Weakly-Supervised Temporal Action Localization with Multi-Modal Plateau Transformers

Xin Hu[1]*, Kai Li[2], Deep Patel[2], Erik Kruus[2], Martin Renqiang Min[2], Zhengming Ding[1]
[1]Tulane University, [2]NEC Labs America
{xhu13, zding1}@tulane.edu {kaili, dpatel, kruus, renqiang}@nec-labs.com

Abstract

Weakly-Supervised Temporal Action Localization (WS-TAL) aims to jointly localize and classify action segments in untrimmed videos with only video-level annotations. To leverage video-level annotations, most existing methods adopt the multiple instance learning paradigm where frame-/snippet-level action predictions are first produced and then aggregated to form a video-level prediction. Although there are trials to improve snippet-level predictions by modeling temporal relationships, we argue that those implementations have not sufficiently exploited such information. In this paper, we propose Multi-Modal Plateau Transformers (M^2PT) for WS-TAL by simultaneously exploiting temporal relationships among snippets, complementary information across data modalities, and temporal coherence among consecutive snippets. Specifically, M^2PT explores a dual-Transformer architecture for RGB and optical flow modalities, which models intra-modality temporal relationship with a self-attention mechanism and inter-modality temporal relationship with a cross-attention mechanism. To capture the temporal coherence that consecutive snippets are supposed to be assigned with the same action, M^2PT deploys a Plateau model to refine the temporal localization of action segments. Experimental results on popular benchmarks demonstrate that our proposed M^2PT achieves state-of-the-art performance.

1. Introduction

Temporal Action Localization (TAL) aims to identify temporal timestamps of action instances and classify their action categories. Most existing works address this problem by training models with fully-annotated data that include both timestamps (start and end positions) and class label of each action instance [4, 20, 42]. TAL is analogous to object detection, but instead localizes and classifies instances in the temporal dimension. Similar to object detection, TAL methods can be categorized as two-stage methods that first generate proposals and then refine them [1, 10], and one-stage methods that directly generate detection results [6, 40].

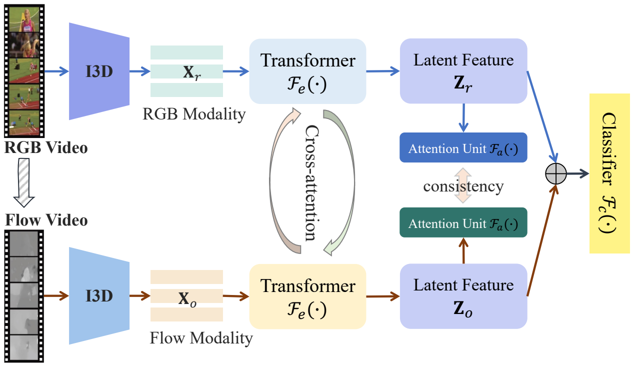

RGB Video
Flow Video
I3D
$\mathbf{X}_r$
RGB Modality
I3D
$\mathbf{X}_o$
Flow Modality
Transformer $\mathcal{F}_e(\cdot)$
Cross-attention
Transformer $\mathcal{F}_e(\cdot)$
Latent Feature $\mathbf{Z}_r$
Attention Unit $\mathcal{F}_a(\cdot)$
consistency
Attention Unit $\mathcal{F}_a(\cdot)$
Latent Feature $\mathbf{Z}_o$
Classifier $\mathcal{F}_c(\cdot)$

Figure 1. Overview of the proposed network, where RGB and optical flow from an untrimmed video are fed into the pre-trained I3D network to extract features $\mathbf{X}_{r/o}$. Then cross-attention Transformer is to generate the refined latent features $\mathbf{Z}_{r/o}$ with complementary information from another modality. Attention units generate attention weights for each branch to filter out background snippets and concatenated features for classification.

While achieving impressive results, TAL methods require fully-annotated data to train models. Collecting fully-annotated training data is expensive and time-consuming, especially for videos that often contain lots of frames. To alleviate this issue, Weakly-Supervised Temporal Action Localization (WS-TAL) is explored to learn from training data with only video-level labels, i.e., action categories within a video; the timestamps of action instances are not provided [5, 8, 9, 26, 37]. Due to the absence of temporal annotations, existing WS-TAL methods usually adopt a Multiple Instance Learning (MIL) paradigm that each video is viewed as a labeled bag consisting of unlabeled snippets (instances) [8, 9, 17]. While it is infeasible to directly learn from predictions of individual snippet, a video-level prediction can be obtained by aggregating the individual predictions.

However, the naive MIL-based solution typically

*This work was done when Xin Hu was an intern at NEC Labs America achieves promising action classification but unsatisfactory localization, because of insufficient action segment boundary supervision. Various efforts have been made to improve this. To name a few, Hong et. al. exploited cross-modal features to reduce task-irrelevant information redundancy from direct feature fusion [11]. He et. al. designed intra- and inter-segment attention modules to explore temporal similarity within and across action segments [9]. With respect to feature learning, contrastive learning was also deployed to refine intermediate features [39]. Along this, Yang et. al. proposed to generate reliable pseudo labels from noisy predictions under an uncertainty-aware mechanism and train the model in a fully-supervised way under such pseudo labels [37]. Despite substantial improvements achieved by existing works, the temporal structures of action videos have not been sufficiently utilized.

