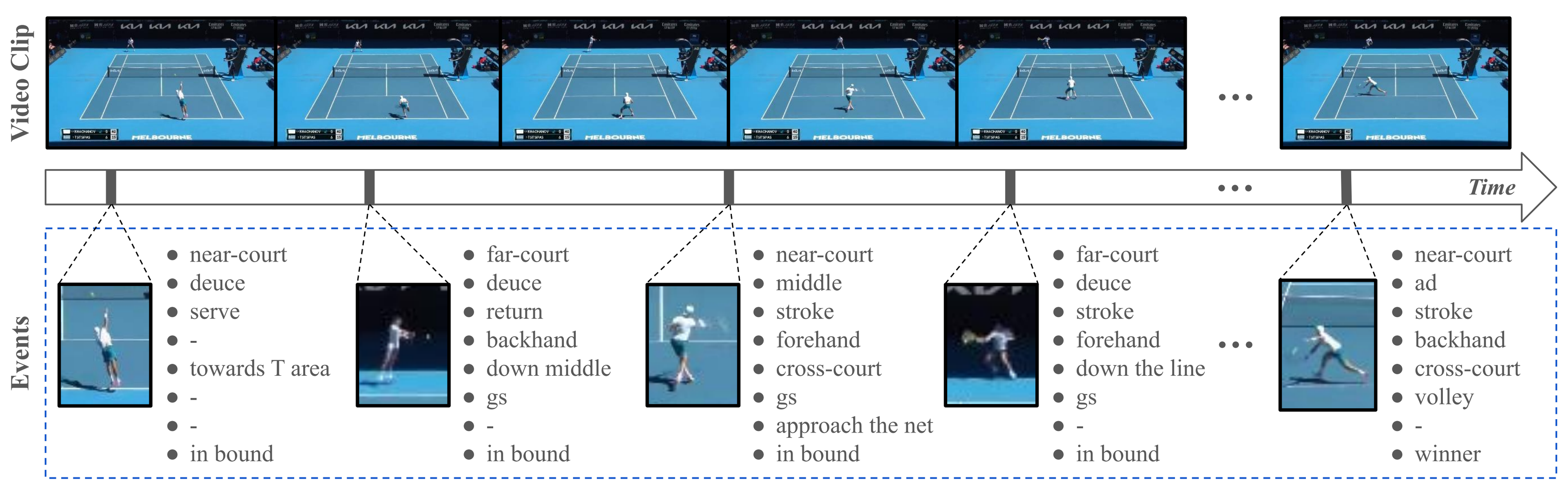
Zhaoyu Liu[1]*, Kan Jiang[1], Murong Ma[1], Zhe Hou[2], Yun Lin[3], Jin Song Dong[1]

