Weakly-Supervised Temporal Action Localization by Inferring Salient Snippet-Feature

Wulian Yun, Mengshi Qi, Chuanming Wang, Huadong Ma*

Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, China
{yunwl,qms,wcm,mhd}@bupt.edu.cn

Abstract

Weakly-supervised temporal action localization aims to locate action regions and identify action categories in untrimmed videos simultaneously by taking only video-level labels as the supervision. Pseudo label generation is a promising strategy to solve the challenging problem, but the current methods ignore the natural temporal structure of the video that can provide rich information to assist such a generation process. In this paper, we propose a novel weakly-supervised temporal action localization method by inferring salient snippet-feature. First, we design a saliency inference module that exploits the variation relationship between temporal neighbor snippets to discover salient snippet-features, which can reflect the significant dynamic change in the video. Secondly, we introduce a boundary refinement module that enhances salient snippet-features through the information interaction unit. Then, a discrimination enhancement module is introduced to enhance the discriminative nature of snippet-features. Finally, we adopt the refined snippet-features to produce high-fidelity pseudo labels, which could be used to supervise the training of the action localization network. Extensive experiments on two publicly available datasets, *i.e.*, THUMOS14 and ActivityNet v1.3, demonstrate our proposed method achieves significant improvements compared to the state-of-the-art methods. Our source code is available at https://github.com/wuli55555/ISSF.

Introduction

Temporal action localization (TAL) (Shou, Wang, and Chang 2016; Zhao et al. 2017; Chao et al. 2018; Huang, Wang, and Li 2022; He et al. 2022) aims to find action instances from untrimmed videos, *i.e.*, predicting the start positions, end positions, and categories of certain actions. It is an important yet challenging task in video understanding and has been widely used in surveillance and video summarization. To achieve accurate localization, most existing methods (Shou, Wang, and Chang 2016; Zhao et al. 2017; Chao et al. 2018; Lin et al. 2018; Long et al. 2019) rely on training a model in a fully supervised manner with the help of human-labeled precise temporal annotations. However, fine-detailed labeling of videos is labor-intensive and expensive. In contrast, weakly-supervised methods recently have gained increasing attention from both academia and industry, since they only utilize video-level labels for temporal action localization, achieving competitive results while reducing the cost of manual annotations.

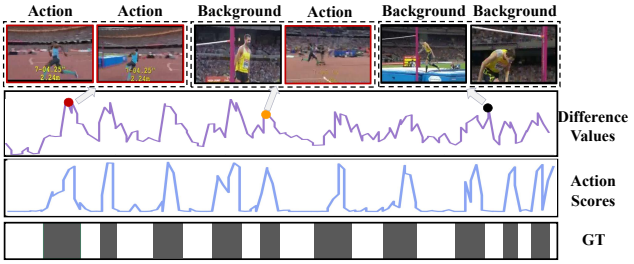

Figure 1: Illustration of difference values among snippets, action scores and Ground-Truth (GT). The action and background snippets are marked as red and black boxes, respectively.

Weakly-supervised TAL methods (Xu et al. 2019; Shi et al. 2020; Qu et al. 2021; Lee et al. 2021; Liu et al. 2021; Narayan et al. 2021) mainly utilize a "localization by classification" framework, where a series of Temporal Class Activation Maps (TCAMs) (Nguyen et al. 2018; Paul, Roy, and Roy-Chowdhury 2018) are obtained by snippet-wise classification, and then TCAMs are used to generate temporal proposals for action localization. However, the classifiers primarily tend to focus on easily distinguishable snippets while ignoring other subtle yet equally important information, so there is a discrepancy between classification and localization. To balance the performance of classification and localization, pseudo label based methods (Huang, Wang, and Li 2022; He et al. 2022; Pardo et al. 2021; Zhai et al. 2020; Luo et al. 2020; Li et al. 2022) have been proposed, which supervises the training of the model mainly by generating snippet-level pseudo label information.

Nevertheless, accurately generating pseudo label remains challenging, since existing methods ignore the important role played in the temporal structure of videos. We observed that neighbor snippets exhibit obvious distinctively difference relationships, which can discover salient features and identify differentiated boundaries. As shown in Figure 1, neighbour snippet-features with substantial varia-

*Corresponding authors: Huadong Ma, Mengshi Qi.

tions (higher difference value) may correspond to the junctions between action and background, alternations between action, or abrupt changes between background. However, how to find these features and refine them into more discriminative features is the key to discover action boundaries and then improve the localization performance.

Inspired by this observation, we propose a novel weakly-supervised TAL method, which takes a new perspective that boosts the generation of high-fidelity pseudo-labels by leveraging the temporal variation. First, we design a saliency inference module to discover significant snippet-feature by leveraging the variation and calculating the difference values of neighbor snippet pairs. However, this process only considers local relationships and ignores the global information in the video. Thus, we propose a boundary refinement module to enhance salient features through information interaction while making the model focus on the entire temporal structure. Subsequently, considering diverse action information can provide additional clues, we propose a discrimination enhancement module to further refine the feature by constructing a memory to introduce the same category of action knowledge. Finally, the output features are fed into the classification head to generate the final refined pseudo labels for supervision.

The contributions can be summarized as follows:

(1) We propose a new pseudo-label generation strategy for weakly-supervised TAL by inferring salient snippet-feature, which can exploit the dynamic variation.

(2) We design a boundary refinement module and a discrimination enhancement module to enhance the discriminative nature of action and background, respectively.

(3) We conduct extensive experiments and the results show our model achieves 46.8 and 25.8 average mAP on THUMOS14 and ActivityNet v1.3, respectively.

