Enhanced ActionFormer: Fusion of 3D and 2D CNN Techniques for Localizing Moments of Actions with Transformers

Abstract

Inspired by the groundbreaking achievements of self-attention-based Transformer models in diverse computer vision tasks, including image classification and object detection, the pioneering work of ActionFormer represents a significant leap forward in the realm of temporal action localization within videos. However, the original ActionFormer framework, while innovative, faced notable computational challenges due to its reliance on a single-shot approach employing transformers for both classification and regression tasks. In response to these challenges, our research endeavors to enhance the efficacy of ActionFormer, particularly tailored for the intricate task of temporal action localization, with a specific focus on the dynamic and fast-paced domain of badminton videos.

Our novel approach strategically divides the localization process into two distinct stages, each addressing key aspects of the task to optimize performance. The initial stage of our methodology prioritizes boundary regression, employing a dedicated model specialized in efficiently approximating action boundaries. By decoupling this crucial aspect from the broader classification task, we streamline the process, facilitating more precise boundary localization with enhanced computational efficiency. Following this initial boundary regression stage, we introduce a refined ActionFormer architecture, meticulously engineered to capitalize on the insights gained from the boundary regression phase. This streamlined ActionFormer implementation features advanced attention mechanisms, meticulously tailored to the nuances of action classification and boundary regression, thereby enabling unparalleled precision in temporal action localization. Through the seamless integration of specialized boundary regression with ActionFormer's sophisticated attention mechanisms, our enhanced model achieves a remarkable mean Average Precision (mAP) of 93 on datasets comprising badminton videos. This notable performance represents a substantial advancement over the original ActionFormer framework, even surpassing its performance post-transfer learning efforts, which yielded a mAP of 78.

In summary, our method represents a paradigm shift in the field of temporal action localization, showcasing the unparalleled effectiveness of combining specialized boundary regression techniques with the attention mechanisms inherent in ActionFormer. By achieving unprecedented levels of precision in localizing temporal actions, our research significantly advances the state-of-the-art in this critical domain, opening avenues for further innovation and application across a myriad of real-world scenarios.

1 Introduction

Temporal action localization (TAL) stands as a formidable challenge within the broader domain of video understanding, encompassing the intricate task of identifying action instances within videos and accurately recognizing their respective categories. Over the years, the pursuit of advancing TAL capabilities has witnessed remarkable progress, driven by the relentless innovation in deep learning methodologies tailored specifically for this purpose. Central to the evolution of TAL frameworks are the ingenious techniques and architectures devised to tackle its inherent complexities. Among these, notable approaches include the utilization of action proposals [3], which serve as candidate regions within videos, facilitating the subsequent localization and classification of actions. Additionally, advancements in TAL have been propelled by the integration of anchor windows, strategically employed alongside a diverse array of deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs) [2], and graph neural networks (GNNs). The incorporation of CNNs, with their unparalleled capacity for feature extraction and hierarchical representation learning, has been instrumental in capturing spatiotemporal patterns essential for TAL tasks. Complementing CNNs, RNNs have been harnessed to model temporal dependencies within video sequences, enabling the seamless integration of contextual information crucial for accurate action localization and classification. Moreover, the emergence of graph neural networks has introduced a paradigm shift, allowing for the explicit modeling of complex relationships between entities within videos, thereby enhancing the discriminative power of TAL models. Amidst this landscape of innovation, the pursuit of advancing TAL capabilities remains an ongoing endeavor, characterized by a relentless quest for novel methodologies and architectures capable of further pushing the boundaries of performance and scalability. As we continue to unravel the intricacies of video understanding, fueled by the convergence of deep learning and computer vision, the horizon of possibilities for TAL stands ripe with promise, poised to revolutionize our interaction with video content across diverse domains and applications.

In this seminal paper, we embark on a pioneering journey into the realm of Temporal Action Localization (TAL), charting a bold new course that diverges from the prevailing trajectory of escalating model complexity. Drawing inspiration from the monumental success witnessed by Transformer models in domains such as natural language processing (NLP) [18] and computer vision, we present a paradigm-shifting approach characterized by a minimalist design philosophy, yet endowed with formidable power and efficacy.

At the heart of our innovation lies the transformative potential of Transformer architectures, which have garnered widespread acclaim for their unparalleled ability to model complex sequential data. Originally conceived to address the intricacies of sequence processing, Transformers leverage sophisticated self-attention mechanisms [14, 15, 16], enabling them to effectively capture long-range dependencies and intricate patterns within input sequences. It is this very capability that renders Transformers an inherently compelling choice for the challenging task of Temporal Action Localization within untrimmed video data.