In this paper, we propose Multi-Modal Plateau Transformers (M^{2}PT) to solve WS-TAL by extensively modeling various temporal structural cues in action videos (shown in Figure 1), namely, temporal relationships among snippets, complementary information across data modalities, and temporal coherence among consecutive snippets. It should be noted that while these temporal structural cues have been modeled individually by the existing methods, M^{2}PT is the very pioneering work to exploit all these cues simultaneously, within a modern Transformer-based framework. Specifically, M^{2}PT takes the RGB and optical flow features extracted from videos as two modalities, and adopts a dual-Transformer structure for the two modalities, respectively. The temporal relationship among continuous snippets is modeled by the self-attention mechanism within each individual Transformer, meanwhile the complementary information between different modalities is modeled by a cross-modality cross-attention mechanism. To encourage the temporal coherence among consecutive snippets, which implies that consecutive snippets are supposed to be assigned with the same label, we explore a Plateau model [22] to refine the temporal localization of action segments and enhance the model with the refined results as the pseudo labels. In fact, to our best knowledge, this is the first dual Transformer-based model for the WS-TAL task. To sum up, our contributions are highlighted as:

- First, we design a multi-modal Transformer model which takes RGB and optical flow as modalities, and models intra-modality temporal relationship with a self-attention mechanism and inter-modality temporal relationship with a cross-attention mechanism.
- Second, we propose to explore Plateau model into weakly-supervised temporal action localization to improve the quality of temporal localized action segments.
- Finally, our proposed M^{2}PT achieves state-of-the-art performance on two popular action benchmarks, i.e., THUMOS14 and ActivityNet1.2 datasets.

2. Related Work

Fully-Supervised Temporal Action Localization (FS-TAL). Different from action recognition, FS-TAL is more difficult and usually processes longer untrimmed videos probably containing multiple action instances. In the fully-supervised setting, models can be trained with data annotated with timestamps and class labels for all instances. In essence, FS-TAL can be analogous to object detection, but rather aims to "detect" instances in the temporal dimension. Similarly, FS-TAL can also be categorized as two-stage methods that first generate action proposals and then refine the boundaries [1, 4, 10, 28, 34, 42], and one-stage methods that directly predict action probability on snippets in videos and use a bottom-up mechanism to group action snippets to action segments [6, 18, 40].

Weakly-Supervised Temporal Action Localization (WS-TAL). WS-TAL only requires video-level labels for training, which has attracted increasing attention from researchers. UntrimmedNet [32] firstly proposes multiple instance learning loss to tackle the classification problem of untrimmed videos. STPN [24] adds a sparsity loss to UntrimmedNet to efficiently separate background snippets and proposes Temporal Class Activation Map (TCAM) to generate action proposals. CO2 [11] proposes a cross-modal network in WS-TAL task which constructs the relation between RGB feature and Optical Flow feature, and filters out task-irrelevant information redundancy. To alleviate the lack of enough temporal annotations, some works utilize pseudo labels to lead full supervision, which causes many false positive action proposals because pseudo labels are not reliable. UGCT [37] introduced an uncertainty loss on pseudo labels, which filters out highly uncertain pseudo labels and decreases false positive samples. RSKP proposed a representative snippet summarization and propagation method which produced pseudo labels only from representative snippets [12]. ASM-Loc used an additional uncertainty prediction module to explicitly output an uncertainty score for weighting each snippet [11].

Transformer Models. Transformer-based architecture [31] has shown excellent performance in modeling long sequence data, especially in Natural Language Processing (NLP). Vision Transformer (ViT) [7] introduces Transformer to computer vision, shifting the backbone from CNN to Transformer. DeiT [30] introduces several training strategies to fit ViT to the smaller ImageNet-1K dataset. CDTrans [35] proposes the cross-attention mechanism of Transformer in domain adaption and achieved state-of-the-art performance.

Differently, our proposed Multi-Modal Plateau Transformers (M^{2}PT) aims to address WS-TAL within a Transformer-based framework, which simultaneously models temporal relationships among snippets, complementary information across data modalities, and temporal coherence among consecutive snippets.

3. The Proposed Method

3.1. Preliminary and Motivation

Given an untrimmed video set $\{\mathcal{V}_i\}_{i=1}^N$, where N is the number of videos and i is the index of the sequence sample. The video-level label is associated as $\{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i \in \mathbb{R}^C$, C is the total number of action categories and $\mathbf{y}_{i,c}$ indicates the existence of action c in $\mathcal{V}_i$. Note that $\mathbf{y}_i$ is a multi-label vector when there are more than one action in the input video and will be normalized with ℓ_1-norm.

Following recent methods [5, 9, 11, 13, 24], each video $\mathcal{V}_i$ first is divided into a sequence of non-overlapping snippets $\{\mathcal{X}_1, \cdots, \mathcal{X}_T\}$, where T is the number of snippets. These snippets are input into a pre-trained I3D network [3] to generate feature representation $\mathbf{X}_r \in \mathbb{R}^{T\times D}$ and $\mathbf{X}_o \in \mathbb{R}^{T\times D}$ for each video in RGB and optical flow modality, respectively. Note that r and o represent RGB modality and optical flow modality, and D denotes the dimension of features.

Existing WS-TAL methods have exploited temporal relationships among snippets [9, 24] and complementary information across data modalities [5, 11] individually, we model them simultaneously with a Transformer model with self-attention and cross-modality cross-attention mechanism. Specially, cross-modal methods are to alleviate the redundant information problem which is first proposed by [11], and [5] expands it with evidential optimization. However, existing cross-modal methods ignore snippet-wise relations. This motivates us to build a Transformer structure for each modality branch and achieve cross-modal by cross-attention mechanism. Plateau models have been exploited in [22, 27], but both works require providing an annotated seed frame for each action instance, which however is not feasible for our weakly-supervised setting where only video-level labels are provided. We instead employ the Plateau model to refine detected action instances for getting more precise pseudo labels.

Thus, we propose a novel Multi-Modal Plateau Transformers (**M^{2}PT**) network shown in Figure 1. The goal is to generate a set of action proposals with each as (t_s, t_e, c), where t_s and t_e are the starting and ending snippet timestamp for action proposal, while c indicates action category.

3.2. Base Model

First of all, we introduce the base model architecture to illustrate the framework overview for WS-TAL. Following existing dual-branch architecture [11, 37], we aim to train a feature embedding architecture $\mathcal{F}_e(\cdot)$ to learn more effective latent features $\mathbf{Z}_r \in \mathbb{R}^{T\times D}$ and $\mathbf{Z}_o \in \mathbb{R}^{T\times D}$ for two modalities, which are further concatenated as $\mathbf{Z}_m = [\mathbf{Z}_r, \mathbf{Z}_o] \in \mathbb{R}^{T\times 2D}$ and fed into the video-level action classifier $\mathcal{F}_c(\cdot)$ to obtain the Temporal Class Activation Map (TCAM) output as:

$$\mathbf{O}_{\text{cam}} = \mathcal{F}_{\text{c}}(\mathbf{Z}_m),$$

where $\mathbf{O}_{\text{cam}} \in \mathbb{R}^{T\times C+1}$ contains $C+1$ dimensions, since we follow existing works [5, 9, 17] and set the last dimension as background.

