

OpenTAD: A Unified Framework and Comprehensive Study of Temporal Action Detection

Shuming Liu* Chen Zhao* Fatimah Zohra Mattia Soldan Alejandro Pardo
 Mengmeng Xu Lama Alssum Merey Ramazanova Juan León Alcázar
 Anthony Cioppa Silvio Giancola Carlos Hinojosa Bernard Ghanem

Video Understanding Group, Image and Video Understanding Lab (IVUL)
 King Abdullah University of Science and Technology (KAUST)

Abstract

*Temporal action detection (TAD) is a fundamental video understanding task that aims to identify human actions and localize their temporal boundaries in videos. Although this field has achieved remarkable progress in recent years, further progress and real-world applications are impeded by the absence of a standardized framework. Currently, different methods are compared under different implementation settings, evaluation protocols, etc., making it difficult to assess the real effectiveness of a specific technique. To address this issue, we propose **OpenTAD**, a unified TAD framework consolidating 16 different TAD methods and 9 standard datasets into a modular codebase. In OpenTAD, minimal effort is required to replace one module with a different design, train a feature-based TAD model in end-to-end mode, or switch between the two. OpenTAD also facilitates straightforward benchmarking across various datasets and enables fair and in-depth comparisons among different methods. With OpenTAD, we comprehensively study how innovations in different network components affect detection performance and identify the most effective design choices through extensive experiments. This study has led to a new state-of-the-art TAD method built upon existing techniques for each component. Our code and models are available at <https://github.com/sming256/OpenTAD>.*

1. Introduction

With the rapid growth of video content across diverse domains—ranging from social media and entertainment to surveillance and autonomous systems—understanding and analyzing videos has become a crucial research focus. While significant progress has been made in classification tasks

such as object and action recognition, temporal reasoning remains a challenging problem [11, 29, 45, 60, 61, 68, 71]. In particular, **temporal action detection (TAD)** [31, 63, 68, 71] has emerged as a fundamental task, serving as a pretext for various downstream applications, including dense video captioning and natural language temporal video grounding [8, 16, 51].

Temporal action detection aims to locate the start and end timestamps of each action instance as well as to identify their categories within a video sequence [31, 63, 68, 71]. Formally, given a video sequence of T frames $\mathcal{V} = \{I_t \in \mathbb{R}^{H \times W \times 3} \}_{t=1}^T$, TAD predicts a set M of action segments: $\{(t_{m,s}, t_{m,e}, s_m)\}_{m=1}^M$, where $t_{m,s}$ and $t_{m,e}$ denote the start and end timestamps of an action, and s_m represents the action label. Conceptually, TAD is analogous to the object detection task [5, 15, 43, 44, 69], which regresses object boundaries as 2D bounding boxes, whereas TAD regresses action boundaries along the temporal dimension of a video.

Various deep neural network-based methods have been proposed for TAD [7, 13, 18, 39, 55, 56, 62, 63, 67, 68], continuously advancing detection performance. Early approaches, such as DAPS [13] and R-C3D [62], pioneered proposal-based methods by adapting techniques from the image domain, particularly the two-stage object detection framework derived from Faster R-CNN [44]. Subsequent works introduced techniques to improve the quality of proposals, focusing on refining boundary localization and exploring alternative loss functions [26]. Meanwhile, other methods have tackled challenges unique to the video domain, such as the large variation in action durations [7]. These variations make it difficult to apply fixed-scale receptive fields, prompting researchers to develop multi-scale modeling techniques that dynamically adapt to different temporal extents.

More recently, recognizing the imbalanced development between video temporal action detection and image object detection techniques, researchers have sought to adapt in-

*Co-First Authors.

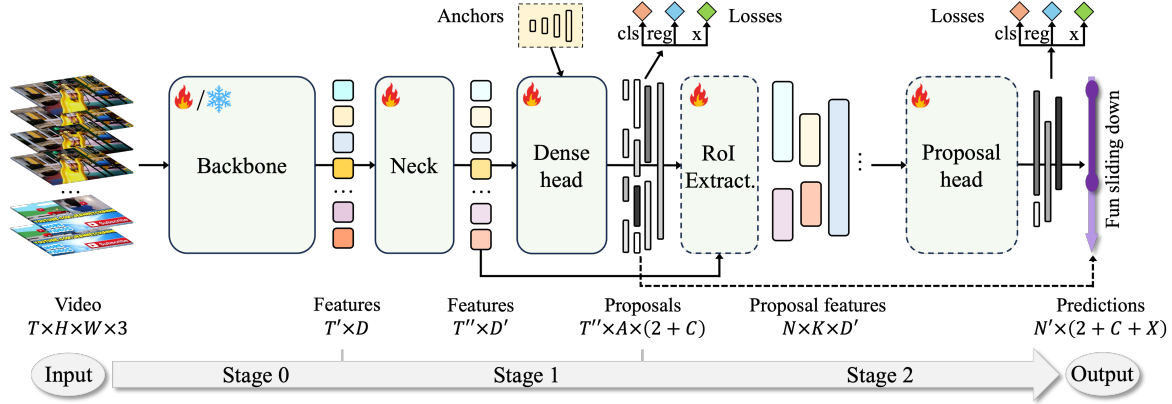


Figure 1. **Unified TAD Pipeline.** Recent TAD methods follow this three-step framework to predict action classes and start/end timestamps from input videos. **1) Stage 0:** Videos are encoded into features using a pretrained video backbone, which may be either fine-tuned or frozen during training. **2) Stage 1:** This stage consists of a neck for temporal aggregation of snippet-level features and a dense head that generates snippet-level predictions. **3) Stage 2 (optional):** This stage further refines action segment proposals using RoI extraction and produces per-proposal predictions.