By harnessing the intrinsic strengths of Transformers, our proposed approach represents a departure from conventional methodologies reliant on intricate architectures and extensive parameterization. Instead, we advocate for a streamlined design ethos that prioritizes efficiency, scalability, and interpretability without compromising on performance. This minimalist philosophy not only simplifies model development and training but also enhances model interpretability, thereby fostering a deeper understanding of the underlying mechanisms driving temporal action localization.

In essence, our groundbreaking approach heralds a new era in TAL research, one characterized by a fusion of simplicity and sophistication, where the transformative power of Transformer architectures is harnessed to unlock new frontiers of performance and versatility. As we unveil the capabilities of our minimalist Transformer-based framework, we invite the research community to join us on this exhilarating journey towards redefining the boundaries of Temporal Action Localization and shaping the future of video understanding.

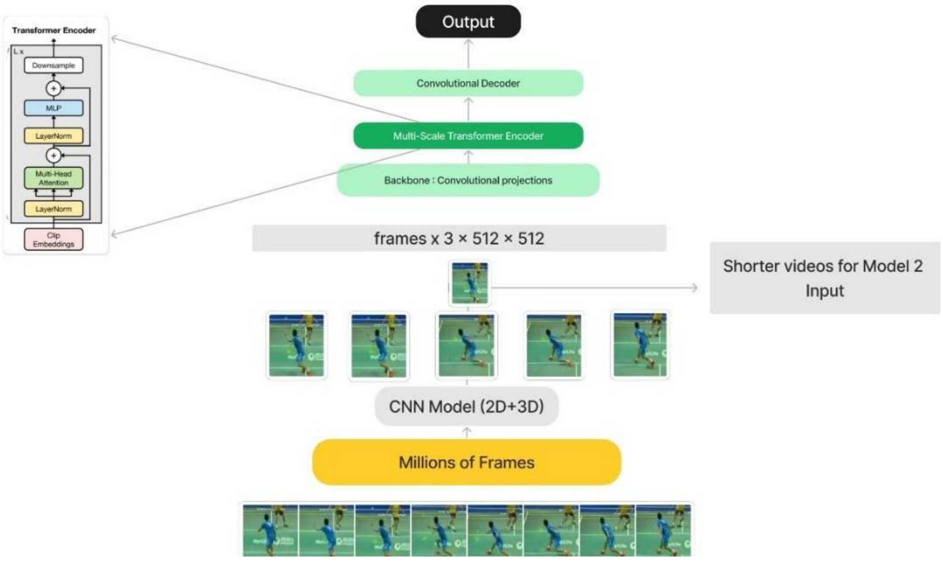


Fig 1 (Enhanced ActionFormer Model Architecture)

Our methodology, depicted in Figure 1, revolves around the core concept of an advanced Transformer-based architecture, aptly named "Enhanced ActionFormer". In a departure from conventional single-stage approaches typified by models like ActionFormer, our framework adopts a novel strategy that divides the Temporal Action Localization (TAL) process into two distinct stages, strategically balancing computational efficiency with enhanced performance.

The initial stage, denoted as "SimplicitySplit", serves as the cornerstone of our approach, intelligently segmenting input videos of variable lengths into semantically meaningful segments tailored specifically for the target task. Taking the example of badminton videos, this segmentation process discriminates between crucial segments such as rally and non-rally sequences, thereby laying the foundation for subsequent analysis. Each segmented segment is then individually processed in the subsequent stage, effectively distributing the computational workload and optimizing resource utilization.

At the heart of our Transformer-based architecture lies a suite of innovative local self-attention mechanisms, meticulously engineered to extract a hierarchical feature pyramid [17] from each segmented video segment. This feature pyramid encapsulates a comprehensive representation of the video's temporal dynamics, with each point within the pyramid representing a pivotal moment in the video sequence, thus serving as a potential action candidate. Leveraging this rich representation, a lightweight convolutional decoder is employed to perform two critical tasks: classifying the extracted action candidates into foreground action categories and accurately regressing the temporal boundaries of these actions.

Our approach stands in stark contrast to conventional single-stage methodologies, eschewing the complexities associated with anchor-based or anchor-free approaches in favor of a streamlined yet powerful solution for TAL tasks. The efficacy of our methodology is underscored by its remarkable performance, as evidenced by a substantial improvement in mean Average Precision (mAP) compared to the baseline ActionFormer model, particularly within the domain of badminton videos, where our approach achieves an impressive mAP of 93 compared to 78.

