Two-Stream Networks for Weakly-Supervised Temporal Action Localization with Semantic-Aware Mechanisms

Yu Wang
Ant Group
Hangzhou, China
yuwangtj@yeah.net

Yadong Li
Ant Group
Hangzhou, China
liyadong.lyd@antgroup.com

Hongbin Wang
Ant Group
Hangzhou, China
hongbin.whb@alibaba-inc.com

Abstract

Weakly-supervised temporal action localization aims to detect action boundaries in untrimmed videos with only video-level annotations. Most existing schemes detect temporal regions that are most responsive to video-level classification, but they overlook the semantic consistency between frames. In this paper, we hypothesize that snippets with similar representations should be considered as the same action class despite the absence of supervision signals on each snippet. To this end, we devise a learnable dictionary where entries are the class centroids of the corresponding action categories. The representations of snippets identified as the same action category are induced to be close to the same class centroid, which guides the network to perceive the semantics of frames and avoid unreasonable localization. Besides, we propose a two-stream framework that integrates the attention mechanism and the multiple-instance learning strategy to extract fine-grained clues and salient features respectively. Their complementarity enables the model to refine temporal boundaries. Finally, the developed model is validated on the publicly available THUMOS-14 and ActivityNet-1.3 datasets, where substantial experiments and analyses demonstrate that our model achieves remarkable advances over existing methods.

1. Introduction

Temporal action localization (TAL) is committed to detecting action intervals in untrimmed videos. It has received increasing popularity recently due to its wide application in surveillance analysis, video summarization and retrieval [39, 44, 47], *etc*. Typically, fully-supervised TAL [51, 58, 59] is prohibitively expensive and unrealistic due to frame-level annotations, thus the weakly-supervised TAL (WS-TAL) [32, 34, 40, 45, 55] that only video-level annotations are required has been advocated recently.

Most WS-TAL methods [8, 9, 31, 43, 45, 46] transform localization into classification tasks that detect temporal regions contributing the most to video-level classification. They divide raw videos into fixed-length non-overlapping snippets, on which snippet-wise attention activations or class activation sequence (CAS) is generated. Temporal regions are detected by thresholding and merging these activations along the time dimension. Specifically, multiple instance learning (MIL) [36, 42] and the attention mechanism [11, 14, 53] are typically employed. The former aggregates snippets that are considered action instances with top-k confidence. However, such a regime over-emphasizes these snippets with top-k confidence, resulting in discarding potential clues in remaining snippets with lower confidence. Besides, MIL chooses the most discriminative snippets ignoring the completeness of action instances, which is incompatible with the localization task. Different from MIL, the attention-based mechanism independently yields class-agnostic confidence for each snippet, which is utilized as a weight to perform temporal pooling over all snippets and generate video-level representations for classification.

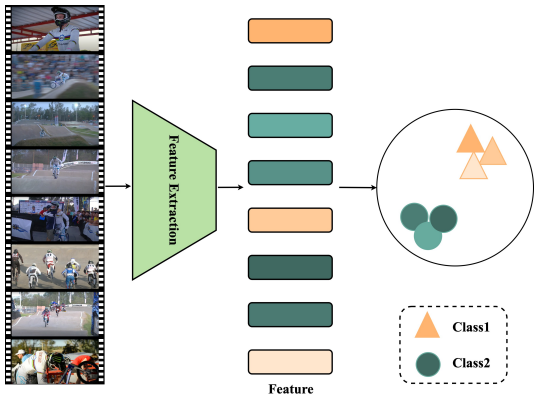

Figure 1. An example contains the action of "*Bicycle Motocross*" and the background. Representations depicted in monochromatic color are similar, which should be regarded to describe the same action (or background) and grouped together. The color brightness indicates the degree of similarity.

Despite the usage of all snippet features for fine-grained patterns, class-agnostic confidence is semantically ambiguous and harmful to precise boundary detection. As a consequence, we design a two-stream network that integrates MIL and attention-based mechanisms to overcome their respective drawbacks. Then a late-fusion operation on the outputs of the two branches is conducted to acquire the final classification results.

Furthermore, a crux of these works lies in accurately predicting confidence scores that each snippet belongs to the foreground or background, which has a nontrivial impact on the subsequent boundary regression. Since the weakly-supervision paradigm does not provide explicit supervision signals, this problem becomes more intractable. A common solution employs temporal class activation map [40] (TCAM) to discover snippets that respond to the video-level classification and assign them high confidence. Other alternatives [11, 14, 45, 53] attempt to mitigate this problem by carefully formulating some attention generation and aggregation mechanisms. Nevertheless, these strategies neglect the semantic consistency between snippets. Intuitively, snippets with similar representations should be considered to be the same class despite the infeasibility of accessing snippet-level annotations. An example is also illustrated in Figure 1. We argue that it is unreasonable that there are no constraints to guarantee such a semantic relation. To address this intractable issue, we set a learnable dictionary where entries are class centroids of the corresponding action categories. The representations of snippets identified as the same action are induced to be close to the same class centroid. In this manner, the semantic relationship of snippets is explicitly explored to encourage a reasonable localization in the weakly-supervised paradigm.

In a nutshell, the main contributions and innovations of this paper are summarized as follows: (1) A novel two-stream network that absorbs the merits of MIL and attention mechanism is proposed to resolve WS-TAL. (2) To perceive semantic information, a learnable dictionary with euclidean constraint is designed to facilitate similar representations to be considered as the same action class. (3) Extensive experiments on THUMOS-14 and ActivityNet-1.3 benchmarks demonstrate that our model acquires remarkable advances. Besides, substantial ablation studies also reveal that the proposed two-stream structure and semantic-aware modules are of effectiveness.