Later on, the latent features $\mathbf{Z}_r$ and $\mathbf{Z}_o$ are input into two attention units $\mathcal{F}_a(\cdot)$ to generate attention weights $\mathbf{a}_r \in \mathbb{R}^T$ and $\mathbf{a}_o \in \mathbb{R}^T$, respectively. To suppress the background parts in $\mathbf{O}_{\text{cam}}$, we integrate the attention weights $\mathbf{a}_m = \frac{1}{2}(\mathbf{a}_r + \mathbf{a}_o)$ and obtain the suppressed TCAM output as:

$$\hat{\mathbf{O}}_{\text{cam}} = \mathbf{a}_m \otimes \mathbf{O}_{\text{cam}},$$

where $\otimes$ denotes element-wise multiplication along the temporal dimension. Following most works [11, 12, 15, 24], the Multi-instance Learning (MIL) loss $\mathcal{L}_{\text{mil}}$ is the fundamental loss function for WS-TAL, which can be derived for video-level classification as follows:

$$\mathcal{L}_{\text{mil}} = \mathcal{L}_{\text{ce}}(\mathbf{y}, \mathbf{p}_{\text{cam}}) + \mathcal{L}_{\text{ce}}(\mathbf{y}, \hat{\mathbf{p}}_{\text{cam}}), \tag{1}$$

where $\mathbf{p}_{\text{cam}}/\hat{\mathbf{p}}_{\text{cam}}$ are the video-level prediction scores with the temporal top-k pooling on $\mathbf{O}_{\text{cam}}$ and $\hat{\mathbf{O}}_{\text{cam}}$. $\mathcal{L}_{\text{ce}}(\cdot,\cdot)$ is defined as the cross-entropy loss function over the video-level ground truth $\mathbf{y}$ and predicted one.

MIL-based methods [24, 32] perform poorly due to weak and implicit supervision on the temporal boundary which can be attributed to the lack of sufficient temporal annotations. To further improve the localization of action segments, we leverage the pseudo-label module and introduce a pseudo-label loss $\mathcal{L}_{\text{pseudo}}$ with uncertainty estimation [9, 37] to explicitly supervise TCAM output $\mathbf{O}_{\text{cam}}$ as:

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{e}^{-u_t}\mathcal{L}_{\text{ce}}(\hat{\mathbf{p}}_t, \mathbf{o}_t) + \tau u_t, \tag{2}$$

where u_t denotes the uncertainty value for each snippet from one convolution layer of $\mathbf{Z}_m$, τ is hyper-parameter, $\mathbf{o}_t \in \mathbb{R}^{C+1}$ is the snippet from $\mathbf{O}_{\text{cam}}$, and $\hat{\mathbf{p}}_t \in \mathbb{R}^{C+1}$ represents snippet-level pseudo labels. Note that we refer to [9, 37]'s method to generate pseudo labels, which is common practice for WS-TAL. Motivated by [11], attention weights are essential to be constrained. So we introduce the mutual learning loss $\mathcal{L}_{\text{ml}}$ to guarantee the consistency between $\mathbf{a}_r$ and $\mathbf{a}_o$:

$$\mathcal{L}_{\text{ml}} = \frac{1}{2}\Big(\mathcal{L}_{\text{mse}}\big(\mathbf{a}_r, \phi(\mathbf{a}_o)\big) + \mathcal{L}_{\text{mse}}\big(\phi(\mathbf{a}_r), \mathbf{a}_o\big)\Big), \tag{3}$$

where $\phi(\cdot)$ represents a function that truncates the gradient of input and $\mathcal{L}_{\text{mse}}(\cdot,\cdot)$ denotes the mean square loss function. Moreover, it is also essential to make attention weights more sparse by using ℓ_1 normalization term as $\mathcal{L}_{\text{norm}}$:

$$\mathcal{L}_{\text{norm}} = \frac{1}{3}(||\mathbf{a}_r||_{\ell_1} + ||\mathbf{a}_o||_{\ell_1} + ||\mathbf{a}_m||_{\ell_1}), \tag{4}$$

where $||\cdot||_{\ell_1}$ is the ℓ_1-norm.

Considering the last vector of $\mathbf{O}_{\text{cam}}$ is the probability distribution of the background class (defined as $\mathbf{p}_b$), it should be opposite to the distribution of attention weight:

$$\mathcal{L}_{\text{oppo}} = \frac{1}{3}\Big(|\mathbf{a}_r + \mathbf{p}_b - 1| + |\mathbf{a}_o + \mathbf{p}_b - 1| + |\mathbf{a}_m + \mathbf{p}_b - 1|\Big), \tag{5}$$

where $|\cdot|$ is the absolute value function.

To sum up, we can obtain the final objective function for our base model as:

$$\mathcal{L} = \mathcal{L}_{\text{mil}} + \mathcal{L}_{\text{ml}} + \lambda_0 \mathcal{L}_{\text{pseudo}} + \lambda_1(\mathcal{L}_{\text{norm}} + \mathcal{L}_{\text{oppo}}), \tag{6}$$

where λ_0 and λ_1 are hyper-parameters for the pseudo label and regularization terms.

3.3. Multi-Modal Plateau Transformers

Unfortunately, pseudo labels and attention weights generated by WS-TAL methods are generally unreliable and contain amounts of noisy patterns. To mitigate this bottleneck, we propose a multi-modal Transformer fusion mechanism to effectively capture more complementary information across two modalities to enhance the feature generalization ability. To further improve the temporal localization segments, we explore the plateau function [22] to refine the temporal attention weight to more continuous patterns along the temporal direction.

