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

1st YICHENG QIU
Department of Informatics
The University of Electro-Communications
Chofu, Tokyo, Japan
qiu-y@mm.inf.uec.ac.jp

2nd Keiji Yanai*
Department of Informatics
The University of Electro-Communications
Chofu, Tokyo, Japan
yanai@cs.uec.ac.jp

***Abstract*—Temporal human action detection aims to identify and localize action segments within untrimmed videos, serving as a pivotal task in video understanding. Despite the progress achieved by prior architectures like CNN and Transformer models, these continue to struggle with feature redundancy and degraded global dependency modeling capabilities when applied to long video sequences. These limitations severely constrain their scalability in real-world video analysis. State Space Models (SSMs) offer a promising alternative with linear long-term modeling and robust global temporal reasoning capabilities. Rethinking the application of SSMs in temporal modeling, this research constructs a novel framework for video human action detection. Specifically, we introduce the Efficient Spatial-Temporal Focal (ESTF) Adapter into the pre-trained layers. This module integrates the advantages of our proposed Temporal Boundary-aware SSM(TB-SSM) for temporal feature modeling with efficient processing of spatial features. We perform comprehensive and quantitative analyses across multiple benchmarks, comparing our proposed method against previous SSM-based and other structural methods. Extensive experiments demonstrate that our improved strategy significantly enhances both localization performance and robustness, validating the effectiveness of our proposed method.**

***Index Terms*—Video processing, Video understanding, Temporal Action Detection, Temporal Action Localization, State Space Model**

I. Introduction

Temporal Action Detection (TAD) constitutes a fundamental and challenging research area within the broader field of video understanding. The primary objective of TAD is to precisely categorize and localize action instances within untrimmed video streams by determining their start and end timestamps. This capability is pivotal for a wide range of real-world applications, including intelligent video surveillance [1,2,3,4], complex human behavior analysis [5,6,7], and embodied intelligence [8,9,10].

In previous research, diverse model architectures have been developed to tackle the TAD task. Earlier works relied on multi-scale feature pyramids [11] and Graph Convolutional Networks (GCNs) [12] to capture temporal contexts. More recently, Transformer-based architectures [13] have set new benchmarks by leveraging self-attention mechanisms to model global dependencies. However, these existing paradigms face inherent limitations when scaling to long video sequences. CNN-based methods often struggle with limited receptive fields, leading to insufficient modeling of long-term dependencies. Conversely, while Transformers enable global context awareness, their quadratic computational complexity $O(N^2)$ creates severe memory bottlenecks and redundancy, making them inefficient for processing high-resolution, untrimmed videos. Consequently, balancing global temporal reasoning with computational efficiency remains an unresolved bottleneck in current TAD research. Recently, the State Space Model (SSM) architecture [14,15], has emerged as a competitive paradigm, offering linear computational complexity $O(N)$ coupled with robust long-sequence modeling capabilities. This presents a significant opportunity to overcome the scalability constraints of Transformers. However, directly applying SSMs to video understanding is not trivial. Purely global modeling in SSMs can sometimes lead to the oversmoothing [16] issue, causing fine-grained boundary details essential for precise action localization to be submerged by global features. Furthermore, retraining large scale video backbones from scratch is computationally expensive. Therefore, a parameter-efficient architecture [17] is needed to leverage Mamba's linear efficiency while retaining discriminative boundary cues for accurate boundary detection.

In this paper, we propose a new TAD framework centered on the Efficient Spatial-Temporal Focal (ESTF) Adapter. Compared with full fine-tuning on pretrained model, our method integrates a spatial feature processing module and a temporal SSM module into a pre-trained video backbone. ESTF improves feature processing efficiency by decoupling spatial and temporal modeling and then fusing them through feature interaction. Specifically, it first applies spatial downsampling to reduce redundant information, and then uses a parameter-efficient Temporal Boundary-aware SSM (TB-SSM) as the core temporal component. We adopt the selective scan mechanism of SSMs to capture long-range dependencies with linear complexity, and further introduce spatial and temporal focal designs to highlight key boundary changes of video human action features. It helps reduce the oversmoothing issue and enables more robust action representation learning under complex backgrounds. We conduct extensive comparative experi-