2. Related Work

2.1. Action Recognition

Action recognition is a fundamental task in video understanding, which is endowed with the responsibility of identifying categories of actions in trimmed videos. Recently, it has made significant progress with advanced deep-learning techniques. Benefiting from this, plenty of off-the-shelf action recognition algorithms are leveraged to abstract video-level representations for complicated downstream tasks. Early studies [2, 6, 21, 48] mainly relied on a two-stream structure, wherein one branch is utilized to encode static appearances and the other branch is designed to capture temporal properties of actions with optical flows. These approaches achieve excellent performance and generalization yet slow speed. To mitigate limitations of speed in using optical flows, follow-up investigations [4, 5, 28] build up lightweight structures to learn useful temporal information. In this paper, we employ I3D [2] as a preliminary representation extractor of videos for subsequent WS-TAL.

2.2. Fully-Supervised Temporal Action Localization

Compare to the action recognition task, TAL not only needs to predict the category of actions but also the temporal intervals from untrimmed videos. Traditionally, fully-supervised TAL adopts frame-level annotations during training. Most existing efforts are primarily divided into two categories: top-down and bottom-up. Inspired by image object detection, top-down approaches [3, 18, 19, 22, 24, 49, 52] transform localization into a detection task in the temporal dimension. In this paradigm, they first generate action proposals and then classify them as well as temporal boundary regression. An advantage of these works is that they can draw on advanced schemes in object detection. On the contrary, the bottom-up methods [23, 27, 57] yield frame-level predictions followed by some well-designed post-processing tricks. Unfortunately, such a fully-supervised paradigm heavily depends on frame-wise annotations, which are prohibitively expensive and unrealistic for much longer videos.

2.3. Weakly-Supervised Temporal Action Localization

WS-TAL merely requires video-level annotations and has received increasing popularity. The attention-based framework has been fully explored. Specifically, UntrimmedNets [50] builds up a soft-attention layer to select relevant segments for boosted performance. HAM-Net [11] devises a hybrid attention mechanism by setting different thresholds to capture both the most salient frames and the full extent of activity. DGAM [45] notices the frequent occurrences of the action-confusion phenomenon in action localization tasks and thus introduces a conditional variational auto-encoder with theoretical proof for the effective separation of action and context instances. LGCA [14] adopts a multi-stage cross-attention strategy to acquire multi-modal representations. However, these attention-based methods produce class-agnostic confidence that is utilized for suboptimal feature combinations.

MIL is another favorite framework that can be regarded as a hard selection mechanism. In specific, ACM-Net [43] combines MIL and a hybrid CAS to distinguish between action instances, context, and non-action instances. W-TALC [42] combines deep metric learning and MIL mechanisms to mine correlations between actions. Similarly, BaS-Net [16] formulates a MIL-based structure and a filtering module to suppress responses from background frames. CoLA [55] also devises a hard snippet mining algorithm to guide the network to precisely perceive temporal boundaries. Nevertheless, MIL-based methods have a major limitation in that it only focuses on the most discriminative frames but ignore the integrity of actions. To address this problem, our method integrates fine-grained clues from an attention-based strategy with salient features from a MIL mechanism to improve the action integrity.

Some other works have also investigated to address the action-context confusion issue. 3C-Net [38] puts forward a formulation with multi-label center loss and action counting loss terms to enhance the feature discriminability and the separability of adjacent action instances. FAC-Net [9] develops a three-branch pipeline to regularize the foreground-action consistency and capture accurate action boundaries. CO_2-Net [8] investigates multimodal feature re-calibration and modal-wise consistency WS-TAL. Besides, pseudo labels are also crafted to guide accurate localization. RefineLoc [41] employs an iterative refinement strategy by estimating snippet-level pseudo labels at each iteration. RSKP [25] uses memory banks to store representative snippets for each class. They are used to generate high-quality pseudo labels, which further generate accurate TCAMs. ASM-Loc [7] leverages a pre-trained teacher model to construct instance-level pseudo labels for more fine-grained supervision. TSCN [54] generates pseudo labels from the late fusion attention sequence at previous iterations, and EM-MIL [32] introduces two pseudo-label generation schemes into an expectation-maximization framework. Nevertheless, they fail to consider the semantic consistency of snippets. Intuitively, snippets with similar representations should be considered to be the same class despite no obvious supervision signal for each snippet. As a result, our model sets a learnable dictionary where entries are the class centroids of the corresponding action categories. The representations of snippets identified as the same action category are induced to be close to the same class centroid. In this manner, the proposed model is semantic-aware and the learned features are discrimination-enhanced.

3. Methodology

3.1. Overview

Given an untrimmed video, which may contain multiple action instances $\{G_i = (g_i^s, g_i^e, y)\}_{i=1}^m$, where g_i^s and g_i^e denote the start and end frame for the i-th action instance G_i respectively, and $m(m >= 1)$ is the number of actions. We propose a two-stream network with a semantic-aware mechanism to detect the temporal intervals of actions with the video-level class label $y \in \{0, 1, ..., C\}$, where C is the number of classes and 0 corresponds to the background. The overview of the architecture is presented in Figure 2. Specifically, I3D [2] is first utilized to extract frozen spatio-temporal representations from raw videos. In order to enhance the expressiveness, these representations are further input into an extra learnable residual block to acquire features $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_T] \in \mathbf{R}^{T \times D}$, where T denotes the number of sampled snippets and D is the dimension of features. Another purpose of introducing the residual block is to guarantee that the proposed semantic-aware mechanism can adjust representations to facilitate semantic coherence of frames with the same action, as described in Section 3.5. Afterwards, $\mathbf{x}$ is fed into the attention-based and MIL-based branches respectively to generate response values for the foreground, background and context from a video-level perspective. Responses of these two branches are further fused to predict the final classification results. In the following sections, we will elaborate on the technical details of each module.

