Temporal Action Localization with Cross Layer Task Decoupling and Refinement

Qiang Li[1,4*], Di Liu[1,2*], Jun Kong[1,3†], Sen Li[1], Hui Xu[4], Jianzhong Wang[1†]

[1]Northeast Normal University
[2]Northeast Electric Power University
[3]KLAS of MOE
[4]Changchun Humanities and Sciences College
{ liq782, kongjun, lis084, wangjz019 }@nenu.edu.cn, 20102313@neepu.edu.cn, xuhui1@ccrw.edu.cn

Abstract

Temporal action localization (TAL) involves dual tasks to classify and localize actions within untrimmed videos. However, the two tasks often have conflicting requirements for features. Existing methods typically employ separate heads for classification and localization tasks but share the same input feature, leading to suboptimal performance. To address this issue, we propose a novel TAL method with Cross Layer Task Decoupling and Refinement (CLTDR). Based on the feature pyramid of video, CLTDR strategy integrates semantically strong features from higher pyramid layers and detailed boundary-aware boundary features from lower pyramid layers to effectively disentangle the action classification and localization tasks. Moreover, the multiple features from cross layers are also employed to refine and align the disentangled classification and regression results. At last, a lightweight Gated Multi-Granularity (GMG) module is proposed to comprehensively extract and aggregate video features at instant, local, and global temporal granularities. Benefiting from the CLTDR and GMG modules, our method achieves state-of-the-art performance on five challenging benchmarks: THUMOS14, MultiTHUMOS, EPIC-KITCHENS-100, ActivityNet-1.3, and HACS.

Code — https://github.com/LiQiang0307/CLTDR-GMG

Introduction

With the aim of accurately identifying action categories and localizing its start and end times within a video, temporal action localization (TAL) has garnered significant attention from the research community due to its wide applications, including intelligent surveillance, video summarization, highlight detection, and visual question answering.

Recent advancements in deep learning techniques have led to remarkable progress in various computer vision fields, including TAL. Numerous deep learning methods have been proposed to address the TAL problem, such as (Shou, Wang, and Chang 2016; Nag et al. 2022; Long et al. 2019). A typical deep learning-based TAL model comprises three main steps: extracting video features using deep learning networks, encoding these features into a multi-layer feature pyramid to capture action information at different temporal scales and using a decoder with classification and regression heads for action recognition and localization. The initial step of video feature extraction is often accomplished using pretrained video classification backbones. Consequently, the encoder and decoder in the subsequent steps become critical for achieving high TAL performance. To model the temporal context of features, various techniques have been employed in TAL, including 1D CNN (Lin et al. 2021), GCN (Xu et al. 2020), xGPN (Zhao, Thabet, and Ghanem 2021) and RNN (Buch et al. 2017). Recently, the Transformer (Vaswani et al. 2017) with self-attention mechanism and its variants have also been adopted to capture long-range temporal dependency between video features (Zhang, Wu, and Li 2022; Tang, Kim, and Sohn 2023; Shi et al. 2023). Nevertheless, the existing TAL methods still encounter two limitations.

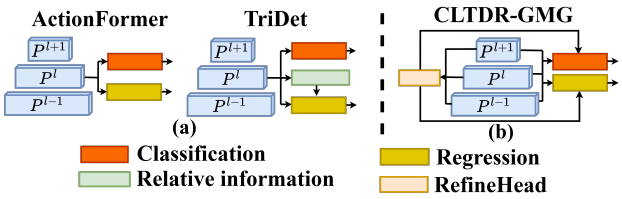

Figure 1: Comparison of different task decoupling. (a) Previous methods use two head branches share the same input feature; (b) Our method uses cross layer features for task decoupling and refinement. Zoom in for better view.

On the one hand, current methods often struggle to balance the trade-off between classification and localization. Typically, these methods adopt two separated convolutional heads at each layer of video feature pyramid to accomplish the category classification and boundary regression of actions. However, this strategy relies heavily on learned parameters in the convolutional networks to decouple tasks because both heads process identical features at each pyramid layer (Zhang, Wu, and Li 2022). Although TriDet (Shi et al. 2023) incorporates relative information to help boundary prediction, this information is still extracted independently at each feature pyramid layer (Fig. 1a). Thus, its overall performance may still be suboptimal. Moreover, existing TAL methods often lack explicit interaction between classi-

*These authors contributed equally.
†Corresponding authors.

fication and localization heads, which would exacerbate the inconsistency between their predictions.

On the other hand, the features encoded by existing methods are incomprehensive since most of them only consider the local temporal information of the video. Though the global temporal context can be captured by self-attention, it is achieved at the cost of high computational overhead. Therefore, a local self-attention strategy is adopted by some studies to restrict the attention within a limited temporal window (Zhang, Wu, and Li 2022; Cheng and Bertasius 2022). While both the local and instant temporal information is aggregated in TriDet (Shi et al. 2023), the global temporal dependencies among the entire video are still neglected.

To alleviate the above limitations and improve the TAL performance, a novel method is presented in this paper.

First, we introduce a Cross-Layer Task Decoupling and Refinement (CLTDR) strategy to disentangle classification and localization tasks and enhance their consistency. Specifically, the CLTDR based decoder incorporates suitable cross layer features to handle the specific tasks in TAL. As shown in Fig. 1b, since the action classification requires rich semantic context information for accurate category inference, we combine the feature at each pyramid layer with temporally-coarse yet semantically-strong feature from higher layer for this task. Conversely, the action localization necessitates highly detailed information for boundary regression. Hence, we fuse the feature of each pyramid layer with a finer feature extracted from lower layer to address this task. During the two cross layer feature fusion, the attention mechanism is utilized to select important information from higher and lower layer features. Furthermore, we also incorporate a refinement head that leverages both higher and lower layer features to harmonize the decoupled classification and localization tasks, thereby the consistency and accuracy of their respective outputs can be enhanced.

