2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-9346-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACV56688.2023.00597

Temporal Feature Enhancement Dilated Convolution Network for Weakly-supervised Temporal Action Localization

Jianxiong Zhou, and Ying Wu
Department of Electrical and Computer Engineering, Northwestern University
jianxiongzhou2026@u.northwestern.edu, yingwu@northwestern.edu

Abstract

Weakly-supervised Temporal Action Localization (WTAL) aims to classify and localize action instances in untrimmed videos with only video-level labels. Existing methods typically use snippet-level RGB and optical flow features extracted from pre-trained extractors directly. Because of two limitations: the short temporal span of snippets and the inappropriate initial features, these WTAL methods suffer from the lack of effective use of temporal information and have limited performance. In this paper, we propose the Temporal Feature Enhancement Dilated Convolution Network (TFE-DCN) to address these two limitations. The proposed TFE-DCN has an enlarged receptive field that covers a long temporal span to observe the full dynamics of action instances, which makes it powerful to capture temporal dependencies between snippets. Furthermore, we propose the Modality Enhancement Module that can enhance RGB features with the help of enhanced optical flow features, making the overall features appropriate for the WTAL task. Experiments conducted on THUMOS'14 and ActivityNet v1.3 datasets show that our proposed approach far outperforms state-of-the-art WTAL methods.

1. Introduction

Temporal action localization (TAL), which is one of the main tasks of video understanding, aims at localizing the start and end timestamps of action instances in an untrimmed video and classifying them. It has been used in various video understanding applications, such as intelligent surveillance analysis [34] and video retrieval [9]. Many works [32, 20, 42, 2, 24, 39] have put their effort into fully-supervised temporal action localization and achieved great localization results. However, fully-supervised methods require a huge amount of fine-grained frame-level annotations, which need manual labeling and have annotation bias of annotators. To address this issue, weakly-supervised temporal action localization (WTAL), which only requires easily collected video-level categorical labels, has gained intensive attention [36, 28, 21, 16, 40, 41, 8] in recent years.

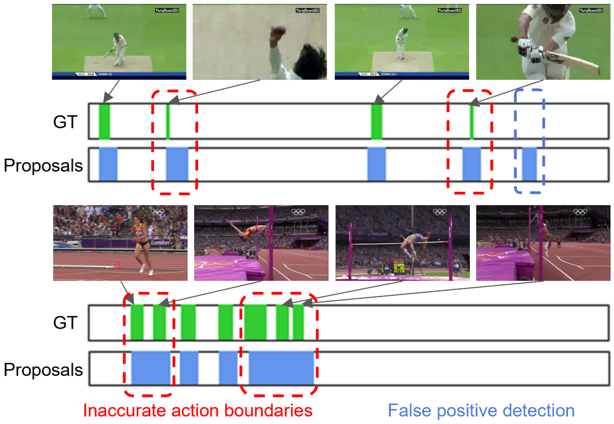

Figure 1. The results of the previous method (BaS-Net [16]) with inaccurate action boundaries and false positive detection.

Though WTAL simplifies the data collection process, it is challenging to do temporal action localization with only video-level annotations, especially for complex action scenes. To tackle the issue, many WTAL methods adopt the multiple-instance learning (MIL) framework [36, 28, 31, 33, 21, 29, 16]. These methods uniformly sample the video into snippets and then generate the Temporal Class Activation Sequence (TCAS), which is the sequence of categorical probabilities over action classes for each snippet. Finally, the top-k mean strategy is used to aggregate TCAS to obtain the final video-level prediction.

While previous methods have achieved significant improvement on WTAL, the performance is still limited. One major problem is the inaccurate predictions of action boundaries. Fig. 1 demonstrates examples of some errors. Though it is challenging to obtain accurate action boundaries with only video-level annotations, we argue that the insufficient use of temporal information is a key reason for the limited results. A complete action instance usually covers a relatively long temporal span, while a snippet is un-

2642-9381/23/$31.00 ©2023 IEEE
DOI 10.1109/WACV56688.2023.00597

able to observe the full dynamics of that action instance. Another reason is that most WTAL methods directly use the RGB and optical flow features extracted by pre-trained models, e.g., I3D [1], which are customized and trained for trimmed video action classification rather than WTAL. Thus, enhancing features with temporal information is a feasible approach to address this issue.

In this paper, we propose the Temporal Feature Enhancement Dilated Convolution Network (TFE-DCN) to address two aforementioned limitations. Inspired by the successful application of temporal convolution network (TCN) on fully-supervised temporal action localization [4], we design a novel Temporal Feature Enhancement Dilated Convolution Module (TFE-DC) with several advantages. First, it enlarges the receptive field, enabling the model to obtain temporal information of complete action instances and eliminating incoherence of temporal information caused by the short temporal span of snippets. Second, it can capture temporal dependencies between snippets in the receptive field, facilitating a snippet to exploit motion clues from other snippets across the entire receptive field to enhance its feature representation, which is powerful for enhancing features and separating action instances from the background.

Though TFE-DC Module extracts temporal information and enhances optical flow features, it is notable that initial RGB features are not enhanced. The inconsistency between the two modalities results in the degradation of performance. Therefore, we propose the Modality Enhancement Module that can enhance RGB features with the help of enhanced optical flow features. In this module, initial RGB features and enhanced optical flow features are fed into a sharing convolution layer to obtain two attention sequences respectively. Then we perform element-wise multiplication on these two attention sequences and initial RGB features to obtain the enhanced RGB features. The Modality Enhancement Module keeps the consistency between the two modalities and also introduces improved optical flow features to enhance RGB features.