3.2. Attention-Based Branch

One branch of the proposed two-stream structure is to learn frame attention by optimizing video-level recognition. In detail, we generate the attention $\mathbf{a}^{fg} = (a_t^{fg})_{t=1}^T$, $\mathbf{a}^{ct} = (a_t^{ct})_{t=1}^T$ and $\mathbf{a}^{bg} = (a_t^{bg})_{t=1}^T$ directly from features, where a_t^{fg}, a_t^{ct} and $a_t^{bg} \in [0, 1]$ are the attentions of frame t, which represent the confidence that the t-th frame belongs to the foreground, context, and background, respectively. These attentions are further utilized as the weights to perform temporal average pooling over all frames and generate video-level foreground features $\mathbf{x}_{fg}$, context features $\mathbf{x}_{ct}$ and background features $\mathbf{x}_{bg}$ by

$$\mathbf{x}_{fg} = \frac{\sum_{t=1}^T a_t^{fg} \mathbf{x}_t}{\sum_{t=1}^T a_t^{fg}}, \quad \mathbf{x}_{ct} = \frac{\sum_{t=1}^T a_t^{ct} \mathbf{x}_t}{\sum_{t=1}^T a_t^{ct}}, \quad \mathbf{x}_{bg} = \frac{\sum_{t=1}^T a_t^{bg} \mathbf{x}_t}{\sum_{t=1}^T a_t^{bg}}. \tag{1}$$

Then, a shared fully-connected layer following a softmax layer is applied on $\mathbf{x}_{fg}$, $\mathbf{x}_{ct}$ and $\mathbf{x}_{bg}$ to produce classification results $\hat{\mathbf{y}}_{fg}^{att}$, $\hat{\mathbf{y}}_{ct}^{att}$ and $\hat{\mathbf{y}}_{bg}^{att}$ for the foreground, context and background, respectively. Notably, this procedure takes full advantage of all frame information without omissions of details, which is beneficial for discovering subtle clues and complementary to salient features extracted by the MIL-based branch described in Section 3.3.

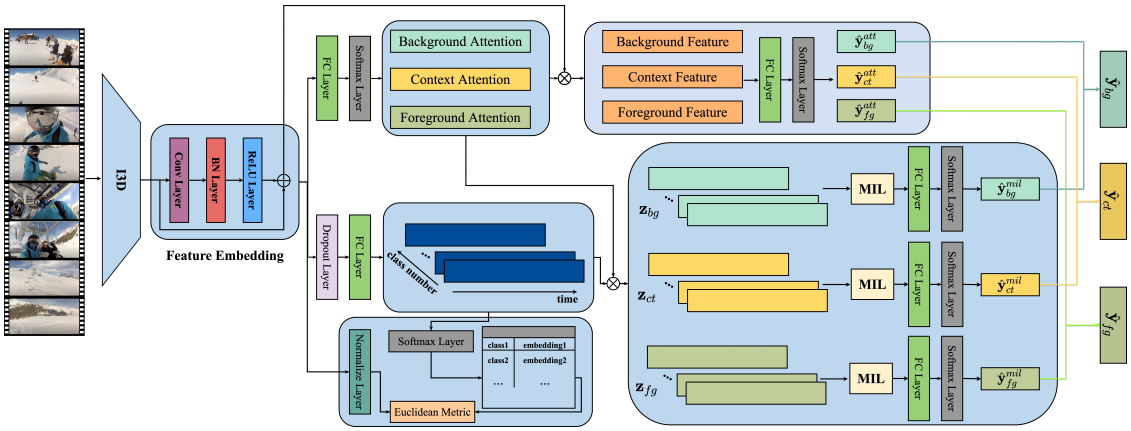

Figure 2. Overview of the proposed model. Our model comprises two branches, one of which is attention-based and does well in encoding local subtle clues, and the other is MIL-based and is an expert on capturing the most discriminative features. The late fusion of class distributions of two branches is employed for video-level recognition. Besides, a learnable dictionary is well-designed for effective semantic awareness. The symbols $\oplus$ and $\otimes$ denote element-wise addition and tensor multiplication respectively.

3.3. MIL-Based Branch

The attention-based branch is a soft combination strategy, which is prone to produce suboptimal combination coefficients that are detrimental to localization, especially for the foreground category due to its sparsity [40]. To this end, a MIL-based branch is introduced to integrate the most discriminative frame-level information into a video-level counterpart. In specific, this branch first utilizes a fully-connected layer with a dropout operation to independently embed the feature of each frame $\mathbf{x}_t$ to category space and thus gets a class activation sequence $\mathbf{p} \in \mathbf{R}^{T\times(C+1)}$. Since the attentions $\mathbf{a}^{fg}$, $\mathbf{a}^{ct}$ and $\mathbf{a}^{bg}$ produced by the attention-based branch are class-agnostic, they have trouble being optimized with video-level supervisory signals, so we further integrate class information $\mathbf{p}$ into them as follows:

$$\mathbf{z}_{fg} = \mathbf{a}^{fg} \times \mathbf{p}, \quad \mathbf{z}_{ct} = \mathbf{a}^{ct} \times \mathbf{p}, \quad \mathbf{z}_{bg} = \mathbf{a}^{bg} \times \mathbf{p}. \quad (2)$$

These calculations are completed by a broadcast mechanism in Python. The advantage of this procedure is that class information is weighted by attention scores, and the generated $\mathbf{z}_{fg}$, $\mathbf{z}_{ct}$, and $\mathbf{z}_{bg}$ are able to focus on the foreground, context and background features respectively. Besides, to capture the most discriminative features, the MIL mechanism views videos as a bag of frames and incorporates respectively the top-k confidence of $\mathbf{z}_{fg}$, $\mathbf{z}_{ct}$ and $\mathbf{z}_{bg}$ by

$$\begin{aligned}
\omega_{fg}^{c} &= \frac{1}{k} \max_{\substack{\mathbf{U}_{fg}^{c} \subset \mathbf{z}_{fg}[:,c], \\ |\mathbf{U}_{fg}^{c}|=k}} \sum_{u\in\mathbf{U}_{fg}^{c}} u, \\
\omega_{ct}^{c} &= \frac{1}{k} \max_{\substack{\mathbf{U}_{ct}^{c} \subset \mathbf{z}_{ct}[:,c], \\ |\mathbf{U}_{ct}^{c}|=k}} \sum_{u\in\mathbf{U}_{ct}^{c}} u, \\
\omega_{bg}^{c} &= \frac{1}{k} \max_{\substack{\mathbf{U}_{bg}^{c} \subset \mathbf{z}_{bg}[:,c], \\ |\mathbf{U}_{bg}^{c}|=k}} \sum_{u\in\mathbf{U}_{bg}^{c}} u,
\end{aligned} \quad (3)$$