novations from object detection to TAD. This has led to the emergence of one-stage methods [21, 27, 36, 42, 48, 52, 64, 68] and DETR-based architectures [22, 23, 39, 73], which leverage direct action localization without requiring predefined proposals. Another line of research has incorporated emerging network architectures, including graph convolutional networks (e.g., GTAD [63], P-GCN [67], VSGN [70]), Transformers (e.g., ActionFormer [68], ViT-TAD [65], LIP [24]), and Mamba models (e.g., VideoMambaSuite [9], S-Temba [50]). These architectures have significantly improved the ability to model long-range temporal dependencies and capture fine-grained action boundaries. Furthermore, recent studies suggest that end-to-end training generally yields higher accuracy than feature-based training. To address the memory constraints when using large-scale video encoding networks, researchers have introduced various optimization strategies, as seen in E2E-TAD [38], Re²TAL [71], and AdaTAD [37].

Throughout the brief history of TAD, different methods have focused on distinct aspects of the TAD pipeline, including video representation backbones, detection heads, loss functions, and training strategies. Consequently, a fair comparison between these methods requires evaluation under a unified TAD framework. Unfortunately, existing methods are often implemented in different frameworks, with varying experimental setups and evaluation protocols. As a result, it becomes difficult to determine whether a reported performance gain originates from the proposed innovations, a more effective implementation framework, or simply hyperparameter tuning. To further advance this field, a unified TAD framework is needed—one that holistically evaluates each individual design choice while ensuring fair and reproducible comparisons across different approaches.

This paper introduces **OpenTAD**, an **Open**-sourced uni-

fied framework for **Temporal Action Detection**, accompanied by a comprehensive study of various innovations across different TAD components through extensive experiments. OpenTAD integrates a diverse suite of TAD methods and datasets within a single framework and codebase, facilitating streamlined implementation, fair comparisons, rigorous analysis, and efficient benchmarking. By leveraging OpenTAD, researchers can systematically identify key factors influencing performance—whether stemming from specific module designs, training strategies, or data processing techniques—thereby enabling the development of more advanced TAD models by combining the strengths of existing approaches. Additionally, OpenTAD provides a scalable foundation for extending existing methods to a broader range of datasets and application scenarios. We believe that OpenTAD will foster faithful and efficient development, evaluation, and assessment of novel module designs for TAD. Our contributions are as follows.

- (I) We introduce **OpenTAD**, a unified framework that modularizes the process of temporal action detection. OpenTAD standardizes the implementation of diverse TAD methods, datasets, and evaluation metrics, enabling systematic assessment, in-depth analysis, and extensive benchmarking. With OpenTAD, researchers can seamlessly switch between feature-based and end-to-end training methods based on computational budgets.
- (II) Based on the OpenTAD framework, we provide an open-source code suite, where we have re-implemented **16 diverse methods**, including one-stage, two-stage, DETR-based, and end-to-end approaches, as well as **9 benchmark datasets**. Thanks to its unified design, OpenTAD can be easily extended to support additional methods and datasets, and we remain committed to continuously expanding its capabilities.

- (III) We conduct extensive experiments within OpenTAD to comprehensively analyze various architectural and methodological innovations across different TAD components and datasets. Our study identifies the most effective designs, and by systematically integrating these into existing methods, we achieve new state-of-the-art performance.

2. A Unified TAD Framework

We unify modern temporal action detection methods within a single framework by analyzing the essential roles of different design choices and modularizing the entire pipeline, as illustrated in Fig. 1. To predict action categories and start/end timestamps, the TAD framework consists of three stages of network components, which will be detailed in Sections 2.1, 2.2, and 2.3. Following the introduction of OpenTAD’s core components, we present our unified data preprocessing and postprocessing pipeline in Sec. 2.4. Finally, we describe how various TAD methods are integrated into OpenTAD, categorizing them into one-stage, two-stage, DETR-based, and end-to-end approaches in Sec. 2.5.

OpenTAD is implemented within a unified codebase, which we have open-sourced along with our trained models, datasets, and configurations for all re-implemented methods. In the OpenTAD codebase, each component is designed as an independent module, allowing it to be instantiated in different variants without affecting—or being affected by—other components. This modular design enables seamless integration of novel techniques into the framework while ensuring fair comparisons under a consistent setup.

For clarity, we consistently adopt the following notations to represent the dimensions of video data throughout the OpenTAD pipeline. We use H and W for frame height and width, respectively, T for sequence length, D for the number of feature channels, C for the number of action classes, A for the number of predefined candidate actions per temporal location (referred to as *anchors*), N for the total number of predicted candidate actions (including *proposals* and final predictions), and K for the temporal length of each extracted proposal. The prime notation ($'$) is used to indicate a value change in a specific dimension.

2.1. Stage 0: Video Feature Extraction

The first computational block of the TAD pipeline encodes raw videos into snippet-level or frame-level features, which serve as generic video representations requiring further processing by subsequent TAD-specific components. Therefore, we designate this step as **Stage 0**.

Backbone: $T \times H \times W \times 3 \rightarrow T' \times D$. The backbone is responsible for feature extraction, encoding an input video sequence $\mathcal{V}$ into a sequence of feature vectors $\mathcal{F}$. It consists of multiple layers that perform spatio-temporal aggregation, such as 3D convolution (e.g., I3D [6], SlowFast [14]) or space-time attention (e.g., VideoSwin [41], VideoMAE [53]).