Related Work

Fully-supervised temporal action localization. Fully-supervised TAL has been an active research area in video understanding (Qi et al. 2021, 2020, 2019, 2018; Liu et al. 2016, 2018) for many years and existing methods are divided into two categories, *i.e.,* one-stage methods and two-stage methods. One-stage methods (Long et al. 2019; Lin, Zhao, and Shou 2017; Yang et al. 2020; Lin et al. 2021) predict action boundaries as well as labels simultaneously. On the contrary, two-stage methods (Shou, Wang, and Chang 2016; Zhao et al. 2017; Chao et al. 2018; Zeng et al. 2019) first find candidate action proposals and then predict their labels. However, these fully-supervised methods are trained with instance-level human annotation, leading to an expensive and time-consuming process.

Weakly-supervised temporal action localization. Weakly-supervised TAL methods (Xu et al. 2019; Min and Corso 2020; Shi et al. 2020; Lee et al. 2021; Liu et al. 2021; Narayan et al. 2021; Zhai et al. 2020; Huang, Wang, and Li 2022; Chen et al. 2022) mainly learn from video-level labels, which avoid labor-intensive annotations compared to the fully-supervised methods. UntrimmedNet (Wang et al. 2017) and STPN (Nguyen et al. 2018) generate class activation sequences by Multiple Instance Learning (MIL) framework and then locate action instances by thresholding processing. RPN (Huang et al. 2020) and 3C-Net (Narayan et al. 2019) use metric learning algorithms to learn more discriminative features. Lee *et al.* (Lee, Uh, and Byun 2020) design a background suppression network to suppress background snippets activation. However, there is still a discrepancy between classification and localization. Recently, numerous methods (Pardo et al. 2021; Luo et al. 2020; Zhai et al. 2020; Yang et al. 2021; Huang, Wang, and Li 2022; He et al. 2022) attempt to generate pseudo labels to supervise the model and thus alleviate the discrepancy. RefineLoc (Pardo et al. 2021) alleviates the discrepancy between classification and localization by extending the previous detection results to generate pseudo labels. Luo *et al.* (Luo et al. 2020) exploit the Expectation–Maximization framework (Moon 1996) to generate pseudo labels by alternately updating the key-instance assignment branch and the foreground classification branch. TSCN (Zhai et al. 2020) generates frame-level pseudo labels by later fusing attention sequences in consideration of two-stream consensus. Li *et al.* (Li et al. 2022) exploit contrastive representation learning to enhance the feature discrimination ability. ASM-Loc (He et al. 2022) generates action proposals as pseudo labels by using the standard MIL-based methods. In contrast, our method exploits the variation between neighbor snippet-features to find salient snippet-features, and further designs a boundary refinement module and a discrimination enhancement module to generate high-fidelity pseudo labels.

Methodology

In this section, we will begin by presenting the problem definition of weakly-supervised TAL and provide an overview of our proposed method. Next, we will describe the different modules of our method in detail, which are designed to generate high-fidelity pseudo labels by utilizing the variation between snippet-features. Finally, we introduce the training details of optimizing the temporal localization model.

Problem definition. Weakly-supervised TAL aims to predict a group of action instances (c, q, t_s, t_e) for each test video with the assistance of a set of untrimmed training videos $\{V_i\}_{i=1}^{N}$ and their corresponding ground-truth labels $\{y_i\}_{i=1}^{N}$. Specifically, $y_i \in \mathbb{R}^C$ is a binary vector indicating the presence/absence of each of C actions. For one action instance, c denotes the action category, q refers to the prediction confidence score, t_s and t_e mean the start time and end time of the action, respectively.

Overview. The overview of our proposed method is shown in Figure 2, which mainly contains four parts: *(a) base branch*, *(b) saliency inference module*, *(c) boundary refinement module*, and *(d) discrimination enhancement module*.

First, in the base branch, we exploit a fixed pre-trained backbone network (e.g., I3D) to extract T snippet-features from both the appearance (RGB) and motion (optical flow) of the input video. Then, a learnable classification head is adopted to classify each snippet and obtain the predicted TCAMs. Second, we utilize the saliency inference module to generate salient snippet-features by calculating the