where $\mathbf{U}_{fg}^{c}$, $\mathbf{U}_{ct}^{c}$ and $\mathbf{U}_{bg}^{c}$ are sets that contain the top-k classification scores over all frames for class c. k is a value proportional to the length of videos and is set as $k = max(\lfloor T/\sigma \rfloor, 1)$, where σ is a hyper-parameter. In this setting, ω_{fg}^{c} represents the confidence that the video contains actions with the class c. Then, a softmax function is applied on ω_{fg}^{c} to normalize probability distribution for each class: $\hat{\mathbf{y}}_{fg}^{mil}(c) = \frac{exp(\omega_{fg}^{c})}{\sum_{\tilde{c}=0}^{C} exp(\omega_{fg}^{\tilde{c}})}$. The same operation is performed on ω_{ct}^{c} and ω_{bg}^{c} to acquire $\hat{\mathbf{y}}_{ct}^{mil}(c)$ and $\hat{\mathbf{y}}_{bg}^{mil}(c)$.

3.4. Late-Fusion of two branches

After we acquire classification results from both attention-based and MIL-based branches, we further integrate them via a late-fusion operation. In detail, we average the predictions of the two-branch as the final results:

$$\begin{aligned}
\hat{\mathbf{y}}_{fg} = (\hat{\mathbf{y}}_{fg}^{att} + \hat{\mathbf{y}}_{fg}^{mil})/2, \quad \hat{\mathbf{y}}_{ct} = (\hat{\mathbf{y}}_{ct}^{att} + \hat{\mathbf{y}}_{ct}^{mil})/2, \\
\hat{\mathbf{y}}_{bg} = (\hat{\mathbf{y}}_{bg}^{att} + \hat{\mathbf{y}}_{bg}^{mil})/2.
\end{aligned} \quad (4)$$

Afterwards, a cross-entropy loss is applied for the foreground, context, and background classification respectively:

$$\begin{aligned}
\mathcal{L}_{cls}^{fg} &= -\sum_{c=0}^{C} \mathbf{y}_{fg}(c) log \hat{\mathbf{y}}_{fg}(c), \\
\mathcal{L}_{cls}^{ct} &= -\sum_{c=0}^{C} \mathbf{y}_{ct}(c) log \hat{\mathbf{y}}_{ct}(c), \\
\mathcal{L}_{cls}^{bg} &= -\sum_{c=0}^{C} \mathbf{y}_{bg}(c) log \hat{\mathbf{y}}_{bg}(c),
\end{aligned} \quad (5)$$

where $\mathbf{y}_{fg}$, $\mathbf{y}_{ct}$ and $\mathbf{y}_{bg}$ denote ground truths that are formulated by a fancy trick. In detail, we set $\mathbf{y}_{fg}(0) = 0$ and

$\mathbf{y}_{fg}(j) = 1$ for $\mathbf{y}_{fg}$, where j denotes ground truth indexes of action instances and $\mathbf{y}_{fg}(0)$ represents the label for the background. For $\mathbf{y}_{bg}$, $\mathbf{y}_{bg}(0) = 1$ and all other class labels are set to 0. Since the context is considered to be related to action instances and backgrounds, we annotate $\mathbf{y}_{ct}(0) = 1$ and $\mathbf{y}_{ct}(j) = 1$, which is more conducive for the model to learn discriminative features.

In summary, the attention- and MIL-based branches are complementary to each other, encoding both all subtle clues and the most salient features, which facilitates subsequent precise action detection.

3.5. Semantic-Aware Mechanism

Despite the absence of explicit supervision signals in WS-TAL, we still expect that the model has the capability of extracting consistent semantic representations. Intuitively, frames predicted to be of the same class tend to have similar representations. Motivated by this, we set a learnable dictionary $Q \in \mathbf{R}^{(C+1)\times D}$ where entries are the class centroids of the corresponding action categories. Given a feature $\mathbf{x}_t$, suppose it is recognized as the class c when generating CAS in the MIL-based branch, then a euclidean distance between the feature $\mathbf{x}_t$ and $Q[c,:]$ is minimized as:

$$\mathcal{L}_{smt} = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{x}_t - Q[\arg\max_c(\mathbf{p}(t)), :]||_2, \quad (6)$$

where : is the slicing operation in Python. In this manner, the representations of frames identified as the same action category are induced to be close to the same class centroid.

3.6. Overall Loss Function

Apart from the aforementioned classification loss and semantic-aware constraint, we further employ a guide loss as used in [11,43] to ensure the consistency of the responses of two branches at frame-level:

$$\mathcal{L}_{gui} = \frac{1}{T}\sum_{t=1}^{T} |1 - a_t^{fg} - \mathbf{z}_{fg}[t,0]|. \quad (7)$$

To sum up, the overall loss function is formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{cls}^{fg} + \lambda_1\mathcal{L}_{cls}^{ct} + \lambda_2\mathcal{L}_{cls}^{bg} + \lambda_3\mathcal{L}_{smt} + \lambda_4\mathcal{L}_{gui} \quad (8)$$

where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters that control the importance of different loss terms.

