DDG-Net: Discriminability-Driven Graph Network for Weakly-supervised Temporal Action Localization

Xiaojun Tang[1], Junsong Fan[23], Chuanchen Luo[2], Zhaoxiang Zhang[234], Man Zhang[1]*, and Zongyuan Yang[1]

[1] Beijing University of Posts and Telecommunications, China

[2] Institute of Automation, Chinese Academy of Sciences, China

[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS, HongKong, China

[4] University of Chinese Academy of Sciences, UCAS, China

{tangxiaojun, zhangman, yangzongyuan0}@bupt.edu.cn,

{junsong.fan, luochuanchen2017, zhaoxiang.zhang}@ia.ac.cn

Abstract

Weakly-supervised temporal action localization (WTAL) is a practical yet challenging task. Due to large-scale datasets, most existing methods use a network pretrained in other datasets to extract features, which are not suitable enough for WTAL. To address this problem, researchers design several modules for feature enhancement, which improve the performance of the localization module, especially modeling the temporal relationship between snippets. However, all of them omit that ambiguous snippets deliver contradictory information, which would reduce the discriminability of linked snippets. Considering this phenomenon, we propose Discriminability-Driven Graph Network (DDG-Net), which explicitly models ambiguous snippets and discriminative snippets with well-designed connections, preventing the transmission of ambiguous information and enhancing the discriminability of snippet-level representations. Additionally, we propose feature consistency loss to prevent the assimilation of features and drive the graph convolution network to generate more discriminative representations. Extensive experiments on THUMOS14 and ActivityNet1.2 benchmarks demonstrate the effectiveness of DDG-Net, establishing new state-of-the-art results on both datasets. Source code is available at `https://github.com/XiaojunTang22/ICCV2023-DDGNet`.

1. Introduction

Temporal action localization (TAL) is a significant yet challenging task in video understanding. It aims to localize the start and end timestamps of the action proposals of interest and recognize their categories in untrimmed

---

*Corresponding author.

Figure 1. An example of ambiguous snippets. The orange and green curves denote the attention weights generated by RGB and optical flow separately. The black horizontal line represents the threshold.

videos [2, 49, 54]. It has attracted great attention in academia and industry due to potential applications for video retrieval, surveillance, and anomaly detection. Full-supervised temporal action localization [6, 31, 32, 46, 50, 56, 66, 68] has made significant progress in recent years. However, these methods require frame-level annotations, which are time-consuming and labor-intensive for large-scale datasets. Therefore, a number of researchers pay essential attention to weakly-supervised temporal action lo-

calization (WTAL).

WTAL learns to localize action instances with only video-level labels (*i.e.* action categories in the video) during the training process. Most existing methods adopt the paradigm of multi-instance learning (MIL), regarding the video as a bag and action regions as instances. However, due to the gap between classification and localization tasks, it is difficult to localize precise action proposals without frame-level annotations. Besides, except for the discriminative snippets (action or background), there are massive ambiguous snippets (*e.g.* action context) in untrimmed videos. For example, as shown in Figure1, during the "Cricket-Bowling" action, action snippets with obvious appearance and quick movement, and background snippets with irrelevant content are easy to be recognized. Instead, the ambiguous snippets (action actually) contain action information (similar appearance) and background information (slow movement) simultaneously, which increases the difficulty of discriminating.

Therefore, it is crucial to localize and recognize ambiguous snippets exactly. Several works [14, 33, 34, 43] are devoted to relieving the interference of action context by separating action and context. Meanwhile, we note that a few works [11, 44, 47, 59] pursue to model the temporal relationship between snippets for receiving more complete information. These methods have achieved certain performances, but they overlook the adverse effects of ambiguous information. When ambiguous snippets spread out information to similar action or background snippets, ambiguity is delivered together. It leads to the result that ambiguous snippets would reduce the discriminability of others when transmitting contradictory information to them.

Inspired by the above research and analysis, we propose a novel graph network, named Discriminablity-Driven Graph Network (DDG-Net), which explicitly separates ambiguous snippets and discriminative snippets to perform more effective graph inference with well-designed connections. Specifically, as shown in Figure 1, we divide firstly the whole video into three types of snippets (*i.e.* pseudo-action, pseudo-background, and ambiguous snippets). Then, we design different connections among them to ensure only the transmission of discriminative information. In this way, the snippets receive complementary information from linked snippets through graph inference. What is more, the discriminability of ambiguous snippets is enhanced while the snippets never do harm to other snippets due to distinctive connections. Besides, we propose feature consistency loss to maintain characteristics of snippet-level representations for localization. Our main contributions can be summarized as follows.

- We propose a novel graph network DDG-Net, which explicitly models ambiguous and discriminative snippets with different types of connections, aiming to spread complementary information and enhance the discriminability of snippet-level features while eliminating the adverse effects of ambiguous information.

- We propose feature consistency loss with DDG-Net, which prevents the assimilation of snippet-level features and drives the graph convolution network to generate more discriminative representations.

- Extensive experiments demonstrate that our method is effective, and establishes new state-of-the-art results on THUMOS14 and ActivityNet1.2 datasets.