Second, we propose a Gated Multi-Granularity (GMG) module at each feature pyramid layer to comprehensively aggregate the information from multiple temporal granularities. The major characteristic of GMG module is that it utilizes three-branch network as a substitution for self-attention in Transformer architecture. In this network, we adopt simple fully-connected operation and 1D depth-wise convolution to obtain the instant and local temporal information. Additionally, a 1D Fast Fourier Transform (FFT) branch with a learnable global filter is employed to capture global temporal dependencies across the video in frequency domain, offering a more efficient alternative to self-attention. To mitigate potential redundancy among features extracted from different temporal granularities, we employ a gated mechanism to selectively fuse features of various branches and improve the discriminative ability of the final representation.

Considering the aforementioned two major contributions, the proposed method is termed CLTDR-GMG. Numerous experiments on five datasets demonstrate that our method achieves state-of-the-art TAL performance.

Related Work

Temporal Action Localization (TAL). The existing TAL methods primarily fall into two categories: two-stage methods and one-stage methods. Two-stage TAL initially generates proposals to localize potential action instances before classifying them into different categories. Therefore, most two-stage methods prioritize proposal generation. Various techniques are employed for this purpose, including anchor window classification (Heilbron, Niebles, and Ghanem 2016), action boundaries detection (Lin et al. 2018), graph based representation learning (Bai et al. 2020), and fine-grained temporal representation design (Qing et al. 2021a). Despite their effectiveness, two-stage methods face significant challenges, including high computational complexity and the inability to be end-to-end optimized. One-stage TAL methods combine action localization and classification into a single framework by eliminating the proposal generation stage. Some previous studies (Lin et al. 2021; Yang et al. 2020; Lin, Zhao, and Shou 2017) accomplished the one-stage TAL by constructing a hierarchical network architecture with CNNs. Subsequently, anchor-free model (Long et al. 2019) has been introduced as a simplified approach for one-stage TAL. More recently, some innovative techniques, such as Transformer (Vaswani et al. 2017), SGP (Shi et al. 2023) and Mamba (Gu and Dao 2024) have also been leveraged to further improve the one-stage TAL performance. However, as mentioned in the Introduction section, there still exist certain limitations in these methods.

Task Decoupling. The conflict between classification and regression tasks, initially observed in object detection, has led to a widespread adoption of decoupled heads in object detectors. Double head RCNN (Wu et al. 2020) introduced separate heads for classification and localization within the RCNN framework. YOLOX (Ge et al. 2021) adopted decoupled heads to address the detrimental impact of coupled classification and localization tasks for YOLO series algorithms. DDOD (Chen et al. 2021) proposed an adaptive disentanglement module that employs deformable convolution to automatically obtain beneficial features for classification and regression. These studies underscore the importance of decoupling classification and localization tasks. However, the task decoupling in them occurs solely at the parameter level with the same input features, leading to an imperfect trade-off between the two tasks. To address this issue, (Zhuang et al. 2023) proposed the task-specific context decoupling (TSCODE) head, which further disentangles the classification and localization tasks at feature level.

Similar to object detection, existing one-stage TAL methods employ separate heads for action classification and localization with shared input feature. Thus, the task decoupling in them is still limited to the parameter level. Additionally, the classification and localization heads in existing methods work independently, which may lead their results to be inconsistent and inaccuracy.

Method

Problem Statement

We assume that an untrimmed video with T frames can be represented by a set of feature vectors $X = \{x_1, \cdots, x_T\}$, where x_t is the feature of frame t. Temporal action localization aims at detecting a set of action instances $Y =$

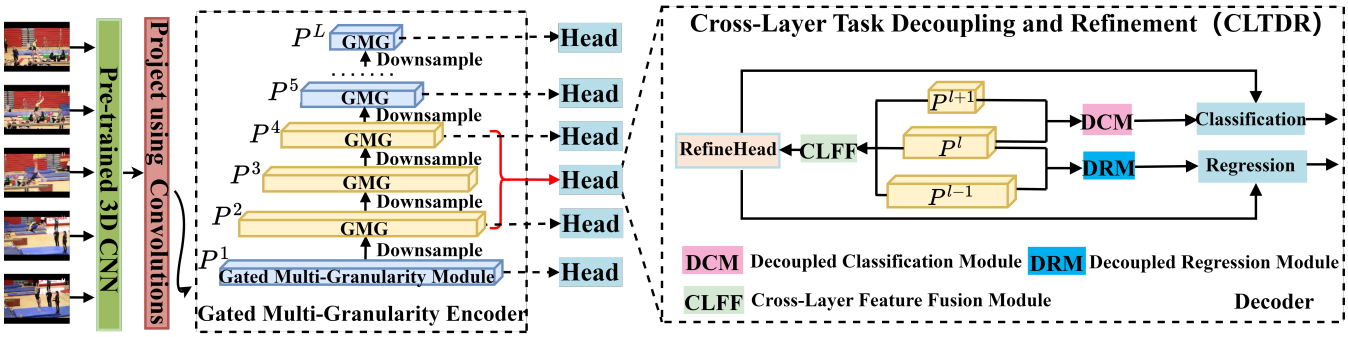

Figure 2: An illustration of our method. We build a feature pyramid with GMG module. The CLTDR decoder at the l-th pyramid layer leverages the features P^{l+1} and P^{l-1} to generate distinct representations for classification and localization tasks, followed by a refinement using RefineHead.

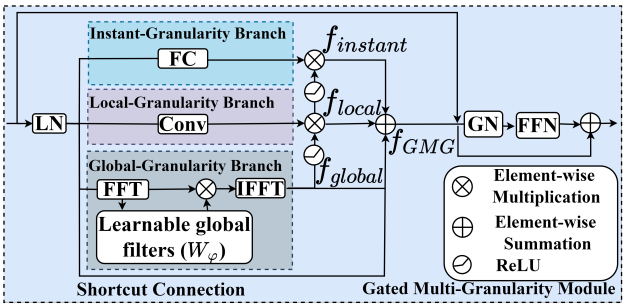

Figure 3: Illustration of GMG. Zoom in for better view.