3.7. Inference

During inference, we feed videos into the network to acquire a video-level class distribution $\hat{\mathbf{y}}_{fg}$ and attention-weighted class activation sequence $\mathbf{z}_{fg}$. We filter out frames with class scores lower than a pre-defined threshold α. For the remaining categories, we extract consecutive segments and generate proposals $(\hat{t}^s, \hat{t}^e, \phi(c))$ for class c by enforcing a threshold η on action-instance activation sequence $\mathbf{z}_{fg}$. By setting different η, the model will generate proposals of various scales. Here $\hat{t}^s$ and $\hat{t}^e$ represent the start and end frames, respectively. $\phi(c)$ is a refined confidence that there exist actions with class c in the proposal. Specifically, $\phi(c)$ absorbs scores of its neighbors and is calculated following the Outer-Inner-Contrastive function of AutoLoc [46]:

$$\begin{aligned} \phi_{in}(c) &= \frac{\int_{\hat{t}^s}^{\hat{t}^e} \mathbf{z}_{fg}[t,c]}{\hat{t}^e - \hat{t}^s}, \\ \phi_{out}(c) &= \frac{\int_{\hat{t}^s-\hat{t}^v}^{\hat{t}^s} \mathbf{z}_{fg}[t,c] + \int_{\hat{t}^e}^{\hat{t}^e+\hat{t}^v} \mathbf{z}_{fg}[t,c]}{2\times\hat{t}^v}, \\ \phi(c) &= \phi_{in}(c) - \phi_{out}(c) + \beta\hat{\mathbf{y}}_{fg}(c). \end{aligned} \quad (9)$$

In fact, $\mathbf{z}_{fg}$ describes frame-level class responses, and $\hat{\mathbf{y}}_{fg}$ is the video-level class responses. Their combinations are leveraged as confidence of action instances. $\hat{t}^v = \frac{\hat{t}^e - \hat{t}^s}{5}$ denotes the inflated contrast area. β is the combination hyper-parameter. Finally, a Non-Maximum Suppression (NMS) mechanism is applied on the refined confidence $\phi(c)$ to remove redundant proposals.

4. Experiments

4.1. Dataset and Setting

THUMOS-14. THUMOS-14 [10] is a challenging action localization dataset that comprises 200 untrimmed videos for training and 213 videos for testing. It contains a total of 20 categories. Each video consists of 15.5 action instances on average and the length of videos varies from a few minutes to tens of minutes.

ActivityNet-1.3. ActivityNet-1.3 [1] is a larger-scale action localization dataset that comprises 200 categories of videos, where each video contains 1.6 action instances on average. ActivityNet-1.3 provides 10024 videos for training, 4926 videos for validation, and 5044 videos for testing. Each video contains approximately 35% frames with the fine-grained distinction between the context and background, and is therefore relatively challenging. We report results on its validation set following the previous work [24,43,45].

Evaluation Protocol. The mean Average Precision (mAP) at different temporal Intersection over Union (t-IoU) thresholds is reported as evaluation criteria. For THUMOS-14, t-IoU thresholds are set to [0.1:0.1:0.7] (from 0.1 to 0.7 in steps of 0.1). For ActivityNet-1.3, t-IoU thresholds are set to [0.5:0.05:0.95] (from 0.5 to 0.95 in steps of 0.05).

Implementation Details. Given a video, we sample continuous non-overlapping 16 frames as a snippet and extract RGB and optical-flow features using I3D framework [2] pre-trained on Kinetics [13]. These two features are further concatenated and form a 2048-dimensional representation. For fair comparisons, we do not finetune the feature

Supervision	Method	mAP@t-IoU(%)							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
Fully Supervised	SSN [58]	66.0	59.4	51.9	41.0	29.8	-	-	-
	BSN [24]	-	-	53.5	45.0	36.9	28.4	20.0	-
	BMN [22]	-	-	56.0	47.4	38.8	29.7	20.5	-
	BSN++ [49]	-	-	59.9	49.5	41.3	31.9	22.8	-
	G-TAD [52]	-	-	66.4	60.4	51.6	37.6	22.9	-
Weakly Supervised †	3C-Net [38]	59.1	53.5	44.2	34.1	26.6	-	8.1	-
	PreTrimNet [56]	57.5	50.7	41.4	32.1	23.1	14.2	7.7	23.7
	SF-Net [33]	71.0	63.4	53.2	40.7	29.3	18.4	9.6	40.8
	Ju *et al.* [12]	72.3	64.7	58.2	47.1	35.9	23.0	12.8	44.9
	LACP [15]	75.7	71.4	64.6	56.5	45.3	34.5	21.8	52.8
Weakly Supervised	MAAN [53]	59.8	50.8	41.1	30.6	20.3	12.0	6.9	31.6
	BasNet [16]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	EM-MIL [32]	59.1	52.7	45.5	36.8	30.5	22.7	16.4	37.7
	DGAM [45]	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
	A2CL-PT [35]	61.2	56.1	48.1	39.0	30.1	19.2	10.6	37.8
	CoLA [55]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9
	HAM-Net [11]	65.4	59.0	50.3	41.1	31.0	20.7	11.4	39.8
	ACSNet [30]	-	-	51.4	42.7	32.4	22.0	11.7	-
	ACM-Net [43]	65.3	59.2	49.5	38.4	27.4	16.4	6.9	37.6
	ASL [34]	67.0	-	51.8	-	31.1	-	-	-
	D2-Net [37]	65.7	60.2	52.3	43.4	36.0	-	-	-
	AUMN [31]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	41.5
	UM [17]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	41.9
	FAC-Net [9]	67.6	62.1	52.6	44.3	33.4	22.5	12.7	42.2
	CO_2-Net [8]	70.1	63.6	54.5	45.7	**38.3**	**26.4**	**13.4**	44.6
	ASM-Loc [7]	71.2	65.5	57.1	46.8	36.6	25.2	**13.4**	45.1
	Ours	**73.0**	**68.2**	**60.0**	**47.9**	37.1	24.4	12.7	**46.2**

Table 1. Quantitative comparisons on THUMOS-14 benchmark. The mAP is used as an evaluation criterion at t-IoU thresholds 0.1:0.1:0.7, and AVG denotes the average of mAP of t-IoU over the interval from 0.1 to 0.7. † means extra training data are used.