2. Related Work

Weakly-supervised temporal action localization. In recent years, Temporal action localization with weak supervision has attracted increasing interest from the community. Weak supervision includes single-frame annotations [7, 22, 25, 36, 38, 57] and video-level labels [19, 23, 40, 55, 60, 63, 65, 67], and we focus on the latter. UntrimmedNet [54] firstly solves the task via a classifier module and a selection module. STPN [41] exploits the sparse loss to suppress unimportant selections. W-TALC [42] punishes the distance between video-level features with the same action class. Bas-Net [26] introduces background class as an auxiliary to propose a two-branch architecture and suppress activation of background snippets. [11, 51] generate more complete snippet-level representations by modeling temporal relationships. [20, 52, 64] learn extra knowledge from trimmed videos. DELU [5] introduces Evidential Learning which effectively alleviates ambiguity between action and background. [33, 34, 43] make efforts to separate action and background by modeling action context. [9, 29, 62] adopt contrastive learning for action-background seperation. [61] proposes a two-stream network to eliminate false positive action proposals. CO2-Net [13] exploits cross-modal channel attention to filter out redundant information. [21, 58] aim to train a stronger model via collaborative learning. These methods exploit the collaboration of RGB and optical flow which is also significant for our model.

As mentioned above, most existing methods have devoted significant efforts to improving localization modules. Different from them, we focus on modeling more suitable features for WTAL. Specifically, we generate more discriminative features by transferring complementary information and eliminating ambiguity via a graph network.

Graph-based Weakly-supervised temporal action localization. Rashid *et al*. [44] constructs sparse action graphs by treating snippets as nodes to connect relevant snippets. DGCNN [47] constructs a dynamic graph with various relations in order to make full use of temporal and contextual relations between pseudo-action snippets. ACGNet [59]

spreads the complementary information through graph inference. Though these works enhance snippet features via graph networks, they ignore the effects of ambiguous information which may reduce the discriminability of snippet-level representations.

Unlike these graph networks, We subdivide the whole video into discriminative and ambiguous snippets to construct different types of connections with the consideration that ambiguous snippets would confuse linked snippets. Specially, we force ambiguous snippets to never spread information to others and only receive relevant information from discriminative snippets. In this way, we enhance the discriminability of ambiguous snippets and prevent the propagation of ambiguity.

3. Base Model

In this section, we introduce the base model based on DELU [5] for WTAL. We first define the problem formulation (Section 3.1). Then we give the pipeline with the base model (Section 3.2).

3.1. Problem Formulation

Given an untrimmed video $\mathbb{V}$ and its corresponding multi-hot action category label $y \in \{0, 1\}^C$ for training, where C is the number of action categories, the goal of temporal action localization is to get a set of action instances $S = (s_i, e_i, c_i, \varphi_i)$ in a testing video, where s_i and e_i is the start and end timestamps of the corresponding action instance, c_i and φ_i is the predicted action category and related confidence score.

3.2. Pipeline

Following previous works [5, 13, 26], we first divide an untrimmed video into T non-overlapping 16-frame snippets. Then, snippet-level features are extracted from RGB and optical flow streams by pretrained networks. For convenience, we denote snippet features as $F^* = [f_1^*, ..., f_T^*] \in \mathbb{R}^{D \times T}$, where D is the dimension of feature and $* \in \{r, f\}$ refers to RGB and optical flow.

We remove the cross-modal consensus module [13] from DELU [5] to emphasize our method because both are modules for feature enhancement. We present extra results combined with the cross-modal consensus module in the supplementary material.

After F^* are extracted, they are put through an attention module to generate action attention weights as follows,

$$\mathbb{A}^* = \Phi^*(F^*) \in \mathbb{R}^{1 \times T} \tag{1}$$

Φ is the attention module consisting of several convolution layers and a sigmoid function. We fuse two attention sequences for use later.

$$\mathbb{A} = \frac{\mathbb{A}^r + \mathbb{A}^f}{2} \tag{2}$$

Then we generate a classification activation sequence (CAS) through a feature fusion module and a classifier, formally,

$$p = \Psi(\Theta(F^r \oplus F^f)) \in \mathbb{R}^{(C+1) \times T} \tag{3}$$

$\oplus$ is a concatenate operator through the channel dimension. Θ is the feature fusion module and Ψ is the classifier include C action classes and background class. We further get the suppressed CAS $\overline{p} = \mathbb{A} \cdot p$, where $\cdot$ denotes element-wise multiplication.

We denote the train objective function of the base model as L_{base}. At the testing stage, we follow the standard process [5] to carry out temporal action localization.

4. Method

In this section, we describe our approach in detail. The architecture of DDG-NeT is shown in Figure 2. We first give definitions of graphs (Section 4.1). Then we introduce the composition principle of graphs (Section 4.2). And the complementary and discriminative information is spread out via graph inference (Section 4.3). To prevent the assimilation of snippet-level features, we propose feature consistency loss which maintains characteristics of discriminative representations (Section 4.4). It is worth mentioning that our method can be applied to any two-stream model [5, 13, 61].