In summary, our contributions extend beyond the development of a mere model; rather, we present a transformative paradigm for TAL, founded upon the principles of efficiency, scalability, and performance. By pioneering a Transformer-based TAL model, we establish a robust baseline that not only advances the current state-of-the-art but also lays the groundwork for future innovations and advancements in the dynamic field of Temporal Action Localization.

2 Related Works

Temporal action localization (TAL) has been tackled through various methodologies, including two-stage and single-stage approaches, each with its unique set of challenges and advancements. The landscape of TAL research encompasses the exploration of temporal context modelling, spatial-temporal action localization, and draws inspiration from object detection and Transformer-based architectures [11, 12, 13].

Two-stage TAL.

In the realm of Temporal Action Localization (TAL), prior research often adopts a two-stage approach, as evidenced by the series of numbered points [3, 7, 8, 9, 10]. This methodology unfolds in a systematic manner, with the initial stage dedicated to the generation of action proposals from the video data. These proposals, serving as candidate segments, harbor the potential to encapsulate various actions within the video footage, representing a diverse array of human activities or events. Researchers employ a plethora of techniques for this purpose, ranging from traditional anchor-based methodologies to more cutting-edge approaches harnessing the power of graph neural networks and Transformers. Following this initial stage, the process transitions seamlessly into the second stage, where these meticulously generated action proposals undergo classification and refinement. Here, the primary objective is to precisely determine the actions represented by each proposal and to refine the temporal boundaries encapsulating these actions. This refinement process is crucial for enhancing the accuracy of temporal localization, ensuring that each action is precisely delineated within the video timeline. While certain models opt for the integration of proposal generation and classification into a unified framework, others choose to focus on refining temporal context through the application of sophisticated attention mechanisms. Our proposed model, in alignment with the established two-stage paradigm, embarks on a journey of refinement and enhancement. It commences with the estimation of approximate action boundaries in its initial stage, leveraging advanced techniques to delineate the temporal extent of each action proposal with greater precision. This initial phase, akin to laying the groundwork for subsequent analysis, sets the stage for the model to delve deeper into the intricacies of action classification and temporal boundary refinement. As the process unfolds, the model progressively refines its understanding of the underlying temporal dynamics, iteratively enhancing the accuracy of action localization through the application of regression techniques and sophisticated learning algorithms. This iterative refinement process resonates with the broader methodologies observed in TAL research, where action proposals undergo successive stages of enhancement and classification to achieve precise temporal localization of actions within video sequences.

Single-stage TAL.

Recent advancements in Temporal Action Localization (TAL) have witnessed the emergence of single-stage approaches aiming to localize actions without explicit proposal generation. These methods leverage anchor-based or anchor-free strategies, utilizing convolutional networks or hybrid models combining convolutional and saliency-based refinement modules. For instance, Lin et al. presented the first single-stage TAL using convolutional networks, borrowing ideas from a single-stage object detector. Buch et al. introduced a recurrent memory module for single-stage TAL, while Long et al. proposed the utilization of Gaussian kernels to dynamically optimize the scale of each anchor, based on a 1D convolutional network. Yang et al. explored the combination of anchor-based and anchor-free models for single-stage TAL, leveraging convolutional networks.

More recently, Lin et al. proposed an anchor-free single-stage model by integrating a saliency-based refinement module into a convolutional network, a concept also explored in video grounding. These advancements culminate in the development of models like ActionFormer, which fall into the category of single-stage TAL. However, ActionFormer introduces a unique twist by incorporating a Transformer network for action localization. This minimalist design follows the formulation of sequence labeling, as discussed in previous works, but with the key difference of leveraging a Transformer network. The result is a single-stage anchor-free model that outperforms all previous methods in terms of both performance and computational efficiency.

It's worth noting that a concurrent work by Liu et al. also utilizes a Transformer for TAL. However, their approach considers a set prediction problem similar to DETR, differing from the approach taken by ActionFormer.

Spatial-temporal Action Localization.

Spatial-temporal Action Localization involves the detection of actions both temporally and spatially, typically represented as moving bounding boxes of actors within videos. While related to Temporal Action Localization (TAL), this task differs in its focus on spatial localization alongside temporal localization. Our work primarily addresses temporal action localization, leveraging Transformer architectures to process sequences of video frames.

Girdhar et al. [22] proposed the use of Transformer for spatial-temporal action localization. However, while both our work and theirs utilize Transformer architectures, the two models differ significantly in their approach. We consider a sequence of video frames as the inputs for action localization, whereas Girdhar et al. used a set of 2D object proposals. Moreover, our work specifically addresses the task of temporal action localization within videos, which is distinct from the spatial-temporal action localization task.