3.3.1 Multi-Modal Attentive Fusion Network

Most recent works directly utilize the channel-wise concatenated feature to do feature refinement and action modeling [8, 9, 17], which overlooks the impact of redundant information and causes extra noise. [11] firstly proposes to construct a cross-modal mechanism to filter out the redundant information, which leverages the performance. However, this method adopts global feature adaptation on the main modal branch and ignores that videos are usually dominated by irrelevant background snippets so it would harm the relevant action features during the global merge. What's more, [11] ignores the temporal correlation between snippets from different modalities.

To efficiently explore the complementary information between RGB and optical flow modality, we develop a multi-modal Transformer architecture with both self-attention Transformer and cross-attention Transformer through two snippet-wise attention maps. In this manner, the most similar parts between static and motion information are addressed and the redundant noise will be filtered out by such mechanism. Specifically, two attention maps $\mathbf{M}_r \in \mathbb{R}^{T\times T}$ and $\mathbf{M}_o \in \mathbb{R}^{T\times T}$ are defined as soft-mask in RGB and optical flow branches, which are then normalized with global softmax and capture the most similar action segments between two modalities.

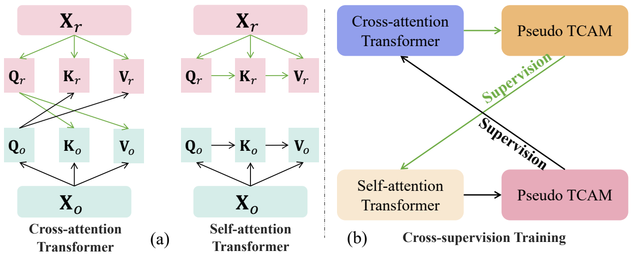


Figure 2. (a) the mechanism of cross-attention and self-attention, and (b) the proposed cross-supervision training strategy.

Following the Transformer design, we define three variables (query, key, value) as $\mathbf{Q}_r = \mathbf{X}_r\mathbf{W}_{qr}$, $\mathbf{K}_r = \mathbf{X}_r\mathbf{W}_{kr}$, $\mathbf{V}_r = \mathbf{X}_r\mathbf{W}_{vr}$, and $\mathbf{Q}_o = \mathbf{X}_o\mathbf{W}_{qo}$, $\mathbf{K}_o = \mathbf{X}_o\mathbf{W}_{ko}$, $\mathbf{V}_o = \mathbf{X}_o\mathbf{W}_{vo}$, where $\mathbf{W}_{qr}$, $\mathbf{W}_{kr}$, $\mathbf{W}_{vr}$, $\mathbf{W}_{qo}$, $\mathbf{W}_{ko}$, $\mathbf{W}_{vo}$ $\in \mathbb{R}^{D\times D}$ are the linear projection matrix for generating query $\mathbf{Q}_{r/o}$, key $\mathbf{K}_{r/o}$ and value $\mathbf{V}_{r/o}$ for RGB and optical flow modality, respectively. Finally, these maps are exploited by each Transformer module to generate the refined latent features:

$$\begin{cases} \mathbf{I}_r = \mathbf{X}_r + \mathcal{F}_{\text{Drop}}(\mathbf{M}_r\mathbf{V}_r\mathbf{W}_r), \\ \mathbf{I}_o = \mathbf{X}_o + \mathcal{F}_{\text{Drop}}(\mathbf{M}_o\mathbf{V}_o\mathbf{W}_o), \\ \mathbf{Z}_r = \mathbf{I}_r + \mathcal{F}_{\text{MLP}}(\mathcal{F}_{\text{LN}}(\mathbf{I}_r)), \\ \mathbf{Z}_o = \mathbf{I}_o + \mathcal{F}_{\text{MLP}}(\mathcal{F}_{\text{LN}}(\mathbf{I}_o)), \end{cases} \tag{7}$$

where $\mathbf{W}_r$ and $\mathbf{W}_o$ are the learnable projections, $\mathcal{F}_{\text{LN}}(\cdot)$ is the layer normalization function, $\mathcal{F}_{\text{Drop}}(\cdot)$ and $\mathcal{F}_{\text{MLP}}(\cdot)$ are the dropout and MLP module separately, and $\mathbf{I}_r$, $\mathbf{I}_o$ are the intermediate latent feature for RGB and optical flow. To build the self-attention and cross-attention Transformers, we can define different $\mathbf{M}_{r/o}$. For self-attention Transformer, $\mathbf{M}_r = \frac{\mathbf{Q}_r\mathbf{K}_r^\top}{\sqrt{T}}$ and $\mathbf{M}_o = \frac{\mathbf{Q}_o\mathbf{K}_o^\top}{\sqrt{T}}$; for cross-attention Transformer, $\mathbf{M}_r = \frac{\mathbf{Q}_o\mathbf{K}_r^\top}{\sqrt{T}}$ and $\mathbf{M}_o = \frac{\mathbf{Q}_r\mathbf{K}_o^\top}{\sqrt{T}}$. Figure 2 (a) shows the comparison of dual-Transformer design.

Moreover, multi-head attention is also introduced to expand the capacity of attention modules [35]. Based on this, we refine the pre-trained cross-modal features with static and motion information from different modalities and find out the most similar segments at the same time, which reduce the impact of background snippets.

3.3.2 Temporal Localization Refinement via Plateau Modelling

Since specific action boundaries are not available for WS-TAL task, existing works [5, 12, 24] typically apply a range of thresholds on $\mathbf{a}_m$ to generate temporal action proposals.

Unfortunately, continuous action patterns are not captured by those methods, that is, an action segment should have very similar attention weights.