Our main contributions are summarized as three-fold:

- We show that TFE-DCN can effectively use temporal features and has accurate predictions of action boundaries. The proposed TFE-DC Module has a novel dilated structure that reflects the influence of temporal information at different receptive field scales on final attention weights, rather than following the common dilated residual layer as MS-TCN [4].
- We propose a Modality Enhancement Module that keeps the consistency between two modalities and recalibrates initial RGB features with the help of enhanced optical flow features, making them more appropriate for the WTAL task.
- Extensive experiments are conducted on THUMOS'14 and ActivityNet v1.3 to demonstrate the effectiveness of our proposed method. Our TFE-DCN outperforms all state-of-the-art WTAL methods.

2. Related Work

Temporal Convolution Network. Temporal Convolution Network is successfully applied in speech synthesis [35] and introduced to temporal action localization by some works [14, 18, 4]. Lea et al. [14] propose an encoder-decoder framework for action segmentation and detection. TDRN [18] uses a residual stream to analyze video information at full temporal resolution. MS-TCN [4] uses dilated convolution residual layer instead of temporal pooling to capture long-range dependencies and gets better results.

Fully-supervised Temporal Action Localization. Fully-supervised TAL requires frame-level annotations of action instances. Most methods [3, 32, 42, 20] generate temporal action proposals and then do classification based on these proposals. CDC [32] performs temporal upsampling and spatial downsampling simultaneously to predict frame-level action proposals. BSN [20] locates temporal boundaries with high probability and then combines these boundaries into proposals. P-GCN [39] uses graph convolution networks to exploit the relation between proposals.

Weakly-supervised Temporal Action Localization. Though some methods [26, 15] use point-level labels, WTAL usually requires only video-level annotations and greatly reduces the workload of labeling. UntrimmedNets [36] formally proposes the WTAL task and tries to address it with Multi-Instance Learning (MIL) method. Sparse Temporal Pooling Network (STPN) [28] introduces an attention mechanism with a proposed sparsity constraint. W-TALC [31] designs a co-activity similarity loss and uses deep metric learning to train the network. However, these early works cannot effectively distinguish action instances and backgrounds and fail to localize the complete action. To tackle the issue, many works [21, 29, 16, 12, 17, 41, 27] improve the attention mechanism to suppress the activation scores of backgrounds and highlight the activation scores of action. BaS-Net [16] introduces an auxiliary class for background and uses a filtering module to suppress the activation of background. Liu et al. [21] develop a parallel multi-branch classification framework to model the complete action. HAM-Net [12] uses a hybrid attention mechanism to localize complete action instances. CoLA [41] utilizes snippet contrastive learning to improve localization results.

Recently, CO_2-Net [8] and ACGNet [38] have all focused on enhancing features for WTAL. CO_2-Net uses a cross-modal consensus module to reduce task-irrelevant information redundancy and make features appropriate for WTAL. ACGNet uses a graph convolutional network to enhance the discriminability of action representations, making

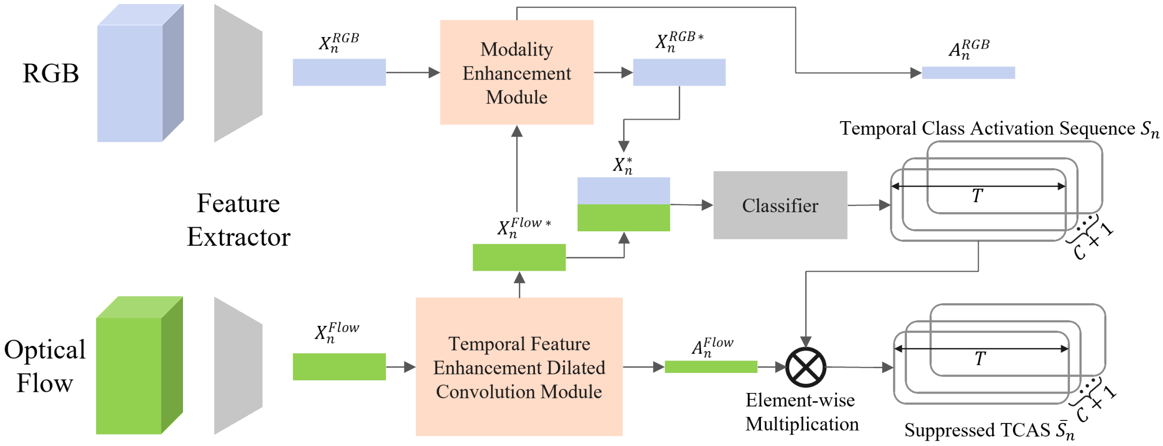
RGB
Feature Extractor
Optical Flow
X_n^{RGB}
Modality Enhancement Module
X_n^{RGB*}
A_n^{RGB}
Temporal Class Activation Sequence S_n
X_n^*
Classifier
T
$C+1$
X_n^{Flow*}
X_n^{Flow}
Temporal Feature Enhancement Dilated Convolution Module
A_n^{Flow}
Element-wise Multiplication
T
$C+1$
Suppressed TCAS $\bar{S}_n$