This backbone is typically pretrained on a large-scale dataset of short video clips, such as Kinetics [74], using classification tasks or self-supervised learning [53].

We identify two primary approaches for encoding video sequences using the backbone: (1) **snippet encoding**. A snippet refers to a short video clip consisting of a small number of frames. In this approach, the video sequence is divided into T' snippets, typically with overlap, and each snippet is processed independently by the backbone to generate a feature vector, i.e., temporal aggregation occurs only within each snippet. Both the spatial and temporal dimensions are globally pooled in the output feature vector for each snippet. (2) **frame encoding**. In contrast, this approach processes the entire video sequence $\mathcal{V}$ as a single long clip, meaning that all frames are temporally-aware within the backbone. The spatial dimension is globally pooled, while the temporal dimension T is preserved in the output feature vectors. Frame encoding offers greater computational and memory efficiency by reducing redundant computations across neighboring snippets, making it particularly suitable for end-to-end training (Sec. 2.5). On the other hand, snippet encoding processes each snippet independently, making it more appropriate for offline feature extraction in feature-based TAD settings. OpenTAD supports both approaches, ensuring flexibility for different experimental setups.

2.2. Stage 1: Initial Prediction

This stage consists of a neck component for temporal aggregation of video features and a dense head that directly predicts candidate actions. It corresponds to the first stage in two-stage methods (e.g., BMN [31], GTAD [63]) and the sole stage in one-stage methods (e.g., ActionFormer [68], TriDet [48]). Therefore, we designate this step as **Stage 1**.

Neck: $T' \times D \rightarrow T'' \times D'$. The neck performs temporal aggregation across the entire feature sequence $\mathcal{F}$. Compared to the backbone, it is more lightweight and can provide either single-scale [31, 63] or multi-scale [68, 70] temporal features. The neck is a crucial component for TAD as it consists of temporal aggregation modules, such as 1D convolution [30, 31], graph networks [63], temporal attention [68], or state-space models (SSM) [9], to capture complex temporal relations. In Sec. 3.1, we analyze the importance of the neck and compare the effectiveness of different designs.

Anchor Generation: Anchors [43] are predefined candidate action segments at each temporal location, allowing the model to predict offsets relative to these anchors instead of directly estimating action boundaries. This strategy simplifies training and improves localization accuracy. Anchors can be defined in different formats, such as fixed temporal segment sizes centered at each location [70] or all possible start-end combinations [63]. Recent one-stage methods have shifted towards an anchor-free approach, where the action start and end offset are directly predicted [68, 70].

Dense Head: $T'' \times D' \rightarrow T'' \times (2A + CA + X)$. The dense head makes predictions at each temporal location after the neck. In one-stage methods, these predictions are directly used for post-processing, whereas in two-stage methods, they serve as *proposals* to be further refined in the next stage. We categorize the predicted outputs at each temporal location as follows: (1) *start/end offsets*: $2A$, where A represents the number of anchors at each temporal location. These offsets define the distances of ground truth actions relative to predefined anchors. (2) *confidence scores of each category*: CA , where C denotes the number of action categories. In some cases, C is set to 2 to represent the presence or absence of an action. (3) *other auxiliary predictions*, which include probabilities of being inside an action segment or located on a start/end boundary [30, 63]. These prediction branches are optimized using different loss functions, *e.g.*, GIoU loss [46] for the start/end offsets regression [70] and cross-entropy loss or focal loss [32] for category prediction [68].

2.3. Stage 2: RoI Extraction and Action Refinement

Given the predicted action candidates from stage 1, *i.e.*, proposals, this stage further refines their boundaries and confidence scores to yield higher-quality action predictions. To achieve this, two key components are required: RoI extraction and the proposal head. This stage is unique to two-stage methods [31, 63] and is therefore designated as **Stage 2**.

RoI Extraction: $T'' \times D' \rightarrow N \times K \times D'$. Given the predicted action candidates, the corresponding proposal features are constructed through the region of interest (RoI) extraction module. Specifically, features within the proposal boundaries are extracted to generate N proposal features, each with a temporal length of K and a feature dimension of D' . Different RoI extraction methods process features differently. RoI Align [7] and SGAlign [63] utilize all features within the boundaries, whereas PBRNet [33] and VSGN [70] extract only center and boundary features. Additionally, the boundary matching mechanism in BMN [31] can also be regarded as a matrix implementation of RoI alignment.

Proposal Head: $N \times K \times D' \rightarrow N \times (2 + C + X')$. The proposal head processes each proposal feature to refine its boundaries and predict the action category [27, 70]. Some methods also include additional predictions, such as completeness scores [31, 63]. To further improve localization accuracy, multiple proposal heads can be stacked sequentially to form a cascade refinement [34]. When dealing with a large number of proposals, the computational complexity of the proposal head can significantly exceed that of the dense head. This is one of the primary reasons why one-stage methods are preferred in scenarios where efficiency is a priority. Similar to the dense head, regression loss and classification loss are used to supervise boundary refinement and action classification, respectively, although different label assignment strategies are often employed.

2.4. Data Pre-processing and Post-processing

Beyond the neural network components, pre-processing of video data and post-processing of proposal predictions play crucial roles in improving TAD performance.

