



# TAL4Tennis: Temporal Action Localization in Tennis Videos Using State Space Models

Ahmed Jouini<sup>1</sup> <sup>a</sup>, Mohamed Ali Lajnef<sup>1</sup>, Faten Chaieb-Chakchouk<sup>1</sup> <sup>b</sup> and Alex Loth<sup>2</sup>

<sup>1</sup>Paris Panthéon-Assas University, EFREI Research Lab, F-94800, Villejuif, France

<sup>2</sup>French Tennis Federation, France

{ahmed.jouini, mohamed-ali.lajnef, faten.chakchouk}@efrei.fr, aloth@fft.fr

**Keywords:** State Space Models, Temporal Action Localization, Tennis Videos.

**Abstract:** Temporal action localization is a classic computer vision problem in video understanding with a wide range of applications. In the context of sports videos, it is integrated into most of the current solutions used by coaches, broadcasters and game specialists to assist in performance analysis, strategy development, and enhancing the viewing experience. This work presents an application study on temporal action localization for tennis broadcast videos. We study and evaluate a foundational video understanding model for identifying tennis actions in match footage. We explore its architecture, specifically the state space model, from video input to the prediction of temporal segments and classification labels. Our experiments provide findings and interpretations of the model's performance on tennis data. We achieved an average mean Average Precision (mAP) of 66.14% over all thresholds on the TenniSet dataset, surpassing the other methods, and 96.16% on our private French Open dataset.

## 1 INTRODUCTION

The use of video analysis in tennis has revolutionized player development, allowing for a deeper understanding of technique, strategy, and overall performance (Jiang and He, 2021; Peng and Tang, 2022). For example, court positioning can significantly impact a player's ability to execute shots and cover the court effectively against technical opponents. By detecting these positions from videos and with advanced data visualisation, players can identify their strengths and weaknesses, understand their opponents' tactics, and develop appropriate tactical plans.


Tennis is a fast racket sport that presents challenges when training a model to localize events over time. For example, in the case of a return from an out-of-bounds serve or a double fault, the model must be able to classify this event as a non-game rather than an exchange. The idea is to be able to localize each fine-grained action in a given input sequence from a tennis match. By doing this, we can extract detailed insights into the game's progress and players' strategies.


Temporal Action Localization (TAL) aims to identify the start and end timestamps of actions in a video

stream (Wang et al., 2024). Real-world applications of TAL remain challenging in computer vision. In this context, several large public datasets that present various challenges, such as fine-grained event detection including actions of varying lengths captured from different viewpoints has been proposed (Idrees et al., 2017; Caba Heilbron et al., 2015; Liu et al., 2022; Damen et al., 2018; Zhao et al., 2019; Grauman et al., 2022).

In this paper, we present a temporal tennis action localization model based on State Space Models. We are interested in localizing Service, Exchange, and Non-Game phases. This model was evaluated on two datasets: TenniSet (Faulkner and Dick, 2017) and a private French Open dataset. Our main contributions could be summarized as follows:

1. A deep understanding of a State Space Model (Chen et al., 2024) and its application to temporal tennis actions localization.
2. Obtained results on TenniSet outperforms the SOTA on a tennis temporal actions localization.
3. A private French Open dataset with three main action phases annotation: Service, Exchange, and Non-Game.

<sup>a</sup>  <https://orcid.org/0009-0000-6491-2875>

<sup>b</sup>  <https://orcid.org/0000-0002-2968-2426>

## 2 RELATED WORK

To the best of our knowledge, TenniSet is the only study that has introduced a comprehensive dataset specifically designed for TAL in broadcast tennis videos, along with proposing models and evaluation frameworks. TAL has evolved significantly with advancements in deep learning and computer vision (Shou et al., 2016; Paul et al., 2018; Nguyen et al., 2019; Shi et al., 2020; Liu et al., 2021b; He et al., 2022; Rizve et al., 2023; Zhang et al., 2022). This section reviews related works in TAL, focusing on early methods and presenting a taxonomy of recently proposed works. Preliminary attempts (Shou et al., 2016; Zelnik-Manor and Irani, 2001; Gorelick et al., 2007; Escorcia et al., 2016) relied on the sliding window approach, where fixed-length windows slide across the video timeline to propose potential action segments. Z. Shou et al. (Shou et al., 2016) utilize these sliding windows to segment videos, creating simple yet effective action proposals. Despite their simplicity, these methods often face inefficiencies due to fixed window lengths and high computational costs from processing numerous overlapping segments.

