Proposal-based Multiple Instance Learning for Weakly-supervised Temporal Action Localization

Huan Ren[1], Wenfei Yang[1], Tianzhu Zhang[1, 2, *], Yongdong Zhang[1]
[1]University of Science and Technology of China, [2]Deep Space Exploration Lab
rh_hr_666@mail.ustc.edu.cn, {yangwf, tzzhang, zhyd73}@ustc.edu.cn

Abstract

Weakly-supervised temporal action localization aims to localize and recognize actions in untrimmed videos with only video-level category labels during training. Without instance-level annotations, most existing methods follow the Segment-based Multiple Instance Learning (S-MIL) framework, where the predictions of segments are supervised by the labels of videos. However, the objective for acquiring segment-level scores during training is not consistent with the target for acquiring proposal-level scores during testing, leading to suboptimal results. To deal with this problem, we propose a novel Proposal-based Multiple Instance Learning (P-MIL) framework that directly classifies the candidate proposals in both the training and testing stages, which includes three key designs: 1) a surrounding contrastive feature extraction module to suppress the discriminative short proposals by considering the surrounding contrastive information, 2) a proposal completeness evaluation module to inhibit the low-quality proposals with the guidance of the completeness pseudo labels, and 3) an instance-level rank consistency loss to achieve robust detection by leveraging the complementarity of RGB and FLOW modalities. Extensive experimental results on two challenging benchmarks including THUMOS14 and ActivityNet demonstrate the superior performance of our method. Our code is available at github.com/RenHuan1999/CVPR2023_P-MIL.

1. Introduction

Temporal Action Localization (TAL) is one of the essential tasks in video understanding, which aims to simultaneously discover action instances and identify their categories in untrimmed videos [9, 18]. TAL has recently received increasing attention from the research community due to its broad application potentials, such as intelligent surveillance [44], video summarization [22], highlight detection [49], and visual question answering [23]. Most existing methods handle this task in a fully-supervised setting [7, 25, 42, 51, 63], which requires instance-level annotations. Despite their success, the requirements for such massive instance-level annotations limit their application in real-world scenarios. To overcome this limitation, Weakly-supervised Temporal Action Localization (WTAL) has been widely studied because it only requires video-level labels [13, 26, 36, 45, 55], which are much easier to collect.

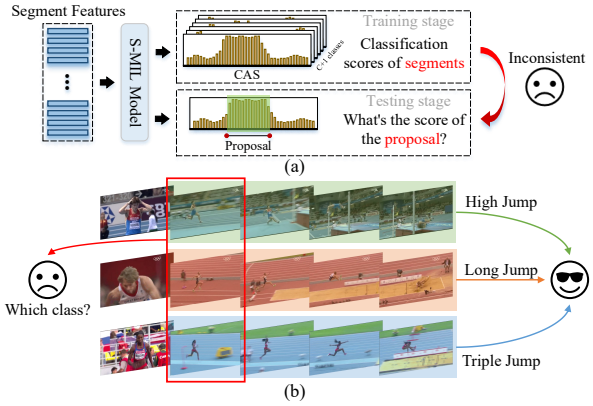

Figure 1. Drawbacks of the Segment-based Multiple Instance Learning framework. (a) The objectives of the training and testing stages are inconsistent. (b) By watching a single running segment in the red box, it is difficult to tell which category it belongs to.

Most existing WTAL methods follow the Segment-based Multiple Instance Learning (S-MIL) framework [16, 32, 45], where the predictions of segments are supervised by the labels of videos. In particular, a class-agnostic attention branch is used to calculate the attention sequence, which indicates the foreground probability of each segment. Meanwhile, a classification branch is used to calculate the Class Activation Sequence (CAS), which indicates the category probability of each segment. In the training stage, the video-level classification scores can be derived by aggregating CAS with the attention sequence, which are then supervised by the video-level category labels. In the testing stage, the candidate proposals are generated by thresholding the attention sequence, and the segment-level CAS corresponding to each proposal is aggregated to score each proposal.

*Corresponding Author

Despite the considerable progress achieved by these methods, the S-MIL framework has two drawbacks. Firstly, the objectives of the training and testing stages are inconsistent. As shown in Figure 1 (a), the target is to score the action *proposals* as a whole in the testing stage, but the classifier is trained to score the *segments* in the training stage. The inconsistent scoring approach can lead to suboptimal results as shown in other weakly-supervised tasks [1,34,43,52,56]. Secondly, it is difficult to classify each segment alone in many cases. As shown in Figure 1 (b), by watching a single running segment, it is difficult to tell whether it belongs to high jump, long jump, or triple jump. Only by watching the entire action instance and using of the contextual information, we can determine which category it belongs to.

Inspired by the above discussions, we propose a novel Proposal-based Multiple Instance Learning (P-MIL) framework, which employs a two-stage training pipeline. In the first stage, an S-MIL model is trained and the candidate proposals are generated by thresholding the attention sequence. In the second stage, the candidate proposals are classified and aggregated into video-level classification scores, which are supervised by video-level category labels. Since the candidate proposals are directly classified in both the training and testing stages, the proposed method can effectively handle the drawbacks of the S-MIL framework. However, there are three issues that need to be considered within the P-MIL framework. First, the model tends to focus on discriminative short proposals. Since the training stage is mainly guided by the video-level classification, the classifier tends to focus on the most discriminative proposals to minimize the classification loss. To solve this problem, we propose a Surrounding Contrastive Feature Extraction module. Specifically, we extend the boundaries of the candidate proposals and then calculate the outer-inner contrastive features of the proposals. By taking surrounding contrastive information into consideration, those discriminative short proposals can be effectively suppressed. Second, the candidate proposals generated by the S-MIL approach may be over-complete, which include irrelevant background segments. In this regard, we present a Proposal Completeness Evaluation module. Concretely speaking, we treat the high-confidence proposals as *pseudo instances*, and then acquire the completeness pseudo label of each proposal by computing the Intersection over Union (IoU) with these pseudo instances. Under the guidance of the completeness pseudo labels, the activation of low-quality proposals can be inhibited. Third, due to the Non-Maximum Suppression (NMS) process in the testing stage, the relative scores of proposals belonging to the same action instance have substantial influences on the detection results. To learn robust relative scores, we design an Instance-level Rank Consistency loss by leveraging the complementarity of RGB and FLOW modalities [55, 60]. Those proposals that overlap with a given candidate proposal are considered as a cluster. By constraining the normalized relative scores within the cluster between RGB and FLOW modalities to be consistent, we can achieve reliable detection by discarding proposals with low relative scores in the NMS process.