Figure 2. An overview of the proposed Temporal Feature Enhancement Dilated Convolution Network (TFE-DCN), which consists of four parts: (1) pre-trained feature extractor that outputs RGB features X_n^{RGB} and optical flow features X_n^{Flow}; (2) Temporal Feature Enhancement Dilated Convolution Module (TFE-DC Module) that generates enhanced optical flow features X_n^{Flow*} and temporal attention weights A_n^{Flow}; (3) Modality Enhancement Module that generates enhanced RGB features X_n^{RGB*} and spatial attention weights A_n^{RGB}; (4) classifier and element-wise multiplication that generate the Temporal Class Activation Sequence (TCAS) S_n and suppressed TCAS $\bar{S}_n$.

it easier to classify hard examples based on enhanced features. Our method is distinct from CO$_2$-Net and ACGNet in two main aspects. (1) TFE-DCN can effectively use temporal information to enhance temporal features and then enhance RGB features with the enhanced temporal features. While CO$_2$-Net does not emphasize temporal information and treats two modality features equally. (2) TFE-DCN uses multi-layer dilated convolutions to capture temporal dependencies between snippets. While ACGNet uses the temporal diffusion graph to obtain temporal dependencies across snippets. Our model achieves much better performance in experiments.

3. Method

In this section, we first present the problem formulation of weakly-supervised temporal action localization (WTAL) and then describe the structure overview of our proposed TFE-DCN. The overall architecture is shown in Fig. 2. The details of the two modules are demonstrated in Section 3.3 and Section 3.4. Finally, we illustrate loss functions and action localization.

3.1. Problem Formulation

Assume we are given a set of N untrimmed videos $\{v_n\}_{n=1}^N$ and the video-level categorical labels $\{y_n\}_{n=1}^N$, where $y_n \in R^C$ is a normalized multi-hot vector and C is the number of action categories. The goal of WTAL is to generate classification and temporal localization results of all action instances as action proposals (t_s, t_e, c, ϕ) for each video, where t_s, t_e, c and ϕ denote the start time, the end time, the predicted action category and the confidence score of the action proposal, respectively.

3.2. Method Overview

3.2.1 Feature Extractor

Following the common practice [28, 16], we first divide each video v_n into 16-frame non-overlapping snippets and sample a fixed number of T snippets to represent the video. The RGB features $X_n^{RGB} = \{x_{n,i}^{RGB}\}_{i=1}^T$ and the optical flow features $X_n^{Flow} = \{x_{n,i}^{Flow}\}_{i=1}^T$ are extracted from the sampled RGB snippets and optical flow snippets respectively with the pre-trained feature extractor, i.e., I3D [1]. $x_{n,i}^{RGB}, x_{n,i}^{Flow} \in R^D$ are features of the i-th RGB snippet and optical flow snippet, and D is the feature dimension.

3.2.2 Structure Overview

The overall framework of our proposed TFE-DCN is demonstrated in Fig. 2. The essential parts of the framework are the Temporal Feature Enhancement Dilated Convolution Module (TFE-DC Module) and Modality Enhancement Module. The TFE-DC Module aims to effectively utilize temporal information and enhance optical flow features. The input of this module is optical flow features X_n^{Flow} and the outputs are enhanced optical flow features X_n^{Flow*} and temporal attention weights $A_n^{Flow} \in R^T$. The Modal-ity Enhancement Module aims to enhance the RGB fea-tures X_n^{RGB} with the help of enhanced optical flow features X_n^{Flow*}. The inputs are X_n^{RGB} and X_n^{Flow*}, and the outputs are enhanced RGB features X_n^{RGB*} and spatial attention weights $A_n^{RGB} \in R^T$. Then X_n^{RGB*} and X_n^{Flow*} are concatenated to obtain $X_n^* \in R^{2D \times T}$.

Given the concatenated features X_n^*, we apply a classi-

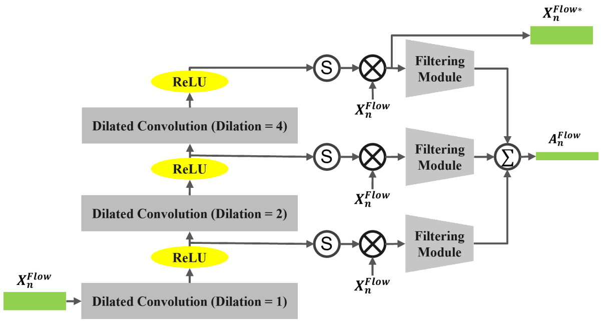

Figure 3. An overview of the proposed Temporal Feature Enhancement Dilated Convolution Module (TFE-DC Module). The module contains a K-layer dilated convolution network ($K = 3$ in this figure) to enlarge the receptive field and capture dependencies between snippets with different temporal scales. It also has an attention weights generation mechanism that averages the attention weights obtained from the outputs of each layer. This allows the final attention weights A_n^{Flow} can cover temporal information of receptive fields with different sizes.

fier to obtain the TCAS S_n.

$$S_n = f_{cls}(X_n^{*}), \tag{1}$$

where f_{cls} is the classifier and $S_n \in R^{(C+1)\times T}$ has $C + 1$ dimensions since we follow the BaS-Net [16] and set one auxiliary class for the background. Then we use temporal attention weights A_n^{Flow} to suppress the activation of backgrounds in S_n and obtain the suppressed TCAS $\bar{S}_n$:

$$\bar{S}_n = A_n^{Flow} \otimes S_n, \tag{2}$$

where $\otimes$ denotes element-wise multiplication over temporal dimension.