This work was supported by JST CRONOS JPMJCS24K4.
* Corresponding author

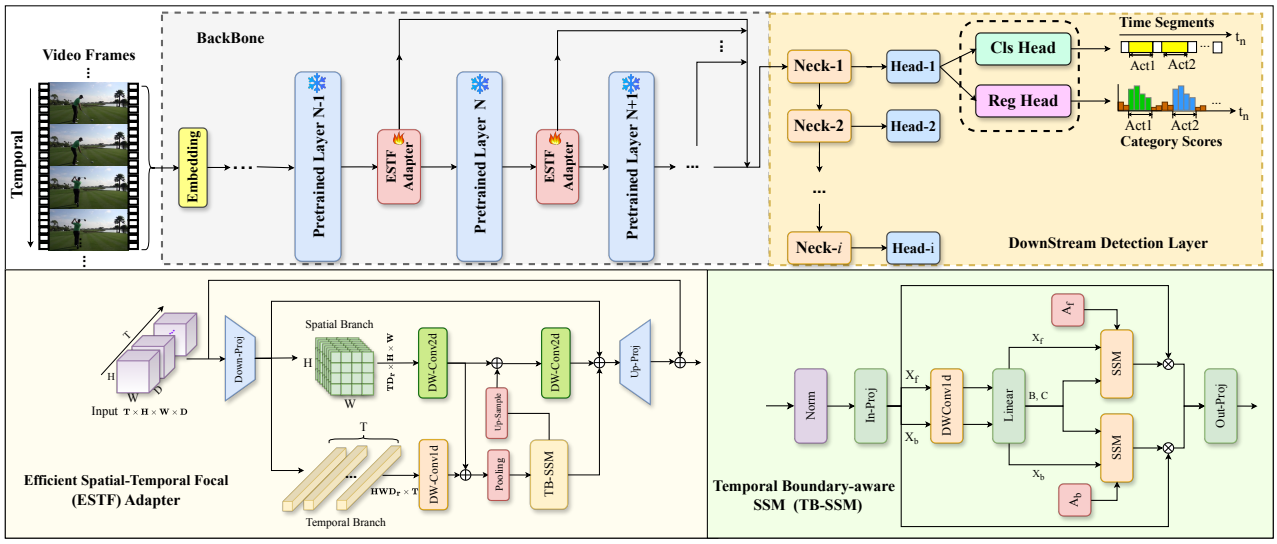


Fig. 1: The architecture of the proposed TAD framework. We integrate ESTF Adapters into the frozen pre-trained backbone layers to adapt representations for temporal action detection. For each adapter, features are processed by two branches: (1) The Spatial Branch that extracts localized structural cues using convolutions, and (2) The Temporal Branch that employs the TB-SSM to capture asymmetric state dynamics for precise boundary regression. Finally, fused features are fed into the detection head.

ments on multiple benchmarks. The results show that, compared with prior SSM-based and Transformer-based methods, our approach consistently improves localization performance and robustness.

The main contributions of this paper are summarized as follows:

- We propose a TAD detection framework equipped with the Efficient Spatial-Temporal Focal (ESTF) Adapter, which integrates the linear-time Mamba architecture into the TAD pipeline and effectively addresses the trade-off between global modeling capability and computational cost.
- Within the ESTF Adapter, we decouple spatial and temporal modeling and perform feature interaction and fusion, enabling the SSM module to better capture long-range dependencies while preserving fine-grained boundary information.
- We propose Temporal Boundary-aware SSM(TB-SSM) structure, where the state transition matrices for the forward and backward directions are parameterized separately to better model the inherent asymmetry of action boundaries.