To sum up, the main contributions of this paper are as follows: (1) We propose a novel Proposal-based Multiple Instance Learning (P-MIL) framework for weakly-supervised temporal action localization, which can handle the drawbacks of the S-MIL framework by directly classifying the candidate proposals in both the training and testing stages. (2) We propose three key designs (Surrounding Contrastive Feature Extraction module, Proposal Completeness Evaluation module, Instance-level Rank Consistency loss), which can deal with the challenges in different stages of the P-MIL framework. (3) Extensive experimental results on THUMOS14 and ActivityNet datasets demonstrate the superior performance of the proposed framework over state-of-the-art methods.

2. Related Work

In this section, we briefly overview methods relevant to fully-supervised and weakly-supervised temporal action localization.

Fully-supervised Temporal Action Localization. Temporal Action Localization (TAL) aims to simultaneously localize and identify action instances in untrimmed videos. Similar to the development of object detection [5, 11, 27, 39], existing fully-supervised approaches can be divided into two categories: two-stage methods [7, 42, 50, 59, 63, 65] and one-stage methods [3, 8, 27, 30, 47, 53, 62]. Two-stage methods first generate the candidate proposals and then feed them into action classifiers, by improving either the quality of proposals [7,42,59,65] or the robustness of classifiers [50,63]. One-stage methods can instead generate the candidate proposals and classify them simultaneously, which have achieved remarkable performance recently by introducing Transformer architecture [8,47,62]. Despite their success, the requirements for massive and expensive instance-level annotations limit their application in real-world scenarios.

Weakly-supervised Temporal Action Localization. To solve the above issue, Weakly-supervised Temporal Action Localization (WTAL) has been widely studied [13, 26, 31, 36, 45, 54, 57, 60], which requires only video-level category labels. UntrimmedNet [45] is the first to introduce a Multiple Instance Learning (MIL) framework [33] to handle the WTAL task by classifying segments and using a selection module to generate the action proposals. However, due to the discrepancy between the classification and localization tasks, the model tends to focus on the most discriminative segments. Step-by-step [64] allows the model to learn more complete localization by gradually erasing the

most discriminative segments during training. WTALC [38] learns compact intra-class feature representations by pulling features of the same category to be closer while pushing those of different categories away. By introducing a class-agnostic attention branch for foreground-background separation, the attention-based approaches [15,21,26,31,36] become mainstream due to their superior performance and the flexibility of the model architectures. STPN [36] presents a sparsity loss on the attention sequence to capture the key foreground segments. CMCS [26] adopts a multi-branch architecture to discover distinctive action parts with a well-designed diversity loss. BaS-Net [21] and WSAL-BM [37] introduce an additional background category for explicit background modeling. CO_2-Net [15] designs a cross-modal attention mechanism to enhance features by filtering out task-irrelevant redundant information. More recently, there are some researches [17, 32, 55, 60] that attempt to generate pseudo labels to guide the model training. In [32], the class-agnostic attention sequence and class activation sequence provide pseudo labels for each other in an expectation-maximization framework. TSCN [60] fuses the attention sequence of RGB and FLOW modalities to generate segment-level pseudo labels, while UGCT [55] uses the RGB and FLOW modalities to generate pseudo labels for each other by leveraging their complementarity. Differently, ASM-Loc [13] uses pseudo labels not only to supervise the model training but also to enhance the segment features by leveraging the action proposals for segment-level temporal modeling. Despite the considerable progress achieved by previous methods, they almost all follow the *Segment*-based MIL framework to achieve temporal action localization, with inconsistent objectives between the training and testing stages. To deal with this issue, we instead propose a novel *Proposal*-based MIL framework to directly classify the candidate proposals in both the training and testing stages.

3. Our Method

In this section, we elaborate on the proposed Proposal-based Multiple Instance Learning framework (P-MIL) for Weakly-supervised Temporal Action Localization (WTAL), as illustrated in Figure 2. Given a video $\mathbf{V}$, the goal of WTAL is to predict a set of action instances $\{(c_i, s_i, e_i, q_i)\}_{i=1}^{M_p}$, where s_i and e_i denote the start time and end time of the i-th action, c_i and q_i represent the action category and the confidence score, respectively. During training, each video $\mathbf{V}$ has its ground-truth video-level category label $\boldsymbol{y} \in \mathbb{R}^C$, where C represents the number of action classes. $\boldsymbol{y}(j) = 1$ if the j-th action presents in the video and $\boldsymbol{y}(j) = 0$ otherwise. The proposed P-MIL framework consists of three steps, including candidate proposal generation (Sec. 3.1), proposal feature extraction and classification (Sec. 3.2), proposal refinement (Sec. 3.3). The details are as follows.