To overcome these limitations, researchers have explored classification-based approaches that focus on identifying actions by analyzing individual frames and subsequently linking these classifications to form continuous action sequences. These models operate by independently classifying each frame and then associating the individual classifications to construct comprehensive action instances. Initially, they handle spatial information by classifying each frame independently and then capture temporal evolution using predefined rules. For instance, certain methods apply a threshold to the classification score of each frame, treating consecutive frames that surpass this threshold as an action instance (Shou et al., 2017), (Piergiovanni and Ryoo, 2019). Additionally, a couple of studies (Lin et al., 2018), (Zhao et al., 2020) assess the likelihood of start and end moments for each frame and associate potential start and end moments to form action instances. While frame classification can achieve accurate action boundaries, they may struggle with background noise and require multiple separated preprocessing (Wang et al., 2024).

Another approach involves proposal-based models, which generate action proposals and subsequently classify each proposal to identify specific actions. To generate proposals, early methods used sliding window strategies as specified in (Shou et al., 2016) (Escorcia et al., 2016) to find the proposals. Alternative methodologies have also been developed, such as temporal action grouping (Zhao et al., 2017; Zhou

et al., 2021), where frames or video segments that likely contain specific actions are grouped together. Another approach is the dense enumeration strategy (Liu et al., 2021a; Lin et al., 2019), which systematically generates a large number of potential action proposals throughout the video, without initially discriminating based on the probability of containing actionable content. Regarding the problem of classification, early explorations (Shou et al., 2016; Escorcia et al., 2016) independently performed classification and regression for each proposal, while more recent approaches (Chao et al., 2018; Zeng et al., 2022) leverage relationships among multiple proposals using graph convolutional networks to enhance detection performance. These models achieve high recall rates and reduce background errors, which are false identifications of actions where none exist. However, they encounter significant computational challenges due to the large number of proposals they generate (Wang et al., 2024).

To address sliding window and frame classification limits, anchor-based methods were introduced. An anchor is a predefined moment in a sequence that serves as a reference point for detecting actions within a video. Gao et al. (Gao et al., 2017) propose an anchor mechanism in their framework to increase the flexibility of proposals and facilitate action detection through regression using default anchors. This approach employs Cascaded Boundary Regression (CBR) to iteratively refine temporal boundaries. The methodology was advanced by TAL-Net (Chao et al., 2018), which enhanced the two-stage anchor-based framework by extending receptive fields and incorporating multi-stream feature fusion. The anchor mechanism ensures more accurate initial proposals for action segments. Given the impressive performance of Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016) on numerous vision and spatiotemporal tasks (Myung et al., 2024; Cheng et al., 2024), AGCN-P-3DCNNs (Li et al., 2020) uses intra-attention mechanisms to capture long-range dependencies within each action proposal, subsequently updating the node matrix of the Intra Attention-based GCN. Additionally, inter-attention mechanisms are employed to learn dependencies between different action proposals, forming the adjacency matrix of the Inter Attention-based GCN. The intra and inter attentions are then fused to simultaneously model both intra long-range dependencies and inter dependencies, enhancing the overall action proposal modeling process. Recent advancements have explored action boundary regression, which bypass the need for predefined anchors (Lin et al., 2021; Tang et al., 2019; Li et al., 2019). Instead, these methods directly predict

action boundaries without relying on reference temporal points for refining output offsets. Lin et al. (Lin et al., 2021) propose to refine initial action proposals, generated by temporal pyramid features, through a saliency-based module, which adjusts boundaries, class scores, and quality scores, all within a fully end-to-end framework without the need for preprocessing. Anchor-free methods have shown to be both robust and straightforward in design, but they can struggle with accurately determining the center of action segments, leading to inaccurate localization (Wang et al., 2024).

### 3 TAL4Tennis ARCHITECTURE

TAL4Tennis is a one-stage anchor-based model whose core element, the Decomposed Bidirectionally Mamba block (DBM), is built upon a state space model (SSM) (Chen et al., 2024). Traditionally, SSMs have been employed to represent the evolution of dynamic systems through state variables. Recent approaches (Gu and Dao, 2024; Gu et al., 2021) have demonstrated that by utilizing SSMs with carefully designed state matrices  $A$ , it is possible to effectively manage long-range dependencies without incurring prohibitive computational costs. In their research (Voelker and Eliasmith, 2018), Voelker and Eliasmith investigated how the brain encodes temporal information, identifying SSMs as highly effective tools for modeling the "time cells" located in regions such as the hippocampus and cortex. Building upon their neuroscience findings, they were pioneers in applying SSMs to the field of deep learning, thereby establishing a novel intersection between these disciplines.

