F³Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos

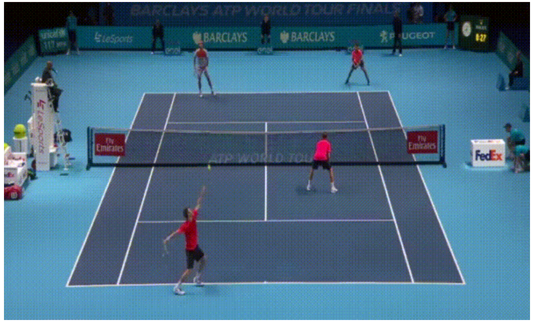
Zhaoyu Liu[1*], Kan Jiang[1], Murong Ma[1], Zhe Hou[2], Yun Lin[3], Jin Song Dong[1]

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

* Email: liuzy@nus.edu.sg

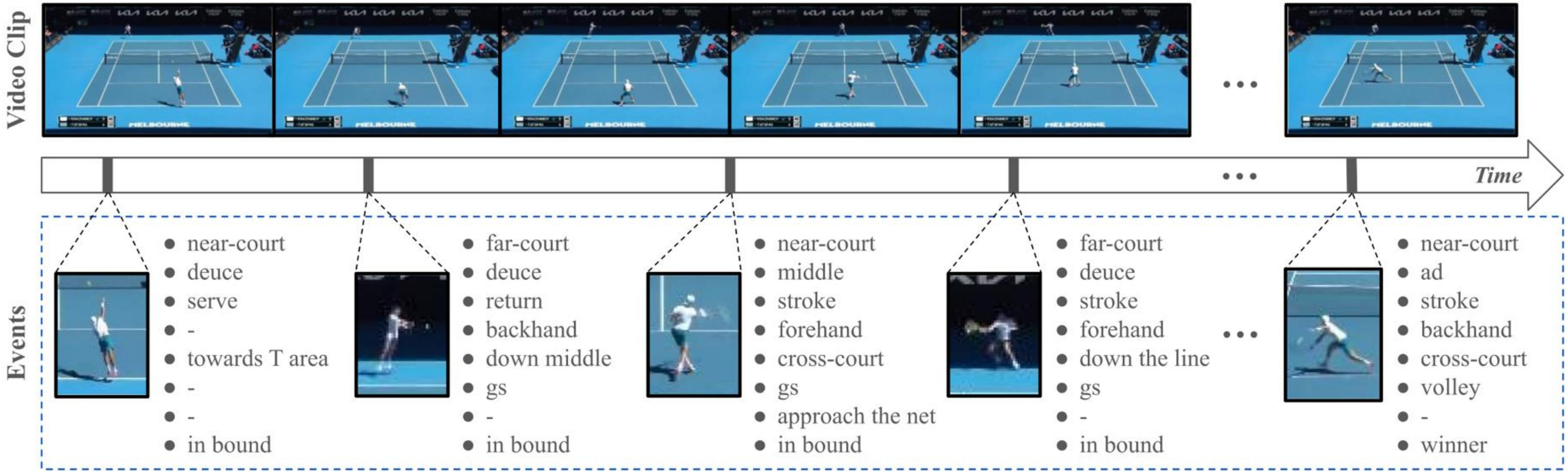
Why F³ Events Matter

- Real-world tasks require precise event detection
- Existing methods and datasets partially address F³
- A new benchmark is needed