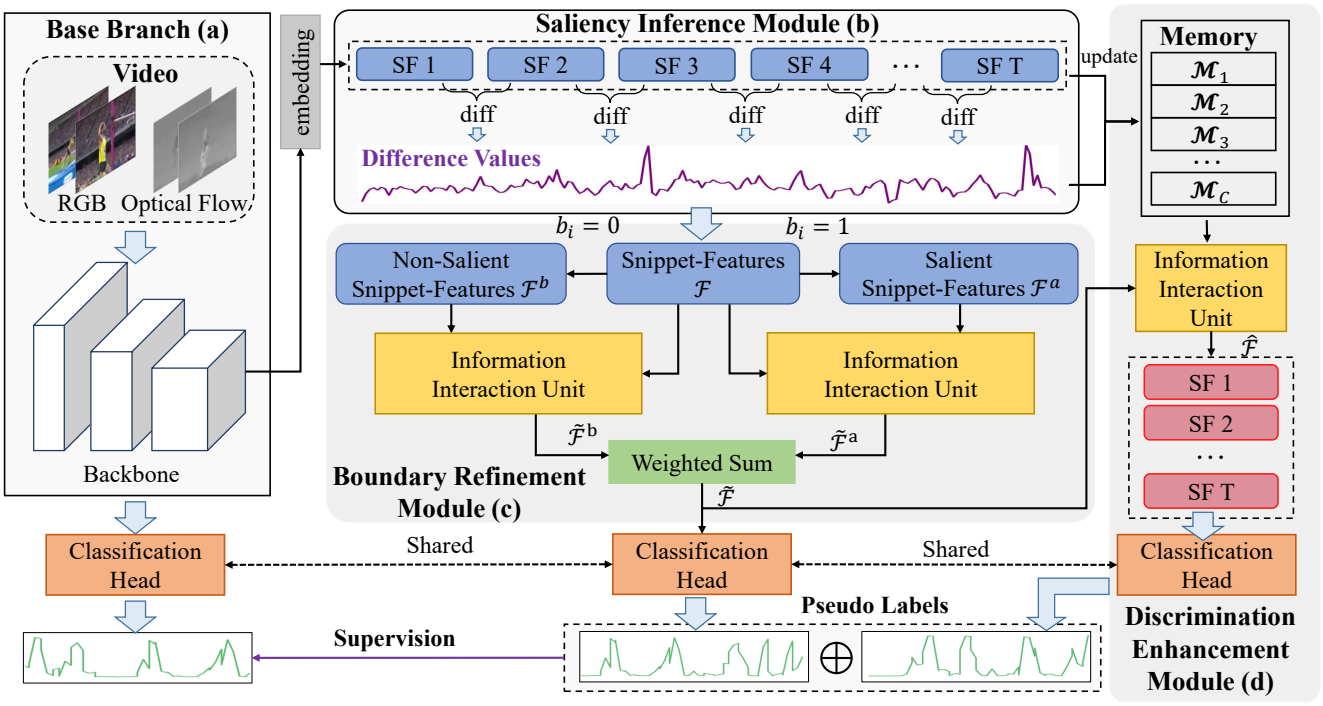

Figure 2: Overview of our model. Firstly, the base branch (a) extracts features from RGB and optical flow in a video and uses the classification head to predict TCAMs. Then, the saliency inference module (b) exploits the variation relationship between snippet-features to discover salient snippet features. Next, the boundary refinement module (c) utilizes the information interaction unit to enhance salient snippet features. Subsequently, the discrimination enhancement module (d) leverages action information stored in memory to enhance the discrimination of action and background. Finally, (c) and (d) generate high-fidelity pseudo labels to supervise the base branch.

difference between adjacent pairs of snippet-features. Subsequently, the boundary refinement module and the discrimination enhancement module both utilize the information interaction unit to refine coarse boundaries by enhancing salient snippet-features and the separability of action snippet-features from those of the background. Finally, the output features are fed into the classification head to generate high-fidelity pseudo labels as a supervised signal for the base branch.

Base Branch

Given an untrimmed video V, we follow (Nguyen et al. 2018; Huang, Wang, and Li 2022) to split it into multiple non-overlapping snippets $\{v_i\}_{i=1}^{T}$, and then we use the I3D (Carreira and Zisserman 2017) network pre-trained on the Kinetics-400 (Kay et al. 2017) dataset to extract features from the RGB and optical flow streams for each snippet. An embedding layer takes the concatenation of these two types of features to fuse them together, and the fused features of all snippets are treated as snippet-features of the video $\mathcal{F} = \{f_1, f_2, \cdots, f_T\} \in \mathbb{R}^{T \times D}$, where T is the number of snippets and D denotes the dimension of one snippet-feature.

Next, we use the classification head to obtain Temporal Class Activation Maps (TCAMs) $\mathcal{T} \in \mathbb{R}^{T \times (C+1)}$, where $C+1$ denotes the number of action categories plus the background class. Specifically, following previous work (Huang, Wang, and Li 2022), the classification head consists of a Class-agnostic Attention (CA) head and a Multiple Instance Learning (MIL) head.

Saliency Inference Module

The significant variation of temporal neighbor snippets can indicate whether each snippet belongs to a salient snippet-feature. Therefore, we propose a saliency inference module that utilizes such variation to explore the difference between neighbor snippet pairs and then use it to identify salient boundaries in the video.

Given a video and its snippet-level representation $\mathcal{F} \in \mathbb{R}^{T \times D}$, we first calculate the difference value $\tau_{(t-1,t)}$ of each pair of temporal adjacent snippet-features $\{f_{t-1}, f_t\}$ in the formulation of:

$$\tau_{(t-1,t)} = \sum_{d=1}^{D} |\mathrm{diff}(f_t, f_{t-1}, d)|, \tag{1}$$

where diff denotes the operation of dimensional-wise subtraction, and $d \in D$ means the element index of the feature. Subsequently, we obtain the difference set τ of the input video by calculating the difference for all pairs:

$$\tau = \{\tau_{(1,2)}, \tau_{(2,3)}, \cdots, \tau_{(t-1,t)}\}. \tag{2}$$

To obtain the salient snippet-features of the video, we first perform a descending sort on the difference set τ, and then assign the initial labels $\mathcal{B} = \{b_i\}_{i=1}^{T}$ to each snippet based on the sorted τ. The snippets with the top K sorted scores are selected as salient snippet-features, while the remaining are selected as non-salient snippet-features, and the process of assigning labels can be formulated as:

$$b_t = \begin{cases} 1, \text{ if } \tau_{(t-1,t)} \in \text{Top}(\text{sorted}(\tau), K) \\ 0, \ otherwise \end{cases}, \tag{3}$$

where $b_t = 1$ denotes that its corresponding snippet f_t belongs to the salient snippet-features, otherwise to the non-salient snippet-features. Finally, salient snippet-features are discovered in a simple manner. However, since these snippet-features cannot be determined as actions or backgrounds, directly using these features to supervise the learning of the base branch may lead to poor performance. Next, we will present how to refine these salient snippet-features.

Boundary Refinement Module

In the saliency inference module, we calculate the difference values between each pair of adjacent snippets, and the operation can be seen as one type of exploiting local relationships, but the relationship among non-local snippets is still underexplored. Therefore, we propose a boundary refinement module to enhance salient snippet-features, where exploring the contextual relationship among the salient snippet-features, non-salient snippet-features, and the same video snippet-features via information interaction unit along the channel and temporal dimensions, respectively.