As shown in Figure 1, input video undergoes feature extraction via a large Vision Transformer (ViT) (Alexey, 2020) model, pretrained on a hybrid dataset and fine-tuned on the Kinetics-710 dataset, and then through a DBM block. The features are further refined by applying normalization without additional transformations. During the forward pass, the generator processes the input feature maps, retrieves corresponding buffered points, and returns a list of points for each Feature Pyramid Network (FPN) level. During training, the model uses ground truth segments and classes to label anchors, creating ground truth offsets and labels. This process involves concatenating points and calculating the distance of each point to segment boundaries. Points within action segments are identified using a center sampling strategy, and the regression range is limited for each location. If multiple actions overlap, the shortest duration segment is

selected. The final classification and regression targets are normalized and used to update the model. During inference, offsets are applied to the points to finalize action localization in the videos. For the classification task, we implement the Sigmoid Focal Loss (Lin et al., 2017), as described by the authors, which modifies the standard binary cross-entropy loss to address class imbalance. The loss is formulated as:

$$\mathcal{L}_{\text{focal}}(p_t) = -\alpha_t \cdot (1 - p_t)^\gamma \cdot \log(p_t), \quad (1)$$

where  $p_t$  is the model's estimated probability for the true class,  $\gamma$  indicates a focusing parameter used to balance easy and hard examples,  $\alpha_t$  represents a weighting factor defined as:

$$\alpha_t = \alpha \cdot y + (1 - \alpha) \cdot (1 - y). \quad (2)$$

For the regression task, we employ the Centered Generalized Intersection over Union (GIoU) Loss (Rezatofghi et al., 2019), specifically adapted for 1D event localization. Given the predicted offsets  $\mathbf{o} = (o_1, o_2)$  and ground truth offsets  $\mathbf{o}^{\text{gt}} = (o_1^{\text{gt}}, o_2^{\text{gt}})$ , the GIoU loss is defined as:

$$\mathcal{L}_{\text{GIoU}} = 1 - \frac{\min(o_1, o_1^{\text{gt}}) + \min(o_2, o_2^{\text{gt}})}{(o_1 + o_2) + (o_1^{\text{gt}} + o_2^{\text{gt}}) - \min(o_1, o_1^{\text{gt}}) - \min(o_2, o_2^{\text{gt}})}. \quad (3)$$

This loss ensures better alignment between predicted and ground truth offsets by penalizing non-overlapping regions and encouraging tighter bounding around the target events.

#### 3.1 State Space Models

To better understand the application of SSMs in our study, we present its continuous representation as a linear time invariant system expressed by the following equations:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \quad (4)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t), \quad (5)$$

where  $\mathbf{u}(t)$  represents the input vector, which is the sequence of features extracted and preprocessed by the backbone network from the input video data. These features act as external stimuli driving changes in the system's internal state. The state vector  $\mathbf{x}(t)$  captures the internal dynamics of the system at time  $t$ . The state transition matrix  $\mathbf{A}(t)$  determines how the current state evolves over time, while the input matrix  $\mathbf{B}(t)$  modulates the influence of the input features on the state vector. An important aspect of the SSM implementation is the discretization of these continuous equations. This enables the transition from the continuous formulation of SSMs to their recursive and convolutive representations. By doing this we can handle