$\{y_1, \cdots, y_N\}$ from X, where N is the number of actions. The n-th action in Y is defined as $y_n = (s_n, e_n, a_n)$, where $s_n \in [1, T]$, $e_n \in [1, T]$ and $a_n \in \{1, \ldots, C\}$ denote its start time, end time and corresponding label, respectively. C is the number of action categories.

Method Overview

As depicted in Figure 2, our CLTDR-GMG architecture comprises three key components: First, a pre-trained 3D CNN network is employed to extract spatiotemporal features from the input video. Subsequently, an encoder formed by the proposed GMG module and downsampling is utilized to construct a feature pyramid that captures actions across various temporal scales. Finally, a CLTDR module decodes the feature within each pyramid layer for action localization. A detailed explanation of our core innovations, i.e., GMG and CLTDR, will be elaborated in the following section.

Encoder with GMG

Our approach begins by encoding the video feature X into a multi-scale feature pyramid $P = \{P^1, P^2, \cdots P^L\}$. The encoder comprises two primary components: a simple 1D CNN for initially feature projection and our proposed GMG module for aggregating information across diverse temporal granularities.

To facilitate local context learning within time series data and promote stable training (Zhang, Wu, and Li 2022), we employ two consecutive 1D CNNs followed by ReLU activation to embed video feature X into a D dimensional space. The projected feature of X is denoted as $P^0 \in \mathbb{R}^{D \times T}$.

Following feature projection, a feature pyramid is constructed using the proposed Gated Multi-Granularity(GMG) module. Briefly, the initial feature P^0 is iteratively downsampled to a series of smaller temporal scales and the feature in each scale are then processed by GMG module to capture instant, local and global temporal information. As shown in Figure 3, GMG utilizes a simple fully-connected (FC) and a 1D depth-wise convolution ($Conv$) network for instant and local feature extraction, respectively. To exploit global feature dependencies, we employ Fast Fourier Transform (FFT), a powerful tool for global feature extraction, as validated by recent research (Lee-Thorp et al. 2022; Rao et al. 2023; Guibas et al. 2021). Given the input sequence x with T time steps, we first transform it to the frequency representations by FFT for global information aggregation:

$$X_f[u] = FFT(x) = \sum_{t=0}^{T-1} x_t e^{-j\frac{2\pi}{N}ut}, 0 < u < T-1 \quad (1)$$

Subsequently a global filter W_φ whose dimension is equivalent to X_f is learned by a convolution, followed by a ReLU activation and another convolution as

$$W_\varphi = Conv\left(ReLU\left(Conv\left(X_f\right)\right)\right) \quad (2)$$

where $Conv$ and $ReLU$ denote the 1D depth-wise convolution and activation function.

Finally, we get the updated feature representations in the original space by

$$f_{global} = IFFT\left(X_f \otimes W_\varphi\right) \quad (3)$$

where $\otimes$ denotes element-wise multiplication (i.e., Hadamard product) and IFFT is inverse FFT. Based on the equivalence between Hadamard product in Fourier domain and convolution in time domain (Oppenheim 1999), the feature f_{global} could be interpreted as the outputs of a

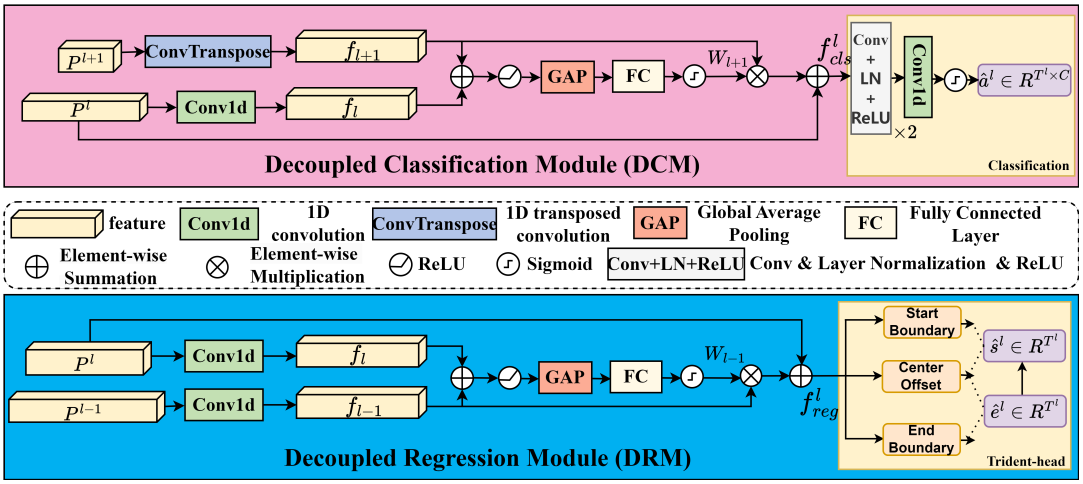

Figure 4: Illustration of Decoupled Classification Module and Decoupled Regression Module.

convolution whose kernel dimension matches the number of feature vectors in x.

Since the feature extracted from larger temporal granularity may inherently contain some information in smaller granularities, we incorporate a gated mechanism with $ReLU$ activation to selectively emphasize discriminative features and suppress redundant information within the instant and local branches. As a result, the feature of GMG can be obtained by combining the original feature x with the features of the three granularity branches as

$$f_{GMG} = x + \underbrace{ReLU(f_{local}) \otimes FC(x)}_{f_{instant}} + \underbrace{ReLU(f_{global}) \otimes Conv(x)}_{f_{local}} + f_{global} \quad (4)$$

After multi-granularity feature extraction, our GMG module incorporates a Group Normalization (GN) and a Feed Forward Network (FFN) with residual connections to further process the feature f_{GMG}. To construct the feature pyramid, 1D max-pooling is employed for downsampling between adjacent layers.