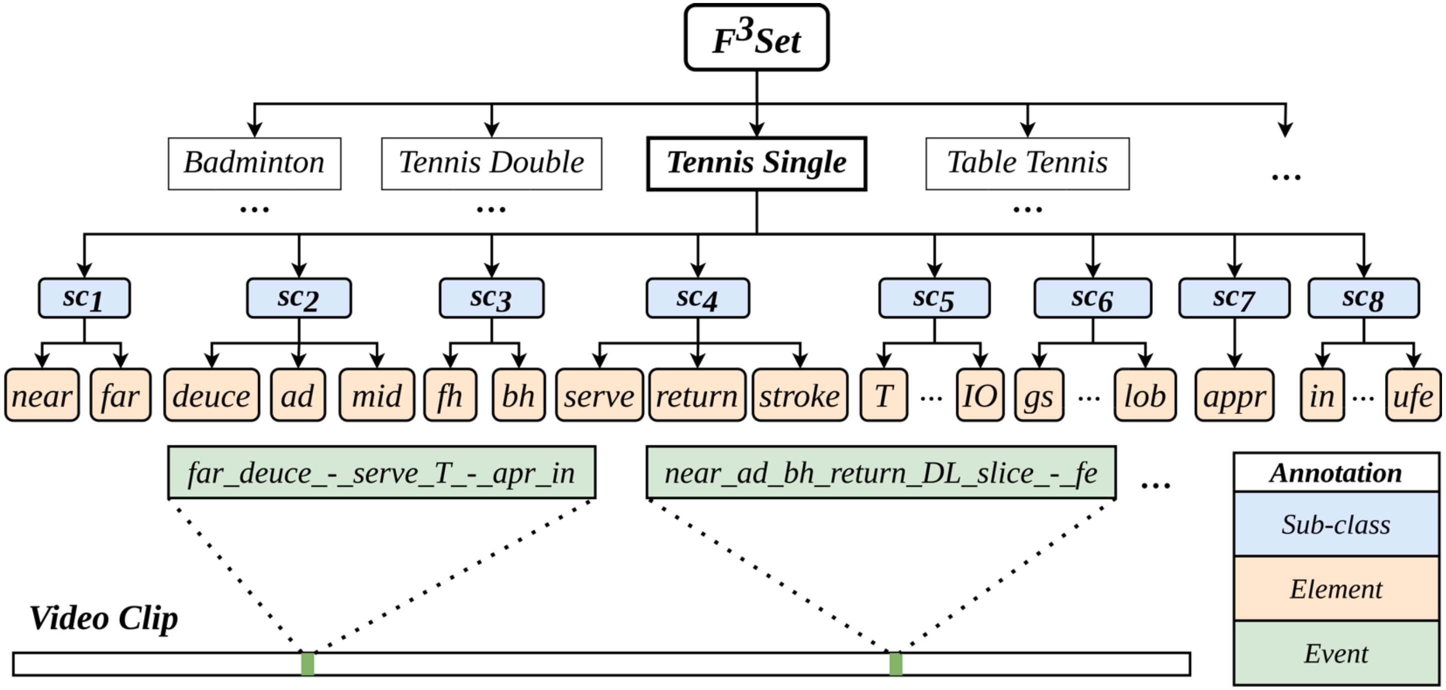


Definition of F^3 Events

- **Fast:** Short actions, 1–2 frames.
- **Frequent:** Multiple events per second.
- **Fine-Grained:** Over 1,000 subtle event types.



F^3Set



F^3Set
Badminton
Tennis Double
Tennis Single
Table Tennis
...
sc_1
sc_2
sc_3
sc_4
sc_5
sc_6
sc_7
sc_8
near
far
deuce
ad
mid
fh
bh
serve
return
stroke
T
IO
gs
lob
appr
in
ufe
far_deuce_-_serve_T_-_apr_in
near_ad_bh_return_DL_slice_-_fe
Video Clip
Annotation
Sub-class
Element
Event