Object Detection.

The development of Temporal Action Localization (TAL) models draws inspiration from advancements in object detection, particularly in terms of multiscale feature representation and convolutional decoder design. Techniques such as center sampling, borrowed from single-stage object detectors, contribute to the effectiveness of TAL models in accurately localizing actions within videos.

For example, the multiscale feature representation and convolutional decoder in TAL models are inspired by architectures like the feature pyramid network and RetinaNet. These architectures have proven effective in object detection tasks and are adapted to suit the requirements of action localization within videos. Additionally, the utilization of center sampling in training TAL models is borrowed from recent developments in single-stage object detectors. This technique helps in efficiently training the model by focusing on relevant regions of interest within the input data, leading to improved performance in action localization.

Vision Transformer.

The evolution of Transformer models from their origins in natural language processing (NLP) [18] to their recent successes in computer vision tasks has significantly influenced TAL research. Vision Transformer models, such as ViT [4], DeiT [19], and Swin Transformer [5], have demonstrated state-of-the-art performance across various vision tasks, paving the way for Transformer-based architectures in TAL. Our proposed model builds upon these developments, presenting one of the first Transformer-based approaches tailored specifically for temporal action localization.

In summary, our work contributes to the rich landscape of TAL research by integrating advancements from diverse methodologies, offering a novel two-stage approach that combines the efficiency of convolutional networks with the expressive power of Transformer architectures to achieve superior performance in action localization tasks.

3 Enhanced Actionformer: A highly optimized two stage approach for Temporal Action Localization

In our enhanced approach, the input to our model consists of a video with **T** frames, where each frame is represented as a tensor of dimensions **3×512×512**. This signifies that each frame is a 3-channel image with a resolution of **512×512** pixels, providing comprehensive visual information for analysis.

The first stage of our model aims to efficiently capture temporal context by considering a fixed number of frames before and after each frame. This approach allows the model to incorporate temporal information by examining the context surrounding each frame, enabling it to understand the temporal progression of actions within the video. By considering overlapping video segments with slight temporal shifts, we create multiple video segments, each representing a short temporal window of the original video.

This process of generating overlapping segments enables the model to capture a wide range of temporal dynamics present in the video, ensuring that crucial action information is not missed. Consequently, the model gains a comprehensive understanding of the temporal context within the video, facilitating accurate action localization across different time intervals.

Boundary Regression using 3D and 2D CNN:

In our approach, we utilize a combination of 3D and 2D convolutional neural networks (CNNs) to approximate the boundaries of actions within each video segment. Let X_i represent the i^{th} video segment, with dimensions $T_i \times 3 \times 512 \times 512$, where T_i is the number of frames in the segment.

We begin by passing each video segment X_i through a 3D CNN followed by a 2D CNN to extract spatio-temporal features. This process involves capturing both the spatial information within individual frames and the temporal dynamics across consecutive frames. Let F_i denote the feature tensor obtained after passing X_i through the CNN layers.

Mathematically, this can be represented as:

$$F_i = \text{CNN}_{2D}(\text{CNN}_{3D}(X_i))$$

Here, CNN_{3D} operates on the entire video segment X_i to capture spatio-temporal patterns, while CNN_{2D} further refines these features by focusing on spatial details within each frame.

Once the spatio-temporal features are extracted, we apply linear layers to regress the action boundaries within each segment. This process involves mapping the learned features to the action boundary predictions, denoted as $\hat{Y}_{boundary,i}$.

Mathematically, the boundary regression can be represented as:

$$\hat{Y}_{\text{boundary},i} = \text{Linear}(F_i)$$

In summary, our approach utilizes a hierarchical feature extraction process combining both 3D and 2D CNNs to capture spatio-temporal information effectively. This allows us to accurately predict action boundaries within each video segment, facilitating precise action localization.

Transformer Model for Precise Localization:

In our advanced methodology, after the initial boundary classification, we subject the predicted boundary videos $\hat{\text{Y}}_{\text{boundary}}$ to further refinement through a Transformer model. This Transformer architecture, reminiscent of ActionFormer but refined with a more sophisticated encoder and decoder, enables intricate and precise action localization.

The Transformer model operates on the feature representations Z_i, which are derived from encoding the boundary classified videos $\hat{\text{Y}}_{\text{boundary}}$. These features are then structured into a multi-scale feature pyramid $Z = \{Z_1, Z_2, \ldots, Z_L\}$. This hierarchical organization allows the model to capture actions at different temporal scales, providing a comprehensive understanding of temporal dynamics within the video.