1. National University of Singapore, 2. Griffith University, 3. Shanghai Jiao Tong University

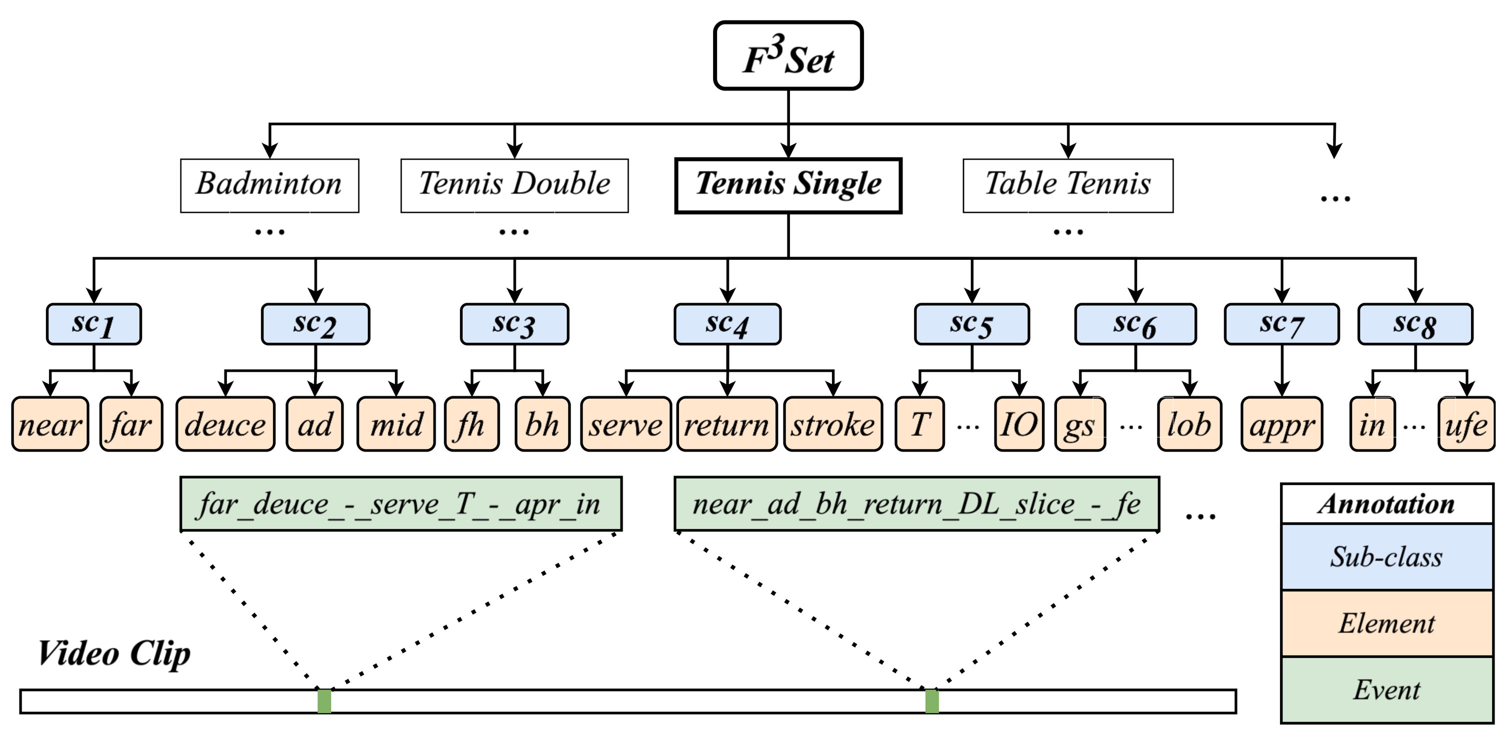
Project Page | Code & Paper | My Homepage

Introduction

- Understanding **fast, frequent, and fine-grained (F^3)** events is crucial for video analytics but remains underexplored.
- Applications: sports analytics, surveillance, autonomous driving…
- Challenges: subtle visual cues, motion blur, dense and precise timing.



F^3Set: A Benchmark Dataset for F^3 Event Detection

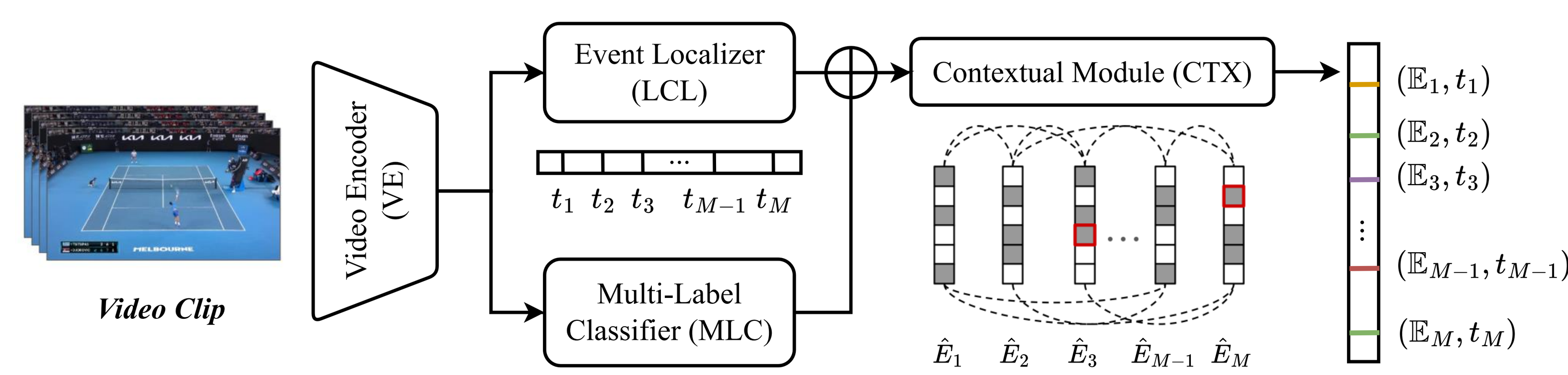


- Over **1,000+ event types**, timestamped at frame-level granularity.
- **Multi-domain**: Tennis, badminton, table tennis, etc.
- Multi-level granularity (G_{low}, G_{mid}, G_{high})