Decoder with CLTDR

The GMG based encoder generates a feature pyramid $P = \{P^1, P^2, \cdots P^L\}$. To decode this feature pyramid into the sequence label $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N\}$, we propose a Cross-Layer Task Decoupling and Refinement (CLTDR) module that disentangles features for classification and localization tasks. The CLTDR consists of three main components: a decoupled classification module, a decoupled regression module, and a Refinehead.

Decoupled Classification Module. Some studies have demonstrated the sufficiency of sparsely distributed key frames for action classification (Yan et al. 2018; Dong et al. 2022). However, existing TAL methods often neglect the redundant and uninformative feature in the classification task. Moreover, the classification accuracy of individual frame benefits from context information from its surrounding frames. For example, the athletic actions like Long Jump and High Jump have the same initial running. Thus, it is hard to identify the categories of the two actions without the information in the following frames. These observations highlight the importance of temporally coarse yet semantically robust features for classification task.

Based on above analysis, we propose a decoupled classification module (DCM) as illustrated in Figure 4. Specifically, the feature at each pyramid layer is combined with its coarse abstraction, i.e., the feature extracted from higher layer, to generate a semantically rich feature for classification. Taking feature $P^l \in R^{D \times T^l}$ at the l-th pyramid layer as an example, we first employ a 1D transposed convolution to upsample the feature P^{l+1} from the l+1-th pyramid layer by a factor of 2 and leverage a 1D convolution to refine P^l to obtain features f_{l+1} and f_l.

Since f_{l+1} may contain some irrelative information for classification, an attention weight W_{l+1} is learned from the element-wise summation of f_l and f_{l+1} as

$$W_{l+1} = Sigmoid\left(FC\left(GAP\left(ReLU\left(f_{l+1} + f_l\right)\right)\right)\right) \quad (5)$$

where $ReLU$, GAP, FC and $Sigmoid$ denote the ReLU activation function, global average pooling, fully connected layer and sigmoid function, respectively.

Then, the important and useful features from f_{l+1} are selected by W_{l+1} and combined with P^l as

$$f^l_{cls} = P_l + f_{l+1} \otimes W_{l+1} \quad (6)$$

At last, the feature f^l_{cls} is fed into a lightweight network composed of three 1D convolutions, layer normalization (LN) and ReLU activation (for the first 2 blocks). A sigmoid function is applied to yield the coarse classification result $\hat{a}^l \in R^{C \times T^l}$, where $\hat{a}^l_t \in R^C \ \left(t = 1, \ldots, T^l\right)$ represents the probabilities of the t-th feature vector belongs to the C action categories in P_l.

Decoupled Regression Module. Different from the classification task, localization requires the feature with temporal details to regress the start and end times of an action.

Nevertheless, most existing TAL methods rely solely on single scale feature from each pyramid layer for localization, which may lose some important information.

To address this issue, our Decoupled Regression Module (DRM) leverages the temporally rich feature from lower pyramid layer, which provides more detailed and boundary-specific information, to help action localization. As depicted in Figure 4, the DRM at the *l*-th pyramid layer first employs a 1D convolution to downsample the feature from the *l*-1-th layer by a factor of 2. Simultaneously, a 1D convolution is also utilized to refine the feature P_l. As a result, we can get two features f_{l-1} and f_l.

Next, similar to the DCM, the features f_{l-1} and f_l are added to learn an attention weight W_{l-1} through a series of operations as

$$W_{l-1} = Sigmoid\left(FC\left(GAP\left(ReLU\left(f_{l-1} + f_l\right)\right)\right)\right) \quad (7)$$

where the meanings of $ReLU$, GAP, FC and $Sigmoid$ are the same as those in Eq. 5.

Based on W_{l-1}, the crucial features for boundary regression in f_{l-1} are selected and combined with P^l as

$$f^l_{reg} = P_l + f_{l-1} \otimes W_{l-1} \quad (8)$$

Our DRM employs the Trident-head (Shi et al. 2023) with relative information to get the coarse boundary regression results from feature f^l_{reg}. The boundary predictions $\hat{s}^l \in R^{T^l}$ and $\hat{e}^l \in R^{T^l}$ represent the start and end times for each time step in P_l, respectively.

RefineHead. While DCM and DRM leverage cross-layer information to generate distinct features for classification and localization tasks, their independent processing may lead to inconsistent predictions. Thus, we propose to refine the outputs of DCM and DRM using the cross layer features. For refining the prediction results of the *l*-th pyramid layer, the features from the *l*+1-th and *l*-1-th layers are employed.

We first combine the features from three pyramid layers by a simple Cross-Layer Feature Fusion (CLFF) module as

$$f^l_c = Conv\left(P^{l-1}\right) + P^l + ConvTransposed\left(P^{l+1}\right) \quad (9)$$

where $ConvTransposed(\cdot)$ and $Conv(\cdot)$ denote the 1D transpose convolution and 1D depth-wise convolution to upsample and downsample P^{l+1} and P^{l-1}, respectively.