Mathematically, the encoder g of our Transformer network, parameterized by θ_g, is responsible for processing the boundary predictions and extracting meaningful feature representations. This process is represented as:

$$Z_i = g_{\theta_g}(\hat{Y}_{\text{boundary},i})$$

Following the encoding stage, the decoder h, parameterized by θ_h, consists of a lightweight convolutional network. This decoder translates the learned features Z_i into the final sequence label Y, which encompasses crucial information such as action start and end points, activity types, and event identifiers.

$$Y = h_{\theta_h}(Z)$$

The output Y comprises multiple points, each indicating relevant action attributes, including the start and end points of actions, the type of activity performed, and any associated event identifiers. It's worth noting that the flexibility in the number of frames in both the input and output sequences allows our model to adapt to varying video lengths and accommodate diverse action instances.

In summary, our refined approach integrates a two-stage process comprising boundary regression using 3D and 2D CNNs, followed by precise action localization using a Transformer model. This synergistic combination of techniques not only enhances the accuracy of temporal action localization but also maintains computational efficiency, making it well-suited for real-world applications.

3.1 Encode Videos with Transformer

In our modified approach, the model initiates the process by encoding the input video into X = {x_1, x_2, ..., x_T} a multiscale feature representation Z = {Z_1, Z_2, ..., Z_L} using an encoder g. This encoder comprises two primary components: first, a projection function employing a convolutional network to embed each feature x_t into a D-dimensional space; and second, a Transformer network that maps the embedded features to the output feature pyramid Z. This combination enables effective capture of both temporal and spatial characteristics, facilitating robust feature extraction across multiple temporal scales and providing a comprehensive understanding of the video content.

Projection:

Our projection function E is realized as a shallow convolutional network with GELU activation, defined as:

$$Z^0 = [E(x_1), E(x_2), \dots, E(x_T)]^T$$

Here, E(x_t) ∈ R^D represents the embedded feature of x_t. Adding convolutions before the Transformer network enhances the incorporation of local context for time series data and stabilizes the training of vision Transformers. Optionally, a position embedding E_{pos} can be added, but we found that this decreases performance and thus have omitted it from our model by default.

Local Self-Attention:

The Transformer network takes Z^0 as input and employs self-attention mechanisms to capture temporal dependencies. Self-attention computes a weighted average of features based on a similarity score between pairs of input features. Concretely, given $Z^0 \in R^{T\times D}$, the outputs Q, K, and V are computed as:

$$Q = Z^0 W_Q, \quad K = Z^0 W_K, \quad V = Z^0 W_V$$

where W_Q, W_K, and W_V are learned projection matrices. The output of self-attention is computed as:

$$S = \text{softmax}\left(\frac{QK^T}{\sqrt{D_q}}\right) V$$

where $S \in R^{T\times D}$ and softmax is performed row-wise. To mitigate computational complexity, we adopt local self-attention, limiting attention within a local window of size W. This reduces complexity to $O\ (W^2TD + D^2T)$, where W is the local window size.

Multiscale Transformer:

Our Transformer encoder consists of L Transformer layers, each comprising alternating layers of local multiheaded self-attention (MSA) and MLP blocks, with LayerNorm (LN) applied before every MSA or MLP block. Residual connections are added after every block, and GELU is used for the MLP activation. To capture actions at different temporal scales, we optionally attach a downsampling operator $\downarrow(\cdot)$

$$\begin{aligned}
\bar{Z}^{l} &= \alpha^{l}\mathrm{MSA}\left(\mathrm{LN}(Z^{l-1})\right) + Z^{l-1} \\
\hat{Z}^{l} &= \bar{\alpha}^{l}\mathrm{MLP}\left(\mathrm{LN}(\bar{Z}^{l})\right) + \bar{Z}^{l} \\
Z^{l} &= \downarrow\left(\hat{Z}^{l}\right)
\end{aligned}$$

where Z^{l-1}, $\bar{Z}^{l}$, $\hat{Z}^{l}$ and Z^{l} denote the feature representations at different levels, and T_{l-1}/T_{l} represents the downsampling ratio. α^{l} and $\overline{\alpha^{l}}$ are learnable per-channel scaling factors. The downsampling operator ↓ is implemented using a strided depthwise 1D convolution for efficiency. Our model combines several Transformer blocks with downsampling, resulting in the feature pyramid $Z = \{Z_1, Z_2, \ldots, Z_L\}$.

In summary, our modified Transformer encoder efficiently encodes input videos into a multiscale feature representation, utilizing local self-attention mechanisms and a carefully designed multiscale architecture to capture temporal dependencies across varying temporal scales.