First, we collect the salient snippet-features (b_i = 1) and non-salient snippet-features (b_i = 0) candidates to form $\mathcal{F}^a \in \mathbb{R}^{T^a \times D}$ and $\mathcal{F}^b \in \mathbb{R}^{T^b \times D}$, respectively, where $\mathcal{F}^a \cup \mathcal{F}^b = \mathcal{F}$, $T^a + T^b = T$, T^a denotes the number of salient snippet-features, and T^b denotes the number of non-salient snippet-features. Then, we leverage a channel-level information interaction unit in the squeeze-and-excitation pattern to generate the feature $\hat{\mathcal{F}}^a \in \mathbb{R}^{T^a \times D}$:

$$\hat{\mathcal{F}}^a = \frac{\exp\left(\theta(\mathcal{F}^a)\right)}{\sum_{d=1}^{D} \exp\left(\theta(\mathcal{F}^a_{\cdot,d})\right)} \otimes \mathcal{F}^a + \mathcal{F}^a, \tag{4}$$

where $\otimes$ denotes the element-wise multiplication. θ is a simple multi-layer perceptron, which is consisted of ***FC-ReLU-FC***. We set the weight of the first *FC* to $\mathbf{W}_1 \in \mathbb{R}^{D \times (D/r)}$ and that of the second *FC* to $\mathbf{W}_2 \in \mathbb{R}^{(D/r) \times D}$, and r is a scaling factor. Residual connection is adopted to maintain the stability of training.

Subsequently, we conduct a temporal-level information interaction unit to capture the global contextual relationships between $\hat{\mathcal{F}}^a$ and $\mathcal{F}$ as the following equation:

$$\tilde{\mathcal{F}}^a = \text{softmax}(\mathcal{F} \odot (\hat{\mathcal{F}}^a)^T) \odot \hat{\mathcal{F}}^a, \tag{5}$$

where $\odot$ denotes the matrix multiplication. By integrating such information, we obtain a set of discriminative snippet-features $\tilde{\mathcal{F}}^a \in \mathbb{R}^{T \times D}$.

However, some information contained in $\mathcal{F}^b$ maybe neglected, which contains some action-related or background-related information. Thus, utilizing the information in $\mathcal{F}^b$ can help boost the diversity of snippet-features, and we also utilize the information interaction unit to generate non-salient enhanced features $\tilde{\mathcal{F}}^b$ between $\mathcal{F}^b$ and $\mathcal{F}$ through Eq.(4) and Eq.(5). Note that the parameters in Eq.(4) are not shared between $\mathcal{F}^a$ and $\mathcal{F}^b$.

Finally, we apply a weighted sum operation to balance the contribution between $\tilde{\mathcal{F}}^a$ and $\tilde{\mathcal{F}}^b$ to obtain the enhanced features $\tilde{\mathcal{F}} \in \mathbb{R}^{T \times D}$ as follows:

$$\tilde{\mathcal{F}} = \text{sum}(\tilde{\mathcal{F}}^a, \tilde{\mathcal{F}}^b, \sigma) = \sigma \tilde{\mathcal{F}}^a + (1 - \sigma)\tilde{\mathcal{F}}^b, \tag{6}$$

where σ denotes a trade-off factor.

Discrimination Enhancement Module

Action information from videos of the same category can provide additional clues to help improve the discriminative nature of the snippet-features and the quality of the generated pseudo-labels. Therefore, we design a discrimination enhancement module that utilizes the correlation among videos to make action and background snippet-features more separable.

First, we introduce a memory bank $\mathcal{M} \in \mathbb{R}^{C \times N \times D}$ as the action knowledge base to store the diverse action information from the entire dataset during training, where C denotes the number of classes, N indicates the number of stored snippets of each class, and D is the dimension number. Initially, we use the classification head to predict the scores of the salient snippet-features and select the snippets with the highest N classification scores to initialize the memory $\mathcal{M}$ along with the scores. At t-th training iteration, we select N snippet-features $\mathcal{F}_{[c]}^{(t)}$ with the high scores for each class to update the memory of last iteration $\mathcal{M}_{[c]}^{(t-1)}$. The process can be formulated as:

$$\mathcal{M}_{[c]}^{(t)} \leftarrow (1 - \eta) \cdot \mathcal{M}_{[c]}^{(t-1)} + \eta \cdot \mathcal{F}_{[c]}^{(t)}, \tag{7}$$

where η denotes the momentum coefficient. To boost the robustness, we exploit the momentum update strategy (He et al. 2020) to update memory $\mathcal{M}$, so η is adjusted by:

$$\eta = \eta_0 \cdot \log\left(\exp\left(e/E\right) + 1\right), \tag{8}$$

where η_0 denotes the initial momentum coefficient, e is the current epoch, E denotes the total epoch, and c is the class index of the current snippet. Meanwhile, we use the temporal-level information interaction unit to implement the interaction between the mixed features $\tilde{\mathcal{F}}$ in the boundary refinement module and memory $\mathcal{M}_{[c]}^{(t)}$ to bring the class information of the entire dataset into $\tilde{\mathcal{F}}$, which can be formulated as:

$$\hat{\mathcal{F}} = \text{softmax}(\tilde{\mathcal{F}} \odot (\mathcal{M}_{[c]}^{(t)})^T) \odot \mathcal{M}_{[c]}^{(t)}. \tag{9}$$

Finally, we get the output features $\tilde{\mathcal{F}}$ and $\hat{\mathcal{F}}$ from the boundary refinement module and discrimination enhancement module. Then, we feed them to the classification head to output two TCAMs, *i.e.,* $\tilde{\mathcal{T}}$ and $\hat{\mathcal{T}}$, which are summed after to obtain $\mathcal{T}^p$ as the pseudo labels to supervise the learning of the base branch.