extractor, *i.e.,* I3D. For the dictionary learning, we first averaged the representations(extracted by the pre-trained I3D) of the same class in the training set. They are initialized to the centroid of the corresponding class and then a warmup operation is conducted. Our model is implemented using the PyTorch framework and runs on NVIDIA Tesla V100 GPUs. Adam with a learning rate of 1e-4 is utilized to optimize the model for 100 epochs. We set $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 =$ 5e-2 and $\lambda_4 =$ 2e-3. The hyper-parameter $\alpha = 0.1$, $\beta = 0.2$, and σ are set to 8, 2, and 2 for the foreground, context, and background classes, respectively. These values are obtained by using the grid search method. In detail, for α, we search from 0.05 to 0.5 in the step of 0.05. For β, we search including two scales: from 0.05 to 0.5 in the step of 0.05 and from 0.1 to 1 in the step of 0.1. For λ, we set initial λ_1 and λ_2 are 0.01, and search λ_3 and λ_4 in two scales: from 0.001 to 0.01 in the step of 0.001, and from 0.01 to 0.1 in the step of 0.01. Then we fix λ_3 and λ_4 and search λ_1 and λ_2 using the same step. The dropout regularization is used with a possibility of 0.5. The learnable dictionary Q is initialized to a uniform distribution between 0 and 1. In order to remove overlap proposals, we perform NMS with a t-IoU threshold of 0.5. For THUMOS-14, the batch size is 16 and the number of snippets T is set as 750. η is set from 0.1 to 0.9 in steps of 0.025. For ActivityNet-1.3, the batch size is 32 and the number of snippets T is set to 75. η are set from 0.005 to 0.025 in steps of 0.005.

4.2. Main Results

We compare our model with state-of-the-art competitors on THUMOS-14 and ActivityNet-1.3 datasets in both fully-supervised and weakly-supervised settings. Some methods [8, 12, 15, 33, 38, 56] utilize extra data or information during training and are also listed for reference. Under the same conditions, our approach achieves remarkable advances.

THUMOS-14. Table 1 illustrates the performance of different competitors on THUMOS-14 dataset. Without additional training data or information accessible, it can observe that our model remarkably outperforms other approaches at most t-IoU thresholds. Also, the average mAP (Avg) from 0.1 to 0.7 is reported for more comprehensive assessments. Results demonstrate that our method achieves the best performance. Especially, we achieve a significant boost over the state-of-the-art methods at AVG (+4.0% for FAC-Net, +1.6% for CO_2-Net and +1.1% for ASM-Loc), indicating that our localization is more precise in general. Further-

Supervision	Method	mAP@t-IoU(%)										
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg
Fully Supervised	SSN [58]	41.3	38.8	35.9	32.9	30.4	27.0	22.2	18.2	13.2	6.1	26.6
	BSN [24]	46.5	-	-	-	-	30.0	-	-	-	8.0	30.0
	G-TAD [52]	50.4	-	-	-	-	34.6	-	-	-	9.0	34.1
Weakly Supervised †	CMCS [26]	36.8	-	-	-	-	22.0	-	-	-	5.6	-
	3C-Net [38]	35.4	-	-	-	-	22.9	-	-	-	8.5	-
	LACP [15]	40.4	-	-	-	-	24.6	-	-	-	5.7	-
Weakly Supervised	UntrimmedNet [50]	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
	AutoLoc [46]	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
	TSM [20]	30.3	-	-	-	-	19.0	-	-	-	4.5	-
	CleanNet [29]	37.1	33.4	29.9	26.7	23.4	20.3	17.2	13.9	9.2	5.0	21.6
	Bas-Net [16]	34.5	-	-	-	-	22.5	-	-	-	5.2	-
	DGAM [45]	40.6	37.0	33.2	29.8	26.6	23.2	19.7	15.1	10.4	5.2	24.1
	EM-MIL [32]	37.4	-	-	-	-	23.1	-	-	-	2.0	-
	TSCN [54]	35.3	-	-	-	-	21.4	-	-	-	5.3	-
	ACM-Net [43]	40.0	36.8	33.9	30.5	27.0	24.0	20.2	15.9	11.0	6.1	24.5
	A2CL-PT [35]	36.8	-	-	-	-	22.0	-	-	-	5.2	-
	AUMN [31]	38.3	-	-	-	-	23.5	-	-	-	5.2	-
	ASM-Loc [7]	41.0	-	-	-	-	24.9	-	-	-	6.2	-
	Ours	**41.8**	**38.5**	**35.8**	**32.6**	**29.2**	**25.7**	**22.7**	**17.5**	**12.6**	**6.5**	**26.3**

Table 2. Quantitative comparisons on ActivityNet-1.3 benchmark. The mAP is used as an evaluation criterion at t-IoU thresholds 0.5:0.05:0.95, and AVG denotes the average of mAP of t-IoU over the interval from 0.5 to 0.95. † means extra training data are used.

more, compared with both fully-supervised methods and weakly-supervised with additional training data, our model can achieve close or even better performance.

ActivityNet-1.3. The comparison results of the state-of-the-art approaches on ActivityNet-1.3 are summarized in Table 2. Since our model is able to learn robust representations and perceive the semantic information of frames with subtle changes, it achieves amazing performance and outperforms all previous WS-TAL methods on all t-IoUs. Specifically, our method exceeds the state-of-the-art method ASM-Loc [7] that designs a complicated multi-step refinement. Surprisingly, the displayed methods utilizing extra training data are also inferior to ours. The overall result proves that our method can not only achieve accurate action localization (THUMOS-14), but also effectively detect boundaries with the fined-grained distinction between the context and background (ActivityNet-1.3).

4.3. Ablation Study

In this section, we conduct ablation studies on different losses and network architectures to prove the effectiveness of these components.

Study on Different Losses. We first analyze the proposed model by experimenting with combinations of different loss terms, and results are displayed in Table 3 and Table 4. For both THUMOS-14 and ActivityNet-1.3, we observe that a remarkable performance advance is obtained when combined with all loss terms, confirming the utility and complementarity of these losses. Furthermore, introducing $\mathcal{L}_{cmt}$ loss can bring more performance improvement when t-IoU thresholds are higher, which reveals the importance of semantic consistency for finer and more precise localization when constraints are tighter.