4.1. Graph Formulation

We denote the complete graph as $G = (V, E)$, and A is its adjacency matrix. In this case, V denotes the set of all snippets in the whole video. We divide it into three subsets V_a, V_b, V_m, which contain pseudo-action, pseudo-background and ambiguous snippets respectively. V_a with symmetrical adjacency matrix A_a constructs a undirected graph G_a, and we regard it as action graph. Background graph G_b is also constructed by V_b with A_b. The nodes of ambiguous graph G_m consisting of V is a directed graph, and the direction of edges is always from nodes of V_a or V_b to V_m. Under such a setting, we construct two subgraphs for interaction between relevant discriminative snippets and a subgraph for receiving discriminative information.

4.2. Graph Generation

We first introduce the pre-classification method for snippets. Without frame-level annotations, we have no supervision to snippet-level categories. Learning from the previous method [33], we consider the action weights of snippets as the judgment basis. We jointly take $\mathbb{A}^r$ and $\mathbb{A}^f$ into consideration. For each timestamp t, when $\mathbb{A}_t^r$ and $\mathbb{A}_t^f$ exceed the action threshold η, it is regarded as a pseudo-action snippet. Also when $\mathbb{A}_t^r$ and $\mathbb{A}_t^f$ be lower than threshold 1-η, it is regarded as a pseudo-background snippet. If a snippet neither belongs to both, we regard it as an ambiguous snippet.
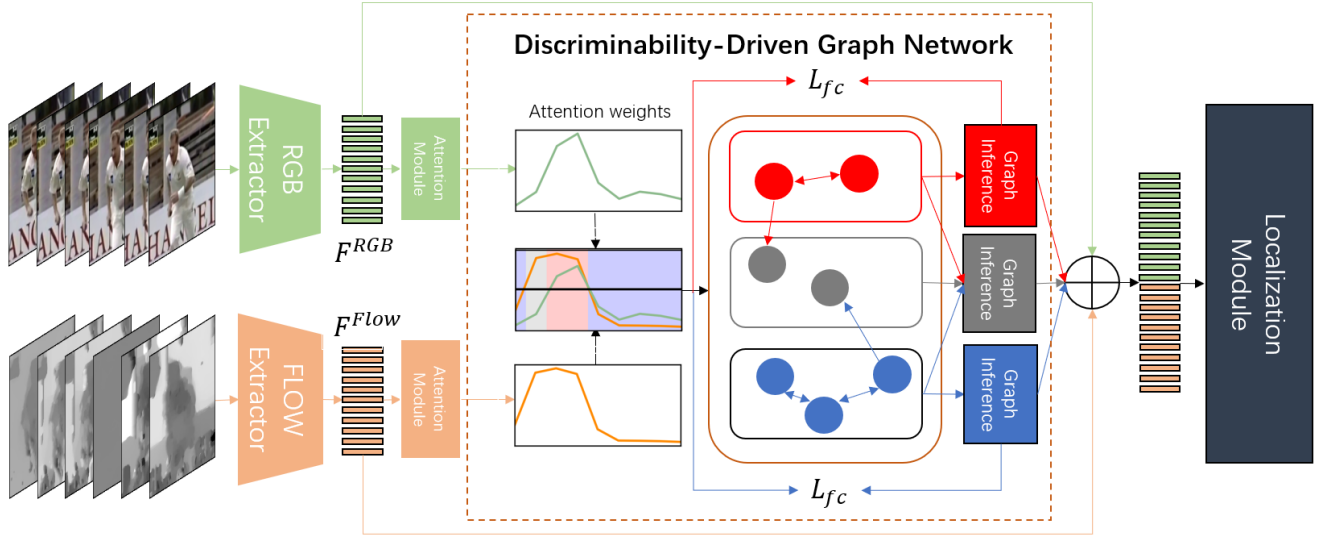
Figure 2. Overview of the proposed DDG-Net. Based on action weights, the snippets are pre-classified as pseudo-action, pseudo-background, and ambiguous snippets respectively. Different connections are linked among them to construct three subgraphs. More discriminative features are generated via graph inference for the localization module.

We simply treat feature similarity as the edge weights between connected nodes. Specially, we construct the graph G^* and compute its adjacency matrix A^* separately for RGB and optical flow features. Exploiting the collaboration of RGB and optical flow, we take the average of their adjacency matrixes instead of the respective adjacency matrix for graph inference. Formally,

$$A_{ij}^* = s(f_i^*, f_j^*) \qquad (4)$$

$$A_{ij} = \frac{A_{ij}^r + A_{ij}^f}{2} \qquad (5)$$

s is the similarity measure, cosine similarity in our experiments.

In order to write concisely, we appropriately omit the logo of RGB and optical flow in the following. We construct three subgraphs according to the division of snippets.

The values of adjacency matrix A_a and A_b are consistent with the values on the corresponding position of A. The ambiguous graph G_m is a directed graph, whose adjacency matrix A_m is masked partly. Formally,

$$A_m(i,j) = \begin{cases} A(i,j), & j \in V_m \quad and \quad i \notin V_m; \\ 1, & i = j \quad and \quad i \in V_m; \\ 0, & otherwise. \end{cases} \qquad (6)$$

Under this treatment, we exploit complementary information of discriminative snippets and abandon confusing information from ambiguous snippets. With different types of connections, the discriminative information is delivered to relevant snippets while the ambiguous is never spread

out. Note that the fusion of cross-modal adjacency matrixes and preventing the propagation of ambiguous information is vital for our model (Section 5.3).