Then, the fused f^l_c, which integrates the features benefit for both classification and localization, is fed into the RefineHead to explicitly refine the coarse predictions of DCM and DRM. For classification, a vector $R^l_{co} \in R^{T^l \times C}$ is obtained from f^l_c by a simple network with three convolutional blocks to adjust the prediction of DCM. The refined result can be denoted by

$$\hat{a}^{rl} = \sqrt{\hat{a}^l \times R^l_{co}} \quad (10)$$

Meanwhile, two offsets $R^l_{so} \in R^{T^l}$ and $R^l_{eo} \in R^{T^l}$ are also learned from f^l_c by a convolutional network to calibrate the regression results of DRM. Specifically, the *t*-th element in R^l_{so} is the offset between time step t and the actual start boundary, given that time step *t* is predicted as the start of an action. Through combining the offset R^l_{so} with the prediction of $\hat{s}^l$, the refined start boundary can be obtained by

$$\hat{s}^{rl} = \hat{s}^l + R^l_{so} \quad (11)$$

Similarly, we can get the refined end boundary as

$$\hat{e}^{rl} = \hat{e}^l + R^l_{eo} \quad (12)$$

Training and Inference

In this study, the varifocal loss (Zhang et al. 2021) and IOU loss (Rezatofighi et al. 2019) are employed to supervise the refined classification and regression outputs respectively. The loss function is defined as follows:

$$\mathcal{L} = \frac{1}{N_{pos}} \sum_{l,t} 1_{\{c^l_t > 0\}} \left(\sigma_{IoU} \mathcal{L}_{VFL} + \mathcal{L}_{IoU}\right) + \frac{1}{N_{neg}} \sum_{l,t} 1_{\{c^l_t = 0\}} \mathcal{L}_{VFL} \quad (13)$$

where σ_{IoU} is the temporal IoU between the predicted action and ground truth. $\mathcal{L}_{VFL}$ is varifocal loss to handle imbalanced samples in different action categories. $\mathcal{L}_{Iou}$ denotes IoU based regression loss and is applied only when the function $l_{\{c^l_t > 0\}}$ indicates that the current time step *t* is a positive sample within an action. N_{pos} and N_{neg} are the numbers of positive and negative samples, respectively. Following other studies (Shi et al. 2023), we utilize the center sampling to identify the positive samples.

At inference time, we first feed the full video sequences into our model to obtain the outputs for each instant *t* across all pyramid layers. Then, the instants with classification scores exceeding a predetermined threshold λ are processed by Soft-NMS (Bodla et al. 2017) to yield the final output.

Experiments

We conduct experiments on five challenging datasets including THUMOS14 (Yeung et al. 2018a), MultiTHUMOS (Yeung et al. 2018b), EPIC-KITCHEN-100 (Damen et al. 2022) ActivityNet-1.3 (Heilbron et al. 2015), and HACS (Zhao et al. 2019) to validate the effectiveness of our method.

Evaluation Metric

To evaluate the performance of our CLTDR-GMG, we employ the widely adopted mAP metric across various temporal IOU (tIoU) thresholds. In line with the prevalent methodologies, the thresholds for THUMOS14, MultiTHUMOS, and EPIC-KITCHENS-100 datasets are set as [0.3:0.1:0.7], [0.1:0.1:0.9], and [0.1:0.1:0.5], respectively. For ActivityNet-1.3 and HACS, we report the results at tIoU thresholds [0.5,0.75,0.95] and the average mAP obtained from thresholds [0.5:0.05:0.95].

Implementation Details

We employ the AdamW with warm-up and a cosine annealing learning rate schedule for model optimization. The number of feature pyramid layers is set to L=6 for THUMOS14, MultiTHUMOS and EPIC-KITCHENS-100, and L=7 for ActivityNet-1.3 and HACS. Given the absence of higher and lower layers for the top and bottom layers in feature pyramid, our CLTDR module is applied exclusively to intermediate feature layers with shared parameters. We conduct our experiments using Python 3.8, PyTorch 2.0, and CUDA 11.8 on a NVIDIA RTX 4090 GPU.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
TALLFormer (Cheng and Bertasius 2022) ‡	76.0	—	63.2	—	34.5	59.2
ActionFormer (Zhang, Wu, and Li 2022)	82.1	77.8	71.0	59.4	43.9	66.8
TemporalMaxer (Tang, Kim, and Sohn 2023)	82.8	78.9	71.8	60.5	44.7	67.7
TFFormer (Yang, Wei, and Zheng 2024)	82.1	78.9	72.0	60.8	44.9	67.8
TransGMC (Yang et al. 2024)	82.3	78.8	71.4	60.0	45.1	67.5
TriDet (Shi et al. 2023)	83.6	80.1	72.9	62.4	47.4	69.3
CLTDR-GMG	**84.1**	**80.3**	**73.6**	**62.4**	**48.2**	**69.9**
ActionFormer (Zhang, Wu, and Li 2022) †	84.0	79.6	73.0	63.5	47.7	69.6
TriDet (Shi et al. 2023) †	84.8	80.0	73.3	63.8	48.8	70.1
CLTDR-GMG †	**85.7**	**81.3**	**75.5**	**65.3**	**51.0**	**71.8**
ActionFormer (Zhang, Wu, and Li 2022) §	82.3	81.9	75.1	65.8	50.3	71.9
TemporalMaxer (Tang, Kim, and Sohn 2023) § ∓	87.4	83.1	76.6	65.7	49.6	72.5
ActionMamba (Chen et al. 2024) §	86.9	83.1	76.9	65.9	50.8	72.7
TriDet (Shi et al. 2023) § ∓	86.9	83.4	76.6	66.3	51.7	73.0
CLTDR-GMG §	**87.0**	**84.0**	**78.4**	**67.9**	**54.0**	**74.3**

Table 1: Performance comparison on THUMOS14 dataset. ∗: TSN features. ‡:Swin Transformer features. †: VideoMAEv2 features. §: InterVideo2-6B features. Others: I3D features. ∓:indicates our implementation.