3.3. Temporal Feature Enhancement Dilated Convolution Module

In this work, we apply the TFE-DC Module to effectively use temporal information and enhance optical flow features to make them more appropriate for the WTAL task. Multi-layer dilated convolution network can enlarge the receptive field and capture long-range dependencies between snippets. These properties are conducive to the model to fully learn the temporal features. Besides, a complete action instance usually spans a relatively long temporal window, while an optical flow snippet only covers 16 frames and is insufficient to observe the full action instance. The TFE-DC Module can enlarge the receptive field to cover the temporal span of complete action instances and observe the full dynamics of that action, which is the embodiment that can make full use of temporal information.

As shown in Fig. 3, this module mainly consists of a K-layer dilated convolution network and an attention weights generation mechanism. In K-layer dilated convolutions, we feed the optical flow features $X_n^{Flow} \in R^{D\times T}$ into the first layer $f_{dilated,1}$ and the dilation value is 1. Then outputs go through a ReLU layer and the intermediate results $M_{n,1}$ are obtained. For the k-th layer $f_{dilated,k}$, the process is formulated as below:

$$\begin{aligned} M_{n,k} &= \mathrm{ReLU}(f_{dilated,k}(M_{n,k-1}, 2^{k-1})), \\ k &= 1, \ldots, K, M_{n,0} = X_n^{Flow}, \end{aligned} \tag{3}$$

where $M_{n,k} \in R^{D\times T}$ is the output of the k-th dilated convolution layer and 2^{k-1} is the dilation value. The receptive field expands to $2^k + 1$ snippets for the k-th layer. Finally, we apply the sigmoid function on $M_{n,K}$, use the outputs to enhance optical flow features, and obtain enhanced optical flow features X_n^{Flow*} as below:

$$X_n^{Flow*} = \sigma(M_{n,K}) \otimes X_n^{Flow}, \tag{4}$$

where $M_{n,K}$ is the final output of the K-layer dilated convolution network, σ is the sigmoid function and $\otimes$ denotes element-wise multiplication.

For attention weights generation, we apply the sigmoid function and element-wise multiplication on each $M_{n,k}$ and use the filtering module $f_{att,k}$ to generate attention weights $A_{n,k}^{Flow} \in R^T$. The filtering module consists of three temporal 1D convolutional layers followed by a sigmoid function. The temporal attention weights A_n^{Flow} are the weighted average of $\{A_{n,k}^{Flow}\}_{k=1}^{K}$. The process is formulated as below:

$$A_{n,k}^{Flow} = f_{att,k}(\sigma(M_{n,k}) \otimes X_n^{Flow}), k = 1, \ldots, K, \tag{5}$$

$$A_n^{Flow} = \sum_{k=1}^{K} a_k A_{n,k}^{Flow}, \tag{6}$$

where $a_k > 0, k = 1, \ldots, K$ are weights and $\sum_{k=1}^{K} a_k = 1$.

3.4. Modality Enhancement Module

After obtaining enhanced optical flow features X_n^{Flow*} and temporal attention weights A_n^{Flow}, the next step is to enhance RGB features X_n^{RGB}. Inspired by the Cross-modal Consensus Module [8], we propose the Modality Enhancement Module that enhances RGB features with the help of enhanced optical flow features. The main difference is that we use a sharing convolution layer to make weights distributions of two modalities more approached. This step does improve performance and is different from existing channel attention methods.

As shown in Fig. 4, we input RGB features X_n^{RGB} and enhanced optical flow features X_n^{Flow*} into a sharing convolution layer and then apply the sigmoid function on the outputs of the convolution layer to obtain two weights. Then

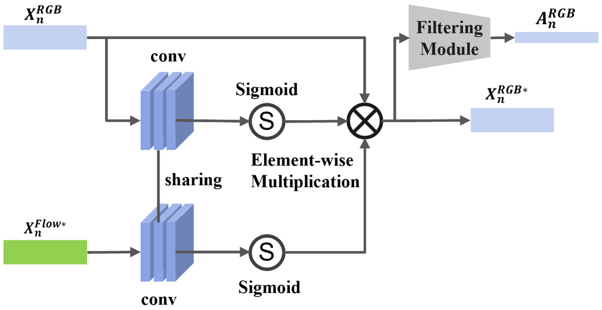

Figure 4. An overview of the proposed Modality Enhancement Module. This module aims to enhance RGB features X_n^{RGB} with the help of enhanced optical flow features X_n^{Flow*}. The sharing convolution layer is beneficial to make weights distributions of the two modalities approached. The enhanced RGB features X_n^{RGB*} are fed into the filtering module to obtain spatial attention weights A_n^{RGB}.

we use these two weights to enhance initial RGB features. The process is expressed as below:

$$X_n^{RGB*} = X_n^{RGB} \otimes \sigma(f_{conv}(X_n^{RGB})) \otimes \sigma(f_{conv}(X_n^{Flow*})), \tag{7}$$

where X_n^{RGB*} is the enhanced RGB features, f_{conv} is the sharing convolution layer, σ is the sigmoid function and $\otimes$ is the element-wise multiplication.

After obtaining X_n^{RGB*}, we feed it into the filtering module to obtain the spatial attention weights A_n^{RGB}:

$$A_n^{RGB} = f_{att}(X_n^{RGB*}), \tag{8}$$

where f_{att} is the filtering module that consists of three temporal 1D convolutional layers followed by the sigmoid function. It is notable that we do not use A_n^{RGB} to suppress background snippets (as shown in Fig. 2).