Pre-processing: Since the TAD task involves long video sequences with variable durations, each input video must be mapped to a fixed length T for batch training. In OpenTAD, we support three pre-processing approaches to accommodate various methods: (1) Rescaling the entire video sequence to a fixed length of T frames/features, *e.g.*, via interpolation. This approach is often used in datasets such as ActivityNet [4] and HACS [72], where action durations vary significantly and may span almost the entire video. (2) Randomly extracting a continuous video segment of T frames from a long video. In this setting, a subset of frames is used for training, as seen in ActionFormer on THUMOS [68]. If the video is shorter than T , zero-padding is applied. (3) Applying a sliding window to partition each video into segments, resulting in multiple training samples from a single video per epoch [31, 63].

Post-processing: Once action predictions are generated from either the dense head or the proposal head—typically numbering in the thousands—redundant proposals must be reduced. To achieve this, techniques such as non-maximum suppression (NMS) or Soft-NMS [3] are commonly employed. If a video has been partitioned into multiple segments using a sliding window during pre-processing, all predictions from these segments are first aggregated before applying NMS or Soft-NMS.

2.5. Methods and Datasets Unified in OpenTAD

We have modularized and re-implemented **16 TAD methods** under the OpenTAD framework and provide a detailed mapping of how each method instantiates OpenTAD components in Tab.14 (supplementary material). Additionally, we compare re-implementation performance to original results in Tab.7, and present benchmarking results on two datasets in Tab.9 and Tab.10, also in the supplementary material.

These methods span various categories, each contributing to different components, as detailed in Table 1. We categorize a TAD method as **end-to-end** if it jointly trains the backbone with other components, *e.g.*, Re²TAL [71] and AdaTAD [37]. Conversely, a method is *feature-based* if it uses pre-extracted features during training, *e.g.*, BMN [31] and ActionFormer [68]. Feature-based methods are further divided into *one-stage* or *two-stage*, depending on whether they incorporate Stage 2 components. We also classify methods as *DETR-based* if they adopt the encoder-decoder architecture of DETR to directly learn action queries (*e.g.*, TadTR [39]). In Column 3 of Table 1, we indicate the OpenTAD component corresponding to each method’s main contribution, while in Column 4, we summarize the claimed innovations related to that component.

Table 1. **Example TAD methods unified in OpenTAD with their main contribution.** **Column 3:** the OpenTAD component corresponding to each paper’s primary contribution. **Column 4:** the claimed innovations related to the target component.

Category	Method	Target	Innovation
Feature-based Approaches			
One-stage	ActionFormer [68]	Neck	Transformer
	TriDet [48]	Neck; Head	SGP; Trident Head
	TemporalMaxer [52]	Neck	MaxPooling
	VideoMambaSuite [9]	Neck	Mamba
	DyFADet [64]	Neck; Head	Dynamic Aggregation
Two-stage	CausalTAD [36]	Neck	Causal Modeling
	BMN [31]	RoI	Boundary Matching
	GTAD [63]	Neck	Graph Network
	TSI [34]	Loss	Scale-Invariant Loss
	VSGN [70]	Neck	Cross-Scale Graph
DETR	TadTR [39]	Architecture	DETR + RoI Extraction
End-to-End Approaches			
Two-stage	AFSD [27]	RoI	Boundary Pooling
DETR	E2E-TAD [38]	Training	Empirical Study
Two-stage	ETAD [35]	Training	Gradient Sampling
One Two	Re ² TAL [71]	Backbone	Reversible Finetuning
One-stage	AdaTAD [37]	Backbone	Adapter Tuning

OpenTAD supports **9 TAD datasets**, including widely used benchmarks such as ActivityNet-v1.3 [4], THUMOS-14 [20], HACS [72], and EPIC-Kitchens 100 [10], as well as recently released datasets like Ego4D-Moment Query [16] and FineAction [40]. Beyond single-label detection, OpenTAD also supports multi-label datasets such as Multi-THUMOS [66] and Charades [49], and extends to audio-based action detection with EPIC-Sounds [19]. Thanks to its modular design, switching between datasets in OpenTAD is straightforward, enabling seamless adaptation of prior methods to all 9 datasets — even those not originally designed for them. In Table 8 (supplementary material), we report ActionFormer’s performance under OpenTAD across all datasets. Our results match or exceed prior performance on benchmarks like THUMOS-14 and EPIC-Kitchens, and set new state-of-the-art on emerging datasets such as FineAction.

3. Experiments: A Comprehensive Study

Despite numerous innovations in TAD, evaluating their generalizability across different methods remains challenging. In this section, we leverage our unified framework, OpenTAD, with its modularized codebase to conduct a comprehensive study, fairly comparing various module designs and systematically analyzing how individual components, model sizes, overall architectures, and other factors affect TAD performance. We perform experiments on two most widely adopted datasets: ActivityNet-v1.3 [4] and THUMOS-14 [20]. We describe the evaluation metrics and implementation details in Sec. B of supplementary material.

3.1. Which design for each component yields the best performance?

As described in Sec. 2.5, various design choices have been proposed for each component in the literature. Among all the variants, is there an absolute winner for each component?