II. Related Work

A. Temporal Action Detection Methods

ActionFormer [13] is built on the transformer architecture. TriDet [11] is constructed based on the multi-scale feature pyramid. TemporalMaxer [18] uses max-pooling to extract information from the video. ActionMamba [19] utilizes the SSM [14] architecture for global extraction of temporal features. BMN [20] introduces a Boundary-Matching mechanism to estimate temporal boundaries. G-TAD [12] employs graph convolutional networks to video actions. AFSD [21] proposes an anchor-free method. TadTR [22] presents an end-to-end model based on the transformer structure, and ETAD [23] introduces sequentialized video encoding and gradient updating to reduce computational resources.

B. State Space Models of Video Understanding

State Space Model(SSM) [14] is a foundational approach to general sequence modeling. Based on SSM, the subsequent work VisionMamba [24], extended the Mamba architecture to the field of Computer Vision. VideoMamba [25] further applies this architecture to video understanding. ActionMamba [19] adopts the bidirectional parameter-sharing strategy, which reduces computational overhead. MS-Temba [26] employs the multi-scale temporal State Space Model (SSM) architecture for video action detection task.

III. Methodology

This section provides an overview of our proposed TAD framework and presents the internal design details of our core Efficient Spatial-Temporal Focal (ESTF) module. We introduce the overall framework of the video action detection task constructed in this paper, as shown in Figure1. We first define the task object setting. Given an untrimmed video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times D}$, where T, H, W, and D denote the number of frames, height, width, and channel dimension respectively. The TAD task aims to accurately and efficiently predict a set of action instances $\Psi_p = \{\phi_i = (t_s^i, t_e^i, c^i, s^i)\}_{i=1}^{M}$, where M is the number of predicted action instances, and t_s^i, t_e^i, c^i, and s^i denote the start time, end time, category, and confidence score of the i-th action instance, respectively. The ground-truth action instances are represented as $\Psi_g = \{\hat{\phi}_i = (\hat{t}_s^i, \hat{t}_e^i, \hat{c}^i)\}_{i=1}^{N}$, where N is the number of ground-truth action instances. The input video is divided into spatio-temporal patches and embedded as $\mathbf{X} = \text{PatchEmbed}(\mathbf{V}) + \mathbf{E}_{pos}$ where $\mathbf{E}_{pos} \in \mathbb{R}^{T \times D}$ denotes the positional embedding and $\mathbf{X} \in \mathbb{R}^{N \times D}$. N is the

number of patches as $N = T/t_p \times H/h_p \times W/w_p$, t_p, h_p, w_p represent the patch size.

We adopt the pretrained backbone with adapter modules. Given an embedded patch sequence $\mathbf{X} \in \mathbb{R}^{N\times D}$, the video backbone consists of a stack of L pretrained main blocks. The main block is augmented with an adapter module. For the l-th block, the feature propagation is defined as

$$\mathbf{X}'_l = \mathbf{X}_{l-1} + \mathbf{W}_l\big(\mathrm{LN}(\mathbf{X}_{l-1})\big), \tag{1}$$

$$\mathbf{X}_l = \mathbf{X}'_l + \mathcal{A}_l\big(\mathbf{X}'_l\big), \tag{2}$$

where $\mathbf{W}_l(\cdot)$ denotes the pretrained main network. $\mathcal{A}_l(\cdot)$ represents the adapter module, which is responsible for efficient temporal feature refinement.

The Neck layer is introduced to transform the backbone features into a temporally organized representation suitable for detection. Temporal downsampling is applied to aggregate temporal information and generate compact temporal feature sequences. At the i-th Neck layer, the temporal features are obtained as:

$$\mathbf{F_i} = \mathrm{TemporalPool}(\mathbf{F}_{i-1}) \in \mathbb{R}^{T_i\times D}, \tag{3}$$

where T_i denotes the temporal length after downsampling. This process yields a sequence of temporally aligned feature maps $\mathbf{F}_i$, enabling the model to capture actions with varying temporal durations.