Method	0.2	0.5	0.7	Avg.
PointTAD (Tan et al. 2022)	39.7	24.9	12.0	23.5
ActionFormer (Zhang, Wu, and Li 2022)	46.4	32.4	15.0	28.6
TemporalMaxer (Tang, Kim, and Sohn 2023)	47.5	33.4	17.4	29.9
TriDet (Shi et al. 2023)	49.1	34.3	17.8	30.7
CLTDR-GMG	**56.7**	**42.2**	**24.1**	**37.1**
TriDet (Shi et al. 2023)∗	55.7	41.0	23.5	36.2
CLTDR-GMG∗	**62.5**	**49.1**	**30.2**	**42.7**
TriDet (Shi et al. 2023)†	57.7	42.7	24.3	37.5
CLTDR-GMG†	**64.9**	**51.4**	**32.9**	**44.9**

Table 2: Performance comparison on MultiTHUMOS dataset. ∗: I3D (RGB+Flow) features. †: VideoMAEv2 features. Others: I3D (only RGB) features.

Results and Analysis

THUMOS14. Following most approaches (Zhang, Wu, and Li 2022; Shi et al. 2023), we leverage the pre-trained two-stream I3D (Carreira and Zisserman 2017) on Kinetics (Kay et al. 2017) to extract features from the THUMOS14 dataset. To ensure a fair comparison with recent state-of-the-art methods, we also employ VideoMAEv2 (Wang et al. 2023) and InterVideo2 (Wang et al. 2024), which are extensively pre-trained large models, for feature extraction. The performance of various TAL methods is shown in Table 1. It is evident that recent methods based on innovative techniques (such as Actionformer, TriDet and ActionMamba) outperform other counterparts. Moreover, the performance of our proposed CLTDR-GMG is superior to all existing methods. Specifically, our method achieves average mAPs of 69.9%, 71.8%, and 74.3% with I3D, VideoMAE V2, and InterVideo2 features, which are at least 0.6% , 1.7%, and 1.3% higher than those of previous SOTA approaches.

MultiTHUMOS. Similar to the experiment on THUMOS14, we employ I3D and VideoMAEv2 to extract video features on this dataset. The results in Table 2 clearly show that our CLTDR-GMG significantly outperforms other methods. Additionally, we find that the more sophisticated

Task	Method	0.1	0.2	0.3	0.4	0.5	Avg.
V.	ActionFormer (Zhang, Wu, and Li 2022)	26.6	25.4	24.2	22.3	19.1	23.5
	TemporalMaxer (Tang, Kim, and Sohn 2023)	27.8	26.6	25.3	23.1	19.9	24.5
	TriDet (Shi et al. 2023)	28.6	27.4	26.1	24.2	20.8	25.4
	TFFormer (Yang, Wei, and Zheng 2024)	28.8	27.7	26.1	24.7	20.5	25.6
	TransGMC (Yang et al. 2024)	27.8	26.7	25.5	23.6	20.7	24.9
	CLTDR-GMG	**29.5**	**28.6**	**27.0**	**24.3**	**20.7**	**26.0**
N.	ActionFormer (Zhang, Wu, and Li 2022)	25.2	24.1	22.7	20.5	17.0	21.9
	TemporalMaxer (Tang, Kim, and Sohn 2023)	26.3	25.2	23.5	21.3	17.6	22.8
	TriDet (Shi et al. 2023)	27.4	26.3	24.6	22.2	18.3	23.8
	TFFormer (Yang, Wei, and Zheng 2024)	27.2	25.9	24.2	21.7	17.9	23.4
	TransGMC (Yang et al. 2024)	26.4	25.2	23.4	21.4	18.1	22.9
	CLTDR-GMG	**28.2**	**26.9**	**25.2**	**22.7**	**19.4**	**24.5**

Table 3: Performance comparison on EPIC-KITCHENS-100 dataset. *V.* and *N.* denote the verb and noun sub-tasks.

Method	tIoU			Avg.
	0.5	0.75	0.95	0.5:0.05:0.95
ReAct (Shi et al. 2022)∗	49.6	33.0	8.6	32.6
TadTR (Liu et al. 2022)∗	51.3	35.0	9.5	34.6
TadTR (Liu et al. 2022)†	53.6	37.5	10.5	36.8
TALLFormer (Cheng and Bertasius 2022)‡	54.1	36.2	7.9	35.6
TFFormer (Yang, Wei, and Zheng 2024)	54.4	36.7	7.5	35.8
ActionFormer (Zhang, Wu, and Li 2022)†	54.7	37.8	8.4	36.6
TransGMC (Yang et al. 2024)†	54.8	37.6	8.5	36.7
TriDet (Shi et al. 2023)†	54.7	38.0	8.4	36.8
CLTDR-GMG†	**55.0**	**38.0**	**8.6**	**37.1**

Table 4: Performance comparison on ActivityNet1.3 dataset. ∗: TSN features. ‡: Swin Transformer features. †: R(2+1)D features. Others: I3D features.

VideoMAEv2 features can boost the TAL performance of our CLTDR-GMG and some other approaches, which is consistent with the results in Table 1. For instance, the VideoMAEv2 features lead to a 2.2% improvement in average mAP for our method.

EPIC-KITCHENS-100. Table 3 shows the performance obtained by different TAL methods using the features extracted by pre-trained SlowFast network (Feichtenhofer et al. 2019). It can be seen that our CLTDR-GMG achieves average mAPs of 26.0% and 24.5% for the verb and noun subtasks, which are both superior to previous methods on these challenging datasets.

ActivityNet-1.3. For the experiment on ActivityNet-1.3, we use TSP R(2+1)D (Alwassel, Giancola, and Ghanem 2021) as the pre-trained model to extract video features, which is consistent with the methodology used in several recent studies (Shi et al. 2023). As shown in Table 4, our method outperforms other methods.

HACS. On HACS dataset, we use the I3D (Carreira and Zisserman 2017) features from RGB stream and the SlowFast (Feichtenhofer et al. 2019) features from TCANet (Qing et al. 2021b) in our experiments. As shown in Table 5, our method achieves average mAPs of 37.2% with I3D features and 39.2% with SlowFast features, which outperforms the most recent TriDet model.