Study on Network Structures. From the results of ablation studies on different losses, we observe that several castrated variants perform better than some prevailing methods, which we attribute to the two-branch network structure. To prove our point, we use castrated counterparts of only a mechanism without a late-fusion operation. Since $\mathcal{L}_{gui}$ and $\mathcal{L}_{smt}$ are only involved when the MIL-based branch is employed, so only $\mathcal{L}_{cls}^{fg}$, $\mathcal{L}_{cls}^{ct}$ and $\mathcal{L}_{cls}^{bg}$ are utilized in each variant for a fair comparison. During inference, temporal regions are detected by thresholding and merging $\mathbf{a}_{fg}$ and $\mathbf{z}_{fg}$ for the attention- and MIL-based variants respectively. As demonstrated in Table in 5 and Table 6, the localization accuracy degrades regardless of whether an attention-based or MIL-based branch is used alone, demonstrating the superiority of the overall architecture. In addition, since the attention-based variant is class-agnostic and lacks discriminability, we observe that its performance is far worse than the MIL-based variant.

4.4. Qualitative Results

To further explore the localization performance of variants with different losses, Figure 3 illustrates qualitative results, including various types of temporal regions. The first example shows a video containing an action of "*Trimming branches or hedges*" with a shot change in the process. The model with an overall loss accurately hits each boundary instance, while predictions of other variants are not desirable. The second example describes the motion of "*Swinging*" and is a more challenging video attributed

$\mathcal{L}_{cls}^{fg}$	$\mathcal{L}_{cls}^{bg}$	$\mathcal{L}_{cls}^{ct}$	$\mathcal{L}_{gui}$	$\mathcal{L}_{smt}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
✓					62.8	56.0	45.8	37.1	26.7	17.4	8.5	36.3
✓	✓				63.4	57.9	47.6	38.7	27.4	18.2	9.0	37.5
✓	✓	✓			67.6	61.7	52.5	42.7	31.9	20.8	10.1	41.0
✓	✓	✓	✓		70.2	65.2	55.6	45.4	32.8	21.3	10.8	43.0
✓	✓	✓	✓	✓	**73.0**	**68.2**	**60.0**	**47.9**	**37.1**	**24.4**	**12.7**	**46.2**

Table 3. Ablation study on the variants of loss function for THUMOS-14 dataset. The mAP is used as an evaluation criterion at t-IoU thresholds 0.1:0.1:0.7, and AVG denotes their average.

$\mathcal{L}_{cls}^{fg}$	$\mathcal{L}_{cls}^{bg}$	$\mathcal{L}_{cls}^{ct}$	$\mathcal{L}_{gui}$	$\mathcal{L}_{smt}$	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg
✓					37.8	35.1	32.2	29.4	25.8	22.8	19.5	15.8	11.0	5.7	23.5
✓	✓				38.4	35.3	32.4	29.7	26.4	23.3	19.8	16.0	11.3	5.9	23.9
✓	✓	✓			39.2	36.3	33.6	30.8	27.5	24.3	20.6	16.4	11.6	6.0	24.6
✓	✓	✓	✓		39.7	36.7	34.1	31.2	28.0	24.6	21.0	16.5	11.7	6.0	25.0
✓	✓	✓	✓	✓	**41.8**	**38.5**	**35.8**	**32.6**	**29.2**	**25.7**	**22.7**	**17.5**	**12.6**	**6.5**	**26.3**

Table 4. Ablation study on the variants of loss function for ActivityNet-1.3 dataset. The mAP is used as an evaluation criterion at t-IoU thresholds 0.5:0.05:0.95, and AVG denotes their average.

Method	mAP@t-IoU(%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
Baseline	**67.6**	**61.7**	**52.5**	**42.7**	**31.9**	**20.8**	**10.1**	**41.0**
Attention-based	57.1	50.5	41.0	30.9	20.6	11.0	4.7	30.8
MIL-based	63.6	57.0	47.3	35.9	24.7	15.5	8.2	36.0

Table 5. Ablation study on network structure for THUMOS-14 dataset. Baseline represents a complete two-stream structure with $\mathcal{L}_{cls}^{fg}$, $\mathcal{L}_{cls}^{ct}$ and $\mathcal{L}_{cls}^{bg}$ losses.

Method	mAP@t-IoU(%)										
	0.5	0.55	0.6	0.65	0.70	0.75	0.8	0.85	0.9	0.95	Avg
Baseline	**39.2**	**36.3**	**33.6**	**30.8**	**27.5**	**24.3**	**20.6**	**16.4**	**11.6**	**6.0**	**24.6**
Attention-based	25.5	23.5	21.5	19.6	17.2	14.9	12.8	10.2	7.1	3.7	15.6
MIL-based	36.4	33.7	31.1	28.3	25.3	22.2	19.0	14.8	10.4	5.4	22.7

Table 6. Ablation study on network structure for ActivityNet-1.3 dataset. Baseline represents a complete two-stream structure with $\mathcal{L}_{cls}^{fg}$, $\mathcal{L}_{cls}^{ct}$ and $\mathcal{L}_{cls}^{bg}$ losses.

to the existence of action-confusion phenomenons. Specifically, contextual actions with high-correlated semantics after "*Swinging*" appear in videos. Anyway, when our model is equipped with $\mathcal{L}_{smt}$ term, it successfully enhances prediction coverage and suppresses the context frames, while the other variants fail to do it. This result also indicates the proposed semantic-aware module indeed perceives subtle semantic discrepancies between frames.

5. Conclusion

In this article, a novel two-stream network for weakly-supervised temporal action localization with a semantic-aware mechanism is proposed. The well-designed two-stream structure absorbs the merits of multiple instance learning and attention-based strategies with a late-fusion operation on the outputs of each branch to acquire classification results. Besides, to mine semantic relationships between snippets, we set a learnable dictionary where entries are the class centroids of the corresponding action categories. The representations of snippets identified as the same action are induced to be close to the same class centroid. Finally, the developed model is evaluated on THUMOS-14 and ActivityNet-1.3. Substantial experiments and analyses proved the effectiveness of our method.