3.5. Loss Functions

To optimize our proposed TFE-DCN framework, we first apply the loss function of BaS-Net [16], which is expressed as:

$$L_{BaS} = L_{base} + L_{supp} + \lambda_1 L_{norm}, \tag{9}$$

where L_{base} and L_{supp} are the top-k multiple-instance learning loss for TCAS S_n and the suppressed TCAS $\bar{S}_n$ respectively and λ_1 is a hyper-parameter. The normalization loss L_{norm} is to make the attention weights sparse:

$$L_{norm} = \frac{1}{2}(\|A_n^{Flow}\|_1 + \|A_n^{RGB}\|_1), \tag{10}$$

where $\|\cdot\|_1$ is the L1-norm function.

To optimize temporal attention weights A_n^{Flow} and spatial attention weights A_n^{RGB}, we apply the L_{guide} [12] to guide the background class activation, which is the last column of TCAS S_n, to follow the opposite of attention weights A_n^{Flow} and A_n^{RGB}:

$$L_{guide} = \sum_{t=1}^{T} [\,|1 - A_n^{Flow}(t) - s_{C+1}(t)| + |1 - A_n^{RGB}(t) - s_{C+1}(t)|\,], \tag{11}$$

where $A_n^{Flow}(t)$, $A_n^{RGB}(t)$ and $s_{C+1}((t)$ are the t-th element of A_n^{Flow}, A_n^{RGB} and background class activation respectively. We also apply mutual learning loss L_{ml} [8] to set A_n^{Flow} and A_n^{RGB} as pseudo-labels of each other to do mutual learning between two modalities.

By aggregating all the above objective functions, we train our proposed TFE-DCN on the final objective function:

$$L = L_{base} + L_{supp} + \lambda_1 L_{norm} + \lambda_2 L_{guide} + \lambda_3 L_{ml}, \tag{12}$$

where λ_1, λ_2 and λ_3 are all hyper-parameters. In experiments, we set $\lambda_1 = 1, \lambda_2 = 1$ and $\lambda_3 = 0.8$ by default.

3.6. Action Localization

Following BaS-Net[16], We first use the top-k strategy on TCAS S_n to obtain top-k scores and calculate video-level categorical probabilities. Then we threshold the activation scores with θ_a to predict action categories in the video. Temporal attention weights A_n^{Flow} are used to discard the background snippets, and the consecutive segments of the remaining snippets become candidate action proposals, i.e., (t_s, t_e, c, ϕ). Then we use suppressed TCAS $\bar{S}_n$ to calculate the confidence score ϕ for each proposal with the Outer-Inner-Contrastive method [33]. Finally, Non-Maximum Suppression (NMS) is used to remove the overlapping proposals.

4. Experiments

4.1. Experiments Setting

Dataset. We conduct experiments on two popular WTAL benchmarks: THUMOS'14 [13] and ActivityNet v1.3 [7]. THUMOS'14 is a widely used benchmark for the WTAL task. It contains 200 validation videos and 213 test videos of 20 sports categories. Following previous works [40, 16, 38], we use 200 validation videos to train our framework and use 213 test videos for evaluation.

ActivityNet v1.3 has 10024 training videos, 4926 validation videos, and 5044 testing videos from 200 action categories. Since annotations for the testing set are not released, we train on the training set and test on the validation set.

Evaluation Metrics. Following the standard evaluation metrics, we evaluate our method with mean Average Precision (mAP) under different Intersection-over-Union (IoU) thresholds. We adopt the official evaluation code provided by ActivityNet to evaluate our method on both datasets.

Supervision (Feature)	Method	Publication	mAP@IoU (%)							AVG	
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.5	0.1:0.7
Fully (-)	SSN [42]	ICCV'17	60.3	56.2	50.6	40.8	29.1	-	-	47.4	-
	TAL-Net [2]	CVPR'18	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	45.1
	GTAN [24]	CVPR'19	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-
	P-GCN [39]	ICCV'19	69.5	67.5	63.6	57.8	49.1	-	-	61.5	-
Weakly (UNT)	Liu et al. [21]	CVPR'19	53.5	46.8	37.5	29.1	19.9	12.3	6.0	37.4	29.3
	BaS-Net [16]	AAAI'20	56.2	50.3	42.8	34.7	25.1	17.1	9.3	41.8	33.6
	TSCN [40]	ECCV'20	58.9		45.0	36.6	27.6	18.8	10.2	44.2	35.7
Weakly (I3D)	Lee et al. [17]	52.9 AAAI'21		67.5	52.3	43.4	33.7	22.9	12.1	51.6	41.9
	CoLA [41]	61.2 CVPR'21		66.2	51.5	41.9	32.2	22.0	13.1	50.3	40.9
	AUMN [25]	59.5 CVPR'21		66.2	54.9	44.4	33.3	20.5	9.0	52.1	41.5
	TS-PCA [22]	61.9 CVPR'21		67.6	53.4	43.4	34.3	24.7	13.7	52.0	42.6
	UGCT [37]	61.1 CVPR'21		69.2	55.5	46.5	35.9	23.8	11.4	54.0	43.6
	FAC-Net [10]	62.9 ICCV'21		67.6	52.6	44.3	33.4	22.5	12.7	52.0	42.2
	CO$_2$-Net [8]	62.1			54.5	45.7	38.3	26.4	13.4	54.4	44.6
	ACGNET [38]	MM'21	70.1	63.6	53.1	44.6	34.7	22.6	12.0	52.6	42.5
	FTCL [5]	AAAI'22	68.1	62.6	55.2	45.2	35.6	23.7	12.2	53.8	43.6
	DCC [19]	CVPR'22	69.6	63.4	55.9	45.9	35.7	24.3	13.7	54.1	44.0
	Huang et al. [11]	CVPR'22	69.0	63.8	55.8	47.5	38.2	25.4	12.5	55.6	45.1
	ASM-Loc [6]	CVPR'22	71.3	65.3	57.1	46.8	36.6	25.2	13.4	55.4	45.1
	TFE-DCN	CVPR'22	71.2	65.5	58.6	49.5	40.7	27.1	13.7	57.5	46.9
		WACV'23	72.3	66.5							