Following [59], we use a similarity threshold θ and a top-K ranking list to filter out nodes with a low contribution. A softmax operator through columns is used to obtain the final adjacency matrix.

4.3. Graph Inference

For action graph and background graph, graph average and graph convolution are directly operated to obtain aggregated features following residual connection. Taking action graph as an example, graph average is applied to action features F_a as follows,

$$F_a^{avg} = F_a A_a \qquad (7)$$

For a graph convolution network (GCN) with L layers, the formula on the l-th layer is presented,

$$F_l^{gcn} = \sigma(W_a^l F_{l-1}^{gcn} A_a) \qquad (8)$$

where F_l^{gcn} is output of the l-th layer, $F_0^{gcn} = F_a$, σ is the activation function LeakyReLU with parameter of 0.2. The aggregated features F_a^g through the graph network is the mean of F_a^{avg} and F_a^{gcn} (i.e. F_L^{gcn}) and enhanced features are got as follows,

$$F_a' = \frac{F_a^g + F_a}{2} \qquad (9)$$

For ambiguous graph, the graph average is applied as above. We never apply a separate GCN for ambiguous features due to the difficulty of training. Instead, we exploit

F_a^{gcn} and F_b^{gcn} to aggregate representations of same forms. We denote $A_{m,a}$ as the partition of A_m which only contain ambiguous columns and action rows. $A_{m,b}$, $A_{m,m}$ is the same. Formally,

$$F_m^{gcn} = F_a^{gcn}A_{m,a} + F_b^{gcn}A_{m,b} + F_mA_{m,m} \qquad (10)$$

Note that $A_{m,m}$ is a diagonal matrix. The subsequent process is consistent with the equation 9. In this case, the GCN only learns to transform discriminative snippet features. Ambiguous snippet representations benefit from the enhanced features via graph inference.

Through graph inference, we get enhanced snippet-level features. We sort updated features together by time and input them instead of original features into the base model to carry out the WTAL task. We find that directly using the GCN is harmful because it drives the model to pursue classification performance, resulting in poor localization with chaotic features. So we propose feature consistency loss to address the problem effectively.

4.4. Feature Consistency Loss

In our experiment, the performance of DDG-Net suffers a sharp decline without any constraint to the GCN. The GCN prefers to transform all snippet-level features close to action characteristics for the classification task, which is harmful to the localization task. Therefore, we propose a feature consistency loss to avoid this case. Intuitively, we hope the most discriminative representations remain unchanged through DDG-Net because they can be recognized easily. However, it is difficult to achieve this goal under the influence of graph interaction. Instead, We punish the distance between graph average features and graph convolution features to realize the same function. Formally,

$$L_{fc} = \frac{1}{|V_a|} \sum_{t \in V_a} w(\mathbb{A}_t) d(F_{a,t}^{gcn}, F_{a,t}^{avg})$$
$$+ \frac{1}{|V_b|} \sum_{t \in V_b} w(1 - \mathbb{A}_t) d(F_{b,t}^{gcn}, F_{b,t}^{avg}) \qquad (11)$$

$$w(x) = exp(-(\frac{1}{x} - 1)/\tau) \qquad (12)$$

where $w(\cdot)$ is a weight function and τ is the temperature parameter. $d(\cdot)$ is the distance measure, euclidean distance in our experiments. Under this constraint, the GCN learns to enhance less discriminative features while maintaining characteristics of the most discriminative representations.

4.5. Train Objective

The complete train objective function is

$$L = L_{base} + \lambda_1 L_{fc} + \lambda_2 L_{cl} \qquad (13)$$

where L_{cl} is the complementary learning loss. [8]

5. Experiments

5.1. Experimental Setup

5.1.1 Datasets

We evaluate our proposed DDG-Net on two public benchmarks, *i.e*., THUMOS14 [17] and ActivityNet1.2 [12].
THUMOS14 [17] is a challenging dataset that contains 200 validation videos for training and 213 test videos for evaluation from 20 action classes. The videos' length varies from a few seconds to several minutes and each video contains about 15 action instances on average.
ActivityNet1.2 [12] is a large-scale dataset that contains 4819 training videos, 2383 validation videos, and 2480 testing videos from 100 action classes. Following previous works, we evaluate our model on the validation set because the ground-truth annotations of the test set are withheld for the challenge.

5.1.2 Evaluation Metrics

Following standard evaluation protocol, we report mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds.

5.1.3 Implement Details

Following existing methods, we use I3D network [3] pretrained on Kinetics dataset [24] to extract both the RGB and optical flow features. The action and similarity thresholds are set to 0.5 and 0.8 respectively in default. The number of convolution layers L is 2. The temperature τ is set to 0.5. The weights λ_1, λ_2 are set to 1, 3.2 for THUMOS14 and 0.5, 1.7 for ActivityNet1.2 dataset. All experiments are run on a single NVIDIA RTX A5000.