To further refine the temporal localization, we propose a plateau refined distribution function on $\mathbf{a}_m$ so that the action probability of each snippet in a segment is evenly distributed. Another desirable property of this distribution is differentiability so that the function can be tuned by the scores in $\mathbf{a}_m$ [22]. Specifically, the plateau fitting function is defined to model the probability density of plateau distribution over the snippets x of an untrimmed video as:

$$\mathcal{F}_{\mathrm{p}}(x|t_c,\omega,\varrho)=\frac{1}{(\mathbf{e}^{\varrho(x-t_c-\omega)}+1)(\mathbf{e}^{\varrho(-x-t_c-\omega)}+1)}, \tag{8}$$

where $\mathcal{F}_{\mathrm{p}}(\cdot|\cdot)$ is the plateau function over the center of the plateau t_c, the width ω and the steepness of the boundary ϱ. The range of the function is [0, 1]. We follow [22] and fit the plateau function on $\mathbf{a}_m$ to obtain the refined t_c, ω, ϱ as follows:

$$t_c,\omega,\varrho=\underset{t_c,\omega,\varrho}{\operatorname{argmin}}\,\mathcal{L}_{\mathrm{mse}}(\mathbf{a}_m,\mathcal{F}_{\mathrm{p}}). \tag{9}$$

As shown in Figure 3, the gray dash line represents a segment of original attention weights $\mathbf{a}_m$ from snippet #48 to snippet #68. The whole $\mathbf{a}_m$ is shown in Figure 4(b). To build the plateau distribution, we firstly use proper thresholds on $\mathbf{a}_m$ to obtain multiple action proposals and their associated attention weights ($\hat{\mathbf{a}}_m$) which is marked in bold black. Each $\hat{\mathbf{a}}_m$ will be input into Eq. (9) to fit $\mathcal{F}_{\mathrm{p}}$ through MSE loss, producing t_c, ω, ϱ. $\mathcal{F}_{\mathrm{p}}$ replaces $\hat{\mathbf{a}}_m$ as new attention weights and refines its time scale. In Figure 3, we use a pink dashed line "foreground plateau" to show the new attention weights, "w" represents ω, and "c" represents t_c. Specially, t_c is around the highest-scoring snippet in $\hat{\mathbf{a}}_m$, ω constrains the width to filter out background snippet and slope - ϱ preserves edge action snippets. However, ω generally is much wider than the real action scale and will include background snippets as "foreground plateau" shows in Figure 3. We introduce background attention weights $\mathbf{b}_m$ (equal to $\mathbf{1}$ - $\mathbf{a}_m$). Same as $\mathbf{a}_m$, we also first apply thresholds on $\mathbf{b}_m$ to obtain background proposals and then build background plateau distribution, marked as "background plateau". Thus there will be two kinds of plateau distribution for each video sample, we filter out the "overlap" area and conclude with refined attention weights (marked as "refined plateau").

Note that we only illustrate one action proposal in Figure 3 to make it straightforward to understand the intuition. This method can also be used in multiple action proposals. Compared to traditional threshold methods, our "dilation-erosion" plateau method is more reasonable since it is based on a center snippet t_c and expands with "ω". This method will filter out negative information while reserving edge action snippets.

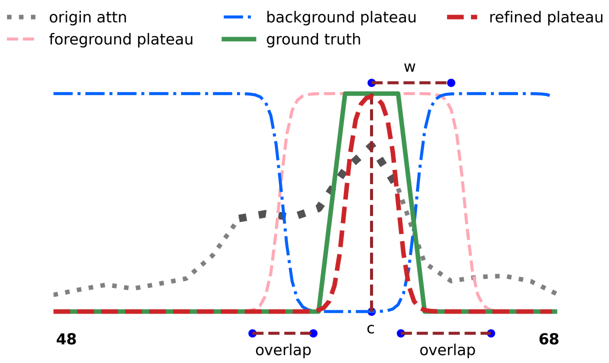


Figure 3. Mechanism of plateau refinement. Compared with traditional threshold methods, our refinement adopts a "dilation-erosion" strategy, which not only filters out background information but also reserves edge action boundary.

3.4. Model Training Strategy

To make the model training converge well, we deploy two-stage training mechanism by first pre-training the transformer block using reconstruction loss, then fine-tuning the whole model with our proposed framework.

Stage 1. Warm-up Stage: The goal of our transformer module is to refine the pre-trained I3D features, although existing works [24, 32] directly adopted $\mathbf{X}_r$ and $\mathbf{X}_o$ and achieved relatively good performance. However, transformer-based architecture needs strong supervision to train well from scratch, which means that scratch-initialized attention block using Eq. (7) probably results in trivial solution to $\mathbf{Z}_r$ and $\mathbf{Z}_o$. To increase the training stability, we introduce a reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ to pre-train the feature embedding function $\mathcal{F}_e(\cdot)$ as follows:

$$\mathcal{L}_{\mathrm{rec}}=\mathcal{L}_{\mathrm{mse}}(\mathbf{Z}_r,\mathbf{X}_r)+\mathcal{L}_{\mathrm{mse}}(\mathbf{Z}_o,\mathbf{X}_o), \tag{10}$$

which can guarantee the learned $\mathbf{Z}_r$ and $\mathbf{Z}_o$ are not far away from $\mathbf{X}_r$ and $\mathbf{X}_o$, so that the effective information existed in pre-trained features would not be destroyed. Note that $\mathbf{Z}_r$ and $\mathbf{Z}_o$ will not be exactly the same as $\mathbf{X}_r$ and $\mathbf{X}_o$, since we only optimize Eq. (10) for warming up with limited training iterations. Details are shown in Sec 4.2. With the pre-trained feature embedding function, we then fine-tune the whole pipeline with Eq. (6) by a small learning rate. To achieve fast convergence, we further apply the extra co-activity similarity loss [25] in this stage.

Stage 2. Optimization Stage: Since we have two kinds of Transformer architectures, we propose a novel optimization strategy with a combination of cross-attention and self-attention. Through pseudo labels generated by these two attention modules, a cross-supervision mechanism is built as shown in Figure 2(b) by optimizing the objective function (Eq. (6)). The plateau refinement (Eq. (9)) is iteratively optimized after the cross-supervision training until the model converges or the maximal iterations reach.

Table 1. Comparison of state-of-the-art methods on THUMOS14 dataset. The average mAP (AVG) is computed under IoU thresholds [0.1:0.1:0.7]. †means additional information is added for training, such as action frequency or human pose.