Method	0.5	0.75	0.95	Avg.
TadTR (Liu et al. 2022)	47.1	32.1	10.9	32.1
TALLFormer (Cheng and Bertasius 2022)‡	55.0	36.1	11.8	36.5
TCANet (Qing et al. 2021b)†	54.1	37.2	11.3	36.8
TriDet (Shi et al. 2023)	54.5	36.8	11.5	36.8
CLTDR-GMG	**55.2**	**37.3**	**11.8**	**37.2**
TriDet (Shi et al. 2023)†	56.7	39.3	11.7	38.6
CLTDR-GMG†	**57.6**	**39.9**	**12.0**	**39.3**

Table 5: Performance comparison on HACS dataset. ‡: Swin Transformer features. †: SlowFast features. Others: I3D features.

Instant	✓	✗	✗	✓	✓	✗	✓
Local	✗	✓	✗	✓	✗	✓	✓
Global	✗	✗	✓	✗	✓	✓	✓
Avg.	73.4	73.4	73.5	73.6	73.8	73.9	74.3

Table 6: The analysis of different temporal granularities.

Ablation Study

We perform various ablation studies on THUMOS14 dataset using InterVideo2-6B features to assess the effectiveness of each component in our CLTDR-GMG.

Ablation on GMG. First, we conduct an ablation experiment to demonstrate the effectiveness of each temporal granularity branch in our GMG module. Table 6 shows that our method does not achieve satisfactory performance when using information from instant, local or global granularity individually. Furthermore, combining global granularity information with either instant or local granularity information yields better results than combining the information of instant and local granularities. This substantiates that global granularity information is crucial for the TAL task. Finally, our method achieves optimal performance when integrating information from all three temporal granularities.

We then explore the impact of gated mechanism in GMG module. As shown in Table 7, our method achieves a performance improvement of 0.4% when the gated mechanism is used. This confirms the effectiveness of gated mechanism for adaptive feature selection and redundancy elimination.

Finally, we replace the FFT-based global feature extractor in our GMG with vanilla self-attention used by Transformer. Table 8 reveals that while vanilla self-attention achieves similar performance as our method, it incurs a significantly higher parameter count, approximately 2.7 times that of our FFT-based approach.

Ablation on CLTDR. To investigate the impact of feature pyramid layers on classification and regression decoupling, we conducted ablation studies as shown in Table 9. Our results indicate that fusing features from higher pyramid layer enhances classification performance due to the richer semantic information in them. For regression tasks, incorporating features from lower layer contributes to more performance gains, as they provide more detailed information. We further find that incorporating both higher and lower pyramid layers leads to a slight performance degradation, which suggests that excessive information may be redundant or even detrimental for task decoupling.

Then, the effectiveness of RefineHead in CLTDR is studied. The results in Table 10 indicates that the performance of our method can be improved by refining the predictions of classification and localization tasks.

At last, the attention based cross layer information fusion strategy in DCM and DRM is compared with addition and concatenation. From Table 11, it is clear that the attention mechanism achieves better performance than other two information fusion manners.

Additional experimental results and analyses of our method are provided in the supplementary material.[1]

Method	0.3	0.5	0.7	Avg.
without Gate	87.0	77.7	53.4	73.9
with Gate	87.3	77.9	53.9	74.3

Table 7: The impact of gated mechanism in GMG.

Global-Level	0.3	0.5	0.7	Avg.	# Params
FFT	87.3	77.9	53.9	74.3	9.5M
Self-Attention	87.5	78.0	53.5	74.2	26.1M

Table 8: The comparison of global feature extractor in GMG.

classification				regression			
P_{l+1}	P_l	P_{l-1}	mAP	P_{l+1}	P_l	P_{l-1}	mAP
	✓		73.3		✓		73.4
✓	✓		74.3		✓	✓	74.3
	✓	✓	73.8	✓	✓		73.6
✓	✓	✓	74.0	✓	✓	✓	73.2

Table 9: Ablation of pyramid layers in CLTDR.

Method	0.3	0.5	0.7	Avg.
without refinement	87.1	77.3	53.1	73.5
only refine classification	87.4	77.8	53.9	73.9
only refine regression	87.6	77.3	52.4	73.8
refine classification and regression	87.3	77.9	53.9	74.3

Table 10: The effectiveness of refinement in CLTDR.

Method	0.3	0.5	0.7	Avg.
Concatenation	86.7	76.7	53.3	73.4
Addition	86.8	77.3	52.8	73.5
Attention	87.3	77.9	53.9	74.3

Table 11: The comparison of fusion strategies in CLTDR.

Conclusion

In this paper, we proposed a GMG based encoder and a CLTDR based decoder for TAL problem. Experiments conducted on five datasets demonstrate that our method outperforms the previous approaches and achieves state-of-the-art performance. Moreover, several ablation studies are also carried out to validate the components in our method.

[1] http://arxiv.org/abs/2412.09202

Acknowledgments

This work is supported by the NSFC under Grant 62272096 and the Jilin Provincial Science and Technology Department under Grant 20230201079GX.

References

Alwassel, H.; Giancola, S.; and Ghanem, B. 2021. TSP: Temporally-Sensitive Pretraining of Video Encoders for Localization Tasks. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, 3166–3176. IEEE.

Bai, Y.; Wang, Y.; Tong, Y.; Yang, Y.; Liu, Q.; and Liu, J. 2020. Boundary Content Graph Neural Network for Temporal Action Proposal Generation. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, 121–137. Springer.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS - Improving Object Detection with One Line of Code. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 5562–5570. IEEE Computer Society.

Buch, S.; Escorcia, V.; Ghanem, B.; Fei-Fei, L.; and Niebles, J. C. 2017. End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4724–4733. IEEE Computer Society.

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding. arXiv:2403.09626.

Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.; and Wu, F. 2021. Disentangle Your Dense Object Detector. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 4939–4948. ACM.