Table 1. Comparisons of our method with state-of-the-art fully-supervised and weakly-supervised TAL methods on the THUMOS'14 testing set. UNT and I3D are abbreviations for UntrimmedNet features and I3D features, respectively. AVG is the average mAP at multiple IoU thresholds, i.e., 0.1:0.1:0.5 and 0.1:0.1:0.7.

Implementation Details. Our proposed TFE-DCN is implemented in PyTorch [30]. We use the I3D network [1] pre-trained on Kinetics [1] to extract both RGB and optical flow features. The extractor is not fine-tuned for fair comparisons. Video snippets are sampled every 16 frames and the feature dimension of each snippet is 1024. During the training, we set the sampling number T to be 320 for THUMOS'14 and 75 for ActivityNet v1.3. All filtering modules that generate attention weights consist of three temporal 1D convolution layers followed by the sigmoid function. The classifier consists of two temporal 1D convolution layers. For TFE-DC Module, we set the number of dilated convolution layers K to be 3.

For optimization, we used Adam optimizer with a learning rate of 5e-4 for both datasets. Training epochs are set to 3000 for THUMOS'14 and 25000 for ActivityNet v1.3. The batch size is set to 10 and 16 for THUMOS'14 and ActivityNet v1.3, respectively. For hyper-parameters, $a_k = \frac{1}{3}(k = 1, 2, 3)$ in TFE-DC Module.

4.2. Comparison with State-of-the-art Methods

In Table 1, we compare our TFE-DCN with state-of-the-art WTAL methods and several fully-supervised methods on THUMOS'14. We observe that our method far outperforms all previous WTAL methods at all IoU thresholds. Especially on the key criterion AVG 0.1:0.5, our method surpasses the state-of-the-art method [11] by 1.9%. When compared with fully-supervised methods, TFE-DCN outperforms SSN [42] and TAL-Net [2] and achieves comparable results with GTAN [24] and P-GCN [39] at low IoU thresholds. The results demonstrate the superior performance of our approach.

We also conduct experiments on ActivityNet v1.3 and the comparison results are shown in Table 2. Our method outperforms all the state-of-the-art WTAL methods and achieves the performance of 25.3% average mAP on ActivityNet v1.3.

4.3. Ablation Study and Analysis

In this work, we propose a TFE-DC Module that enlarges the receptive field and captures temporal dependencies between snippets, and a Modality Enhancement Module to recalibrate initial RGB features with the help of enhanced optical flow features. Also, the final objective function consists of several components. Therefore, we first verify the effectiveness of each component. Then we analyze the efficacy of each module in TFE-DCN. All ablation studies are conducted on the THUMOS'14 testing set.

Ablation Study on Final Objective Function. In Table 3, we conduct an ablation study to investigate the contribution of each component in the final objective function (Eq.12). We do not test L_{base} and L_{supp} because they

Method	mAP@IoU (%)			
	0.5	0.75	0.95	AVG
BaS-Net [16], AAAI'20	34.5	22.5	4.9	22.2
TSCN [40], ECCV'20	35.3	21.4	5.3	21.7
ACSNet [23], AAAI'21	36.3	24.2	5.8	23.9
AUMN [25], CVPR'21	38.3	23.5	5.2	23.5
TS-PCA [22], CVPR'21	37.4	23.5	5.9	23.7
UGCT [37], CVPR'21	39.1	22.4	5.8	23.8
FAC-Net [10], ICCV'21	37.6	24.2	6.0	24.0
FTCL [5], CVPR'22	40.0	24.3	6.4	24.8
DCC [19], CVPR'22	38.8	24.2	5.7	24.3
Huang et al. [11], CVPR'22	40.6	24.6	5.9	25.0
ASM-Loc [6], CVPR'22	41.0	24.9	6.2	25.1
TFE-DCN, WACV'23	41.4	24.8	6.4	25.3

Table 2. Comparison of our method with state-of-the-art WTAL methods on the ActivityNet v1.3 validation set. AVG is the average mAP at the IoU threshold 0.5:0.05:0.95.

Exp	L_{base}	L_{supp}	L_{norm}	L_{guide}	L_{ml}	AVG
1	✓	✓			29.5 2	✓
✓	✓			36.6 3	✓	✓
✓			44.1 4	✓		✓
✓	41.6 5	✓	✓		✓	✓
46.5 6	✓	✓	✓		✓	43.6 7
✓	✓		✓ ✓	44.3 8	✓	✓
✓	✓	✓	46.9			

Table 3. Ablation studies of different components of final loss function on the THUMOS'14 testing set. AVG is the average mAP at the IoU threshold 0.1:0.1:0.7.

are basic objective functions of the framework and should not be removed. We observe that L_{norm}, L_{guide}, and L_{ml} all contribute to the final performance. Among them, L_{guide} largely enhances the performance since it encourages the background class activation to be opposite of attention weights A_n^{Flow} and A_n^{RGB}, and therefore improves the action-background separation [12].