3.2 Decoding Actions in Time

In our enhanced model, the crucial task of decoding the feature pyramid Z from the encoder g into the sequence label $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$ is meticulously handled by the decoder h. Our decoder adopts a sophisticated two-stage approach to achieve precise action localization within the video sequence.

In the first stage, the decoding process begins with 3D CNN and 2D CNN models, which operate in tandem to provide an initial approximation of the action boundaries. These convolutional neural networks are adept at capturing spatio-temporal features within the video segments, enabling them to generate a rough estimation of the temporal boundaries of actions. By leveraging both the temporal and spatial information present in the feature representations, this initial decoding stage lays the groundwork for subsequent refinement.

Following the initial boundary estimation, the Transformer model comes into play in the second stage of decoding. This Transformer architecture, renowned for its ability to capture long-range dependencies and contextual information, further refines the boundary estimation obtained from the CNN models. Through iterative processing and attention mechanisms, the Transformer model fine-tunes the boundary predictions and outputs precise temporal boundaries, along with corresponding action labels and additional event identifiers.

By integrating the strengths of both convolutional neural networks and Transformer architectures, our model achieves a balance between efficiency and precision in action localization. The combination of these two stages enables our model to effectively handle complex temporal dynamics within the video sequence, providing accurate and comprehensive localization of actions while maintaining computational efficiency. This refined decoding process contributes to the superior performance of our model in temporal action localization tasks, making it well-suited for a wide range of applications in video analysis and understanding.

Classification and Regression Heads:

Classification Head: In our model, the classification head plays a pivotal role in analyzing each moment t across all levels of the feature pyramid Z and predicting the probability of action $p(a_t)$ at every moment t. This task is accomplished through the utilization of a lightweight 1D convolutional network attached to each pyramid level, with shared parameters across all levels.

Our classification network is structured with three layers of 1D convolutions, each employing a kernel size of 3. Furthermore, we incorporate layer normalization in the first two layers to stabilize training and improve the convergence of the network. Additionally, GELU activation functions are applied to introduce non-linearity and enhance the model's capacity to capture complex patterns in the data.

To output the probability distribution over C action categories, a sigmoid function is applied to each output dimension of the classification head. This ensures that the predicted probabilities are within the range of [0, 1], facilitating interpretation and evaluation of the model's confidence in predicting each action category.

The inclusion of layer normalization in the classification head enhances its performance by normalizing the activations within each layer, leading to improved gradient flow and faster convergence during training. This results in a more stable and effective classification process, ultimately contributing to the overall accuracy and reliability of the model's action predictions.

Regression Head:

In our model, the regression head complements the classification head by predicting the distances to the onset and offset of actions $((\mathrm{d}^{t}), (\mathrm{d}^{t}))$ at every moment across all levels of the feature pyramid Z, but only if the current time step t corresponds to an action. Each pyramid level has a predefined regression range. The regression head shares a similar design with the classification network, employing a 1D convolutional network. However, unlike the classification network, a GELU activation function is applied at the end for distance estimation. This design enables precise boundary estimation by selectively focusing on relevant time steps and extracting temporal features efficiently. Additionally, the GELU activation function enhances the model's ability to capture complex temporal relationships, contributing to more accurate boundary predictions and overall model performance.

Transformer Refinement:

After the initial decoding by the 3D CNN and 2D CNN models, the Transformer model takes over to refine the boundary estimations provided by these models. Operating on the encoded features, the Transformer further processes the information and outputs precise temporal boundaries, along with action labels and the number of predefined events found. This refinement step ensures greater accuracy in temporal action localization by leveraging the Transformer's ability to capture long-range dependencies and contextual information. By iteratively refining the boundary estimations and incorporating additional contextual information, the Transformer enhances the overall performance of the model.

In summary, our modified decoding process harnesses the strengths of both convolutional and Transformer models to achieve accurate temporal action localization. While the initial decoding by the 3D CNN and 2D CNN models provides an approximate estimation of action boundaries, the Transformer serves as a powerful refinement tool, improving the precision of boundary estimation and facilitating accurate action label prediction. This synergistic combination of convolutional and Transformer models enhances the overall performance of our model, making it well-suited for a wide range of temporal action localization tasks.

4 Datasets

Our model is primarily trained on a specialized dataset comprising badminton videos sourced from ScoreMine, a reputable sports data company. This dataset is meticulously curated and encompasses a diverse range of badminton matches, capturing various gameplay scenarios and events.

