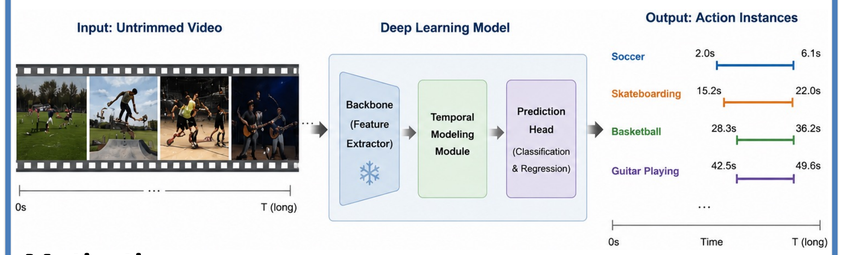
Efficient Spatial-Temporal Focal Adapter with SSM for Temporal Action Detection

YICHENG QIU, Keiji Yanai
The University of Electro-Communications, Tokyo

ICME 2026
Bangkok, Thailand | 5 - 9 July 2026

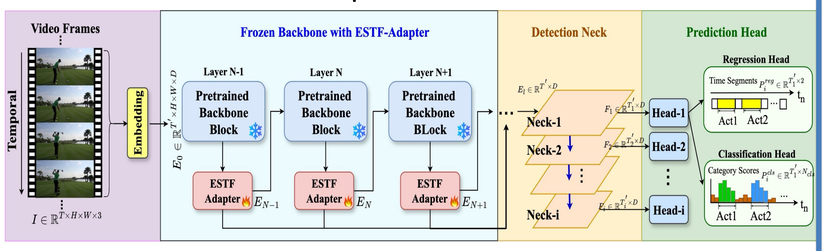
Introduction



Motivations:

We aim to efficiently localize action instances in long, original videos. Given an untrimmed video, Temporal Action Detection (TAD) predictions:

- action categories
- start and end timestamps



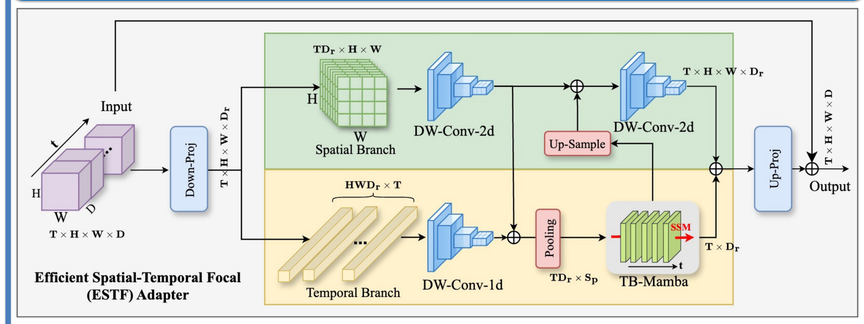
Challenges:

- Redundant background frames in long untrimmed videos
- High computational cost of dense temporal modeling
- Limited long-range dependency modeling in local operators
- Ambiguous start and end boundaries caused by feature over smoothing

Contributions:

- ESTF Adapter for parameter-efficient temporal action detection
- Decoupled spatial-temporal modeling for redundancy reduction, TB-SSM for asymmetric action onset/offset modeling
- Better localization accuracy with lower memory and computation

Method



Frozen Backbone with ESTF Adapter:

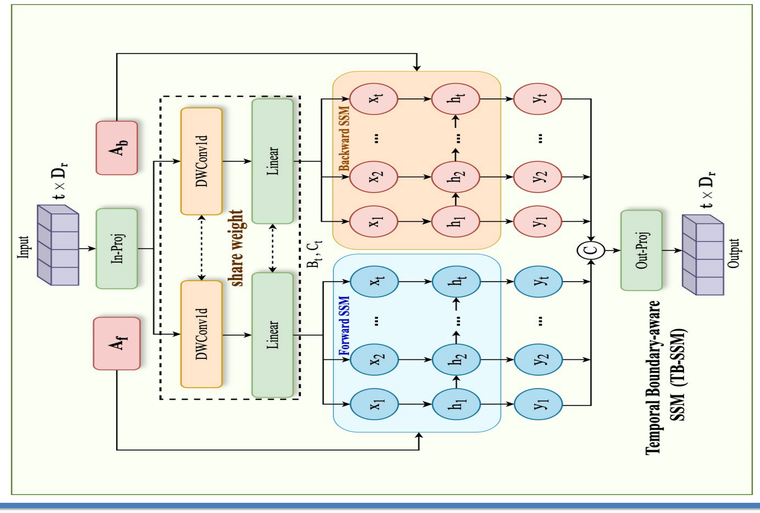
Given video features, we insert lightweight ESTF adapters into frozen pre-trained video blocks. Only adapters and detection heads are trainable, reducing training cost while preserving strong video representations.