5.2. Comparisons with State-of-the-Art Methods

In Table 1, we compare the performance of DGG-Net with existing fully-supervised and weakly-supervised TAL methods on the THUMOS14 dataset. This table shows that our approach outperforms the baseline at all IoU thresholds. Especially mAP has a significant increase of 2.3% at the IoU threshold of 0.5, and the average of mAP exceeds the baseline of 1.5%. DGG-Net performs better than all state-of-the-art methods at most IoU thresholds except 0.7 and achieves a new state-of-the-art result with a 0.9% improvement in the average of mAP. What is more, compared with some fully-supervised and single-frame methods, our approach has competitive and even better performance.

Table 2 presents the comparison results with existing WTAL methods on the AcitivityNet1.2 dataset. The dataset is more simple with an average of 1.5 action instances per video. As shown, our base model has achieved a competitive performance. With DDG-Net, the performance ex-

Supervision	Method	mAP(%)@IoU							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
Fully	S-CNN [48]	47.7	43.5	36.3	28.7	19.0	10.3	5.3	27.3
	TAL-Net [4]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1
	DBS [10]	56.7	54.7	50.6	43.1	34.3	24.4	14.7	39.8
	AGCN [28]	59.3	59.6	57.1	51.6	38.6	28.9	17.0	44.6
Weakly single-frame	SF-Net [36]	68.3	62.3	52.8	42.2	30.5	20.6	12.0	41.2
	Ju et al. [22]	72.8	64.9	58.1	46.4	34.5	21.8	11.9	44.3
	Lee and Byun [25]	75.7	71.4	64.6	56.5	45.3	34.5	21.8	52.8
Weakly video-level	STPN [41]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	27
	BaS-Net [26]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	ACM-Net [43]	68.9	62.7	55.0	44.6	34.6	21.8	10.8	42.6
	CoLA [62]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9
	CO2-Net [13]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	44.6
	CSCL [21]	68.0	61.8	52.7	43.3	33.4	21.8	12.3	41.9
	FAC-Net [15]	67.6	62.1	52.6	44.3	33.4	22.5	12.7	42.2
	ACGNet [59]	68.1	62.6	53.1	44.6	34.7	22.6	12.0	42.5
	ASM-Loc [11]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	45.1
	FTCL [9]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	43.6
	DCC [29]	69.0	63.8	55.9	45.9	35.7	24.3	13.7	44.0
	Huang et al. [16]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	45.1
	DGCNN [47]	66.3	59.9	52.3	43.2	32.8	22.1	13.1	41.3
	DELU [5]	71.5	66.2	56.5	47.7	40.5	27.2	**15.3**	46.4
	baseline	71.5	65.9	56.5	47.3	39.1	26.0	14.3	45.8
	+DDG-Net	**72.5**$^{+1.0}$	**67.7**$^{+1.8}$	**58.2**$^{+1.7}$	**49.0**$^{+1.7}$	**41.4**$^{+2.3}$	**27.6**$^{+1.6}$	14.8$^{+0.5}$	**47.3**$^{+1.5}$

Table 1. Comparison results with existing methods on THUMOS14 dataset.

Method	mAP(%)@IoU			
	0.5	0.75	0.95	Avg
DGAM [45]	41.0	23.5	5.3	24.4
RefineLoc [1]	38.7	22.6	5.5	23.2
ACSNet [34]	40.1	26.1	**6.8**	26.0
HAM-Net [18]	41.0	24.8	5.3	25.1
Lee et al. [27]	41.2	25.6	6.0	25.9
ASL [37]	40.2	-	-	25.8
CoLA [62]	42.7	25.7	5.8	26.1
AUMN [35]	42.0	25.0	5.6	25.5
UGCT [58]	41.8	25.3	5.9	25.8
D2-Net [39]	42.3	25.5	5.8	26.0
ACM-Net[43]	43.0	25.8	6.4	26.5
CO2-Net [13]	43.3	26.3	5.2	26.4
CSCL [21]	43.8	**26.9**	5.6	26.9
ACGNet [59]	41.8	26.0	5.9	26.1
Li et al. [30]	41.6	24.8	5.4	25.2
DGCNN [47]	42.0	25.8	6.0	26.2
DELU [5]	44.2	26.7	5.4	26.9
baseline	43.9	26.6	5.4	26.8
+DGG-Net	**44.3**$^{+0.4}$	**26.9**$^{+0.3}$	5.5$^{+0.1}$	**27.0**$^{+0.2}$

Table 2. Comparison results on ActivityNet1.2 dataset.

ceeds the base model completely and existing methods at most Iou thresholds. It is worth noting that the improvement on ActivityNet 1.2 is slight compared with THU-MOS14. This phenomenon is also observed in related work like ACGNet [59]. We attribute the major reason to the difference in the dataset characteristics. THUMOS14 contains 15 action instances per video on average, while ActivityNet1.2 contains only 1.5 action instances. Therefore, THUMOS14 generally contains more ambiguous snippets. According to the result of pre-classification, the proportion of ambiguous snippets is 18.9% on THUMOS14 and 14.9% on ActivityNet1.2. Our method contributes to eliminating the adverse effects of ambiguous information, therefore, improves more on THUMOS14. In addition, the performance of pre-classification for pseudo-action and pseudo-background snippets on ActivityNet1.2 is worse, leading to less effective interaction between snippets, which also results in less improvement.