Neck. This is an essential component present in almost every TAD method. As shown in Tab. 1, many methods propose innovative designs in the neck to enhance the temporal aggregation of video sequence features. Examples include the simple MaxPooling in TemporalMaxer [52], the convolutional layer in BSN [30] and BMN [31], the GCN layer in GTAD [63] and VSGN [70], the Transformer block in ActionFormer [68], and the Mamba block in VideoMambaSuite [9]. We incorporate these neck designs into four different TAD methods: ActionFormer [68] and TriDet [48], which produce multi-scale features, and GTAD [63] and BMN [31], which produce single-scale features. We then compare their contributions to the final performance for each method, as shown in Tab. 2 and Tab. 11 in the supplementary material. *Macro Block* in Column 1 refers to the highest-level repetitive building blocks, such as the following: (a) Transformer block, which contains a sequential modeling module and an MLP module, each with a residual connection; (b) Mamba block, which consists of a sequential modeling branch combined with a gated MLP branch; and (c) Block mix, which is a combination of the above two types of blocks. We partition Tab. 2 and Tab. 11 into **four regions** based on the *macro blocks*. Besides, *Sequential Module* in Column 2 refers to the core sequential modeling module within the macro block, such as self-attention and SSM.

From the top region in Tab. 2, we identify two winners: the Transformer block and the Mamba block, both of which feature more sophisticated macro architectures than the others. The Transformer block achieves the best performance for the single-scale methods BMN (50.36%) and GTAD (50.97%), while the Mamba block performs best for the multi-scale methods ActionFormer (68.28%) and TriDet (68.64%). To further investigate which sequential modeling module within each macro architecture is the best choice, we replace the self-attention module in the Transformer block and the SSM module in the Mamba block with one of five sequential modeling modules: convolution, graph convolution, LSTM, self-attention, and SSM. We present the results in the *Transformer Block* and *Mamba Block* regions of Tab. 2 and Tab. 11. SSM demonstrates a clear advantage for the multi-scale methods ActionFormer and TriDet, while LSTM and self-attention exhibit superior performance for the single-scale methods BMN and GTAD. We attribute the strong performance of these three modules to their ability to perform temporal aggregation across the entire sequence, in contrast to the others.

Can we identify a neck design that performs best across all methods? Considering the outstanding performance of

Table 2. **Analysis of the neck design choices**, measured by average mAP(%) on THUMOS-14. We ran all experiments 5 times and report their mean and standard deviation. The **4 macro-block regions** mean the following respectively. **Top region**: macro blocks with their original sequential modules are adopted as a whole; **Transformer block**: self-attention modules in Transformer blocks are replaced with different sequential modules; **Mamba block**: SSM modules in Mamba blocks are replaced with different sequential modules; **Bottom region**: a combination of two blocks.

Neck		Method			
Macro Block	Sequential Module	ActionFormer	TriDet	BMN	GTAD
Convolution Block	Convolution [31]	52.58±1.54	67.92±0.22	48.43±0.38	49.62±0.22
GCN Block	Graph convolution [63]	67.48±0.22	68.32±0.39	48.84±0.26	50.36±0.29
Transformer Block	Self-attention [68]	67.93±0.19	68.58±0.20	50.36±0.56	50.97±0.31
Mamba Block	SSM [9]	68.28±0.48	68.64±0.18	49.39±0.33	50.22±0.35
Transformer Block	Convolution	68.11±0.36	68.46±0.52	48.73±0.15	50.31±0.16
	Graph convolution	67.85±0.26	68.51±0.11	48.84±0.26	50.33±0.18
	Self-attention	67.93±0.19	68.58±0.20	50.36±0.56	50.97±0.31
	LSTM	68.10±0.33	68.53±0.27	51.66±0.35	53.31±0.27
	SSM	68.23±0.31	68.79±0.37	49.23±0.28	50.09±0.35
Mamba Block	Convolution	68.25±0.34	68.61±0.41	48.40±0.20	50.11±0.40
	Graph convolution	68.02±0.29	68.50±0.45	48.73±0.28	49.84±0.33
	Self-attention	68.24±0.22	68.63±0.26	49.57±0.72	50.89±0.70
	LSTM	67.82±0.29	68.85±0.56	49.20±0.27	49.87±0.31
	SSM	68.28±0.48	68.64±0.18	49.39±0.33	50.22±0.35
Mamba + Mamba	SSM + SSM	68.03±0.38	68.09±0.87	49.66±0.17	50.40±0.48
Transf. + Transf.	LSTM + Self-attention	67.78±0.31	68.59±0.62	50.91±0.27	53.46±0.56
Mamba + Transf.	SSM + Self-attention	68.37±0.22	68.41±0.31	49.76±0.46	50.60±0.39
Mamba + Transf.	SSM + LSTM	68.44±0.41	68.90±0.60	52.19±0.28	53.70±0.26

Table 3. **Analysis of the backbone design choices**, measured by average mAP (%). Please note that all experiments use only the RGB modality, excluding the optical flow modality. Backbone models with * have been pre-trained using self-supervised learning. **Top region**: Resnet-based backbone models; **bottom region**: Transformer-based backbone models including VideoSwin and ViT architectures. ‘Param.’: number of model parameters; K400: Kinetics-400; Top-1: top-1 accuracy (%). We highlight three backbones that don’t have consistent action recognition and TAD performance.