3.1. Candidate Proposal Generation

In order to generate the candidate proposals, an S-MIL model [15] is trained. We first divide each untrimmed video into non-overlapping 16-frame segments, and then apply the pretrained feature extractor (*e.g.* I3D [6]) to extract segment-wise features $\mathbf{X}_S \in \mathbb{R}^{T\times D}$ for both RGB and FLOW modalities, where T indicates the number of segments in the video and D is the feature dimension. Following the typical two-branch architecture, a category-agnostic attention branch is utilized to calculate the attention sequence $\mathbf{A} \in \mathbb{R}^{T\times 1}$ and a classification branch is used to predict the base Class Activation Sequence (CAS) $\mathbf{S}_{base} \in \mathbb{R}^{T\times(C+1)}$, where the $(C+1)$-th class indicates the background [21, 37]. By multiplying $\mathbf{S}_{base}$ with $\mathbf{A}$ in the temporal dimension, we can obtain the background suppressed CAS $\mathbf{S}_{supp} \in \mathbb{R}^{T\times(C+1)}$. After that, the predicted video-level classification scores $\hat{\boldsymbol{y}}_{base}, \hat{\boldsymbol{y}}_{supp} \in \mathbb{R}^{C+1}$ are derived by applying a temporal top-k aggregation strategy [13, 19, 21] to $\mathbf{S}_{base}$ and $\mathbf{S}_{supp}$, respectively, followed by a softmax operation.

Guided by the video-level category label $\boldsymbol{y}$, the classification loss is formulated as

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C+1}\Big(\boldsymbol{y}_{base}(c)\log\hat{\boldsymbol{y}}_{base}(c) + \boldsymbol{y}_{supp}(c)\log\hat{\boldsymbol{y}}_{supp}(c)\Big), \tag{1}$$

where $\boldsymbol{y}_{base} = [\boldsymbol{y}, 1] \in \mathbb{R}^{C+1}$ and $\boldsymbol{y}_{supp} = [\boldsymbol{y}, 0] \in \mathbb{R}^{C+1}$, based on the assumption that background is present in all videos but filtered out by the attention sequence $\mathbf{A}$. Furthermore, a sparsity loss [36] $\mathcal{L}_{norm} = \frac{1}{T}\sum_{t=1}^{T}|\mathbf{A}(t)|$ is also employed on the attention sequence $\mathbf{A}$ to focus on the key foreground segments. Overall, the training objectives are as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{norm}\mathcal{L}_{norm}, \tag{2}$$

where λ_{norm} is a balancing factor.

With the trained S-MIL model, we apply multiple thresholds θ_{act} on the attention sequence $\mathbf{A}$ to generate the candidate action proposals $P_{act} = \{(s_i, e_i)\}_{i=1}^{M_1}$. To enable our P-MIL model to learn better foreground-background separation in the training stage, we also apply extra thresholds θ_{bkg} to generate the background proposals $P_{bkg} = \{(s_i, e_i)\}_{i=1}^{M_2}$, where the attention sequence $\mathbf{A}$ is below θ_{bkg}. Thus, the final candidate proposals for training are formulated as

$$P = P_{act} + P_{bkg} = \{(s_i, e_i)\}_{i=1}^{M}, \tag{3}$$

where $M = M_1 + M_2$ indicates the total number of the candidate proposals. Note that we only use the action proposals P_{act} for inference.

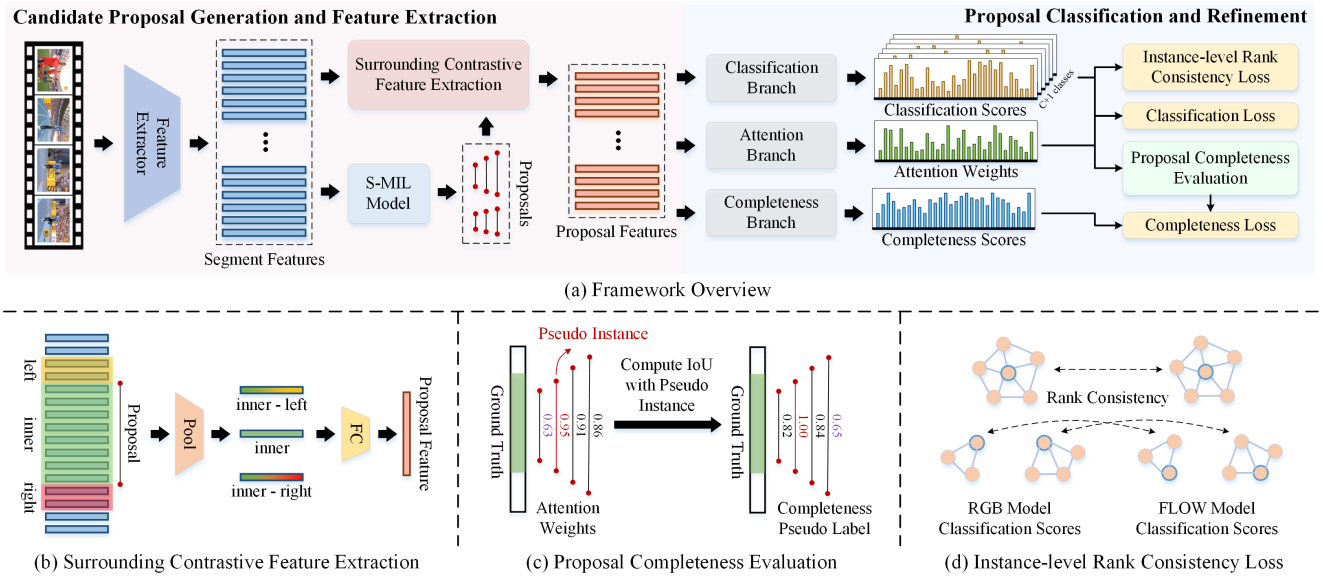

Figure 2. (a) Overview of the proposed Proposal-based Multiple Instance Learning framework, which consists of candidate proposal generation, proposal feature extraction, proposal classification and refinement. (b) The Surrounding Contrastive Feature Extraction (SCFE) module extends the boundaries of the candidate proposals and then calculates the outer-inner contrastive features of the candidate proposals. (c) The Proposal Completeness Evaluation (PCE) module generates the completeness pseudo label by computing the IoU with the selected pseudo instances. (d) The Instance-level Rank Consistency (IRC) loss constrains the normalized relative classification scores within the cluster between RGB and FLOW modalities to be consistent.