Ablation Study on TFE-DC Module. The TFE-DC Module is to generate enhanced optical flow features and temporal attention weights. Its key component is the K-layer dilated convolution network, which enlarges the receptive field and captures the temporal dependencies between snippets. However, if the receptive field is too large, it may cover too many irrelevant background snippets, resulting in performance degradation. To verify the effectiveness of the TFE-DC Module with different numbers of dilated convolution layers, we conducted related ablation studies.

Table 4 lists the detailed performance comparison among the model with different numbers of dilated convolution layers. Here $K = 0$ means the module outputs the initial optical flow features without any enhancement and directly feeds initial features into the filtering module to obtain temporal attention weights. The results show that performance first increases with the number of dilated convolution layers and then decreases. The best average performance is achieved when $K = 3$. This is because when $K = 3$, the receptive field covers 9 snippets. Since each snippet contains 16 frames and the frame rate of samples is 25, the receptive field covers temporal information within $\frac{9\times16}{25} = 5.76$ sec. The average duration of all action instances in the THUMOS'14 testing set is about 4.49 sec. If K is lower than 3, the receptive field cannot completely cover the temporal span of most action instances. If K is higher than 3, the receptive field may cover too many background snippets and reduce the impact of action instance snippets. This trade-off of covering complete action instances while reducing background snippets makes $K = 3$ the optimal value. The variation trend presented in Table 4 demonstrates the effectiveness of our TFE-DC Module.

K	mAP@IoU (%)				AVG	
	0.1	0.3	0.5	0.7	0.1:0.5	0.1:0.7
0	70.2	55.0	38.1	14.4	54.6	44.8
1	71.3	56.4	38.4	12.9	55.6	45.3
2	71.6	57.3	39.2	13.1	56.3	45.8
3	72.3	58.6	40.7	13.7	57.5	46.9
4	71.7	57.2	38.8	13.1	56.1	45.7

Table 4. Ablation studies of our model with different numbers of dilated convolution layers K on the THUMOS'14 testing set.

Ablation Study on Modality Enhancement Module. In our proposed Modality Enhancement Module, RGB features are enhanced with the help of enhanced optical flow features. As shown in Fig. 4, we utilize a sharing convolution layer on initial RGB features X_n^{RGB} and enhanced optical flow features $X_n^{Flow'}$ to generate two weights. Then we enhance initial RGB features X_n^{RGB} with these two weights by element-wise multiplication. To verify the efficacy of our Modality Enhancement Module, we evaluate the different kinds of modality combinations.

Table 5 lists the performance comparison between models with different kinds of modality combinations. From top to bottom, "Original RGB" means the module directly outputs initial RGB features X_n^{RGB} without any enhancement. "RGB Only" means that the module uses RGB self attention weights to enhance RGB features, i.e. $X_n^{RGB'} = \sigma(f_{conv}(X_n^{RGB})) \otimes X_n^{RGB}$. "Flow Only" means that the module only uses enhanced optical flow to enhance RGB features, i.e. $X_n^{RGB'} = \sigma(f_{conv}(X_n^{Flow'})) \otimes X_n^{RGB}$. "Not Sharing" means that we employ convolution layer f_{conv1} on X_n^{RGB} and f_{conv2} on $X_n^{Flow'}$, and these two convolution layers do not share parameters. "Exchange Modalities" on the last row means we exchange X_n^{RGB} and X_n^{Flow} shown

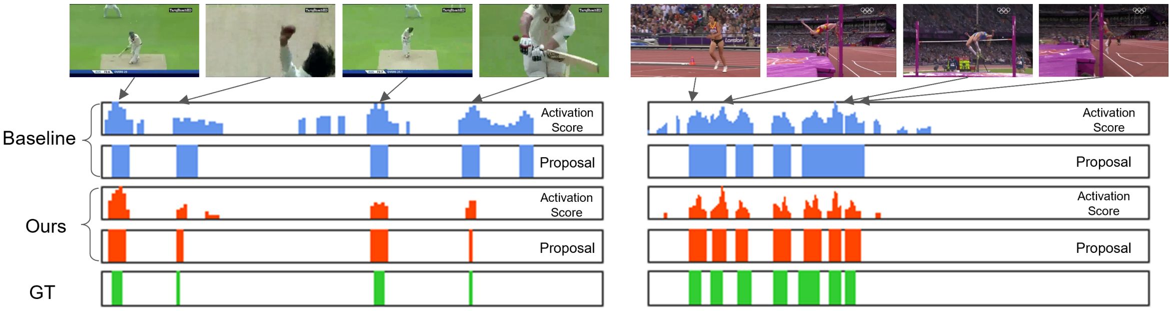

Figure 5. Qualitative visualization of two typical video examples from THUMOS'14. The results of BaS-Net (baseline), our method, and ground truth (GT) are shown in blue, red, and green, respectively. Since we introduce an auxiliary class for background in TCAS, we set the activation score to be 0 if the background class of this snippet gains the highest activation score among all classes.