Setting	**Method**	**mAP@IoU (%)**							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
FS-TAL	SSN [42]	60.3	56.2	50.6	40.8	29.1	-	-	-
	TAL-Net [4]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1
	P-GCN [38]	69.5	67.5	63.6	57.8	49.1	-	-	-
	GTAN [20]	69.1	63.7	57.8	47.2	38.8	-	-	-
WS-TAL†	CMCS [19]	57.4	50.8	41.2	32.1	23.1	15.0	7.0	32.4
	STAR [36]	68.8	60.0	48.7	34.7	23.0	-	-	-
	3C-Net [23]	59.1	53.5	44.2	34.1	26.6	-	8.1	-
	PreTrimNet [41]	57.5	50.7	41.4	32.1	23.1	14.2	7.7	23.7
	SF-Net [21]	71.0	63.4	53.2	40.7	29.3	18.4	9.6	40.8
WS-TAL	CoLA [39]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9
	UGCT [37]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	43.6
	CSCL [13]	68.0	61.8	52.7	43.3	33.4	21.8	12.3	41.9
	CO$_2$-Net [11]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	44.6
	FTCL [8]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	43.6
	RSKP [12]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	45.1
	ASM-Loc [9]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	45.1
	DGCNN [26]	66.3	59.9	52.3	43.2	32.8	22.1	13.1	41.3
	Li *et al.* [17]	69.7	64.5	58.1	**49.9**	39.6	27.3	14.2	46.1
	DELU [5]	71.5	66.2	56.5	47.7	40.5	27.2	**15.3**	46.4
	TFE-DCN [43]	72.3	66.5	58.6	49.5	40.7	27.1	13.7	46.9
	Two-Stream [33]	73.0	68.2	60.0	47.9	37.1	24.4	12.7	46.2
	Boosting [16]	-	-	56.2	47.8	39.3	27.5	15.2	-
	DDG-Net [29]	72.5	67.7	58.2	49.0	**41.4**	27.6	14.8	47.3
	Ours	**74.1**	**69.2**	**60.0**	49.8	41.1	**28.0**	15.1	**48.2**

4. Experimental Results

4.1. Datasets

Our work is highly motivated by [5, 11]. To make fair comparison, we mainly evaluate our proposed method on two public video datasets: THUMOS14 [14] and ActivityNet1.2 [2].

THUMOS14 consists of 200 untrimmed validation videos and 213 untrimmed test videos. These videos have various lengths from tens of seconds to several minutes. There are 20 action categories distributed in these videos, in which one video might contain multiple actions. Following previous work [9, 11, 12], we adopt the 200 validation videos for training and 213 test videos for evaluation.

ActivityNet1.2 is a large action localization dataset with 100 daily action classes. It includes 4,819 training videos and 2,383 videos for validation. Considering the annotations for test videos in this dataset are not released, we deploy the training videos for model optimization and validation videos for performance evaluation, by following the protocol in previous work [5, 11].

4.2. Implementation Details

We utilize the two-stream I3D [3] pre-trained on Kinetics-400 [3] to extract two-stream features. The optical flow is extracted by TV-L1 algorithm. The features output by I3D network is a sequence of snippets, in which each snippet consists of 16 non-overlapping frames sampled from the original video. One snippet is a 2,048-dimension vector: the first 1,024 is RGB feature and another 1,024 is optical flow feature. In the training phase, we fix the number of snippets T as 512 for THUMOS14 dataset and 60 for ActivityNet1.2, while the original length is retained during testing. In view of fair comparisons, we do not fine-tune the I3D feature extractor. For the Transformer model, we adopt a multi-head design and the number of heads is 4. Every Transformer only has 1 attention module in consideration of simple structure. The attention unit is constructed with 3 convolution layers, whose output dimensions are 512, 512, and 1 while kernel sizes are 3, 3, and 1. Two dropout layers with a rate of 0.5 are also embedded into each attention unit. The classification module contains 3 temporal convolution layers, between which two dropout layers with a rate of 0.7 are added to regularize intermediate features. Note that "self" and "cross" attention share the same model weight.

Table 2. Comparison of state-of-the-art methods on ActivityNet1.2 dataset. The average mAP (AVG) is computed under IoU thresholds [0.5:0.05:0.95].

Setting	Method	mAP@IoU (%)			
		0.5	0.75	0.95	AVG
FS-TAL	SSN [42]	41.3	27.0	6.1	26.6
WS-TAL†	CMCS [19]	35.4	22.9	8.5	21.1
	3C-Net [23]	36.8	22.0	5.6	22.4
WS-TAL	CoLA [39]	42.7	25.7	5.8	26.1
	UGCT [37]	41.8	25.3	5.9	25.8
	CO$_2$-Net [11]	43.3	26.3	5.2	26.4
	DGCNN [26]	42.0	25.8	**6.0**	26.2
	Li *et al.* [17]	41.6	24.8	5.4	25.2
	DELU [5]	44.2	26.7	5.4	26.9
	DDG-Net [29]	44.3	26.9	5.5	27.0
	Ours	**45.1**	**27.7**	5.5	**27.6**

Table 3. Contribution analysis of each component.

Model	Module			mAP@IoU (%)				
	$\mathcal{L}_{\text{pseudo}}$	**CS**	**Π**	0.1	0.3	0.5	0.7	AVG
Baseline				64.7	50.3	34.5	11.9	42.2
	✓			66.2	48.5	30.5	9.1	38.8
	✓		✓	64.6	50.7	36.6	13	41.0
M^{2}T				69.8	54.3	36.7	11.7	43.5
	✓			70.4	54.5	37.3	12.5	44.1
	✓	✓		70.8	56.5	39.4	14.3	45.7
	✓		✓	71.6	57.8	39.4	14.2	46.6
	✓	✓	✓	74.1	60.0	41.1	15.1	**48.2**

For the hyper-parameters, we set batch size as 10 for one iteration, 20 iterations as one step for model evaluation and pseudo label generation on THUMOS14 and 500 iterations on ActivityNet1.2. We warm up M^{2}T with Eq. (10) in 2,000 iterations and fine-tune it with 20,000 iterations. For the plateau refinement stage threshold, we select 0.4 for action proposal generation and 0.45 for background proposal generation. We choose $\lambda_0 = 10$ for pseudo label loss, while $\lambda_1 = 0.8$ for the last two terms of loss function [11]. For uncertainty, we set $\tau = 0.2$ for THUMOS14 and $\tau = 0.001$ for ActivityNet1.2, respectively. For top-k pooling of TCAM, we select $k = 7$ for THUMOS14 and $k = 5$ for ActivityNet1.2. For the model optimization, we employ Adam optimizer with a learning rate as 5e^{-5} and the weight decay rate is 0.001 for THUMOS14, while 3e^{-5} and 5e^{-4} for ActivityNet1.2. All experiments are run on a single NVIDIA RTX 3090 GPU.