3.2. Proposal Feature Extraction and Classification

Given the candidate proposals P, previous S-MIL methods use the CAS to calculate the confidence score (*e.g.* Outer-Inner Score [41]) of each proposal. However, these indirect scoring approaches can lead to suboptimal results. To address this problem, we propose to directly classify the candidate proposals and aggregate them into video-level classification scores, which are supervised by video-level category labels.

Surrounding Contrastive Feature Extraction. For the given candidate proposals P, we first extract corresponding proposal features $\mathbf{X}_P \in \mathbb{R}^{M\times D}$. Since the training stage is mainly guided by the video-level classification, the classifier tends to focus on discriminative short proposals to minimize the classification loss. To address this issue, we propose a Surrounding Contrastive Feature Extraction (SCFE) module. Specifically, given a candidate proposal $P_i = (s_i, e_i)$, we first extend the boundary by α of its length on both the left and right sides, yielding three regions: *left*, *inner*, and *right*. For each region, we then employ RoIAlign [14] followed by max-pooling on the segment features $\mathbf{X}_S$ to extract an associated D-dimensional feature vector, indicated by $\mathbf{X}_i^l$, $\mathbf{X}_i^n$ and $\mathbf{X}_i^r$, respectively. An intuitive way to obtain the proposal feature is to directly concatenate the three feature vectors and feed them into a fully connected layer. However, inspired by AutoLoc [41], we take a more effective approach that calculates the outer-inner contrastive features of the proposal followed by a fully connected layer, which is written as:

$$\mathbf{X}_i = FC\big(\text{Cat}(\mathbf{X}_i^n - \mathbf{X}_i^l, \mathbf{X}_i^n, \mathbf{X}_i^n - \mathbf{X}_i^r)\big), \quad (4)$$

where Cat denotes the concatenate operation. By taking the surrounding contrastive information into consideration, those discriminative short proposals can be effectively suppressed.

Classification Head. Similar to the pipeline of the S-MIL framework, given the proposal features $\mathbf{X}_P$, a category-agnostic attention branch is then used to predict the attention weights $\mathbf{A} \in \mathbb{R}^{M\times 1}$, which indicate the foreground probability of each proposal. Meanwhile, a classification branch is used to predict the base classification scores $\mathbf{S}_{base} \in \mathbb{R}^{M\times(C+1)}$ of the proposals. By multiplying $\mathbf{S}_{base}$ with $\mathbf{A}$, we obtain the background suppressed classification scores $\mathbf{S}_{supp} \in \mathbb{R}^{M\times(C+1)}$. Finally, the predicted video-level classification scores $\hat{\boldsymbol{y}}_{base}, \hat{\boldsymbol{y}}_{supp} \in \mathbb{R}^{C+1}$ are derived by applying a top-k pooling followed by softmax to $\mathbf{S}_{base}$ and $\mathbf{S}_{supp}$, respectively, which are supervised by the video-level category labels.

3.3. Proposal Refinement

Proposal Completeness Evaluation. The candidate proposals generated by the S-MIL method may be over-complete, which include irrelevant background segments.

In this regard, we present a Proposal Completeness Evaluation (PCE) module. Given the candidate proposals, we use the attention weights to select the high-confidence proposals as *pseudo instances*, and then obtain the completeness pseudo label of each proposal by computing the Intersection over Union (IoU) with these pseudo instances. Formally, we first apply a threshold $\gamma \cdot max(\mathbf{A})$ (γ is set to 0.8 in our case) to the attention weights $\mathbf{A}$ of proposals to select a set of high-confidence proposals Q. Then, following the Non-Maximum Suppression (NMS) process, we select the proposal with the highest attention weight as the pseudo instance, remove the proposals that overlap with it from Q, and repeat the process until Q is empty. After that, we acquire a set of pseudo instances $G = \{(s_i, e_i)\}_{i=1}^{N}$. By computing the IoU between the candidate proposals P and the pseudo instances G, we can obtain an IoU matrix of $M \times N$ dimensions. We assign the pseudo instance with the largest IoU to each proposal via taking maximum for the N dimension, and then we obtain the completeness pseudo labels $\boldsymbol{q} \in \mathbb{R}^M$ for the candidate proposals. Under the guidance of $\boldsymbol{q}$, a completeness branch is introduced to predict the completeness scores $\hat{\boldsymbol{q}} \in \mathbb{R}^M$ in parallel with the attention branch and the classification branch, which can help to inhibit the activation of low-quality proposals.