Datasets	# Vid.	# Clips.	Avg. Clip Len.	# Classes	Evt. Len.	# Evt. / sec
(a) Fine-grained				**Fine-grained**	**Fast**	**Frequent**
FineAction [41]	-	16,732	149.5s	101	6.9s	0.3
ActivityNet [4]	-	19,994	116.7s	200	49.2s	0.01
FineGym [58]	303	32,697	50.3s	530	1.7s	0.3
(b) Fast						
CCTV-Pipe [42]	575	575	549.3s	16	< 0.1s	0.02
SoccerNetV2 [11]	9	9	99.6min	12	< 0.1s	0.3
(c) Frequent						
FineDiving [69]	135	3,000	4.2s	29	1.1s	~1
(d) Fast & Frequent						
ShuttleSet [66]	44	3,685	10.9s	18	< 0.1s	~1
P^2ANet [3]	200	2,721	360.0s	14	< 0.1s	~2
(d) Fast & Frequent & Fine-grained						
F^3Set	114	11,584	8.4s	**1,108**	**< 0.1s**	**~1**

Our Proposed Approach: F^3ED

- **Video Encoder (VE)**
  Extracts frame-wise features: $F = VE(X)$, where $X \in \mathbb{R}^{H\times W\times 3\times N}$
- **Event Localizer (LCL)**
  Predicts event probabilities per frame: $(\hat{p}_1, \hat{p}_2, \dots, \hat{p}_N) = \sigma(LCL(F))$
- **Multi-label Classifier (MLC)**
  Predicts event types: $\hat{E}_i = \sigma(MLC(f_i)) = [\hat{e}_{i,1}, \hat{e}_{i,2}, \dots, \hat{e}_{i,K}],\ \hat{e}_{i,j} \in [0,1]$
- **Contextual Module (CTX)**
  Refines event sequence: $(\mathbb{E}_1, \mathbb{E}_2, \dots, \mathbb{E}_M) = CTX(\hat{E}_1, \hat{E}_2, \dots, \hat{E}_M)$



Experimental Results

Evaluation Metrics

- F1 score (event & element level): Measures precision and recall of predicted events.
- Edit score: Sequence-level similarity using Levenshtein distance.
- Temporal tolerance: F1 scores computed with ±1 frame tolerance for precise localization.

Baseline Models

- Video encoders: TSN, I3D, VTN, SlowFast, TSM.
- Head architectures: MS-TCN, ASFormer, G-TAD, ActionFormer, E2E-Spot.

Results & Analysis

- Capturing fine-grained temporal cues matters more than encoder complexity.
- E2E-Spot (GRU) head module outperforms, offering efficient long-term modeling.
- F^3ED consistently outperforms all baselines across F1 and Edit scores.

		F^3Set (G_{high})			F^3Set (G_{mid})			F^3Set (G_{low})		
Video encoder	Head arch.	$F1_{evt}$	$F1_{elm}$	Edit	$F1_{evt}$	$F1_{elm}$	Edit	$F1_{evt}$	$F1_{elm}$	Edit
TSN [64]	MS-TCN [19]	15.9	59.8	53.5	23.2	60.9	65.8	45.7	70.4	72.8
	ASformer [71]	11.9	54.3	49.8	17.3	56.1	62.5	40.3	67.3	70.3
	G-TAD [70]	6.0	47.5	24.7	14.1	52.1	48.6	19.9	57.4	44.7
	ActionFormer [72]	18.4	60.6	55.2	24.8	61.9	67.3	48.7	70.6	72.2
	E2E-Spot [24]	24.7	65.3	60.1	31.5	66.2	71.0	53.5	73.6	75.0
SlowFast [20]	MS-TCN [19]	17.2	63.1	56.2	24.3	65.5	70.3	47.4	73.1	73.5
	ASformer [71]	14.1	60.8	55.3	20.3	62.8	69.4	44.8	72.9	71.9
	G-TAD [70]	23.0	66.1	64.0	29.6	66.5	74.2	53.3	76.0	77.9
	ActionFormer [72]	28.7	70.0	67.6	35.5	70.9	76.4	59.3	77.1	81.5
	E2E-Spot [24]	25.9	69.4	65.7	33.8	70.4	75.4	55.5	76.5	79.5
I3D [5]	E2E-Spot [24]	22.7	59.7	68.7	27.1	60.7	74.2	51.9	67.7	78.3
VTN [52]	E2E-Spot [24]	14.8	58.3	56.7	20.0	59.4	68.2	39.7	63.1	73.1
TSM [35]	MS-TCN [19]	21.7	67.3	58.6	30.4	69.5	73.0	50.2	74.0	75.3
	ASformer [71]	17.6	61.9	57.5	25.5	64.0	74.2	46.0	72.9	74.0
	G-TAD [70]	16.9	62.5	55.2	29.8	66.9	74.8	39.8	70.1	67.2
	ActionFormer [72]	22.4	65.7	60.3	31.0	68.2	74.7	52.4	73.8	74.9
	E2E-Spot [24]	31.4	71.4	68.7	39.5	72.3	77.9	60.6	78.4	82.1
TSM[35]	F^3ED	**40.3**	**75.2**	**74.0**	**48.0**	**76.5**	**82.4**	**68.4**	**80.0**	**87.2**