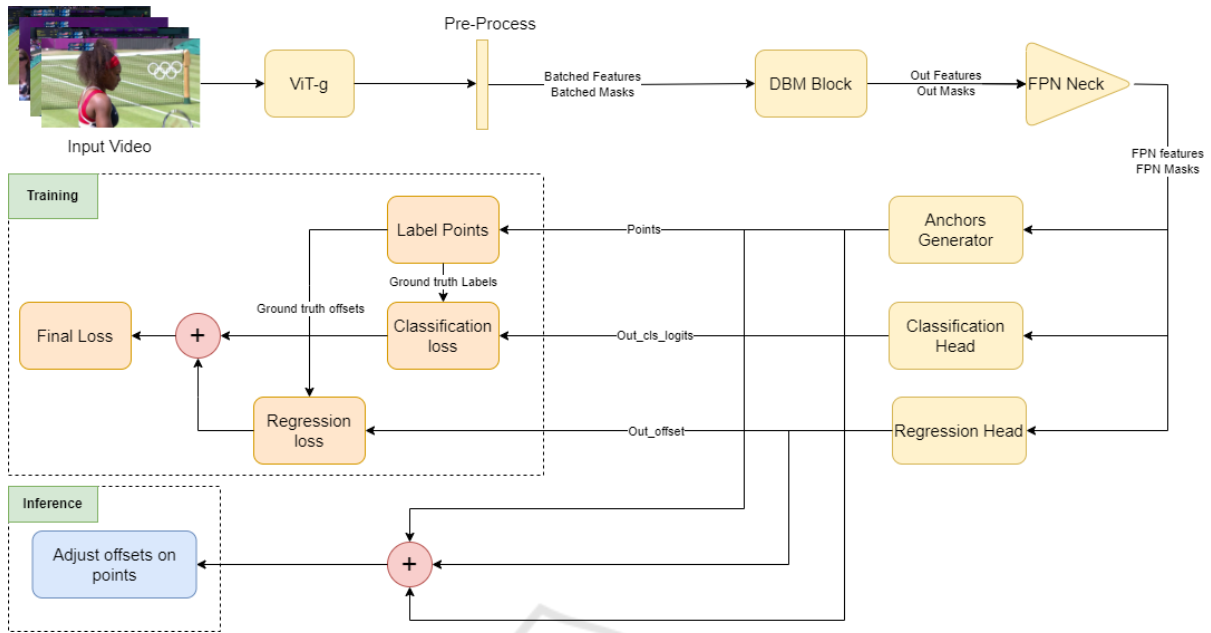


Figure 1: Overview of the TAL4Tennis architecture.

the temporal dynamics in discrete time steps, which align with the sequential nature of video data, and update the state vector  $\mathbf{x}$ . The direct transmission matrix  $\mathbf{D}$  allows for any immediate influence of the input features  $\mathbf{u}$  on the output  $\mathbf{y}$ . In deep learning SSMs,  $\mathbf{D}$  is often set to zero to simplify the model without affecting performance. The final output  $\mathbf{y}$  represents the processed features that are subsequently fed into the FPN neck for further refinement and action localization.

### 3.2 Classification Logits and Regression Offsets

Classification logits and regression offsets are computed for each feature map at different stages of the FPN. The feature maps are processed through a series of one-dimensional convolutional layers, followed by an activation function and a normalization layer. For classification, the model produces logits for each point in the feature map through a sequence of 1D convolutional layers. The output tensor is three-dimensional: the first dimension is the batch size, the second is the length of the temporal sequence in the feature map, and the third is the number of classes.

Similar to classification, regression is performed by shared 1D convolutional heads with a stride of 1. In the forward pass, FPN features and masks are passed to generate output offsets with shape  $[B, 2, T_{\max}]$ , where  $B$  is the batch size, 2 represents the start and end offsets for each action, and  $T_{\max}$  is the

maximum number of actions per batch, with padding applied as needed. The final convolutional layer reduces the feature dimension to two values for each point.

## 4 EXPERIMENTAL RESULTS

### 4.1 Datasets

#### The TenniSet Dataset

It consists of five singles matches from the 2012 London Olympics, all played on grass courts (Faulkner and Dick, 2017). These matches were sourced from YouTube, with each video having a resolution of  $1280 \times 720$  and a frame rate of 25 fps. The matches are annotated with sequences related to specific tennis events, each belonging to a distinct event type. Eleven temporal event categories are introduced. Each frame in a match is assigned a label corresponding to one of the eleven distinct classes as shown in Figure 2. Frames labeled as "other" correspond to non-game footage, including instances such as replays, crowd shots, pauses, and other similar content. The terms "far" and "near" describe the player's distance from the camera, indicating whether they are far away or close. The label "in" denotes that the ball has landed within the boundaries of the court, whereas "fault" indicates that the ball has landed outside. The term "let" is used when the ball makes contact with the net dur-