Finally, the Head layer is applied to predict temporal action instances from the temporal features $\{\mathbf{F}_i\}$. The head jointly performs action classification and temporal boundary regression, and the final prediction set is formulated as

$$\Psi_p = \mathrm{Head}(\{\mathbf{F}_i\}) = \{\phi_m = (t_s^m, t_e^m, c^m, s^m)\}_{m=1}^{M}, \tag{4}$$

where m indexes the predicted action instances and M denotes the total number of predictions. Predictions Ψ_p from multiple temporal scales are aggregated and further refined by Non-Maximum Suppression (NMS) [27] in post-processing, which removes redundant and highly overlapping proposals while preserving high-confidence detections.

A. Efficient Spatial-Temporal Focal Adapter

As illustrated in the lower-left part of Figure 1, the Efficient Spatial-Temporal Focal (ESTF) Adapter is inserted into selected pretrained backbone blocks as a lightweight adapter to enhance temporal modeling capability with minimal additional computation.

Given the intermediate feature representation $\mathbf{X}'_l \in \mathbb{R}^{N\times D}$ from the l-th pretrained block, the ESTF module first applies a channel down-projection to reduce computational cost:

$$\mathbf{Z} = \mathbf{X}'_l\mathbf{W}_{down}, \quad \mathbf{W}_{down} \in \mathbb{R}^{D\times D_r}, \tag{5}$$

where $D_r < D$ denotes the rank as the reduced channel dimension.

The projected features are then reshaped into spatio-temporal form $\mathbf{Z} \in \mathbb{R}^{T\times H\times W\times D_r}$. The spatial branch focuses on local spatial context modeling by applying depthwise separable convolutions across spatial dimensions:

$$\mathbf{Z}'_s = \mathrm{DWConv}_{2d}(\mathbf{Z}), \tag{6}$$

where $\mathbf{Z}'_s$ serves as the intermediate spatial feature. We combine it with the temporal features for fusion spatial-temporal modeling. After upsampling, the temporal features $\mathbf{Z}'_t$ are added to obtain the result, which is further refined through a second 2D convolution to obtain the final spatial feature representation $\mathbf{Z}_s$:

$$\mathbf{Z}_s = \mathrm{DWConv}_{2d}(\mathbf{Z}'_s + \mathrm{UpSampling}(\mathbf{Z}'_t)). \tag{7}$$

For the temporal branch, we first apply a 1D depthwise convolution. The resulting temporal features are then combined with the spatial features, followed by average pooling to reduce spatial dimensions. Finally, the SSM module is employed to capture global temporal dynamics, as formulated below:

$$\mathbf{Z}'_t = \big(\mathrm{DWConv}_{1d}(\mathbf{Z})\big), \tag{8}$$

$$\mathbf{Z}_t = \text{TB-SSM}\big(\mathrm{AvgPool}(\mathbf{Z}'_t + \mathbf{Z}'_s)\big). \tag{9}$$

This design allows ESTF to efficiently encode both local and long-range temporal information without relying on quadratic self-attention.

The spatial and temporal features are then adaptively fused and projected back to the original channel dimension:

$$\mathcal{A}_l(\mathbf{X}'_l) = \big(\mathbf{Z}_s + \mathbf{Z}_t + \mathbf{Z}\big)\mathbf{W}_{up}. \tag{10}$$

where $\mathbf{W}_{up} \in \mathbb{R}^{D_r\times D}$ denotes the channel up-projection. By decoupling spatial refinement and temporal focal modeling, the ESTF module achieves an effective balance between representation capacity and computational efficiency.

B. Temporal Boundary-aware SSM

The temporal modeling in ESTF is performed by our proposed Temporal Boundary-aware SSM (TB-SSM). We build upon the basic Mamba [14], which efficiently captures long-range dependencies through linear state updates. TB-SSM adopts independent state transition parameters for temporal forward and backward scans, while sharing the input-dependent projection parameters that generate $\mathbf{B}_t, \mathbf{C}_t$.