5.3. Ablation Study

Effects of components. Table 3 presents the results of adding components on the baseline. The improvement of the average of mAP is 0.6% when only applying graph average during graph inference. The performance has a sharp decline of 10.3% because the GCN prefers to generate features attached to action for classification tasks, which is harmful to localization tasks. The assimilation of features causes massive false action snippets, which link separate action instances, leading to the diminution of the number of

Exp	AVG	GCN	L_{fc}	Number	mAP Avg
1				38118	45.8
2	✓			34961	46.4
3	✓	✓		26755	36.1
Ours	✓	✓	✓	34794	**47.3**

Table 3. Ablation study of the effectiveness of components on THUMOS14. AVG and GCN denote graph average and graph convolution respectively. Number denotes the number of predicted action instances.

Exp	D	FAD	mAP(%)@IoU			
			0.1	0.4	0.7	Avg
4	✓		70	46.5	13.4	44.9
5		✓	71.3	47.4	13.6	45.8
Ours	✓	✓	**72.5**	**49.0**	**14.8**	**47.3**

Table 4. Verification of ideas on THUMOS14. D and FAD denote the disconnection of ambiguity and the fused adjacency matrix.

predicted instances and localization performance. With the constraint of feature consistency loss, the model is forced to generate consistent and discriminative features, which significantly improve the performance of the localization module. In a word, the result of the localization module is significantly improved with enhanced features, which are generated by DDG-Net equipped with feature consistency loss. It is worth noting that the number of parameters increases from 21.5 M to 29.9 M and the number of floating point operations increases from 6.89 G to 13.39 G when equipping the base model with DDG-Net, which is acceptable due to consistent performance gains.

Analysis of insight. Reviewing our idea, ambiguous snippets only should receive information from discriminative snippets and never confuse others with their ambiguity. We denote this strategy in DDG-Net as the disconnection of ambiguity. Additionally, we find it is also essential for keeping collaboration between RGB and Flow features, which refers to the fused adjacency matrix in equation 5. Table 4 demonstrates our experimental results. In Exp 4, we take the respective adjacency matrix used for graph inference. The performance suffers a serious decline because the consistency of the cross-modal features is destroyed after graph inference. Exp 5 shows almost the same performance as the baseline when allowing the propagation of information from ambiguous snippets. In this case, though the complementary information is spread out to relevant snippets, the Ambiguous information reduces the discriminability of features. Only when both the disconnection of ambiguity and the fused adjacency matrix are applied, does our approach generate consistent and discriminative features that improve the performance of the localization module effectively. Note that our idea can be utilized by other models
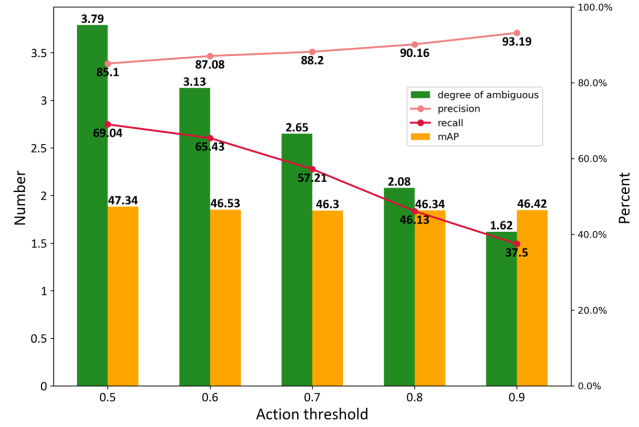

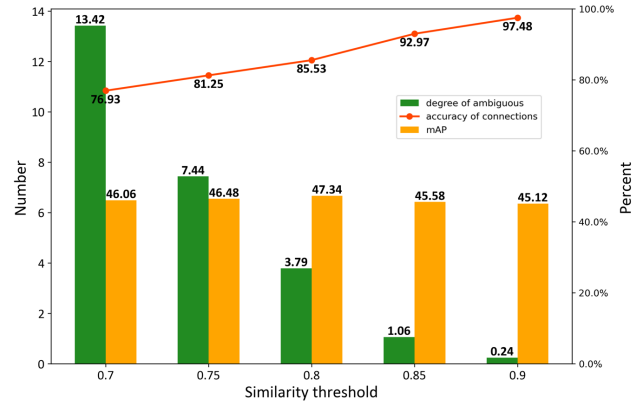Figure 3. Effects of action threshold.

Figure 4. Effects of similarity threshold.

for modeling relationships between snippets, such as the attention mechanism [53].