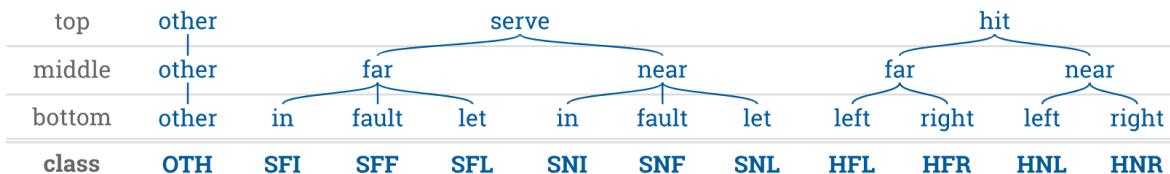


Figure 2: Overview of all of the eleven classes (Faulkner, 2024).

ing a serve but remains in play. "Left" and "right" are used to specify the side of the court where the player hits the ball, according to the camera's perspective and the side of the player's body that makes contact with the ball. These terms are chosen instead of "forehand" and "backhand" to equally apply to both left- and right-handed players.

**Original Split.** The authors divided the five videos into training, validation, and test sets. We notice that some classes have very low occurrences. This imbalanced split poses problems in class-wise evaluation. In the validation set, we have only one sample for the SFL and SNL categories comparing to the 160 samples for the OTH class. The imbalanced nature of the dataset poses substantial challenges for class-wise evaluation and overall model performance. Specifically, the model may develop a bias towards the majority class, resulting in poor performance on minority classes, as it struggles to effectively learn their patterns.

**Games-Based Split.** We propose a different split that includes only training and validation sets, as presented in Table 1. Each match video is segmented into tennis games, and randomly assigned to either the training set or the validation set. By doing this, we ensure that the model will process full actions which were previously cut before its end, therefore maintaining the continuity of spatiotemporal information.

Table 1: Label counts in the modified TenniSet.

	Train		Validation		
	#	%	#	%	
<b>Games</b>	82	69.5	36	30.5	
<b>Events</b>	OTH	2094	71.0	856	29.0
	SFI	269	70.4	113	29.6
	SFF	93	75.0	31	25.0
	SFL	20	76.9	6	23.1
	SNF	97	75.8	31	24.2
	SNL	10	83.3	2	16.7
	HFL	443	73.8	157	26.2
	HFR	438	66.8	218	33.2
	HNL	456	68.1	214	31.9
	HNR	452	72.3	173	27.7

## French Open Dataset

It consists of 19 games from French Open broadcast matches with high resolution delivered by the french federation of tennis, which were semi-automatically annotated using our pre-trained model, followed by a manual correction of false predictions. The dataset contains 161 instances of Service (34.11%), 136 instances of Exchange (28.81%), and 175 instances of Other (37.08%). This dataset will be used exclusively for the test phase to evaluate the proposed model's ability to generalize to clay courts.

## 4.2 Experiments Details

In our experiments, we analyzed the influence of different optimizers and hyper-parameter settings on model performance, including the maximum sequence length per video, the training batch size, as well as the type and feature stride of the backbone network. All details and results are presented in Table 2. 'Max Seq Len' refers to the maximum number of feature vectors (frames) the model processes in a single pass. For videos exceeding this length, the data loader divides them into manageable chunks, each processed independently. This approach enables the model to efficiently handle long videos without exceeding memory or computational constraints. Temporal Intersection over Union (tIoU) is a metric commonly used to evaluate the accuracy of temporal action localization. It is defined as the ratio of the intersection duration (overlap) of the predicted and ground truth intervals to their union duration. The fine-tuned model used in our study was initially pre-trained on the Thumos dataset (Ildrees et al., 2017), with all layers frozen except the classification and regression heads. For the optimizer, AdamW's decoupled weight decay improves generalization performance of the model (Loshchilov and Hutter, 2017). Additionally, we found that reducing the feature stride from 4 to 2 slightly increases the average mAP. Figure 4 shows the evolution of average mAP during validation for 3 different backbones. We found that Mamba proves to be more effective than ConvTransformer and convolution-based backbones in capturing and modeling visual features throughout the training epochs.

Table 2: mAP results of TAL4Tennis with different backbones and hyper-parameters on TenniSet-Games-based split dataset.