Sup	Method	Feature	mAP@IoU(%)							AVG	AVG	AVG
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.5)	(0.3:0.7)	(0.1:0.7)
Full	S-CNN (Shou, Wang, and Chang 2016)	-	47.7	43.5	36.3	28.7	19.0	10.3	5.3	35.0	19.9	27.3
	SSN (Zhao et al. 2017)	-	66.0	59.4	51.9	41.0	29.8	-	-	49.6	-	-
	TAL-Net (Chao et al. 2018)	-	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
	GTAN (Long et al. 2019)	-	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-	-
Weak*	STAR (Xu et al. 2019)	I3D	68.8	60.0	48.7	34.7	23.0	-	-	47.0	-	-
	3C-Net (Narayan et al. 2019)	I3D	59.1	53.5	44.2	34.1	26.6	-	8.1	43.5	-	-
Weak	STPN (Nguyen et al. 2018)	I3D	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	18.5	27.0
	RPN (Huang et al. 2020)	I3D	62.3	57.0	48.2	37.2	27.9	16.7	8.1	46.5	27.6	36.8
	BaS-Net (Lee, Uh, and Byun 2020)	I3D	58.2	52.3	44.6	36.0	27.0	18.6	10.4	43.6	27.3	35.3
	DGAM (Shi et al. 2020)	I3D	60.0	56.0	46.6	37.5	26.8	17.6	9.0	45.6	27.5	37.0
	TSCN (Zhai et al. 2020)	I3D	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	28.8	37.8
	A2CL-PT (Min and Corso 2020)	I3D	61.2	56.1	48.1	39.0	30.1	19.2	10.6	46.9	29.4	37.8
	UM (Lee et al. 2021)	I3D	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	32.9	41.9
	CoLA (Zhang et al. 2021)	I3D	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	AUMN (Luo et al. 2021)	I3D	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4	41.5
	UGCT (Yang et al. 2021)	I3D	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	D2-Net (Narayan et al. 2021)	I3D	65.7	60.2	52.3	43.4	36.0	-	-	51.5	-	-
	FAC-Net (Huang, Wang, and Li 2021)	I3D	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	33.1	42.2
	DCC (Li et al. 2022)	I3D	69.0	63.8	55.9	45.9	35.7	24.3	13.7	54.1	35.1	44.0
	RSKP (Huang, Wang, and Li 2022)	I3D	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	ASM-Loc (He et al. 2022)	I3D	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	DELU (Chen et al. 2022)	I3D	71.5	66.2	56.5	47.7	40.5	**27.2**	**15.3**	56.5	37.4	46.4
	FBA-Net (Moniruzzaman and Yin 2023)	I3D	71.9	65.8	56.7	48.6	39.3	26.4	14.2	56.5	37.0	46.1
	Ours	I3D	**72.4**	**66.9**	**58.4**	**49.7**	**41.8**	25.5	12.8	**57.8**	**37.6**	**46.8**

Table 1: Comparison with state-of-the-art methods on THUMOS14 dataset. The AVG columns show average mAP under IoU thresholds of 0.1:0.5, 0.3:0.7 and 0.1:0.7. I3D denotes the utilization of the I3D network as the feature extractor, respectively. * indicates the methods use extra information. The best results are highlighted in bold. Sup means supervision manner.

Training loss

Following previous methods, the whole learning process is jointly driven by video-level classification loss $\mathcal{L}_{\rm cls}$, knowledge distillation loss $\mathcal{L}_{\rm kd}$ and attention normalization loss $\mathcal{L}_{\rm att}$ (Zhai et al. 2020). The total loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\rm cls} + \mathcal{L}_{\rm kd} + \lambda \mathcal{L}_{\rm att}, \tag{10}$$

where λ denotes trade-off factors. The knowledge distillation $\mathcal{L}_{\rm kd}$ in (Huang, Wang, and Li 2022) is used to implement the process of $\mathcal{T}^p$ supervising $\mathcal{T}$ for training. The video-level classification loss is the combination of two losses calculated from the CA head and MIL head, which can be formulated as:

$$\mathcal{L}_{\rm cls} = \mathcal{L}_{\rm CA} + \theta \mathcal{L}_{\rm MIL}, \tag{11}$$

where θ is a hyper-parameter. More details about each loss function please refer to the corresponding references.

Experiments

Datasets and Evaluation Metrics.

We conduct our experiments on the two commonly-used benchmark datasets, including THUMOS14 (Jiang et al. 2014) and AcitivityNet v1.3 (Heilbron et al. 2015). Following the general weak-supervised setting, we only use the video-level category labels in the training process.

THUMOS14 includes 200 untrimmed validation videos and 212 untrimmed test videos, where videos are collected from 20 action categories. Following the previous work (Wang et al. 2017; He et al. 2022; Huang, Wang, and Li 2021), we use the validation videos to train our model and test videos for evaluation.

ActivityNet v1.3 contains 10,024 training videos, 4,926 validation videos, and 5,044 testing videos of 200 action categories. Following (Lee et al. 2021; Huang, Wang, and Li 2022), we use the training videos to train our model and validation videos for evaluation.

Evaluation metrics. We evaluate the performance of our method with the standard evaluation metrics: mean average precise (mAP) under different intersection over union (IoU) thresholds. For THUMOS14 dataset, we report the mAP under thresholds IoU=$\{$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7$\}$. For ActivityNet v1.3 dataset, we report the mAP under thresholds [0.5:0.05:0.95]. At the same time, we also calculate the average mAP for different IoU ranges on the two datasets.