Ablation Studies

- **Frame-wise > clip-wise**: dense sampling is crucial for fast actions.
- **Multi-label > multi-class**: better handles long-tail event combinations.
- **CTX (BiGRU)** improves sequence validity and accuracy.
- **Longer clips** help but with diminishing returns.
- **Stride size matters**: larger strides hurt performance.

	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
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 [5] (clip-wise)	22.7	59.7	68.7	27.1	60.7	74.2	51.9	67.7	78.3
VTN [52] (video transformer)	14.8	58.3	56.7	20.0	59.4	68.2	39.7	63.1	73.1
ST-GCN++ [17] (skeleton-based)	25.4	62.1	56.1	32.4	63.9	63.5	55.1	69.4	73.2
PoseConv3D [18] ((skeleton-based))	20.1	54.5	53.2	26.0	55.4	61.9	48.8	63.0	69.7
(b) Stride size = 4	25.9	69.2	62.7	33.4	69.9	73.0	60.0	77.9	78.8
Stride size = 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 = 32	26.3	67.4	54.5	35.5	69.4	71.8	53.2	75.1	68.9
Clip Length = 64	30.7	71.2	67.4	38.6	72.4	77.5	58.4	77.9	81.1
Clip 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**

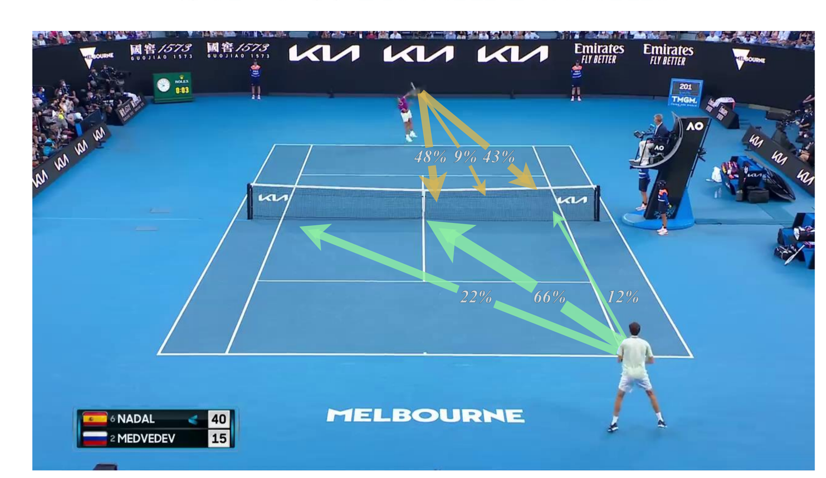
Generalizability to "Semi-F^3" Data

- F^3ED performs well on other domains: **badminton, diving, gymnastics, soccer, pipe inspection**.

	ShuttleSet [66]		FineDiving [69]		FineGym [58]		SoccerNetV2 [11]		CCTV-Pipe [42]	
Head arch.	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit	$F1_{evt}$	Edit
MS-TCN [19]	70.3	74.4	65.7	92.2	57.6	65.3	43.4	74.5	25.8	31.3
ASformer [71]	55.9	70.6	49.9	87.6	53.6	66.3	46.3	76.1	15.4	33.4
G-TAD [70]	48.2	61.1	52.1	82.6	45.8	51.4	42.3	72.3	31.3	33.6
ActionFormer [72]	62.1	67.5	68.3	92.4	54.0	59.7	43.0	64.6	18.8	29.5
E2E-Spot [24]	70.2	75.0	75.8	93.7	62.1	65.4	46.2	72.9	27.2	35.2
F^3ED	**70.7**	**77.1**	**77.6**	**95.1**	**70.9**	**70.7**	**48.1**	**76.6**	**37.0**	**39.5**

Real-World Applications

Serve & Return Patterns



On-the-run returning an approach shot