4.3. Comparison with the State-of-the-Art

We compare our proposed method with state-of-the-art WS-TAL methods and several fully-supervised TAL methods to examine the validness of our model.

In Table 1, we report the compared results on THUMOS14 dataset. We observe that our method establishes a state-of-the-art result on THUMOS14 with 48.2% average mAP for IoU thresholds 0.1:0.7. In particular, our model achieves better performance for all IoU thresholds than CO$_2$-Net [11] and most IoU than DELU [5], both of which also apply cross-modal features for action localization. And for those models with pseudo labels and uncertainty to guide model training, i.e., ASM-Loc [9] and UGCT [37], our model outperforms their methods for most IoU thresholds. Our model also performs better than all Transformer-based methods[9, 16]. Even compared with fully supervised methods, our model performs better than SSN and TAL-Net, and is comparable with GTAN and P-GCN at low IoU threshold. The results validate the superior effectiveness of our proposed method.

We also conduct experiments on larger dataset ActivityNet1.2 and the results are reported in Table 2. Our method obtains the performance of 27.6% average mAP on ActivityNet1.2. The reason for slightly lower performance compared to THUMOS14 is that there are only 1.5 action instances per video in ActivityNet1.2, while THUMOS14 contains 15 action segments per video [11]. Also, the annotations are coarser than THUMOS14. These factors lead to limited improvements on ActivityNet1.2 for all existing methods. Following most works [5, 8, 12, 13, 17], we prefer to conduct ablation studies on THUMOS14 dataset.

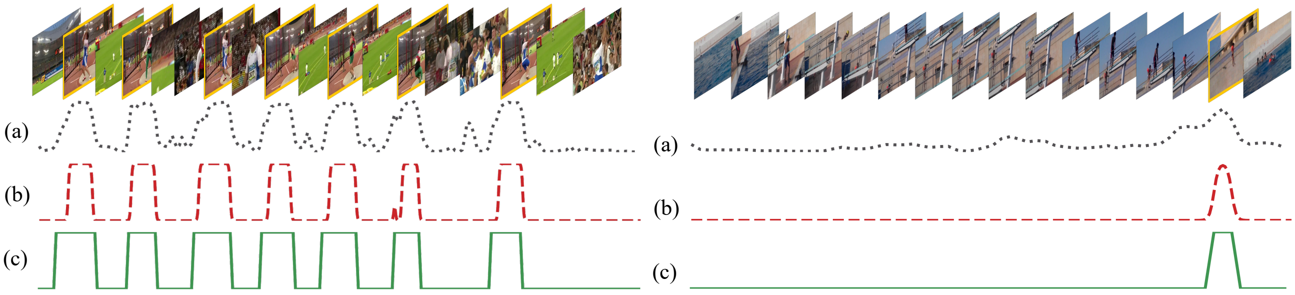
4.4. Ablation Study on THUMOS14

4.4.1 Influence of Model Variants

In Table 3, we conduct ablation study to investigate the contribution of each component in our model. To efficiently verify the effect of each component, we design a **Baseline** using one convolution layer as $\mathcal{F}_e(\cdot)$ to refine the input feature with necessary constraints, i.e., $\mathcal{L}_{\text{mil}}$, $\mathcal{L}_{\text{ml}}$, $\mathcal{L}_{\text{norm}}$ and $\mathcal{L}_{\text{oppo}}$.

Based on this baseline, we design several variants of our M^{2}PT by including more components. First, "M^{2}T" denotes that we construct $\mathcal{F}_e(\cdot)$ with the multi-modal Transformer feature embedding model. "**Π**" means we deploy the plateau function to refine the attention weights. We also examine the pseudo label loss $\mathcal{L}_{\text{pseudo}}$ and the cross-supervision training mechanism (**CS**).

From the results in Table 3, it is obvious that a consistent gain is achieved when adding modules to the proposed M^{2}PT. In particular, compared with **Baseline**, our multi-modal Transformers design significantly increases the performance by 1.3% on average. The two branches complement each other and filter out redundant information. When incorporating all introduced modules and adopting cross-supervision training, our framework boosts the final performance from 42.2% to 48.2%. Regarding the performance drop for baseline with pseudo labels, we notice that noisy pseudo labels impair the base model as depicted in Figure 5, which further verifies that the simple feature encoding of baseline leads to weak robustness to negative samples.



(a) Hammer Throwing

(b) Diving

Figure 4. Illustration of two qualitative cases from action "Hammer Throwing" and "Diving" in THUMOS14. In each figure, row-(a) denotes the original attention weights, row-(b) shows the output of our plateau refinement and row-(c) means the ground-truth temporal localization.

Table 4. Comparison of different plateau settings.