Finetune	Backbone	Feature Stride	Max Seq Len	Batch	Opt	tIoU					
						0.3	0.4	0.5	0.6	0.7	Avg
✓	Mamba	4	4608	2	AdamW	59.00	54.15	42.92	28.94	13.17	39.64
×	Mamba	4	4608	2	AdamW	80.73	80.46	79.95	77.92	72.25	78.36
×	Mamba	4	4608	2	SGD	40.78	37.87	31.58	21.51	10.28	28.40
×	Mamba	4	9216	2	AdamW	81.28	80.92	80.43	78.52	72.52	78.68
×	Mamba	2	9216	2	AdamW	80.78	80.61	80.26	78.45	<b>74.51</b>	78.93
×	Mamba	2	9216	1	AdamW	<b>81.76</b>	<b>81.50</b>	<b>81.15</b>	<b>79.00</b>	<b>74.51</b>	<b>79.26</b>
×	Convolution	2	9216	1	AdamW	75.04	74.89	74.59	72.74	65.81	72.62
×	ConvTransformer	2	9216	1	AdamW	76.99	76.75	76.47	74.23	66.29	74.15

### 4.3 Performance Comparison

Table 3 reports The results with 11 classes on the original split, including comparative evaluation with results from TenniSet’s original paper and ActionFormer (Zhang et al., 2022), a transformer based model for TAL. the data used for training validation and test are the same. TAL4Tennis achieves a higher average mAP of 66% across various tIoU thresholds, compared to 62% for the Bi-Directional RNN model with window of size  $w_{rnn} = 40$  and 61% for ActionFormer. At lower thresholds, we notice that TAL4Tennis performs slightly below the referenced model. However, at much more challenging thresholds, it maintains better temporal precision. At tIoU thresholds of 0.7 and 0.9, a decrease in the precision of the other models is observed, as these thresholds require a higher degree of overlap and impose stricter constraints on the boundaries for positive localization.

Table 3: Comparisons for event detection on TenniSet-Original split: mAP over different IoU thresholds  $\alpha$ .

Model	$\alpha$					
	0.1	0.3	0.5	0.7	0.9	Avg
<b>CNN+Pooling</b>	0.89	0.86	0.78	0.52	0.04	0.62
<b>Bi-D RNN</b> ( $w_{rnn} = 25$ )	0.89	0.87	0.79	0.52	0.03	0.62
<b>Bi-D RNN</b> ( $w_{rnn} = 40$ )	<b>0.90</b>	<b>0.88</b>	<b>0.81</b>	0.48	0.02	0.62
<b>ActionFormer</b>	0.75	0.74	0.73	0.66	0.17	0.61
<b>TAL4Tennis</b>	0.80	0.79	0.76	<b>0.71</b>	<b>0.22</b>	<b>0.66</b>

### 4.4 Coarse-Grained Experiment

The results in Table 3 show that the model struggles to accurately identify the 11 detailed classes, particularly those that are rare and have fewer examples in the dataset. Thus, we decided to experiment with a coarse-grained labeling strategy. We annotated the videos into three main classes: "Other", "Service", and "Exchange" inspired by the logical flow of a tennis match, as shown in Figure 3. If the serve is successful and the ball lands within the court (labeled as "in"), the match transitions to the exchange phase,

where the rally between players takes place. However, if the serve is unsuccessful ("fault") or if the ball goes out of bounds during an exchange, the match shifts to a different phase, labeled as "other." This phase represents non-play activities such as pauses, replays, or transitions between points. Following this, the sequence loops back to the service phase, signaling the start of a new point or game.

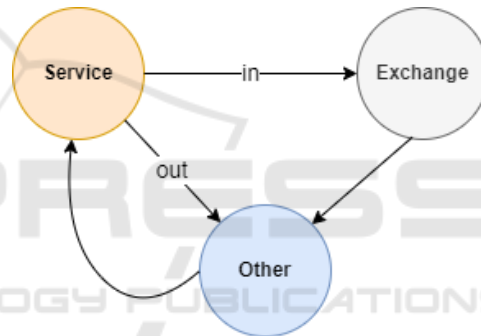


Figure 3: Coarse-Grained classes.

We first trained the TAL4Tennis model on the TenniSet (Games-based split) with three classes (Service, Exchange, Other). Then, evaluation is carried out on it on TenniSet (Games-based split) and French Open dataset. The model had not been exposed to sequences from clay courts during training, making this evaluation particularly challenging as it tests the model’s ability to generalize from grass to clay surfaces. Results are shown in Table 4.