Given the input sequence $\mathbf{X} \in \mathbb{R}^{HW\times T\times D_r}$, we first apply layer normalization and an input projection:

$$\mathbf{X}^f, \mathbf{X}^b = \mathrm{Split}(\mathbf{W}_{in}(\mathrm{Norm}(\mathbf{X}))), \tag{11}$$

where $\mathbf{W}_{in} \in \mathbb{R}^{D_r\times 2D_r}$ denotes the input projection of the TB-SSM, and two streams $\mathbf{X}^f \in \mathbb{R}^{HW\times T\times D_r}, \mathbf{X}^b \in \mathbb{R}^{HW\times T\times D_r}$ for forward and backward temporal scans split from $\mathbf{X} \in \mathbb{R}^{HW\times T\times 2D_r}$.

$$\mathbf{X}^f = \mathbf{X}, \quad \mathbf{X}^b = \mathbf{X}.\mathrm{Flip}(), \tag{12}$$

For the forward stream, TB-SSM performs selective state updates with an independent transition matrix $\mathbf{A}^f$:

$$\mathbf{h}_t^f = \mathbf{A}^f\mathbf{h}_{t-1}^f + \mathbf{B}_t\mathbf{X}_t^f, \tag{13}$$

$$\mathbf{Y}_t^f = \mathbf{C}_t\mathbf{h}_t^f, \tag{14}$$

TABLE I: Comparison with previous methods on THUMOS14 and ActivityNet-1.3 Datasets.

Method	Backbone	E2E	Mem	THUMOS14						ActivityNet-1.3			
				0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
BSFTAL [28]	InternVideo2 [19]	✗	–	84.2	79.5	73.5	61.8	46.1	69.0	54.8	37.5	9.1	37.1
CLTDR-GMG [29]	InternVideo2 [19]	✗	–	85.3	80.6	74.8	63.2	48.0	70.4	58.3	40.2	9.4	39.5
VideoMamba [25]	InternVideo2 [19]	✗	–	86.5	82.1	76.3	65.1	50.2	71.8	62.1	43.2	9.5	41.6
ActionMamba [19]	InternVideo2 [19]	✗	–	87.1	82.7	76.5	65.6	49.9	72.3	62.2	37.1	9.6	41.7
BDRC-Net [30]	InternVideo2 [19]	✗	–	87.3	83.7	78.1	67.4	53.1	73.9	59.6	38.9	9.8	41.9
AdaTAD [31]	VideoMAEv2 [32]	✓	29.9G	86.8	82.3	76.1	65.3	49.9	72.1	59.8	42.0	9.7	40.8
AdaTAD++ [33]	VideoMAEv2 [32]	✓	30.9G	86.5	81.9	75.6	64.8	49.1	71.6	60.3	43.2	10.1	41.0
MambaTAD [34]	InternVideo2 [35]	✓	30.7G	87.5	83.8	78.3	67.3	52.9	74.2	63.1	44.2	11.0	43.5
ESTF-SSM (Ours)	VideoMAEv2 [32]	✓	28.6G	**89.5**	**85.3**	**78.9**	**69.0**	**54.1**	**75.3**	**63.8**	**44.6**	**11.2**	**43.9**

where $\mathbf{X}_t^f$ denotes the token at time step t. For the backward stream, we apply the same update rule but with a different transition matrix $\mathbf{A}^b$:

$$\mathbf{h}_t^b = \mathbf{A}^b \mathbf{h}_{t-1}^b + \mathbf{B}_t \mathbf{X}_t^b, \tag{15}$$

$$\mathbf{Y}_t^b = \mathbf{C}_t \mathbf{h}_t^b, \tag{16}$$

Finally, we concatenate the bidirectional outputs and apply an output projection to obtain the temporal representation:

$$\mathbf{Z}_t = \mathbf{W}_{out}(\mathbf{Y}^f \| \mathbf{Y}^b.\text{Flip}()), \tag{17}$$

where $\mathbf{W}_{out} \in \mathbb{R}^{2D_r \times D_r}$ denotes the output projection. This design enables direction-specific state dynamics for modeling asymmetric action boundaries, while keeping the overall temporal modeling linear in sequence length.

TABLE II: Comparison with previous methods on Charades Dataset.