Setting	**mAP@IoU (%)**				
	0.1	0.3	0.5	0.7	AVG
self-attention	70.5	55.8	39.7	14.6	45.8
cross-attention	71.4	56.9	40.0	13.8	46.1
both	74.1	60.0	41.1	15.1	**48.2**

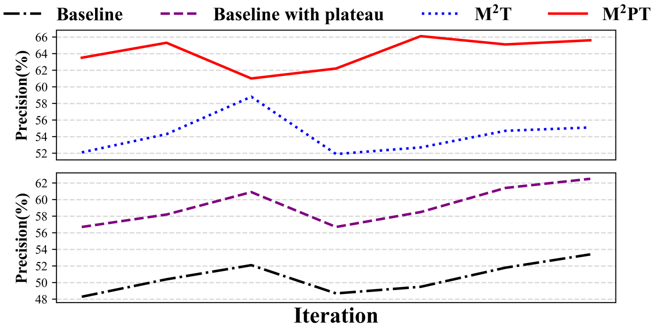


Figure 5. Comparison on pseudo label precision without and with plateau refinement. The upper figure shows the comparison of M^{2}T and M^{2}PT, and the bottom one lists the comparison on Baseline and Baseline with Plateau refinement.

4.4.2 Impact of plateau function

We explore the validness of the plateau function in refining the noisy pseudo labels and attention weights, which mitigates the impact of false positive samples on pseudo-label loss. Table 4 reports the influence of different plateau settings. Because we adopt a cross-supervision strategy, it is helpful to investigate which pseudo label generation scheme is noisier. According to the results, we conclude that pseudo labels by self-attention Transformers contain more negative samples.

Figure 4 illustrates the visualization comparisons among original temporal localization, our plateau-refined localization and ground-truth. We observe that original attention weights are usually noisy and contain many negative samples, while most existing works heavily rely on attention weights to generate action proposals, which hurts their performance. To handle this, we propose our plateau function to efficiently suppress these background parts and make the pseudo label more similar to the ground truth, such as the "Hammer Throwing" action localization in Figure 4 [Left]. As shown in Figure 4 [Right], the plateau function refines more precise localization on short action "Diving", which is generally the weakness for most MIL-based methods.

Figure 5 shows the improvement in the quality of pseudo labels. We compare "Baseline" with "Baseline with plateau" and "M^{2}T" with "M^{2}PT", separately. The average precision of "Baseline with plateau" is 8.7% higher than "Baseline", and "M^{2}PT" outperforms 10% than "M^{2}T". These results substantially verify the effectiveness of plateau functions. Additionally, pseudo label precision of "M^{2}T" achieves a 3.6% improvement over the baseline model, which proves feature enhancement of multi-modal Transformers.

More qualitative results and effects of hyperparameters are in supplementary materials.

5. Conclusion

In this work, we propose a novel Multi-Modal Plateau Transformers (M^{2}PT) for weakly-supervised temporal action localization (WS-TAL) problem. To efficiently exploit temporal structure within videos and simultaneously reduce redundant information caused by the pre-trained I3D network, we construct cross-attention and self-attention modules to conduct feature embedding on two-stream features. To enhance the performance of action localization, we also utilize pseudo labels to iteratively refine latent features, and introduce plateau functions to refine the temporal localization and then improve the precision of pseudo labels. Experiments on two popular action benchmarks verify the effectiveness of our proposed model when compared with the state-of-the-art methods.

References

[1] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017. 1, 2

[2] 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*, pages 961–970, 2015. 6

[3] 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*, pages 6299–6308, 2017. 3, 6

[4] 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*, pages 1130–1139, 2018. 1, 2, 6

[5] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 1, 3, 4, 6, 7

[6] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. *arXiv preprint arXiv:2204.01680*, 2022. 1, 2

[7] 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 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022. 1, 4, 6, 7

[9] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022. 1, 2, 3, 4, 6, 7

[10] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1914–1923, 2016. 1, 2

[11] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021. 2, 3, 4, 6, 7

[12] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022. 2, 3, 4, 6, 7

[13] Yuan Ji, Xu Jia, Huchuan Lu, and Xiang Ruan. Weakly-supervised temporal action localization via cross-stream collaborative learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 853–861, 2021. 3, 6, 7

[14] 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://crcv.ucf.edu/THUMOS14/, 2014. 6

[15] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11320–11327, 2020. 3

[16] Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Boosting weakly-supervised temporal action localization with text information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10657, 2023. 6, 7

[17] Ziqiang Li, Yongxin Ge, Jiaruo Yu, and Zhongming Chen. Forcing the whole video as background: An adversarial learning strategy for weakly temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5371–5379, 2022. 1, 3, 4, 6, 7

[18] 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*, pages 11499–11506, 2020. 2

[19] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 6, 7

[20] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019. 1, 6

[21] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. Sf-net: Single-frame supervision for temporal action localization. In *European conference on computer vision*, pages 420–437. Springer, 2020. 6

[22] Davide Moltisanti, Sanja Fidler, and Dima Damen. Action recognition from single timestamp supervision in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9915–9924, 2019. 2, 3, 4, 5

[23] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8687, 2019. 6, 7

[24] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 2, 3, 4, 5

[25] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 5

[26] Haichao Shi, Xiao-Yu Zhang, Changsheng Li, Lixing Gong, Yong Li, and Yongjun Bao. Dynamic graph modeling for weakly-supervised temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3820–3828, 2022. 1, 6, 7

[27] Yooju Shin, Susik Yoon, Sundong Kim, Hwanjun Song, Jae-Gil Lee, and Byung Suk Lee. Coherence-based label propagation over time series for accelerated active learning. In *International Conference on Learning Representations*, 2022. 3

[28] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058, 2016. 2

[29] Xiaojun Tang, Junsong Fan, Chuanchen Luo, Zhaoxiang Zhang, Man Zhang, and Zongyuan Yang. Ddg-net: Discriminability-driven graph network for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6622–6632, 2023. 6, 7

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[32] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 2, 3, 5

[33] Yu Wang, Yadong Li, and Hongbin Wang. Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18878–18887, 2023. 6

[34] 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*, pages 5783–5792, 2017. 2

[35] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 2, 4

[36] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9070–9078, 2019. 6

[37] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021. 1, 2, 3, 6, 7

[38] 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*, pages 7094–7103, 2019. 6

[39] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 2, 6, 7

[40] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2

[41] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, and Peng Li. Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12886–12893, 2020. 6

[42] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 1, 2, 6, 7

[43] Jianxiong Zhou and Ying Wu. Temporal feature enhancement dilated convolution network for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6028–6037, 2023. 6