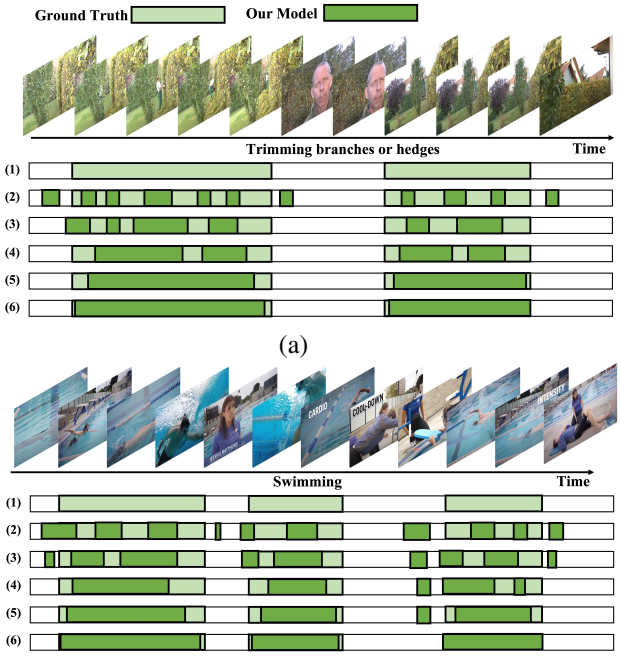

Figure 3. Qualitative results and comparisons of variants with different loss terms. (a) describes the actions of "*Trimming branches or hedges*" . (b) shows a video containing actions of "*Swinging*". (1) Ground Truth (2) $\mathcal{L}_{cls}^{fg}$ (3) $\mathcal{L}_{cls}^{fg} + \mathcal{L}_{cls}^{bg}$ (4) $\mathcal{L}_{cls}^{fg} + \mathcal{L}_{cls}^{bg} + \mathcal{L}_{cls}^{ct}$ (5) $\mathcal{L}_{cls}^{fg} + \mathcal{L}_{cls}^{bg} + \mathcal{L}_{cls}^{ct} + \mathcal{L}_{gui}$ (6) $\mathcal{L}_{cls}^{fg} + \mathcal{L}_{cls}^{bg} + \mathcal{L}_{cls}^{ct} + \mathcal{L}_{gui} + \mathcal{L}_{smt}$

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 3, 5

[3] 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 *CVPR*, pages 1130–1139, 2018. 2

[4] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, pages 203–213, 2020. 2

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 2

[7] 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 *CVPR*, pages 13915–13925, 2022. 3, 6, 7

[8] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *ACM MM*, pages 1591–1599, 2021. 1, 3, 6

[9] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *ICCV*, pages 8002–8011, 2021. 1, 3, 6

[10] 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". *CVIU*, 155:1–23, 2017. 5

[11] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *AAAI*, pages 1637–1645, 2021. 1, 2, 5, 6

[12] Chen Ju, Peisen Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Point-level temporal action localization: Bridging fully-supervised proposals to weakly-supervised losses. *arXiv preprint arXiv:2012.08236*, 2020. 6

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[14] Jun-Tae Lee, Sungrack Yun, and Mihir Jain. Leaky gated cross-attention for weakly supervised multi-modal temporal action localization. In *WACV*, pages 3213–3222, 2022. 1, 2

[15] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *ICCV*, pages 13648–13657, 2021. 6, 7

[16] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI*, pages 11320–11327, 2020. 3, 6, 7

[17] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, pages 1854–1862, 2021. 6

[18] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*, pages 11499–11506, 2020. 2

[19] 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 *CVPR*, pages 3320–3329, 2021. 2

[20] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 7

[21] Jia-Ming Lin, Kuan-Ting Lai, Bin-Ray Wu, and Ming-Syan Chen. Efficient two-stream action recognition on fpga. In *CVPR*, pages 3076–3080, 2021. 2

[22] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *CVPR*, pages 3889–3898, 2019. 2, 6

[23] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, pages 988–996, 2017. 2

[24] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018. 2, 5, 6, 7

[25] Huang Linjiang, Wang Liang, and Li Hongsheng. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *CVPR*, pages 3272–3281, 2022. 3

[26] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, pages 1298–1307, 2019. 7

[27] 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. 2

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2

[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 *ICCV*, pages 3899–3908, 2019. 7

[30] 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 *AAAI*, pages 2233–2241, 2021. 6

[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 *CVPR*, pages 9969–9979, 2021. 1, 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 *ECCV*, pages 729–745, 2020. 1, 3, 6, 7

[33] 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 *ECCV*, pages 420–437, 2020. 6

[34] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *CVPR*, pages 7587–7596, 2021. 1, 6

[35] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *ECCV*, pages 283–299, 2020. 6, 7

[36] Md. Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C. Leu. Action completeness modeling with background aware networks for weakly-supervised temporal action localization. In *ACM MM*, pages 2166–2174, 2020. 1

[37] 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 *ICCV*, pages 13608–13617, 2021. 6

[38] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *ICCV*, pages 8679–8687, 2019. 3, 6, 7

[39] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1):393–402, 2019. 1

[40] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, pages 6752–6761, 2018. 1, 2, 4

[41] Alejandro Pardo, Humam Alwassel, Fabian Caba, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In *WACV*, pages 3319–3328, 2021. 3

[42] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, pages 563–579, 2018. 1, 3

[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. 1, 3, 5, 6, 7

[44] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, pages 347–363, 2018. 1

[45] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*, pages 1009–1019, 2020. 1, 2, 5, 6, 7

[46] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, pages 154–171, 2018. 1, 5, 7

[47] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *CVPR*, pages 20020–20029, 2022. 1

[48] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[49] Haisheng Su, Weihao Gan, Wei Wu, Yu Qiao, and Junjie Yan. Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. In *AAAI*, volume 35, pages 2602–2610, 2021. 2, 6

[50] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 2, 7

[51] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. Learning to refactor action and co-occurrence features for temporal action localization. In *CVPR*, pages 13884–13893, 2022. 1

[52] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, pages 10156–10165, 2020. 2, 6, 7

[53] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *ICLR*, 2019. 1, 2, 6

[54] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *ECCV*, pages 37–54, 2020. 3, 7

[55] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, pages 16010–16019, 2021. 1, 3, 6

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

[57] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, pages 539–555, 2020. 2

[58] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 1, 6, 7

[59] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *ICCV*, pages 13516–13525, 2021. 1