Effects of thresholds. We first study the effects of the action threshold. Figure 3 shows the result and statistical information at different action thresholds. From the figure, we can find when the action threshold increases, the precision of pre-classification for action and background snippets increases slowly but the recall has a quick decline, which leads to a decrease in the mean degree of ambiguous snippets. Our approach achieves the best balance at the threshold of 0.5. Then we explore the effects of the similarity threshold in Figure 4. when the similarity threshold increases, the connections for ambiguous snippets are more accurate (the connection is true while linked snippets both belong to action or background) but the mean degree of ambiguous snippets reduces quickly. When the similarity threshold is 0.9, ambiguous snippets receive little discriminative information from few connected snippets, which causes the significant degradation of mAP. Optimal performance is achieved at the threshold of 0.8.
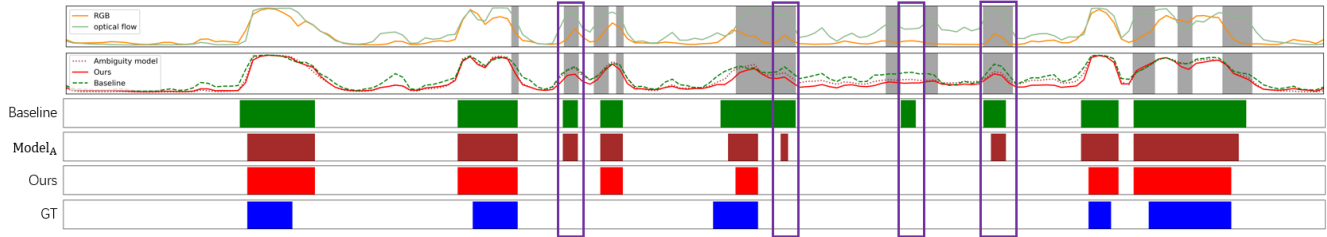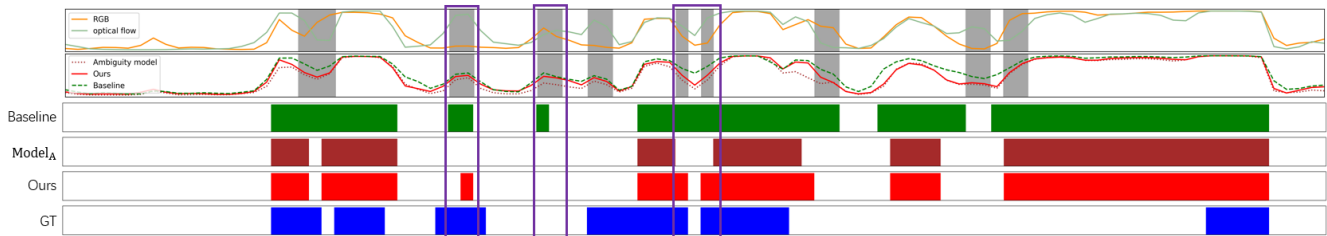
Figure 5. Visualization results on THUMOS14 dataset. The first and second rows show the original two-stream attention weights and the final attention weights of different models. Following rows present localization results and ground truth (GT). Model_A denote the ambiguity model. The purple boxes show our improvements.
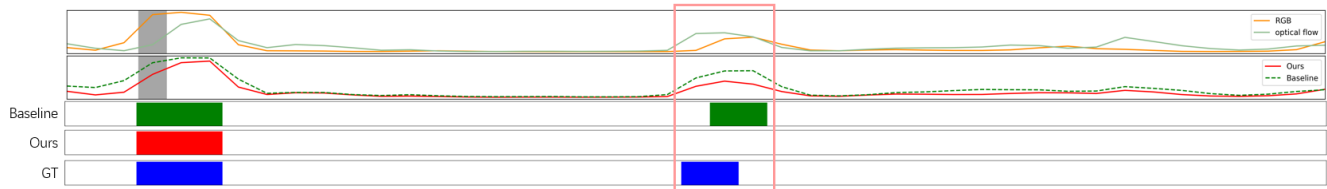

Figure 6. A bad case on THUMOS14 dataset. The pink box presents a false suppression while corresponding snippets are pre-classify as pseudo-background wrongly.

5.4. Visualization Results

Figure 5 visualizes the qualitative comparisons between the baseline and our method. Additionally, we show the results of Exp 5 in Table 4, denoted as the ambiguity model. The curves denote the attention weights and shaded parts represent ambiguous snippets. The blocks represent action proposals at the IoU threshold of 0.5. In Figure 5a, we can find both baseline and ambiguity model classify several ambiguous snippets as action snippets which are background actually. Our approach effectively enhances the discriminability of ambiguous snippets and generates more accurate action instances by suppressing the action weights. In Figure 5b, while the ambiguity model suppresses the ambiguous action regions wrongly, DDG-Net has better robustness. In addition, we show a bad result compared with the base model in Figure 6. The action snippets are pre-classified as pseudo-background, which suppresses their attention weights due to receiving background information. It indicates the performance of pre-classification is essential to our method, as mentioned in Section 5.2.

6. Conclusions

In this paper, we present the DDG-Net to explore a novel graph network for effectively enhancing the discriminability of snippet-level representations for WTAL. Concretely, we divide snippets using a simple yet valid method and design different types of connections for discriminative and ambiguous snippets. In this way, we significantly improve the discriminability of snippet-level representations, especially for ambiguous snippets via graph inference. Our method improves the performance of the base model by a large margin and achieves state-of-the-art results. Extensive experiments verify the superiority of our method.