Instance-level Rank Consistency. Due to the NMS process in the testing stage, the relative scores of the candidate proposals belonging to the same action instance have a significant impact on the detection results. In order to learn robust relative scores, we design an Instance-level Rank Consistency (IRC) loss by leveraging the complementarity of RGB and FLOW modalities. In detail, we first apply a threshold $mean(\mathbf{A})$ to the attention sequence $\mathbf{A}$ to eliminate the low-confidence proposals, and the remaining proposals are denoted as R. For each proposal r in R, those candidate proposals that overlap with it are considered as a cluster Ω_r, where $|\Omega_r| = N_r$. The classification scores $\mathbf{S}_{base}$ corresponding to this cluster are indexed from the RGB and FLOW modalities, given as $p_{r,c}^{RGB}$ and $p_{r,c}^{FLOW}$, respectively, where c indicates one of the ground truth categories. Then the normalized relative scores within the cluster are formulated as

$$D_{r,c}^{*} = \mathrm{softmax}(p_{r,c}^{*}), \forall * \in \{RGB, FLOW\}. \quad (5)$$

The Kullback-Leibler (KL) divergence is used to constrain the consistency between RGB and FLOW modalities, defined as:

$$\begin{aligned}\mathcal{L}_{IRC} = \frac{1}{|R|}\sum_{r\in R}\Big(&\mathrm{KL}(D_{r,c}^{FLOW}||D_{r,c}^{RGB}) \\ &+\mathrm{KL}(D_{r,c}^{RGB}||D_{r,c}^{FLOW})\Big),\end{aligned} \quad (6)$$

$$\mathrm{KL}(D_{r,c}^{t}||D_{r,c}^{s}) = -\sum_{i=1}^{N_r} D_{r,c}^{t}(i)\log\frac{D_{r,c}^{s}(i)}{D_{r,c}^{t}(i)}. \quad (7)$$

With the IRC loss, we can achieve reliable detection by discarding proposals with low relative scores in the NMS process.

3.4. Network Training and Inference

Network Training. In the training stage, guided by the video-level category label $\boldsymbol{y}$, the main classification loss is formulated as

$$\begin{aligned}\mathcal{L}_{cls} = -\sum_{c=1}^{C+1}\big(&\boldsymbol{y}_{base}(c)\log\hat{\boldsymbol{y}}_{base}(c) \\ &+\boldsymbol{y}_{supp}(c)\log\hat{\boldsymbol{y}}_{supp}(c)\big),\end{aligned} \quad (8)$$

where $\boldsymbol{y}_{base} = [\boldsymbol{y}, 1] \in \mathbb{R}^{C+1}$ and $\boldsymbol{y}_{supp} = [\boldsymbol{y}, 0] \in \mathbb{R}^{C+1}$. Moreover, with the PCE module, a completeness loss is defined as the Mean Square Error (MSE) between the completeness pseudo labels $\boldsymbol{q}$ and the predicted completeness scores $\hat{\boldsymbol{q}}$:

$$\mathcal{L}_{comp} = \frac{1}{M}\sum_{i=1}^{M}\big(\boldsymbol{q}(i) - \hat{\boldsymbol{q}}(i)\big)^2. \quad (9)$$

Overall, the training objective of our model is

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{comp}\mathcal{L}_{comp} + \lambda_{IRC}\mathcal{L}_{IRC}, \quad (10)$$

where λ_{comp} and λ_{IRC} are balancing hyper-parameters.

Inference. In the testing stage, we first apply the threshold θ_{cls} to the video-level classification scores $\hat{\boldsymbol{y}}_{supp}$ and neglect those categories below θ_{cls}. For each remaining category c, we score the i-th candidate proposal as

$$\boldsymbol{s}(i) = \mathbf{S}_{supp}(i, c) * \hat{\boldsymbol{q}}(i). \quad (11)$$

Finally, the class-wise soft-NMS [2] is employed to remove the duplicate proposals.

3.5. Discussions

In this section, we discuss the differences between the proposed method and several relevant methods, including AutoLoc [41] and CleanNet [29]. To deal with the inconsistency between the *localization* objective of the testing stage and the *classification* objective of the training stage, AutoLoc and CleanNet propose to directly predict the temporal boundaries of action instances, with the supervision provided by the Outer-Inner-Contrastive loss and the temporal contrast loss, respectively. Different from these approaches, we concentrate on another inconsistency in the S-MIL framework about what to score between the training and testing stages. The candidate *proposals* need to be scored in the testing stage, while the S-MIL classifier is trained to score the *segments* during training. To resolve this inconsistency, we propose a novel Proposal-based Multiple Instance Learning framework that directly classifies the candidate proposals in both the training and testing stages.

Table 1. Detection performance comparison with state-of-the-art methods on the THUMOS14 test set. TSN, UNT and I3D represent TSN [46], UntrimmedNet [45] and I3D [6] features, respectively. $*$ means fusing the detection results of the S-MIL and our P-MIL model.

Supervision	Methods	Feature	mAP@IoU (%)							AVG mAP (%)		
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.5	0.3:0.7	0.1:0.7
Fully	TAL-Net [7], CVPR2018	I3D	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
	BMN [25], ICCV2019	TSN	-	-	56.0	47.4	38.8	29.7	20.5	-	38.5	-
	GTAD [51], CVPR2020	TSN	-	-	54.5	47.6	40.3	30.8	23.4	-	39.8	-
	ContextLoc [65], ICCV2021	I3D	-	-	68.3	63.8	54.3	41.8	26.2	-	50.9	-
	RefactorNet [48], CVPR2022	I3D	-	-	70.7	65.4	58.6	47.0	32.1	-	54.8	-
Weakly	AutoLoc [41], ECCV2018	UNT	-	-	35.8	29.0	21.2	13.4	5.8	-	21.0	-
	CleanNet [29], ICCV2019	UNT	-	-	37.0	30.9	23.9	13.9	7.1	-	22.6	-
	STPN [36], CVPR2018	I3D	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	18.5	27.0
	WTALC [38], ECCV2018	I3D	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-	-
	CMCS [26], CVPR2019	I3D	57.4	50.8	41.2	32.1	23.1	15.0	7.0	40.9	23.7	32.4
	WSAL-BM [37], ICCV2019	I3D	60.4	56.0	46.6	37.5	26.8	19.6	9.0	45.5	27.9	36.6
	DGAM [40], CVPR2020	I3D	60.0	54.2	46.8	38.2	28.8	19.8	11.4	45.6	29.0	37.0
	EM-MIL [32], ECCV2020	I3D	59.1	52.7	45.5	36.8	30.5	22.7	**16.4**	44.9	30.4	37.7
	TSCN [60], ECCV2020	I3D	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	28.8	37.8
	CoLA [61], CVPR2021	I3D	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	AUMN [31], CVPR2021	I3D	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4	41.5
	UGCT [55], CVPR2021	I3D	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	CO_2-Net [15], MM2021	I3D	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
	D2-Net [35], ICCV2021	I3D	65.7	60.2	52.3	43.4	36.0	-	-	51.5	-	-
	FAC-Net [16], ICCV2021	I3D	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	33.1	42.2
	FTCL [10], CVPR2022	I3D	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
	RSKP [17], CVPR2022	I3D	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	ASM-Loc [13], CVPR2022	I3D	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	DCC [24], CVPR2022	I3D	69.0	63.8	55.9	45.9	35.7	24.3	13.7	54.1	35.1	44.0
	ours	I3D	70.9	66.6	57.8	48.6	39.8	27.1	14.4	56.8	37.5	46.5
	ours*	I3D	**71.8**	**67.5**	**58.9**	**49.0**	**40.0**	**27.1**	15.1	**57.4**	**38.0**	**47.0**