Backbone	Param.	K400	ActivityNet-v1.3			THUMOS-14		
Model	(M)	Top-1	ActionFormer	GTAD	VSGN	ActionFormer	GTAD	VSGN
TSN-R50 [57]	24	74.12	34.64±0.07	34.09±0.09	33.36±0.18	49.79±0.37	35.18±0.32	32.22±0.79
TSM-R50 [28]	24	75.12	34.51±0.08	33.70±0.09	34.35±0.21	50.94±0.30	34.59±0.15	35.09±0.77
R(2+1)D-R34 [54]	64	75.46	35.06±0.09	34.11±0.06	34.83±0.22	48.39±0.39	34.12±0.23	39.08±0.76
SlowFast-R101 [14]	63	78.65	35.95±0.12	35.28±0.12	35.98±0.19	63.04±0.33	47.08±0.28	50.54±0.22
VideoSwin-S [41]	50	80.54	35.60±0.13	35.24±0.10	32.85±0.22	56.72±0.70	38.99±0.32	42.14±0.56
VideoSwin-L [41]	197	83.46	35.91±0.03	35.73±0.10	32.93±0.32	58.78±0.38	41.55±0.16	45.66±0.59
MViTv2-L [25]	213	85.40	35.96±0.18	35.30±0.12	34.23±0.18	47.62±0.48	31.27±0.51	41.78±0.47
VideoMAE-H* [53]	633	87.40	36.96±0.07	36.57±0.07	35.07±0.32	67.23±0.32	48.16±0.37	53.93±0.34
VideoMAEv2-G* [58]	1B	88.40	-	-	-	71.48±0.12	54.16±0.59	54.94±0.64
InternVideo2-6B* [59]	6B	92.10	38.95±0.13	37.35±0.12	38.63±0.09	72.36±0.13	56.32±0.16	58.16±0.66

the Transformer and Mamba blocks, as well as the LSTM module, we propose a novel macro block that combines these two different blocks, as shown in the bottom region of Tab. 2 and Tab. 11. By mixing Mamba and Transformer macro blocks—where the self-attention in the Transformer block is replaced with LSTM—we achieve the highest performance across all four methods on both datasets, leading to an additional 0.8% improvement on THUMOS-14 over the best-performing macro block with the optimal sequential module.

RoI Extraction. This component is utilized in two-stage methods and determines the quality of the proposal features used in Stage 2. Keypoint sampling, employed in PBR-Net [33] and VSGN [70], extracts only the boundary and center features within a proposal. RoI Align [17] utilizes all features, while SGAlign [63], which builds upon RoI Align, further enhances it by incorporating a GCN layer. Boundary matching, introduced in BMN [31], densely extracts proposals based on every pair of start and end boundaries. We evaluate these four RoI strategies using three two-stage

Table 4. **Ablation study of Stage 2** on one-stage, two-stage, and DETR-based methods, measured by average mAP (%). N/A means training details are not provided in the original paper or code.

Category	Method	ActivityNet-v1.3	THUMOS-14	HACS
One-stage	ActionFormer [68]	37.00±0.05	67.93±0.19	37.69±0.05
	ActionFormer + Stage 2	37.29 ±0.03	68.61 ±0.29	38.25 ±0.12
Two-stage	VSGN - Stage 2	36.18±0.07	49.71±0.38	36.02±0.20
	VSGN [70]	36.71 ±0.13	52.42 ±0.31	37.19 ±0.12
DETR-based	TadTR - Stage 2	N/A	52.75±0.55	N/A
	TadTR [39]	N/A	54.42 ±0.72	N/A

Table 5. **Progressive modification of components in ActionFormer [68]**, performance measured by average mAP (%). **Row 1**: baseline ActionFormer with TSP features for ActivityNet-v1.3 and I3D features for THUMOS-14; **Top region**: progressive adding or removing a component or a technique; **Bottom region**: replacing the backbone and applying all the modifications in the top region.

Setting	ActivityNet-v1.3	THUMOS-14
ActionFormer (TSP backbone I3D backbone): baseline	37.00±0.05	67.93±0.19
- Remove the transformer block	36.04±0.33	67.52±0.36
+ Add block mix	37.50±0.07	68.44±0.41
+ Add the second stage to refine proposals	37.59±0.08	68.63±0.33
+ Use input channel dropout, longer training epochs	37.62 ±0.02	68.81 ±0.43
Replace the backbone with InternVideo2	38.95±0.13	72.36±0.13
+ All the above modifications	39.51 ±0.09	73.42 ±0.40

methods, with the results presented in Tab. 12 in the supplementary material.

For both BMN and GTAD, boundary matching achieves the highest performance, while keypoint sampling yields the lowest. This is because both methods use a limited number of features due to their single-scale nature and the relatively small number of temporal length (e.g., 100 for ActivityNet). Given these constraints, the computational cost of using boundary matching to extract the largest number of proposals is acceptable. In contrast, for VSGN, which employs multi-scale feature maps and a larger number of input frames (e.g., 1280 frames for ActivityNet), applying dense boundary matching is infeasible, and keypoint sampling proves to be as effective as RoI Align. Therefore, the choice of RoI extraction method depends on specific properties of each approach, such as feature temporal scales. Since multi-scale, long-input methods like VSGN generally outperform conventional methods such as BMN and GTAD, boundary matching is generally not preferred.

Backbone. It produces video features, which are to be further processed by the subsequent components of the TAD framework (e.g., the neck and the heads) to make predictions. Therefore, the quality of the video features and the backbone is critical, significantly influencing the detection performance. A backbone network is usually pretrained on a larger-scale action recognition dataset such as Kinetics [74] before being trained for TAD, and example backbones are TSN [57], SlowFast [14], and VideoSwin [41]. Recently, ViT [12] is adopted as video backbone network,

pre-pretrained with self-supervised learning, multi-modal joint learning, or other methods (e.g., VideoMAE [53], InternVideo2 [59]), and has shown remarkable performance on TAD. In Tab. 3, we present experiments across three TAD methods and two datasets, evaluating different backbone models. These models vary in architecture, including ResNet-based and Transformer-based designs, and in pre-training strategies, including action recognition pretraining and additional pre-pretraining.