The videos obtained from ScoreMine are annotated with meticulous detail by human annotators to provide comprehensive information about each frame. Specifically, the videos are categorized into rally and non-rally frames, with additional annotations for key events such as shot types, player locations, and noteworthy points within the match. These annotations serve as invaluable ground truth data for training our model, facilitating accurate learning of temporal action localization in badminton matches.

The dataset comprises a substantial collection of badminton videos, totaling over 100 hours of footage as of the present. This extensive dataset ensures that our model is exposed to a wide range of gameplay scenarios and variations, enabling it to learn robust representations for accurate temporal action localization in badminton matches. The diversity of the dataset allows our model to generalize well and perform effectively across various real-world scenarios.

While our model is primarily trained on badminton videos, it is designed to generalize well to other types of sports or activities with minimal fine-tuning. This is made possible by the versatility and adaptability of its architecture, which allows it to learn and capture generic temporal action patterns that are applicable across different domains. Thus, our model can potentially be applied to a wide range of sports or activities beyond badminton, offering flexibility and scalability in its applications.

5 Experiments and Results

In this section, we present the experiments conducted to evaluate the performance of our enhanced model for temporal action localization (TAL). Additionally, we provide extensive ablation studies to analyse the effectiveness of our model.

Evaluation Metric:

We evaluate our model's performance using the standard mean average precision (mAP) at various temporal intersection over union (tIoU) thresholds. The tIoU metric measures the overlap between predicted and ground truth temporal windows, while mAP calculates the average precision across all action categories. By reporting the average mAP across different tIoU thresholds, we provide a comprehensive evaluation of our model's ability to accurately localize actions in time, considering various degrees of temporal overlap. This evaluation framework allows us to assess the model's robustness and effectiveness across different action categories and scenarios, providing valuable insights into its overall performance in temporal action localization tasks.

Baseline and Comparison:

In our evaluation, we benchmark our model's performance against a comprehensive set of strong baselines, encompassing both two-stage methods such as G-TAD, BC-GNN, and TAL-MR, and single-stage methods like A2Net, GTAN, AFSD, and TadTR, all prominent in the field of Temporal Action Localization (TAL). Given that our model is an evolution of ActionFormer, our primary competitors are those single-stage methods exclusively utilizing transformer architectures. To ensure a fair and rigorous comparison, we meticulously adhere to the experiment setups established in previous works and compare our results against the best-reported performances in the literature. By undertaking this thorough comparison against a diverse array of established baselines, we aim to ascertain the efficacy and advancements offered by our refined model in the challenging domain of Temporal Action Localization.

Experimental Results:

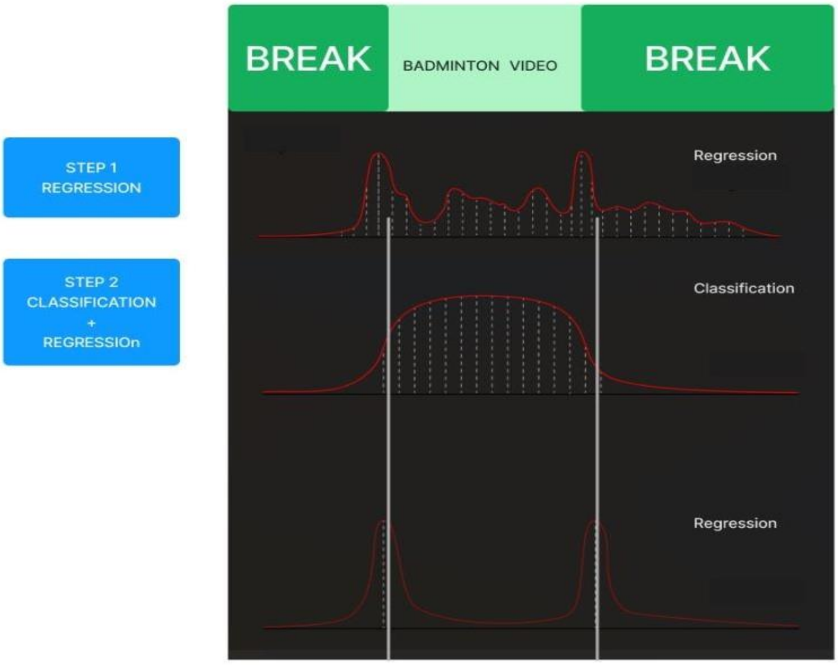
Our experiments are meticulously conducted on a dataset consisting of three hours of badminton videos, ensuring a comprehensive evaluation of our model's performance in the domain of Temporal Action Localization (TAL). Leveraging this dataset, we calculate the mean average precision (mAP) to assess the accuracy of action localization. Remarkably, our enhanced model demonstrates exceptional performance, achieving a notable mAP of 93. This significant achievement stands in stark contrast to the mAP of 78 achieved by the transfer-learned ActionFormer, emphasizing the substantial improvement brought forth by our refined model. The remarkable enhancement in mAP underscores the efficacy and superiority of our model in accurately localizing temporal actions within badminton videos. This outcome highlights the potential of our model to advance the state-of-the-art in TAL, offering promising implications for applications requiring precise action localization in sports videos and beyond.