ESTF decouples feature adaptation:

Spatial Branch: captures local spatial cues using depth-wise 2D convolution. Temporal Branch: models long-range temporal dynamics after spatial pooling.

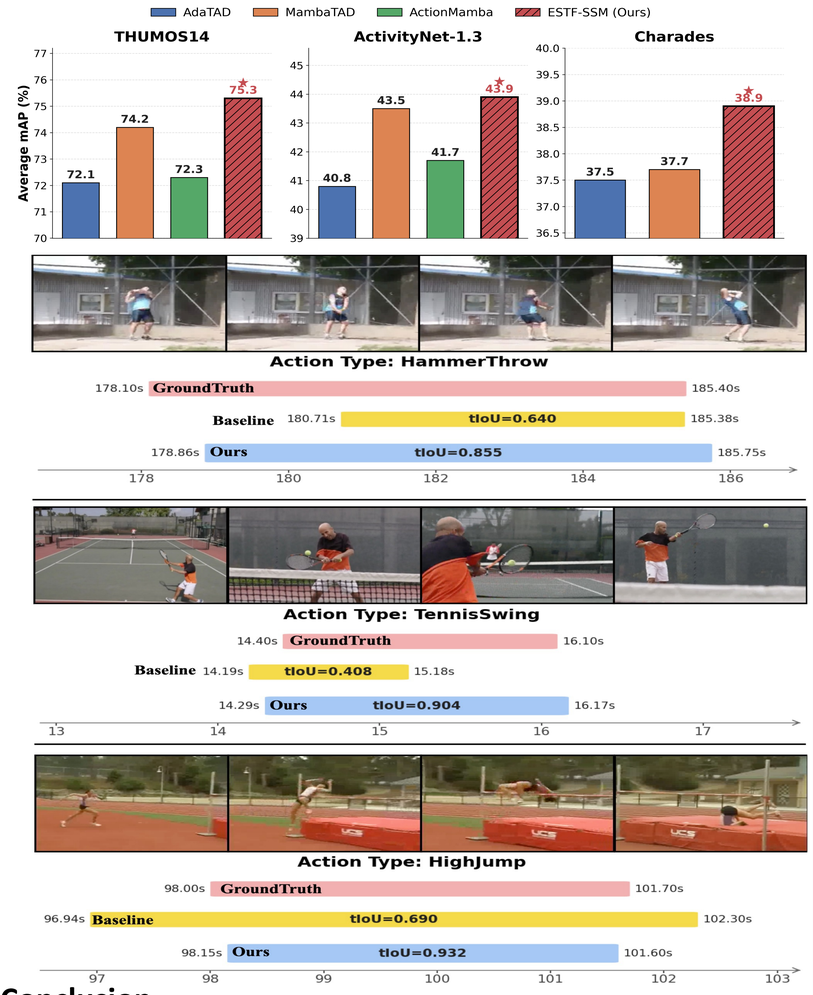
Temporal Boundary-aware SSM:

TB-SSM action onset and offset with asymmetric bidirectional state dynamics. Forward SSM focuses on action start dynamics.Backward SSM focuses on action end dynamics.



Experiment

- Strong accuracy-efficiency trade-off, our method obtains higher mAP with lower memory usage.
- Better temporal boundary localization. Our ESTF-SSM predicts align more closely with ground-truth.



Conclusion

- We propose ESTF, a parameter-efficient adapter that decouples spatial and temporal modeling for temporal action detection.
- TB-SSM captures asymmetric action onsets/offsets via independent forward/backward state dynamics, giving sharper boundary localization.