Modality	mAP@IoU (%)			
	0.3	0.5	0.7	AVG
Original RGB	54.9	37.5	13.2	44.1
RGB Only	56.6	39.1	13.3	45.4
Flow Only	57.7	38.8	13.4	45.9
RGB + Flow (Not Sharing)	57.9	39.7	13.1	46.5
RGB + Flow (Sharing Conv)	58.6	40.7	13.7	46.9
Exchange Modalities	53.3	36.5	12.2	43.2

Table 5. Ablation studies of our model with different kinds of modality combinations on the THUMOS'14 testing set. AVG is the average mAP at the IoU threshold 0.1:0.1:0.7.

in Fig. 2 while keeping other settings the same with "Sharing Conv".

We can find that X_n^{Flow*} does enhance the RGB features, improving the average mAP (0.1:0.7) from 44.1% ("Original RGB") to 45.9% ("Flow Only"). This is because the initial RGB features contain task-irrelevant information that hinders performance and X_n^{Flow*} can help filter out task-irrelevant information. But only using X_n^{Flow*} to coordinate X_n^{RGB} may lose spatial information. Therefore, using both modalities to enhance RGB features achieves the best results. When it comes to the performance difference between "Sharing Conv" and "Not Sharing", it is because the inconsistency between the two modalities will cause the degradation of performance, while the sharing convolution layer is beneficial to make weights distributions of the two modalities more approached. The performance degradation caused by "Exchange Modalities" shows that temporal modeling (TFE-DC Module) should be applied to optical flow features rather than RGB features. According to the above analysis, our proposed Modality Enhancement Module is reasonable for enhancing RGB features.

4.4. Qualitative Results

To illustrate the efficacy of our proposed method, we demonstrate the detected results of two typical video samples in Fig. 5. These two samples are representative because the first example contains category 'CricketBowling' and 'CricketShot' and each action instance of these two categories is extremely short (about 0.6 sec). While the second example contains the category 'HighJump' and each action instance of this class is relatively long (about 6.1 sec). BaS-Net is used as the baseline because our model follows its background suppression structure and uses its loss functions as basic loss functions during optimization. It can be observed that our method has more accurate localization proposals than the baseline, indicating our method effectively utilizes temporal information. For instance, in the second example, the baseline method incorrectly combines several action instances into one. While our method can localize every action instance very clearly. Meanwhile, the activation scores of background snippets are quite low, showing that our method can successfully suppress the activation scores of background snippets and separate action instances from backgrounds. These two typical samples fully demonstrate the superiority of our method.

5. Conclusions

In this paper, we explore how to effectively use temporal information and enhance features to improve temporal action localization results. We propose a novel WTAL framework named TFE-DCN to tackle the issue. We use the TFE-DC Module to enlarge the receptive field and capture long-range dependencies between snippets to enhance optical flow features. We also propose a Modality Enhancement Module to enhance RGB features with the help of enhanced optical flow features. Experiments on two datasets demonstrate that our TFE-DCN outperforms current state-of-the-art methods, and validate our idea that the efficient use of temporal information can significantly improve the performance of temporal action localization.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017.

[2] 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1130–1139, 2018.

[3] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 5727–5736, 2017.

[4] Yazan Abu Farha and Jürgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3570–3579, 2019.

[5] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19967–19977, 2022.

[6] 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13915–13925, 2022.

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 961–970, 2015.

[8] 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, MM '21, page 1591–1599, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6):797–819, 2011.

[10] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 7982–7991, 2021.

[11] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3262–3271, 2022.

[12] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence, 35(2):1637–1645, May 2021.

[13] 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.

[14] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1003–1012, 2017.

[15] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 13628–13637, 2021.

[16] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):11320–11327, Apr 2020.

[17] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3):1854–1862, May 2021.

[18] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6742–6751, 2018.

[19] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19882–19892, 2022.

[20] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, pages 3–21, Cham, 2018. Springer International Publishing.

[21] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1298–1307, 2019.

[22] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6172–6181, 2021.

[23] Ziyi Liu, Le Wang, Qilin Zhang, Wei Tang, Junsong Yuan, Nanning Zheng, and Gang Hua. Acsnet: Action-context separation network for weakly supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3):2233–2241, May 2021.

[24] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 344–353, 2019.

[25] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9964–9974, 2021.

[26] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 420–437, Cham, 2020. Springer International Publishing.

[27] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 13588–13597, 2021.

[28] Phuc Nguyen, Bohyung Han, Ting Liu, and Gautam Prasad. Weakly supervised action localization by sparse temporal pooling network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6752–6761, 2018.

[29] Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Weakly-supervised action localization with background modeling. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 5501–5510, 2019.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[31] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, pages 588–607, Cham, 2018. Springer International Publishing.

[32] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1417–1426, 2017.

[33] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, pages 162–179, Cham, 2018. Springer International Publishing.

[34] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6479–6488, 2018.

[35] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In 9th ISCA Speech Synthesis Workshop, pages 125–125, 2016.

[36] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6402–6411, 2017.

[37] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and FengWu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 53–63, 2021.

[38] Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3):3090–3098, Jun. 2022.

[39] Runhao Zeng, Wenbing Huang, Chuang Gan, Mingkui Tan, Yu Rong, Peilin Zhao, and Junzhou Huang. Graph convolutional networks for temporal action localization. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 7093–7102, 2019.

[40] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 37–54, Cham, 2020. Springer International Publishing.

[41] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16005–16014, 2021.

[42] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2933–2942, 2017.