## 5 CONCLUSION

In this work, we have presented a temporal tennis action localization model based on SSMs. The model predicts temporal segments with their respective labels, and use the tIoU metric for evaluation. Experiments carried on two splits of the TenniSet dataset show that TAL4Tennis achieves competitive performance compared to state-of-the-art approaches, out-

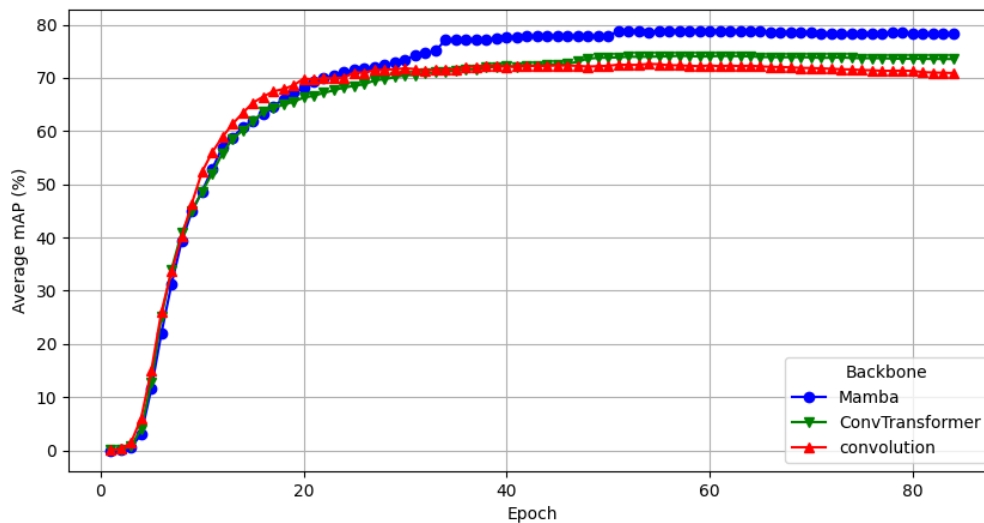


Figure 4: Evolution of Average Validation mAP of TAL4Tennis over all tIoU Thresholds of different backbones.

Table 4: mAP for different court types using different IoU thresholds  $\alpha$ .

Class	$\alpha$				
	0.3	0.4	0.5	0.6	0.7
<b>TenniSet -Games-based split Dataset</b>					
Other	0.991	0.989	0.983	0.976	0.976
Service	0.999	0.996	0.996	0.969	0.876
Exchange	0.981	0.970	0.959	0.929	0.863
<b>French Open Dataset</b>					
Other	0.936	0.925	0.905	0.865	0.799
Service	0.991	0.991	0.988	0.977	0.900
Exchange	0.937	0.932	0.904	0.860	0.798

performing them at higher tIoU thresholds, as shown in Table 3. Additionally, we introduce a new dataset for Tennis Action Localization (TAL), derived from French Open clay court footage. This dataset includes annotations for three primary action phases: Serve, Rally, and Non-Game. In our future work, we propose to create action-tubes of each player in order to better understand fine-grained events (Rajasegaran et al., 2023). In fact, in tennis the main object leading the change in events is the player.

## REFERENCES

- Alexey, D. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.
- Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970.
- Chao, Y.-W., Vijayanarasimhan, S., Seybold, B., Ross, D. A., Deng, J., and Sukthankar, R. (2018). Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139.
- Chen, G., Huang, Y., Xu, J., Pei, B., Chen, Z., Li, Z., Wang, J., Li, K., Lu, T., and Wang, L. (2024). Video mamba suite: State space model as a versatile alternative for video understanding. *ArXiv*, abs/2403.09626.
- Cheng, T., Bi, T., Ji, W., and Tian, C. (2024). Graph convolutional network for image restoration: A survey. *Mathematics*, 12(13).
- Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736.
- Escorcia, V., Caba Heilbron, F., Niebles, J. C., and Ghanem, B. (2016). Daps: Deep action proposals for action understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 768–784. Springer.
- Faulkner, H. (2024). Tennis. <https://github.com/HaydenFaulkner/Tennis>.
- Faulkner, H. and Dick, A. (2017). Tenneset: a dataset for dense fine-grained event recognition, localisation and description. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.
- Gao, J., Yang, Z., and Nevatia, R. (2017). Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*.
- Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE*