Cheng, F.; and Bertasius, G. 2022. TallFormer: Temporal Action Localization with a Long-Memory Transformer. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIV*, volume 13694 of *Lecture Notes in Computer Science*, 503–521. Springer.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1): 33–55.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2022. Identifying the key frames: An attention-aware sampling method for action recognition. *Pattern Recognition*, 130: 108797.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6201–6210. IEEE.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. In *International Conference on Learning Representations*.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 961–970. IEEE Computer Society.

Heilbron, F. C.; Niebles, J. C.; and Ghanem, B. 2016. Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 1914–1923. IEEE Computer Society.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontañón, S. 2022. FNet: Mixing Tokens with Fourier Transforms. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 4296–4313. Association for Computational Linguistics.

Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. Learning Salient Boundary Feature for Anchor-free Temporal Action Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 3320–3329. Computer Vision Foundation / IEEE.

Lin, T.; Zhao, X.; and Shou, Z. 2017. Single Shot Temporal Action Detection. In Liu, Q.; Lienhart, R.; Wang, H.; Chen, S. K.; Boll, S.; Chen, Y. P.; Friedland, G.; Li, J.; and Yan, S., eds., *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 988–996. ACM.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, 3–21. Springer.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-End Temporal Action Detection With Transformer. *IEEE Trans. Image Process.*, 31: 5427–5441.

Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; and Mei, T. 2019. Gaussian Temporal Awareness Networks for Action Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 344–353. Computer Vision Foundation / IEEE.

Nag, S.; Zhu, X.; Song, Y.; and Xiang, T. 2022. Semi-supervised Temporal Action Detection with Proposal-Free Masking. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, volume 13663 of *Lecture Notes in Computer Science*, 663–680. Springer.

Oppenheim, A. V. 1999. *Discrete-time signal processing*. Pearson Education India.

Qing, Z.; Su, H.; Gan, W.; Wang, D.; Wu, W.; Wang, X.; Qiao, Y.; Yan, J.; Gao, C.; and Sang, N. 2021a. Temporal Context Aggregation Network for Temporal Action Proposal Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 485–494. Computer Vision Foundation / IEEE.

Qing, Z.; Su, H.; Gan, W.; Wang, D.; Wu, W.; Wang, X.; Qiao, Y.; Yan, J.; Gao, C.; and Sang, N. 2021b. Temporal Context Aggregation Network for Temporal Action Proposal Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 485–494. Computer Vision Foundation / IEEE.

Rao, Y.; Zhao, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. GFNet: Global Filter Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9): 10960–10973.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 658–666. Computer Vision Foundation / IEEE.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Lit, J.; and Tao, D. 2023. TriDet: Temporal Action Detection with Relative Boundary Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 18857–18866. IEEE.

Shi, D.; Zhong, Y.; Cao, Q.; Zhang, J.; Ma, L.; Li, J.; and Tao, D. 2022. ReAct: Temporal Action Detection with Relational Queries. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, volume 13670 of *Lecture Notes in Computer Science*, 105–121. Springer.

Shou, Z.; Wang, D.; and Chang, S. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 1049–1058. IEEE Computer Society.

Tan, J.; Zhao, X.; Shi, X.; Kang, B.; and Wang, L. 2022. PointTAD: Multi-Label Temporal Action Detection with Learnable Query Points. In *NeurIPS*.

Tang, T. N.; Kim, K.; and Sohn, K. 2023. TemporalMaxer: Maximize Temporal Context with only Max Pooling for Temporal Action Localization. arXiv:2303.09055.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 14549–14560. IEEE.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Wang, C.; Chen, G.; Pei, B.; Yan, Z.; Zheng, R.; Xu, J.; Wang, Z.; Shi, Y.; Jiang, T.; Li, S.; Zhang, H.; Huang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2024. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. arXiv:2403.15377.

Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; and Fu, Y. 2020. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10186–10195.

Xu, M.; Zhao, C.; Rojas, D. S.; Thabet, A. K.; and Ghanem, B. 2020. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10153–10162. Computer Vision Foundation / IEEE.

Yan, X.; Gilani, S. Z.; Qin, H.; Feng, M.; Zhang, L.; and Mian, A. 2018. Deep Keyframe Detection in Human Action Videos. arXiv:1804.10021.

Yang, J.; Wei, P.; Ren, Z.; and Zheng, N. 2024. Gated Multi-Scale Transformer for Temporal Action Localization. *IEEE Transactions on Multimedia*, 26: 5705–5717.

Yang, J.; Wei, P.; and Zheng, N. 2024. Cross Time-Frequency Transformer for Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 4625–4638.

Yang, L.; Peng, H.; Zhang, D.; Fu, J.; and Han, J. 2020. Revisiting Anchor Mechanisms for Temporal Action Localization. *IEEE Trans. Image Process.*, 29: 8535–8548.

Yeung, S.; Russakovsky, O.; Jin, N.; Andriluka, M.; Mori, G.; and Fei-Fei, L. 2018a. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *Int. J. Comput. Vis.*, 126(2-4): 375–389.

Yeung, S.; Russakovsky, O.; Jin, N.; Andriluka, M.; Mori, G.; and Fei-Fei, L. 2018b. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *Int. J. Comput. Vis.*, 126(2-4): 375–389.

Zhang, C.; Wu, J.; and Li, Y. 2022. ActionFormer: Localizing Moments of Actions with Transformers. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, 492–510. Springer.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8514–8523.

Zhao, C.; Thabet, A. K.; and Ghanem, B. 2021. Video Self-Stitching Graph Network for Temporal Action Localization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 13638–13647. IEEE.

Zhao, H.; Torralba, A.; Torresani, L.; and Yan, Z. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 8667–8677. IEEE.

Zhuang, J.; Qin, Z.; Yu, H.; and Chen, X. 2023. Task-Specific Context Decoupling for Object Detection. arXiv:2303.01047.