Implementation Details

We implement our model with the PyTorch framework and train the model with Adam optimizer (Kingma and Ba 2015). The scaling factor r is set to 4. The hyper-parameter θ and λ are set to 0.2 and 0.1, respectively. The feature is extracted using the I3D (Carreira and Zisserman 2017), which is pre-trained on the Kinetics-400 (Kay et al. 2017) dataset. For THUMOS14 dataset, we train 180 epochs with a learning rate of 0.00005, the batch size is set to 10, σ is set to 0.88, and K is set to $\lfloor 50\% * T \rfloor$, where T is the number of video snippets. For ActivityNet v1.3 dataset, we train 100 epochs with a learning rate of 0.0001, the batch size is set to 32, σ is set to 0.9, and K is set to $\lfloor 90\% * T \rfloor$.

Comparison with State-of-the-Art Methods

THUMOS14. We first compare our method with the state-of-the-art (SOTA) methods on THUMOS14 dataset. These SOTA methods contain fully-supervised methods and weakly-supervised methods, the results are shown in Table 1. We can observe that our proposed model outperforms the SOTA weakly-supervised temporal action localization methods. Our proposed method reaches 46.8 at average mAP for IoU thresholds 0.1:0.7. Meanwhile, our result can reach 41.8 at mAP@0.5. The reasons for the improved performance stem from 1) our method uses the variation relationships between snippet-features to generate salient snippet-features and then considers contextual information to enhance salient snippet-features, thereby improving the discriminative of snippet-features; 2) we introduce additional clues to leverage the relationships between videos, improving the discriminative nature of the action and background snippet-features. Thus, generating more high-fidelity pseudo labels can significantly improve the performance.

ActivityNet v1.3. Table 2 shows the evaluation results in terms of mAP@IoU on ActivityNet v1.3 dataset. From the table, our model achieves competitive performance compared to other SOTA methods. In addition, our method achieves 25.8 for average mAP, which is 0.7 higher than ASM-Loc, demonstrating the superiority of our method.

Method	mAP@IoU(%)			
	0.5	0.75	0.95	AVG
STPN (Nguyen et al. 2018)	29.3	16.9	2.6	16.3
CMCS (Liu, Jiang, and Wang 2019)	34.0	20.9	5.7	21.2
BaS-Net (Lee, Uh, and Byun 2020)	34.5	22.5	4.9	22.2
TSCN (Zhai et al. 2020)	35.3	21.4	5.3	21.7
A2CL-PT (Min and Corso 2020)	36.8	22.0	5.2	22.5
TS-PAC (Liu et al. 2021)	37.4	23.5	5.9	23.7
UGCT (Yang et al. 2021)	39.1	22.4	5.8	23.8
AUMN (Luo et al. 2021)	38.3	23.5	5.2	23.5
FAC-Net (Huang, Wang, and Li 2021)	37.6	24.2	6.0	24.0
DCC (Li et al. 2022)	38.8	24.2	5.7	24.3
RSKP (Huang, Wang, and Li 2022)	40.6	24.6	5.9	25.0
ASM-Loc (He et al. 2022)	**41.0**	24.9	6.2	25.1
Ours	39.4	**25.8**	**6.4**	**25.8**

Table 2: Comparison with state-of-the-art methods on ActivityNet v1.3 dataset. The AVG column shows the averaged mAP under the IoU thresholds [0.5:0.05:0.95].

Method	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
random	68.6	53.1	34.4	11.7	42.3
classification	67.7	53.4	36.9	11.7	43.0
cosine distance	68.6	53.5	36.6	12.3	43.3
L_2 distance	70.7	55.5	36.6	12.3	44.2
L_1 distance (Ours)	**72.4**	**58.4**	**41.8**	**12.8**	**46.8**

Table 3: Ablation studies about different strategies of detecting salient snippet-feature on THUMOS14 dataset.

Ablation Study

We conduct experiments to demonstrate the impact of different components in our method on THUMOS14 dataset.

Impact of Saliency Inference Module. To find a proper function in Eq.(1), we explore several strategies to calculate the difference between each pair of neighbor snippets, including cosine distance, L_1 distance, and L_2 distance, and the results are reported in Table 3. In addition, we explore other ways of generating salient snippet-features, such as random assignment and classification. Among them, random assignment means randomly assigning salient or non-salient labels to each snippet, and classification uses the pre-trained classification head of the base model to classify snippets into salient and non-salient. The results show that L_2 distance can achieve higher mAP than cosine distance, and L_1 distance yields the best results compared to other methods, so we adopt it as the default *diff* function. The reason is that L_1 focuses on the subtle variations between features by computing the absolute differences, which is important for TAL. Whereas cosine distance calculates relative differences and L_2 may suppress these subtle differences by squaring and then taking the square root.

Impact of Different Modules. We evaluate the effect of the boundary refinement module and discrimination enhancement module. The results are presented in Table 4. We set the base branch as *Base* and progressively add the boundary refinement module and discrimination enhancement module to Base, and the performances are continuously improved by 3.4 and 6.3 for average mAP.

Impact of Boundary Refinement Module. We evaluate the impact of different variants of boundary refinement module. The results are reported in Table 5, in which 1) *self* denotes utilize temporal-level information interaction unit to interact information between video snippet-features $\mathcal{F}$ and itself; 2) *w/o salient* and *w/o non-salient* denote removing the salient and non-salient snippet-features, respectively; 3) *salient + non-salient* denotes directly adding the two types of features together; 4) *weighted sum* denotes using weighted sum operation to fuse two types of features; 5) *temporal* denotes enhancing snippet-features only using the temporal-level information interaction unit. We can observe that 1) the salient snippet-features have a significant influence on the performance, removing them will lead to a significant drop in the performance; 2) weighted sum is more effective compared to directly adding, which can assist the utilizing of the infor-

Method	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
Base	62.7	45.5	29.3	10.4	37.1
Base + BRM	66.3	49.7	32.6	11.3	40.5
Base + BRM + DEM	**72.4**	**58.4**	**41.8**	**12.8**	**46.8**

Table 4: The effects of different modules on THUMOS14 dataset. BRM and DEM denote boundary refinement module and discrimination enhancement module, respectively.