Method	Backbone	E2E	0.2	0.5	0.7	Avg.
MS-TST [36]	CLIP [37]	✗	50.1	37.9	19.2	31.9
AAN [38]	CLIP [37]	✗	51.2	38.4	20.7	32.0
MT-Temba [26]	CLIP [37]	✗	51.2	38.4	20.7	32.3
AdaTAD [31]	VideoMAEv2 [32]	✓	53.7	42.9	26.7	37.5
MambaTAD [34]	VideoMAEv2 [32]	✓	54.1	43.2	26.6	37.7
ESTF-SSM (Ours)	VideoMAEv2 [32]	✓	**57.2**	**45.5**	**27.1**	**38.9**

IV. Experiments

A. Datasets and Metrics

We conduct experiments on three datasets for temporal action detection. THUMOS14 [39] contains 413 untrimmed videos with temporal annotations over 20 action categories, and is widely used for evaluating precise temporal localization. ActivityNet-1.3 [40] is a larger benchmark with 19,994 videos and 200 categories, featuring diverse daily activities and long untrimmed videos. Charades [41] focuses on daily indoor activities with frequent action co-occurrence and complex temporal structures, providing a challenging testbed for long-range temporal reasoning.

We use the mean Average Precision (mAP) as the primary metric to evaluate the prediction results. The mAP is calculated by averaging the precision results across different temporal Intersection over Union (tIoU) thresholds. The tIoU is defined as the ratio of the intersection to the union of the predicted and ground-truth time intervals.

B. Implementation Details

We implemented our method on NVIDIA A100 GPUs, utilizing CUDA 12.8 within the PyTorch 2.8 environment. Throughout the model training process, we adopted the AdamW [42] optimizer and employed a scheduler that integrates linear warm-up and cosine annealing to adjust the learning rate, which was set to 1e-4, with a batch size of 8. For video data, the sampling length of video frames was fixed at 768. Videos exceeding this length were subsampled to the maximum length, while those shorter were padded. Each frame was resized to a resolution of 224×224 pixels.

C. Evaluation and Comparison

We evaluate our method on the THUMOS14 and ActivityNet-1.3 benchmarks, comparing against both end-to-end (E2E) and non-E2E baselines. The quantitative results are presented in Table I. On THUMOS14, ESTF-SSM achieves superior average mAP, maintaining a consistent lead across all evaluated tIoU thresholds. This trend is particularly evident at high tIoU, indicating that our boundary-aware temporal modeling significantly refines boundary regression precision rather than merely improving coarse detection. On ActivityNet-1.3 dataset, ESTF-SSM demonstrates strong generalization ability by outperforming competitive baselines. The improvements are consistent across various overlap thresholds, verifying that our method effectively handles long-range temporal reasoning. Furthermore, regarding computational efficiency, our approach maintains a lower memory footprint compared to other E2E baselines, striking a favorable balance between performance and resource consumption.

We further extend our evaluation to the Charades dataset, as shown in Table II. ESTF-SSM consistently surpasses previous methods across all metrics. The performance gap is most distinct at tighter tIoU thresholds, which supports our motivation that modeling asymmetric boundary dynamics is crucial for preserving discriminative cues in scenarios involving complex activities and heavy background clutter.

D. Qualitative Analyses

Figure 2 shows the qualitative visualization results for the video action detection task, where the horizontal axis represents time. The colored bars represent the predicted action segments with their corresponding categories, compared alongside the Ground Truth intervals representing the actual

TABLE III: Ablation study on the effect of each component in the proposed method.

Components				THUMOS14	ActivityNet-1.3
Spatial	Temporal	S-T Fusion	TB-SSM	mAP Avg.	mAP Avg.
✗	✓	✗	✓	74.4	43.0
✓	✗	✗	✗	72.3	41.8
✓	✓	✗	✓	73.9	42.9
✓	✓	✓	✗	73.7	42.6
✓	✓	✓	✓	**75.3**	**43.9**

actions. Each action category is assigned a distinct color. Compared with the previous method [34], our TB-SSM method demonstrates superior capability in recognizing action start and end boundaries.