In Tab. 3, the backbone model sizes are arranged in non-decreasing order from top to bottom within each group: ResNet-based backbones in the top region and Transformer-based backbones in the bottom region. We observe that action recognition performance on Kinetics-400 consistently improves as model sizes increase (Column 2), and that all the Transformer-based models perform better than the ResNet-based models. However, this trend does not always hold for TAD. Though VideoSwin-S [41] and MViTv2-L [25] have significantly higher action recognition accuracy than SlowFast-101 [14], their mAPs are lower for most TAD experiments. Similarly, while MViTv2-L outperforms VideoSwin-L on Kinetics, it performs worse on THUMOS-14 for TAD. We attribute these inconsistencies across tasks to the architectural differences and suggest that ViT-based models are more effective for TAD. Notably, the largest ViT-based model, InternVideo2-6B [59] achieves the highest performance across all methods and datasets.

Table 6. **Results on ActivityNet-v1.3 and THUMOS-14**, measured by mAP (%) at different tIoU thresholds. TSP [1] features are used for ActivityNet, and I3D [6] features are used for THUMOS for all methods except VideoMambaSuite, which uses the InternVideo2 [59] features for both datasets.

Method	ActivityNet-v1.3				THUMOS-14					
	0.5	0.75	0.95	Avg.	0.3	0.4	0.5	0.6	0.7	Avg.
GTAD [63]	52.33	37.58	8.42	36.20	64.35	59.07	51.76	42.65	31.66	49.70
GTAD +	52.46	37.86	9.55	36.67	67.08	62.87	56.81	46.68	35.74	53.83
BMN [31]	52.90	37.30	9.67	36.40	64.99	60.70	54.54	44.11	34.16	51.80
BMN +	53.82	38.14	9.92	37.04	66.82	61.94	55.22	45.16	34.11	52.65
ActionFormer [68]	55.08	38.27	8.91	37.07	83.78	80.06	73.16	60.46	44.72	68.44
ActionFormer +	55.76	39.13	7.43	37.68	83.05	79.20	72.20	62.73	47.25	68.89
TriDet [48]	54.84	37.46	7.98	36.51	84.46	81.05	73.41	62.58	46.51	69.60
TriDet +	55.49	38.65	8.95	37.47	84.49	80.77	73.71	62.22	47.16	69.67
VideoMambaSuite [9]	63.13	44.36	10.36	42.80	87.30	82.95	77.17	67.06	51.74	73.24
VideoMambaSuite +	63.22	45.01	8.79	43.02	86.79	83.11	77.54	67.21	53.66	73.66

3.2. Is the optional Stage 2 necessary?

Stage 2 is optional in the TAD framework, as it is only used in two-stage methods. While one-stage methods were originally introduced to improve efficiency in object detection [47], the question remains: is Stage 2 necessary for achieving higher accuracy?

To answer this question, we compare the performance of methods with and without Stage 2 across three datasets in Tab. 4, including originally one-stage, two-stage, and DETR-based methods. For one-stage methods, we incorporate Stage 2 by using RoI Align [17] for RoI extraction and applying the same classification and regression losses as in Stage 1. We define positive samples as proposals with tIoU > 0.7 relative to the ground-truth actions. For two-stage methods, we entirely remove Stage 2, using Stage 1 predictions as the final output for ablation. Additionally, since the DETR-based method TadTR [39] includes a second-stage refinement, we also evaluate its performance without this stage for comparison. As shown in Tab. 4, adding Stage 2 improves the performance of the one-stage method ActionFormer[68], while removing Stage 2 reduces the performance of the two-stage method VSGN [70] and the DETR-based method TadTR. These results indicate that Stage 2 can further enhance one-stage methods, and is essential for two-stage methods.

3.3. What makes an optimal detector?

Through the above analysis within our unified OpenTAD framework, we have identified the optimal design choices for the neck, RoI extraction, and backbone components. This raises an intriguing question: Can we further enhance existing TAD methods by applying these winning techniques?

In Tab. 5, we use ActionFormer as the baseline (Row 1) and progressively apply the winning techniques—excluding backbone modifications—across different components (Rows 2–5). We evaluate the impact of these changes on two datasets. As expected, performance

improves progressively as more techniques are applied, culminating in a model with mAP improvements of 0.62/0.88 over the original. Next, we separately replace the backbone with the winning InternVideo2 model and then apply all other techniques (Rows 6–7). While backbone replacement alone substantially boosts performance, incorporating additional modifications further enhances the results, achieving an overall improvement of 0.56/1.06 mAP.

In Tab. 6, we apply all the modifications—except for the backbone replacement—to five different TAD methods across two datasets, consistently improving their performance. Notably, for the recent method VideoMambaSuite [9], which utilizes the powerful InternVideo2 backbone, our modifications further enhance its results, achieving state-of-the-art performance on both ActivityNet-v1.3 and THUMOS-14.

4. Conclusion

In this work, we introduced OpenTAD, a unified framework that standardizes and modularizes temporal action detection across diverse methods and datasets, providing the most extensive TAD codebase available, to the best of our knowledge. OpenTAD facilitates fair comparisons across different methods and enables seamless plug-in implementations for new innovations within its modular components. Leveraging OpenTAD, we conducted a comprehensive analysis and identified key factors that significantly impact TAD performance, such as second-stage processing and neck architectures. Based on these insights, we developed a new state-of-the-art TAD method. We invite the research community to engage with OpenTAD, which is publicly available along with all associated resources, to drive further innovation in video understanding. Through collaborative efforts, we envision OpenTAD advancing not only action detection but also broader video understanding tasks.