Method	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
self	65.8	48.1	30.6	10.9	39.2
w/o salient	49.9	34.8	20.3	5.9	27.6
w/o non-salient	71.7	56.9	39.9	12.4	45.9
salient + non-salient	71.0	54.1	35.0	10.3	43.1
weighted sum	**72.4**	**58.4**	**41.8**	**12.8**	**46.8**
temporal-level	71.4	57.4	39.5	12.8	46.0

Table 5: The effect of different components in boundary refinement module on THUMOS14 dataset.

mation in non-salient snippet-features; 3) information interaction unit at both the channel-level and temporal-level can enhance the discriminative nature of features better.

Impact of Memory Update Strategies. We explore the impact of different memory update strategies in the discrimination enhancement module. The evaluation results are shown in Table 6. We evaluate two variants of memory update strategy, *i.e.*, only using the high-confidence action snippet-features to direct update memory, and only using the momentum update strategy. From the table, we can see that our method obtains better performance than only using the momentum update strategy, because the momentum update strategy will include many noisy features and impair the learning of intra-video relation. The results indicate that our method effectively incorporates more action information compared to the direct update strategy.

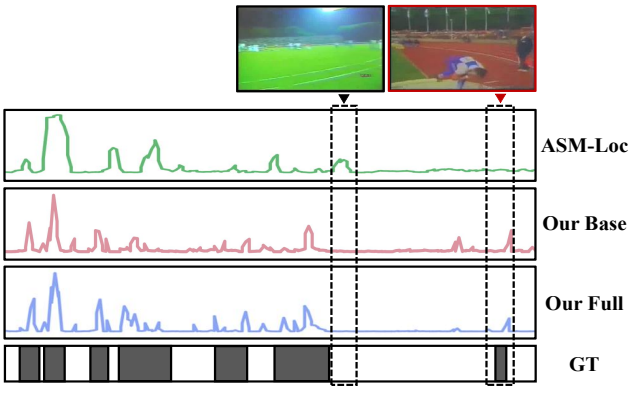

Figure 3: Qualitative comparisons of our method, our Base, and ASM-Loc on "Shotput" on THUMOS14.

Method	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
direct update	71.9	58.1	40.3	12.3	46.4
momentum update	71.0	55.7	37.4	11.3	44.5
Ours	**72.4**	**58.4**	**41.8**	**12.8**	**46.8**

Table 6: The effect of different memory update strategies on THUMOS14 dataset.

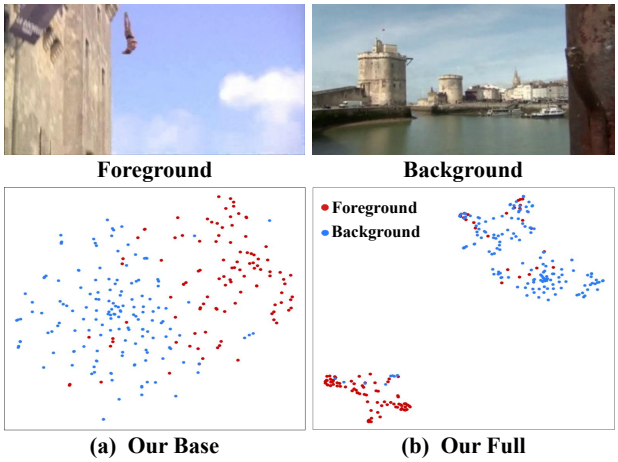

Figure 4: T-SNE visualization of foreground and background features on example "CliffDiving" on THUMOS14.

Qualitative results

To help understand the effect of our proposed method, we present some qualitative results in this subsection. First, we show one case selected from THUMOS14 dataset in Figure 3, and we observe that our method can locate more accurate action and background regions than our Base (base branch) and ASM-Loc. Meanwhile, we adopt t-SNE technology to project the embedding features in THUMOS14 dataset into a 2-dimensional features space, and the results are shown in Figure 4. We observe that our method can accurately bring the embedding features of foregrounds together, and make them away from the background.

Conclusion

In this paper, we propose a novel weakly-supervised TAL method by inferring salient snippet-feature, of which several modules are designed to assist pseudo label generation by exploring the information variation and interaction. Comprehensive experiments demonstrate the effectiveness and superiority of our proposed method.

Acknowledgements

This work was partly supported by the Funds for Innovation Research Group Project of NSFC under Grant 61921003, NSFC Project under Grant 62202063, and 111 Project under Grant B18008.

References

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4724–4733.

Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1130–1139.

Chen, M.; Gao, J.; Yang, S.; and Xu, C. 2022. Dual-Evidential Learning for Weakly-supervised Temporal Action Localization. In *Proceedings of the European Conference on Computer Vision*, 192–208.

He, B.; Yang, X.; Kang, L.; Cheng, Z.; Zhou, X.; and Shrivastava, A. 2022. ASM-Loc: Action-Aware Segment Modeling for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13925–13935.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 961–970.

Huang, L.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Relational Prototypical Network for Weakly Supervised Temporal Action Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11053–11060.

Huang, L.; Wang, L.; and Li, H. 2021. Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision*, 8002–8011.

Huang, L.; Wang, L.; and Li, H. 2022. Weakly Supervised Temporal Action Localization via Representative Snippet Knowledge Propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3272–3281.

Jiang, Y.-G.; Liu, J.; Zamir, A. R.; Toderici, G.; Laptev, I.; Shah, M.; and Sukthankar, R. 2014. THUMOS challenge: Action recognition with a large number of classes. https://www.crcv.ucf.edu/THUMOS14/.