E. Ablation Studies

We conduct comprehensive ablation studies on THUMOS14 and ActivityNet-1.3 to validate the contribution of each component in ESTF. The results are summarized in Table III. First, removing the temporal branch results in the most significant performance degradation, confirming that temporal modeling is the primary driver for long-range action reasoning. Second, discarding the spatial refinement branch also leads to consistent drops, suggesting that lightweight spatial processing remains essential for retaining localized appearance cues that aid in temporal boundary discrimination. Third, employing a simple parallel structure without the spatial-temporal fusion mechanism yields suboptimal results, indicating that explicit feature interaction is necessary to effectively couple spatial evidence with temporal dynamics. Finally, replacing TB-SSM with a variant lacking the proposed boundary-aware asymmetric state dynamics leads to inferior performance. This validates that direction-specific temporal state transitions are vital for capturing the intrinsic asymmetry between action onsets and offsets. Overall, the full model delivers the best results, demonstrating that these components contribute complementarily to accurate and robust temporal localization.

TABLE IV: Ablation study on different temporal modeling strategies.

Temporal Module Strategy	THUMOS14	ActivityNet-1.3
	mAP Avg.	mAP Avg.
w/o	72.3	41.8
Transformer [43]	72.4	42.0
VideoMamba [25]	73.6	42.4
ActionMamba [19]	73.2	42.7
TB-SSM(Ours)	**75.3**	**43.9**

In Table IV, we further investigate various temporal modeling strategies. Our experiments show that Transformer-based modules offer only marginal gains compared to the baseline without temporal modeling, whereas SSM-based designs consistently yield better performance. Among them, linear-time state-space models like VideoMamba and ActionMamba improve detection accuracy, confirming their suitability for long untrimmed sequences. However, our TB-SSM achieves the highest performance, demonstrating that introducing boundary-aware asymmetric temporal dynamics provides a stronger inductive bias for temporal localization than conventional bidirectional or shared-parameter SSM designs.



Fig. 2: Qualitative results of our proposed method and previous method on THUMOS14 Dataset for Video Action Detection task.

V. Conclusion

We propose a novel temporal action detection framework, ESTF-SSM, which integrates the ESTF Adapter into a frozen video backbone. The ESTF module effectively combines the strengths of the SSM for long-range temporal modeling with efficient spatial refinement. By decoupling spatial and temporal processing and introducing the TB-SSM with asymmetric state dynamics, our approach successfully addresses the challenges of feature redundancy and boundary ambiguity in long video sequences. Extensive experiments on THUMOS14, ActivityNet-1.3, and Charades benchmarks demonstrate that our method significantly enhances localization performance and robustness compared to state-of-the-art methods. Future work will explore extending this efficient adaptation strategy to online action detection and multi-modal video understanding tasks.

References

[1] J. Yang, K. Liu, M. Zhao, and S. Li, "Video process detection for space electrostatic suspension material experiment in china's space station," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107804, 2024.

[2] F. Hamann, S. Ghosh, I. J. Martinez, T. Hart, A. Kacelnik, and G. Gallego, "Low-power continuous remote behavioral localization with event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 612–18 621.

[3] T. Geng, T. Wang, Y. Zhang, J. Duan, W. Guan, and F. Zheng, "Uniav: Unified audio-visual perception for multi-task video localization," *arXiv preprint arXiv:2404.03179*, 2024.

[4] Y. Qiu, F. Sha, L. Niu, and G. Zhang, "Fire anomaly detection based on low-rank adaption fine-tuning and localization using gradient filtering," *Applied Soft Computing*, p. 112782, 2025.

[5] A. Xarles, S. Escalera, T. B. Moeslund, and A. Clapés, "Astra: An action spotting transformer for soccer videos," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023, pp. 93–102.

[6] Y. Qiu, L. Niu, and F. Sha, "Multipath 3d-conv encoder and temporal-sequence decision for repetitive-action counting," *Expert Systems with Applications*, vol. 249, p. 123760, 2024.