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our method on two benchmark datasets including THUMOS14 [18] and ActivityNet [4]. **THUMOS14** dataset contains 200 validation videos and 213 testing videos from 20 categories. Following previous works [21, 36, 38], we use the validation set for training and the testing set for evaluation. **ActivityNet** dataset includes two versions, ActivityNet1.2 and ActivityNet1.3, with 9,682 videos from 100 categories and 19,994 videos from 200 categories, respectively. The training, validation and testing sets are divided from ActivityNet dataset by the ratio of 2:1:1. Following previous works [29, 31, 55], we use the training set for training and the validation set for evaluation.

Evaluation Metrics. In this work, we evaluate the localization performance with the mean Average Precision (mAP) values at different Intersection over Union (IoU) thresholds, following the standard evaluation protocol [1].

4.2. Implementation Details

Network Architecture. We employ the I3D [6] networks pretrained on Kinetics-400 [6] for feature extraction. The dimension D of the extracted segment-wise features is 1024. Optical flow frames are extracted from RGB frames using the TV-L1 [58] algorithm. The category-agnostic attention branch is implemented by two fully-connected layers followed by a sigmoid activation function, which is the same as the completeness branch. The classification branch consists of two fully-connected layers.

Hyper-parameters Setting. Our method is trained using the Adam [20] optimizer with the learning rate of 5×10^{-5} and the mini-batch size of 10. The extended ratio α is set to 0.25. The RoI size of RoIAlign [14] is 2, 8, 2 for the left, inner and right region, respectively. The loss function weights λ_{comp} = 20, λ_{IRC} = 2 in Equation (10). Since the attention weights are less reliable in the early training stage, we multiply the loss function $\mathcal{L}_{comp}$ and $\mathcal{L}_{IRC}$ with a time-varying parameter as employed in UGCT [55], which is gradually increased to 1. For the candidate proposal generation, the thresholds θ_{act} = [0.1:0.1:0.9] and θ_{bkg} = [0.3:0.2:0.7]. During inference, the video-level classification threshold θ_{cls} is set to 0.2.

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

Results on THUMOS14. Table 1 shows the comparison of our method with weakly-supervised and several fully-supervised methods on the THUMOS14 dataset. From the results we can see that our method outperforms the prior

[1] http://github.com/activitynet/ActivityNet

Table 2. Detection performance comparison with state-of-the-art methods on the ActivityNet1.2 validation set. AVG represents the average mAP at IoU thresholds 0.5:0.05:0.95. $*$ means fusing the detection results of the S-MIL model and our P-MIL model.

Methods	mAP@IoU (%)			
	0.5	0.75	0.95	AVG
WTALC [38], ECCV2018	37.0	14.6	-	18.0
CMCS [26], CVPR2019	36.8	22.0	5.6	22.4
BaS-Net [21], AAAI2020	38.5	24.2	5.6	24.3
TCAM [12] CVPR2020	40.0	25.0	4.6	24.6
DGAM [40], CVPR2020	41.0	23.5	5.3	24.4
EM-MIL [32], ECCV2020	37.4	23.1	2.0	20.3
TSCN [60], ECCV2020	37.6	23.7	5.7	23.6
CoLA [61], CVPR2021	42.7	25.7	5.8	26.1
AUMN [31], CVPR2021	42.0	25.0	5.6	25.5
UGCT [55], CVPR2021	41.8	25.3	**5.9**	25.8
CO_2-Net [15], MM2021	43.3	**26.3**	5.2	26.4
D2-Net [35], ICCV2021	42.3	25.5	5.8	26.0
ours	42.2	25.0	4.9	25.5
ours*	**44.2**	26.1	5.3	**26.5**

Table 3. Detection performance comparison with state-of-the-art methods on the ActivityNet1.3 validation set. AVG represents the average mAP at IoU thresholds 0.5:0.05:0.95. $*$ means fusing the detection results of the S-MIL model and our P-MIL model.