- transactions on pattern analysis and machine intelligence*, 29(12):2247–2253.
- Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.
- Gu, A. and Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces.
- Gu, A., Goel, K., and R’e, C. (2021). Efficiently modeling long sequences with structured state spaces. *ArXiv*, abs/2111.00396.
- He, B., Yang, X., Kang, L., Cheng, Z., Zhou, X., and Shrivastava, A. (2022). Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13925–13935.
- Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., and Shah, M. (2017). The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23.
- Jiang, W. and He, G. (2021). Study on the effect of shoulder training on the mechanics of tennis serve speed through video analysis. *Molecular & Cellular Biomechanics*, 18(4):221.
- Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Li, J., Liu, X., Zong, Z., Zhao, W., Zhang, M., and Song, J. (2020). Graph attention based proposal 3d convnets for action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4626–4633.
- Li, L., Kong, T., Sun, F., and Liu, H. (2019). Deep pointwise prediction for action temporal proposal. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26*, pages 475–487. Springer.
- Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., and Fu, Y. (2021). Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3320–3329.
- Lin, T., Liu, X., Li, X., Ding, E., and Wen, S. (2019). Bmn: Boundary-matching network for temporal action proposal generation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3888–3897.
- Lin, T., Zhao, X., Su, H., Wang, C., and Yang, M. (2018). Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.
- Liu, X., Hu, Y., Bai, S., Ding, F., Bai, X., and Torr, P. H. (2021a). Multi-shot temporal event localization: a benchmark. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12591–12601.
- Liu, Y., Wang, L., Wang, Y., Ma, X., and Qiao, Y. (2022). Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950.
- Liu, Z., Wang, L., Zhang, Q., Tang, W., Yuan, J., Zheng, N., and Hua, G. (2021b). Acsnet: Action-context separation network for weakly supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2233–2241.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Myung, W., Su, N., Xue, J.-H., and Wang, G. (2024). Degen: Deformable graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33:2477–2490.
- Nguyen, P. X., Ramanan, D., and Fowlkes, C. C. (2019). Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5502–5511.
- Paul, S., Roy, S., and Roy-Chowdhury, A. K. (2018). Wtalc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579.
- Peng, X. and Tang, L. (2022). Biomechanics analysis of real-time tennis batting images using internet of things and deep learning. *The Journal of Supercomputing*, 78(4):5883–5902.
- Piergiovanni, A. and Ryoo, M. (2019). Temporal gaussian mixture layer for videos. In *International Conference on Machine Learning*, pages 5152–5161. PMLR.
- Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., and Malik, J. (2023). On the benefits of 3d pose and tracking for human action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 640–649.
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666.
- Rizve, M. N., Mittal, G., Yu, Y., Hall, M., Sajeed, S., Shah, M., and Chen, M. (2023). Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002.
- Shi, B., Dai, Q., Mu, Y., and Wang, J. (2020). Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1009–1019.



- Shou, Z., Chan, J., Zareian, A., Miyazawa, K., and Chang, S.-F. (2017). Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5734–5743.
- Shou, Z., Wang, D., and Chang, S.-F. (2016). Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058.
- Tang, Y., Niu, C., Dong, M., Ren, S., and Liang, J. (2019). Afo-tad: Anchor-free one-stage detector for temporal action detection. *arXiv preprint arXiv:1910.08250*.
- Voelker, A. R. and Eliasmith, C. (2018). Improving spiking dynamical networks: Accurate delays, higher-order synapses, and time cells. *Neural Computation*, 30:569–609.
- Wang, B., Zhao, Y., Yang, L., Long, T., and Li, X. (2024). Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2171–2190.
- Zelnik-Manor, L. and Irani, M. (2001). Event-based analysis of video. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II.
- Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., and Gan, C. (2022). Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6209–6223.
- Zhang, C.-L., Wu, J., and Li, Y. (2022). Actionformer: Localizing moments of actions with transformers. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 492–510, Cham. Springer Nature Switzerland.
- Zhao, H., Torralba, A., Torresani, L., and Yan, Z. (2019). Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678.
- Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., and Tian, Q. (2020). Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer.
- Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., and Lin, D. (2017). Temporal action detection with structured segment networks. *International Journal of Computer Vision*, 128:74 – 95.
- Zhou, Y., Wang, R., Li, H., and Kung, S.-Y. (2021). Temporal action localization using long short-term dependency. *IEEE Transactions on Multimedia*, 23:4363–4375.