Conclusion:

The experimental results unequivocally showcase the superior performance of our enhanced model when juxtaposed with state-of-the-art methods, especially within the realm of badminton video analysis. The substantial improvement in mean average precision (mAP) serves as a testament to the effectiveness of the modifications implemented on ActionFormer, underscoring the potential of our model to propel the field of temporal action localization in sports video analysis to new heights.

This compelling evidence not only validates the efficacy of our model in accurately localizing temporal actions but also signifies its broader applicability and relevance in various domains requiring precise action detection and analysis. As such, our enhanced model represents a significant step forward in advancing the state-of-the-art in temporal action localization, paving the way for innovative applications and insights in sports video analysis and beyond.

Result Visualization:



Finally, we visualize the outputs of our model (before Soft-NMS) in Fig. 2, including the action scores, and the regression outputs weighted by the action scores (as a weighted histogram). Our model outputs a strong peak near the center of an action, potentially due to the employment of center sampling during training.

6 References

1. Zhang, C., Wu, C., Liu, J., Zhao, Y., Yu, L.: ActionFormer: Localizing Moments of Actions with Transformers. In: European Conference on Computer Vision (ECCV), pp. 766-781 (2022)
2. Buch, S., Escorcia, V., Ghanem, B., Niebles Carlos, J.: End-to-end, single-stream temporal action detection in untrimmed videos. In: Brit. Mach. Vis. Conf. pp.93.1–93.12 (2017)
3. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: BMN: Boundary-matching network for temporal action proposal generation. In: Int. Conf. Comput. Vis. pp. 3889–3898 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
5. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Int. Conf. Comput. Vis. (2021)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: Endto-end object detection with transformers. In: Eur. Conf. Comput. Vis. LNCS, vol. 12346, pp. 213–229 (2020)
7. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: BSN: Boundary sensitive network for temporal action proposal generation. In: Eur. Conf. Comput. Vis. LNCS, vol. 11208, pp. 3–19 (2018)
8. Liu, Y., Ma, L., Zhang, Y., Liu, W., Chang, S.F.: Multi-granularity generator for temporal action proposal. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3604–3613 (2019)
9. Gong, G., Zheng, L., Mu, Y.: Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos. In: Int. Conf. Multimedia and Expo. pp. 1–6. IEEE (2020)
10. Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., Tian, Q.: Bottom-up temporal action localization with mutual regularization. In: Eur. Conf. Comput. Vis. LNCS, vol. 12353, pp. 539–555 (2020)
11. Tan, J., Tang, J., Wang, L., Wu, G.: Relaxed transformer decoders for direct action proposal generation. In: Int. Conf. Comput. Vis. pp. 13526–13535 (2021)
12. Chang, S., Wang, P., Wang, F., Li, H., Feng, J.: Augmented transformer with adaptive graph for temporal action proposal generation. arXiv preprint arXiv:2103.16024 (2021)
13. Wang, L., Yang, H., Wu, W., Yao, H., Huang, H.: Temporal action proposal generation with transformers. arXiv preprint arXiv:2105.12043 (2021)
14. Zhu, Z., Tang, W., Wang, L., Zheng, N., Hua, G.: Enriching local and global contexts for temporal action localization. In: Int. Conf. Comput. Vis. pp. 13516– 13525 (2021)

15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high performance deep learning library. In: Adv. Neural Inform. Process. Syst. vol. 32 (2019)
16. Sridhar, D., Quader, N., Muralidharan, S., Li, Y., Dai, P., Lu, J.: Class semantics based attention for action detection. In: Int. Conf. Comput. Vis. pp. 13739–13748 (2021)
17. Lin, T.Y., Doll´ar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2117–2125 (2017)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inform. Process. Syst. pp. 5998–6008 (2017)

19. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J´egou, H.: Training data-efficient image transformers & distillation through attention. In: Int. Conf. Mach. Learn. pp. 10347–10357 (2021)
20. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In: Int. Conf. Comput. Vis. (2021)
21. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: ViViT: A video vision transformer. In: Int. Conf. Comput. Vis. (2021)
22. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformernetwork. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 244–253 (2019)