[7] Y. Qiu, F. Sha, and L. Niu, "Efficient temporal attention with state space model for temporal action localization," in *International Conference on Neural Information Processing*. Springer, 2024, pp. 183–197.

[8] S. Cheng, K. Fang, Y. Yu, S. Zhou, B. Li, Y. Tian, T. Li, L. Han, and Y. Liu, "Videgothink: Assessing egocentric video understanding capabilities for embodied ai," *arXiv preprint arXiv:2410.11623*, 2024.

[9] B. Zhao, J. Fang, Z. Dai, Z. Wang, J. Zha, W. Zhang, C. Gao, Y. Wang, J. Cui, X. Chen *et al.*, "Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces," *arXiv preprint arXiv:2503.06157*, 2025.

[10] A. Suglia, C. Greco, K. Baker, J. L. Part, I. Papaioannou, A. Eshghi, I. Konstas, and O. Lemon, "Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding," *arXiv preprint arXiv:2406.13807*, 2024.

[11] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, "Tridet: Temporal action detection with relative boundary modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 857–18 866.

[12] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 156–10 165.

[13] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 492–510.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[15] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.

[16] A. Behrouz and F. Hashemi, "Graph mamba: Towards learning on graphs with state space models," in *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 2024, pp. 119–130.

[17] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[18] T. N. Tang, K. Kim, and K. Sohn, "Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization," *arXiv preprint arXiv:2303.09055*, 2023.

[19] G. Chen, Y. Huang, J. Xu, B. Pei, Z. Chen, Z. Li, J. Wang, K. Li, T. Lu, and L. Wang, "Video mamba suite: State space model as a versatile alternative for video understanding," *arXiv preprint arXiv:2403.09626*, 2024.

[20] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3889–3898.

[21] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Learning salient boundary feature for anchor-free temporal action localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3320–3329.

[22] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022.

[23] S. Liu, M. Xu, C. Zhao, X. Zhao, and B. Ghanem, "Etad: Training action detection end to end on a laptop," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4524–4533.

[24] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[25] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2025, pp. 237–255.

[26] A. Sinha, M. S. Raj, P. Wang, A. Helmy, and S. Das, "Ms-temba: Multi-scale temporal mamba for efficient temporal action detection," *arXiv preprint arXiv:2501.06138*, 2025.

[27] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[28] J. Xu, Y. Zhang, W. Zhou, and H. Liu, "Bfstal: Bidirectional feature splitting with cross-layer fusion for temporal action localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[29] Q. Li, D. Liu, J. Kong, S. Li, H. Xu, and J. Wang, "Temporal action localization with cross layer task decoupling and refinement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4878–4886.

[30] Z. Fang, J. Yu, and R. Hong, "Boundary discretization and reliable classification network for temporal action detection," *IEEE Transactions on Multimedia*, 2025.

[31] S. Liu, C.-L. Zhang, C. Zhao, and B. Ghanem, "End-to-end temporal action detection with 1b parameters across 1000 frames," *arXiv preprint arXiv:2311.17241*, 2023.

[32] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 549–14 560.

[33] T. Agrawal, A. Ali, A. Dantcheva, and F. Bremond, "Scaling action detection: Adatad++ with transformer-enhanced temporal-spatial adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 12 222–12 231.

[34] H. Lu, Y. Yu, S. Lu, D. Rajan, B. P. Ng, A. C. Kot, and X. Jiang, "Mambatad: When state-space models meet long-range temporal action detection," *arXiv preprint arXiv:2511.17929*, 2025.

[35] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang *et al.*, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv preprint arXiv:2212.03191*, 2022.

[36] R. Dai, S. Das, K. Kahatapitiya, M. S. Ryoo, and F. Brémond, "Ms-tct: Multi-scale temporal convtransformer for action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 041–20 051.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[38] R. Dai, S. Das, M. S. Ryoo, and F. Brémond, "Attributes-aware network for temporal action detection," in *BMVC*, 2023.

[39] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," 2014.

[40] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[41] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European conference on computer vision*. Springer, 2016, pp. 510–526.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.