Methods	mAP@IoU (%)			
	0.5	0.75	0.95	AVG
STPN [36], CVPR2018	29.3	16.9	2.6	16.3
CMCS [26], CVPR2019	34.0	20.9	5.7	21.2
WSAL-BM [37], ICCV2019	36.4	19.2	2.9	19.5
TSCN [60], ECCV2020	35.3	21.4	5.3	21.7
TS-PCA [28], CVPR2021	37.4	23.5	5.9	23.7
AUMN [31], CVPR2021	38.3	23.5	5.2	23.5
UGCT [55], CVPR2021	39.1	22.4	5.8	23.8
FAC-Net [16], ICCV2021	37.6	24.2	6.0	24.0
FTCL [10], CVPR2022	40.0	24.3	**6.4**	24.8
RSKP [17], CVPR2022	40.6	24.6	5.9	25.0
ASM-Loc [13], CVPR2022	41.0	24.9	6.2	25.1
DCC [24], CVPR2022	38.8	24.2	5.7	24.3
ours	39.5	23.6	4.9	23.9
ours*	**41.8**	**25.4**	5.2	**25.5**

state-of-the-art weakly-supervised methods, and by fusing the detection results of the S-MIL model and our P-MIL model, we can even achieve better performance. Specifically, our method surpasses the previous best performance by 1.5% and 1.4% in terms of the mAP@0.5 and the average mAP@0.1:0.7, respectively, and further widens the gap to 1.7% and 1.9% after fusion. Even when compared to certain fully-supervised methods (*e.g.* BMN [25] and GTAD [51]), our model can achieve comparable results at low IoU thresholds.

Results on ActivityNet. Table 2 and Table 3 show the performance comparison on the larger ActivityNet1.2 and ActivityNet1.3 datasets, respectively. The experimental results are consistent with those on the THUMOS14 dataset and we achieve state-of-the-art performance. Specifically, after fusion, we achieve 26.5% on the ActivityNet1.2 dataset and 25.5% on the ActivityNet1.3 dataset in terms of the average mAP.

4.4. Ablation Studies

To analyze the impact of each design, we conduct a series of ablation studies on the THUMOS14 dataset, as detailed below.

Proposal Generation. Table 4 shows the impact of different candidate proposal generation methods on the final detection performance. To enable our P-MIL model to learn better foreground-background separation during training, we generate additional background proposals to fill the candidate proposals in Equation (3). In order to validate the effectiveness of introducing the background proposals in the training stage, we keep the candidate proposals used for testing to be consistent, which consist of only the action proposals. From Table 4, we can see that when only the action proposals are used for training, the average mAP is 41.2%. After incorporating the background proposals into the training stage, the average mAP increases by 5.3% to 46.5%, which significantly demonstrates the effectiveness of this design.

Proposal Scoring. Table 5 shows the impact of different proposal scoring approaches on the detection performance. To evaluate the upper bound of the detection performance, we use the IoU with the ground truth to score the candidate proposals. The results indicate that the localization quality of the candidate proposals is already high enough and the bottleneck of the detection performance lies in the scoring of the candidate proposals. To evaluate the effectiveness of our P-MIL method compared to the S-MIL method, we apply different scoring approaches to the same set of candidate proposals. It can be observed that when the S-MIL method is used to score the candidate proposals, the average mAP is 43.6%. When we utilize our P-MIL method to score the candidate proposals, the average mAP increases by 2.9%. The results show that the direct scoring of proposals by our P-MIL method is better than the indirect scoring of proposals by the S-MIL method. Note that after fusing the detection results of the S-MIL model and our P-MIL model, the performance can be further improved to 47.0% in terms of the average mAP, indicating that the two methods can complement each other.

Proposal Feature Extraction. Table 6 shows the impact of different variants of proposal feature extraction on the detection performance. From the experimental results, it can be seen that the average mAP is only 41.0% when the boundaries of the candidate proposals are not extended. After extending the left and right boundaries of the candidate proposals, we can obtain the feature vectors of the three regions. However, simply concatenating these three feature vectors increases the average mAP by just 0.9%. When

Table 4. Ablation studies about the candidate proposal generation methods. Action and background indicate that the candidate proposals are generated by θ_{act} and θ_{bkg}, respectively.

Proposal Generation	mAP@IoU (%)				
	0.1	0.3	0.5	0.7	AVG
action	62.6	51.1	34.9	12.7	41.2
action + background	**70.9**	**57.8**	**39.8**	**14.4**	**46.5**

Table 5. Ablation studies about different proposal scoring approaches. GT denotes using the IoU with the ground truth to score the candidate proposals. And FUSE means fusing the detection results of the S-MIL model and our P-MIL model.

Proposal Scoring	mAP@IoU (%)				
	0.1	0.3	0.5	0.7	AVG
GT	83.3	76.3	64.7	41.7	67.4
S-MIL	68.1	54.1	37.2	12.4	43.6
P-MIL	70.9	57.8	39.8	14.4	46.5
FUSE	**71.8**	**58.9**	**40.0**	**15.1**	**47.0**

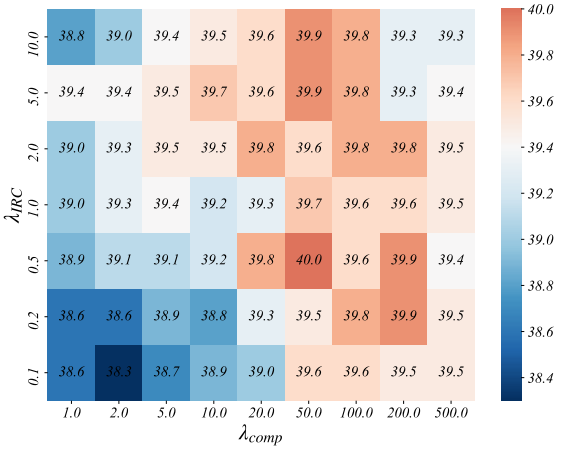

Figure 3. The affection of the coefficients for the completeness loss and the Instance-level Rank Consistency (IRC) loss. We show the mAP at the IoU threshold 0.5.

calculating the outer-inner contrastive features in Equation (4), the performance is significantly improved by 5.5% to 46.5%. These results validate the effectiveness of the Surrounding Contrastive Feature Extraction (SCFE) module.