Acknowledgement

References

[1] Humam Alwassel, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. *arXiv preprint arXiv:1904.00227*, 2(6), 2019.

[2] Meng Cao, Can Zhang, Long Chen, Mike Zheng Shou, and Yuexian Zou. Deep motion prior for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 31:5203–5213, 2022.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018.

[5] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 192–208. Springer, 2022.

[6] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13053–13064, 2021.

[7] Xinpeng Ding, Nannan Wang, Jie Li, and Xinbo Gao. Weakly supervised temporal action localization with segment-level labels. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part I*, pages 42–54. Springer, 2021.

[8] Jia-Run Du, Jia-Chang Feng, Kun-Yu Lin, Fa-Ting Hong, Xiao-Ming Wu, Zhongang Qi, Ying Shan, and Wei-Shi Zheng. Weakly-supervised temporal action localization by progressive complementary learning. *arXiv e-prints*, pages arXiv–2206, 2022.

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

[10] Zhanning Gao, Le Wang, Qilin Zhang, Zhenxing Niu, Nanning Zheng, and Gang Hua. Video imprint segmentation for temporal action detection in untrimmed videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8328–8335, 2019.

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

[12] 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. IEEE, 2015.

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

[14] Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Relational prototypical network for weakly supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11053–11060, 2020.

[15] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8002–8011, 2021.

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

[17] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[18] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1637–1645, 2021.

[19] Ashraful Islam and Richard Radke. Weakly supervised temporal action localization using deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 547–556, 2020.

[20] Mihir Jain, Amir Ghodrati, and Cees GM Snoek. Actionbytes: Learning from trimmed videos to localize actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1171–1180, 2020.

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

[22] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13455–13464, 2021.

[23] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Xiaoyun Zhang, Yanfeng Wang, and Qi Tian. Adaptive mutual supervision for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia*, 2022.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu-

man action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[25] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13648–13657, 2021.

[26] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11320–11327, 2020.

[27] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1854–1862, 2021.

[28] Jin Li, Xianglong Liu, Zhuofan Zong, Wanru Zhao, Mingyuan Zhang, and Jingkuan Song. Graph attention based proposal 3d convnets for action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4626–4633, 2020.

[29] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19914–19924, 2022.

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

[31] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2021.

[32] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[33] Ziyi Liu, Le Wang, Wei Tang, Junsong Yuan, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through learning explicit subspaces for action and context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2242–2250, 2021.

[34] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2233–2241, 2021.

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9969–9979, 2021.

[36] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 420–437. Springer, 2020.

[37] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7596, 2021.

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

[39] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13608–13617, 2021.

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

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

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

[43] Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021.

[44] Maheen Rashid, Hedvig Kjellstrom, and Yong Jae Lee. Action graphs: Weakly-supervised action localization with graph convolution networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 615–624, 2020.

[45] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1019, 2020.

[46] Dingfeng Shi, Yujie Zhong, Qiong Cao, Jing Zhang, Lin Ma, Jia Li, and Dacheng Tao. React: Temporal action detection with relational queries. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 105–121. Springer, 2022.

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

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

[49] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017.

[50] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13739–13748, 2021.

[51] Haisheng Su, Xu Zhao, and Tianwei Lin. Cascaded pyramid mining network for weakly supervised temporal action localization. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 558–574. Springer, 2019.

[52] Haisheng Su, Xu Zhao, Tianwei Lin, Shuming Liu, and Zhilan Hu. Transferable knowledge-based multi-granularity fusion network for weakly supervised temporal action detection. *IEEE Transactions on Multimedia*, 23:1503–1515, 2020.

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

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

[55] Hongsheng Xu, Zihan Chen, Yu Zhang, Xin Geng, Siya Mi, and Zhihong Yang. Weakly supervised temporal action localization with proxy metric modeling. *Frontiers of Computer Science*, 17(2):172309, 2023.

[56] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020.

[57] Le Yang, Junwei Han, Tao Zhao, Tianwei Lin, Dingwen Zhang, and Jianxin Chen. Background-click supervision for temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9814–9829, 2021.

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

[59] Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3090–3098, 2022.

[60] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 28(12):5797–5808, 2019.

[61] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 37–54. Springer, 2020.

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

[63] Chengwei Zhang, Yunlu Xu, Zhanzhan Cheng, Yi Niu, Shiliang Pu, Fei Wu, and Futai Zou. Adversarial seeded sequence growing for weakly-supervised temporal action localization. In *Proceedings of the 27th ACM international conference on multimedia*, pages 738–746, 2019.

[64] Xiao-Yu Zhang, Changsheng Li, Haichao Shi, Xiaobin Zhu, Peng Li, and Jing Dong. Adapnet: Adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization. *IEEE transactions on neural networks and learning systems*, 2020.

[65] Xiao-Yu Zhang, Hai-Chao Shi, Chang-Sheng Li, and Li-Xin Duan. Twinnet: twin structured knowledge transfer network for weakly supervised action localization. *Machine Intelligence Research*, 19(3):227–246, 2022.

[66] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021.

[67] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 35–44, 2018.

[68] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525, 2021.