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. *CoRR*, abs/1705.06950.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*.

Lee, P.; Uh, Y.; and Byun, H. 2020. Background Suppression Network for Weakly-Supervised Temporal Action Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11320–11327.

Lee, P.; Wang, J.; Lu, Y.; and Byun, H. 2021. Weakly-supervised Temporal Action Localization by Uncertainty Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1854–1862.

Li, J.; Yang, T.; Ji, W.; Wang, J.; and Cheng, L. 2022. Exploring Denoised Cross-Video Contrast for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19914–19924.

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

Lin, T.; Zhao, X.; and Shou, Z. 2017. Single Shot Temporal Action Detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, 988–996.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Liu, D.; Jiang, T.; and Wang, Y. 2019. Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1298–1307.

Liu, K.; Liu, W.; Gan, C.; Tan, M.; and Ma, H. 2018. T-C3D: Temporal Convolutional 3D Network for Real-time Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 7138–7145.

Liu, X.; Liu, W.; Ma, H.; and Fu, H. 2016. Large-scale Vehicle Re-identification in Urban Surveillance Videos. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 1–6.

Liu, Y.; Chen, J.; Chen, Z.; Deng, B.; Huang, J.; and Zhang, H. 2021. The Blessings of Unlabeled Background in Untrimmed Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6176–6185.

Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; and Mei, T. 2019. Gaussian Temporal Awareness Networks for Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 344–353.

Luo, W.; Zhang, T.; Yang, W.; Liu, J.; Mei, T.; Wu, F.; and Zhang, Y. 2021. Action Unit Memory Network for Weakly Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9969–9979.

Luo, Z.; Guillory, D.; Shi, B.; Ke, W.; Wan, F.; Darrell, T.; and Xu, H. 2020. Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning. In *Proceedings of the European Conference on Computer Vision*, 729–745.

Min, K.; and Corso, J. J. 2020. Adversarial Background-Aware Loss for Weakly-Supervised Temporal Activity Localization. In *Proceedings of the European Conference on Computer Vision*, 283–299.

Moniruzzaman, M.; and Yin, Z. 2023. Collaborative Foreground, Background, and Action Modeling Network for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11): 6939–6951.

Moon, T. 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6): 47–60.

Narayan, S.; Cholakkal, H.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. D2-Net: Weakly-Supervised Action Localization via Discriminative Embeddings and Denoised Activations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13608–13617.

Narayan, S.; Cholakkal, H.; Khan, F. S.; and Shao, L. 2019. 3C-Net: Category Count and Center Loss for Weakly-Supervised Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision*, 8679–8687.

Nguyen, P.; Han, B.; Liu, T.; and Prasad, G. 2018. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6752–6761.

Pardo, A.; Alwassel, H.; Heilbron, F. C.; Thabet, A.; and Ghanem, B. 2021. RefineLoc: Iterative Refinement for Weakly-Supervised Action Localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3319–3328.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-TALC: Weakly-supervised Temporal Activity Localization and Classification. In *Proceedings of the European Conference on Computer Vision*, 563–579.

Qi, M.; Li, W.; Yang, Z.; Wang, Y.; and Luo, J. 2019. Attentive Relational Networks for Mapping Images to Scene Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3957–3966.

Qi, M.; Qin, J.; Li, A.; Wang, Y.; Luo, J.; and Van Gool, L. 2018. stagNet: An Attentive Semantic RNN for Group Activity Recognition. In *Proceedings of the European Conference on Computer Vision*, 104–120.

Qi, M.; Qin, J.; Yang, Y.; Wang, Y.; and Luo, J. 2021. Semantics-Aware Spatial-Temporal Binaries for Cross-Modal Video Retrieval. *IEEE Transactions on Image Processing*, 30: 2989–3004.

Qi, M.; Wang, Y.; Li, A.; and Luo, J. 2020. STC-GAN: Spatio-Temporally Coupled Generative Adversarial Networks for Predictive Scene Parsing. *IEEE Transactions on Image Processing*, 29: 5420–5430.

Qu, S.; Chen, G.; Li, Z.; Zhang, L.; Lu, F.; and Knoll, A. 2021. ACM-Net: Action Context Modeling Network for Weakly-Supervised Temporal Action Localization. *arXiv preprint arXiv:2104.02967*.

Shi, B.; Dai, Q.; Mu, Y.; and Wang, J. 2020. Weakly-Supervised Action Localization by Generative Attention Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1009–1019.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1049–1058.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. UntrimmedNets for Weakly Supervised Action Recognition and Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4325–4334.

Xu, Y.; Zhang, C.; Cheng, Z.; Jianwen, X.; Niu, Y.; Pu, S.; and Wu, F. 2019. Segregated Temporal Assembly Recurrent Networks for Weakly Supervised Multiple Action Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9070–9078.

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

Yang, W.; Zhang, T.; Yu, X.; Qi, T.; Zhang, Y.; and FengWu. 2021. Uncertainty Guided Collaborative Training for Weakly Supervised Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 53–63.

Zeng, R.; Huang, W.; Gan, C.; Tan, M.; Rong, Y.; Zhao, P.; and Huang, J. 2019. Graph Convolutional Networks for Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision*, 7094–7103.

Zhai, Y.; Wang, L.; Tang, W.; Zhang, Q.; Yuan, J.; and Hua, G. 2020. Two-Stream Consensus Network for Weakly-Supervised Temporal ActionLocalization. In *Proceedings of the European Conference on Computer Vision*, 37–54.

Zhang, C.; Cao, M.; Yang, D.; Chen, J.; and Zou, Y. 2021. CoLA: Weakly-Supervised Temporal Action Localization With Snippet Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16010–16019.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal Action Detection with Structured Segment Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2914–2923.