Proposal Refinement. Table 7 shows the impact of the two designs of proposal refinement on the detection performance, including the Proposal Completeness Evaluation (PCE) module and the Instance-level Rank Consistency (IRC) loss. It can be observed that both designs can bring performance gain. Specifically, the PCE module and the IRC loss boost performance by 0.7% and 0.8% in terms of the average mAP, respectively, and when used together, the performance increases by 1.3%. The experimental results demonstrate the effectiveness of both designs.

Table 6. Ablation studies about different variants of proposal feature extraction. The results demonstrate the effectiveness of the Surrounding Contrastive Feature Extraction module.

Proposal Feature Extraction	mAP@IoU (%)				
	0.1	0.3	0.5	0.7	AVG
w/o extending	64.5	52.0	34.3	11.1	41.0
simply concatenate	65.8	53.1	35.1	11.2	41.9
outer-inner contrast	**70.9**	**57.8**	**39.8**	**14.4**	**46.5**

Table 7. Ablation studies about the two designs of proposal refinement. PCE and IRC denote the Proposal Completeness Evaluation module and the Instance-level Rank Consistency loss, respectively.

Proposal Refinement		mAP@IoU (%)				
PCE	IRC	0.1	0.3	0.5	0.7	AVG
		70.2	57.1	37.7	13.4	45.2
✓		70.4	57.6	38.7	14.5	45.9
	✓	70.6	58.0	39.0	13.8	46.0
✓	✓	**70.9**	**57.8**	**39.8**	**14.4**	**46.5**

Hyper-parameters Sensitivity Analysis. There are two hyper-parameters in our P-MIL method, including the coefficients λ_{comp} and λ_{IRC} of the loss function in Equation (10). To analyse the sensitivity of these hyper-parameters, we evaluate the performance change in terms of the mAP@0.5 for different combinations of λ_{comp} and λ_{IRC}. As shown in Figure 3, our model is not very sensitive to these two hyper-parameters, and the performance fluctuations in terms of mAP@0.5 are less than 2%. We set a moderate value for each of these two hyperparameters. Specifically, with $\lambda_{comp} = 20$ and $\lambda_{IRC} = 2$, our method achieves 39.8% in terms of the mAP@0.5.

5. Conclusion

In this paper, we propose a novel Proposal-based Multiple Instance Learning (P-MIL) framework for weakly-supervised temporal action localization, which can achieve the unified objectives of the training and testing stages by directly classifying the candidate proposals. We introduce three key designs to deal with the challenges in different stages of the P-MIL framework, including the surrounding contrastive feature extraction module, the proposal completeness evaluation module and the instance-level rank consistency loss. Extensive experimental results on two challenging benchmarks demonstrate the effectiveness of our method.

6. Acknowledgement

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62121002), and National Defense Basic Scientific Research Program of China (Grant JCKY2021130B016).

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 2

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5561–5569, 2017. 5

[3] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Proceedings of the British Machine Vision Conference*, volume 2, page 7, 2017. 2

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 6

[7] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 1, 2, 6

[8] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. In *Proceedings of the European Conference on Computer Vision*, pages 503–521, 2022. 2

[9] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35(11):2782–2795, 2013. 1

[10] 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. 6, 7

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2015. 2

[12] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian. Learning temporal co-attention models for unsupervised video action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2020. 7

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

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 4, 6

[15] 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. 3, 6, 7

[16] 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. 1, 6, 7

[17] 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. 3, 6, 7

[18] 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. 1, 6

[19] 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. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation Conference on Learning Representations*, 2014. 6

[21] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11320–11327, 2020. 3, 6, 7

[22] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012. 1

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018. 1

[24] 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. 6, 7

[25] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action pro-

posal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 1, 6, 7

[26] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 1, 2, 3, 6, 7

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016. 2

[28] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6176–6185, 2021. 7

[29] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3899–3908, 2019. 5, 6

[30] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019. 2

[31] 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. 2, 3, 6, 7

[32] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *European Conference on Computer Vision*, pages 729–745, 2020. 1, 3, 6, 7

[33] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, 10, 1997. 2

[34] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 2

[35] 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. 6, 7

[36] 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. 1, 2, 3, 6, 7

[37] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5502–5511, 2019. 3, 6, 7

[38] 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*, pages 563–579, 2018. 3, 6, 7

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 2

[40] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1009–1019, 2020. 6, 7

[41] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision*, pages 154–171, 2018. 4, 5, 6

[42] 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. 1, 2

[43] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017. 2

[44] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013. 1

[45] 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. 1, 2, 6

[46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36, 2016. 6

[47] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *European Conference on Computer Vision*, pages 358–375. Springer, 2022. 2

[48] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. Learning to refactor action and co-occurrence features for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2022. 6

[49] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection

from video duration. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1258–1267, 2019. 1

[50] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5783–5792, 2017. 2

[51] 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. 1, 6, 7

[52] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8381, 2019. 2

[53] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020. 2

[54] Wenfei Yang, Tianzhu Zhang, Zhendong Mao, Yongdong Zhang, Qi Tian, and Feng Wu. Multi-scale structure-aware network for weakly supervised temporal action detection. *IEEE Transactions on Image Processing*, 30:5848–5861, 2021. 2

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

[56] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021. 2

[57] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised and unsupervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[58] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223, 2007. 6

[59] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7094–7103, 2019. 2

[60] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus networks for weakly-supervised temporal action localization. In *Proceedings of the European Conference on Computer Vision*, August 2020. 2, 3, 6, 7

[61] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogintion*, 2021. 6, 7

[62] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 2022. 2

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

[64] 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 ACM Multimedia Conference on Multimedia Conference*, pages 35–44, 2018. 2

[65] 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. 2, 6