References

- [1] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. TSP: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8, 15
- [2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 12
- [3] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV)*, 2017. 4
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4, 5, 13
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 1
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 8
- [7] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. Rethinking the Faster R-CNN architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 4
- [8] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. InternVideo-Ego4D: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022. 1
- [9] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 2, 3, 5, 6, 8, 13, 14, 16
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5, 13
- [11] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2694–2703, 2023. 1
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 7
- [13] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1
- [14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International conference on computer vision (ICCV)*, 2019. 3, 6, 7
- [15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, and et al. Liu, Xingyu. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 5, 13
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017. 6, 8, 15
- [18] Kai Hu, Chaowen Shen, Tianyan Wang, Keer Xu, Qingfeng Xia, Min Xia, and Chengxue Cai. Overview of temporal action detection based on deep learning. *Artificial Intelligence Review*, 2024. 1, 12
- [19] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2023. 5, 13
- [20] 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. <http://cvcv.ucf.edu/THUMOS14/>, 2014. 5, 13
- [21] Tae-Kyung Kang, Gun-Hee Lee, and Seong-Whan Lee. HT-Net: Anchor-free temporal action localization with hierarchical transformers. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 365–370. IEEE, 2022. 2
- [22] Jihwan Kim, Miso Lee, and Jae-Pil Heo. Self-feedback DETR for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10286–10296, 2023. 2
- [23] Jihwan Kim, Miso Lee, Cheol-Ho Cho, Jihyun Lee, and Jae-Pil Heo. Prediction-feedback DETR for temporal action detection. *arXiv preprint arXiv:2408.16729*, 2024. 2
- [24] Jihwan Kim, Miso Lee, and Jae-Pil Heo. Long-term pre-training for temporal action detection with transformers. *arXiv preprint arXiv:2408.13152*, 2024. 2
- [25] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViT2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7

- [26] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1
- [27] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 5, 13, 14, 16
- [28] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6
- [29] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wen-zhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1
- [30] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3, 4, 5, 14
- [31] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *Proceedings of The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4, 5, 6, 8, 12, 13, 14, 15, 16
- [32] Tsung-Yi Lin, Priyal Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42, 2020. 4, 14
- [33] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 4, 6, 14
- [34] Shuming Liu, Xu Zhao, Haisheng Su, and Zhilan Hu. Tsi: Temporal scale invariant network for action proposal generation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 4, 5, 12, 13, 14, 16
- [35] Shuming Liu, Mengmeng Xu, Chen Zhao, Xu Zhao, and Bernard Ghanem. ETAD: A unified framework for efficient temporal action detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2023. 5, 13, 14, 16
- [36] Shuming Liu, Lin Sui, Chen-Lin Zhang, Fangzhou Mu, Chen Zhao, and Bernard Ghanem. Harnessing temporal causality for advanced temporal action detection. *arXiv preprint arXiv:2407.17792*, 2024. 2, 5, 13, 14, 16
- [37] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4, 5, 12, 13, 16
- [38] Xiaolong Liu, Song Bai, and Xiang Bai. An empirical study of end-to-end temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5, 13, 16
- [39] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing (TIP)*, 2022. 1, 2, 4, 5, 7, 8, 12, 13, 14, 16
- [40] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing (TIP)*, 31, 2022. 5, 13
- [41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6, 7
- [42] Ranyu Ning, Can Zhang, and Yuexian Zou. SRF-Net: Selective receptive field network for anchor-free temporal action detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2460–2464. IEEE, 2021. 2
- [43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3
- [44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NeurIPS)*, 2015. 1
- [45] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023. 1
- [46] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4
- [47] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 8
- [48] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5, 8, 12, 13, 14, 16
- [49] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 5, 13
- [50] Arkaprava Sinha, Monish Soundar Raj, Pu Wang, Ahmed Helmy, and Srijan Das. MS-Temba: Multi-scale temporal mamba for efficient temporal action detection. *arXiv preprint arXiv:2501.06138*, 2025. 2

- [51] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1
- [52] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 2, 5, 13, 14, 16
- [53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3, 6, 7
- [54] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6
- [55] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 1, 12
- [56] Binglu Wang, Yongqiang Zhao, Le Yang, Teng Long, and Xuelong Li. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 1, 12
- [57] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6, 7
- [58] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6
- [59] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 6, 7, 8
- [60] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1
- [61] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 1
- [62] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region convolutional 3D network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1
- [63] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 4, 5, 6, 8, 12, 13, 14, 15, 16
- [64] Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. Dyfadet: Dynamic feature aggregation for temporal action detection. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 5, 13, 16
- [65] Min Yang, Huan Gao, Ping Guo, and Limin Wang. Adapting short-term transformers for action detection in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18570–18579, 2024. 2
- [66] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision (IJCV)*, 2018. 5, 13
- [67] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2
- [68] Chenlin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 13, 14, 16
- [69] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [70] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16
- [71] Chen Zhao, Shuming Liu, Karttikeya Mangalam, and Bernard Ghanem. Re²TAL: Rewiring pretrained video backbones for reversible temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4, 5, 12, 13, 16
- [72] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4, 5, 13
- [73] Yuhao Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detr for multi-label temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18559–18569, 2024. 2
- [74] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 7