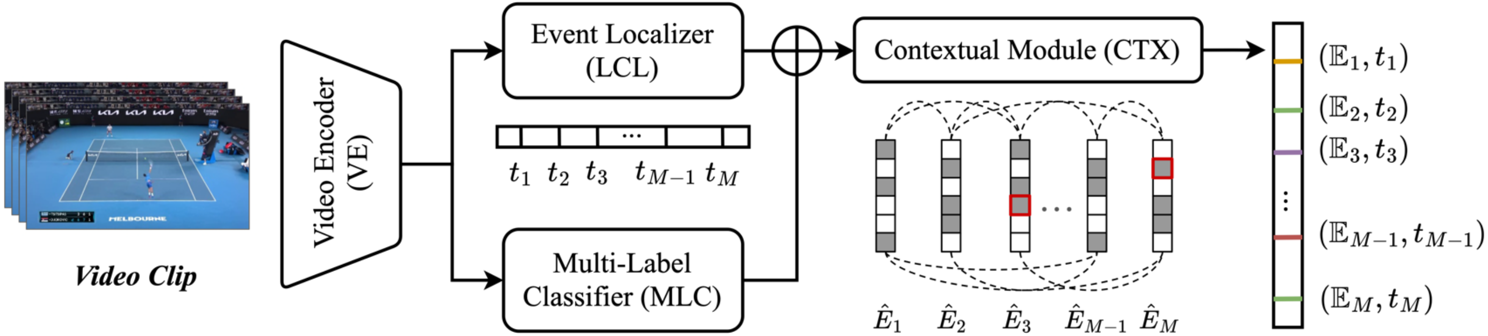
How F^3Set Stands Out

				Fine-grained	Fast	Frequency
Datasets	# Vid.	# Clips.	Avg. Clip Len.	# Classes	Evt. Len.	# Evt. / sec
(a) Fine-grained						
FineAction [39]	-	16,732	149.5s	101	6.9s	0.3
ActivityNet [4]	-	19,994	116.7s	200	49.2s	0.01
FineGym [51]	303	32,697	50.3s	530	1.7s	0.3
(b) Fast						
CCTV-Pipe [40]	575	575	549.3s	16	< 0.1s	0.02
SoccerNetV2 [12]	9	9	99.6min	12	< 0.1s	0.3
(c) Frequent						
FineDiving [62]	135	3,000	4.2s	29	1.1s	$\sim$1
(d) Fast & Frequent						
ShuttleSet [59]	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

Why Existing Methods Struggle

- Coarse frame sampling misses fine-grained cues
- Motion blur obscures fast events
- Pretrained encoders lack domain-specific nuance

F³ED: Our End-to-End Model



- Frame-wise features via TSM
- Event localizer + multi-label classifier
- GRU-based context refinement

Performance on F³Set

Video encoder	Head arch.	F^{3}Set (G_{high})			F^{3}Set (G_{mid})			F^{3}Set (G_{low})		
		F1$_{evt}$	F1$_{elm}$	Edit	F1$_{evt}$	F1$_{elm}$	Edit	F1$_{evt}$	F1$_{elm}$	Edit
TSN [57]	MS-TCN [18]	15.9	59.8	53.5	23.2	60.9	65.8	45.7	70.4	72.8
	ASformer [64]	11.9	54.3	49.8	17.3	56.1	62.5	40.3	67.3	70.3
	G-TAD [63]	6.0	47.5	24.7	14.1	52.1	48.6	19.9	57.4	44.7
	ActionFormer [65]	18.4	60.6	55.2	24.8	61.9	67.3	48.7	70.6	72.2
	E2E-Spot [23]	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 [64]	14.1	60.8	55.3	20.3	62.8	69.4	44.8	72.9	71.9
	G-TAD [63]	23.0	66.1	64.0	29.6	66.5	74.2	53.3	76.0	77.9
	ActionFormer [65]	28.7	70.0	67.6	35.5	70.9	76.4	59.3	77.1	81.5
	E2E-Spot [23]	25.9	69.4	65.7	33.8	70.4	75.4	55.5	76.5	79.5
I3D [5]	E2E-Spot [23]	22.7	59.7	68.7	27.1	60.7	74.2	51.9	67.7	78.3
VTN [45]	E2E-Spot [23]	14.8	58.3	56.7	20.0	59.4	68.2	39.7	63.1	73.1
TSM [33]	MS-TCN [18]	21.7	67.3	58.6	30.4	69.5	73.0	50.2	74.0	75.3
	ASformer [64]	17.6	61.9	57.5	25.5	64.0	74.2	46.0	72.9	74.0
	G-TAD [63]	16.9	62.5	55.2	29.8	66.9	74.8	39.8	70.1	67.2
	ActionFormer [65]	22.4	65.7	60.3	31.0	68.2	74.7	52.4	73.8	74.9
	E2E-Spot [23]	31.4	71.4	68.7	39.5	72.3	77.9	60.6	78.4	82.1
TSM[33]	F^{3}ED	**40.3**	**75.2**	**74.0**	**48.0**	**76.5**	**82.4**	**68.4**	**80.0**	**87.2**

Generalizing to Other “Semi-F^3” Datasets

	ShuttleSet [59]		FineDiving [62]		FineGym [51]		SoccerNetV2 [12]		CCTV-Pipe [40]	
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 [64]	55.9	70.6	49.9	87.6	53.6	66.3	46.3	76.1	15.4	33.4
G-TAD [63]	48.2	61.1	52.1	82.6	45.8	51.4	42.3	72.3	31.3	33.6
ActionFormer [65]	62.1	67.5	68.3	92.4	54.0	59.7	43.0	64.6	18.8	29.5
E2E-Spot [23]	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



THANK YOU

Refer to our paper for full experimental results and details

Code and data are available on our project website