Action-semantic Consistent Knowledge for Weakly-Supervised Action Localization

Yu Wang, *Member, IEEE*, Shengjie Zhao, *Senior Member, IEEE*, and Shiwei Chen

***Abstract*—Weakly-supervised temporal action localization aims to detect temporal intervals of actions in arbitrarily long untrimmed videos with only video-level annotations. Owing to label sparsity, learning action consistency is intractable. In this paper, we assume that frames with similar representations in a given video should be considered as the same action. To this end, we develop a query-based contrastive learning paradigm to ensure action-semantic consistency. This mechanism encourages normalized embeddings with the same class to be pulled closer together, while embeddings from different classes are repelled apart. Besides, we design a two-branch framework, consisting of a class-aware branch and a class-agnostic branch, to learn salient features and fine-grained clues respectively. To further guarantee the action-semantic consistency of the two branches, unlike previous methods that handle each branch independently, we model the relationship between the two branches to avoid unreasonable predictions. Finally, the proposed model demonstrates superior performance over existing methods on the publicly available THUMOS-14 and ActivityNet-1.3 datasets. Substantial experiments and ablation studies also demonstrate the effectiveness of our model.**

***Index Terms*—Temporal action localization, Contrastive learning, Action-semantic consistency**

I. Introduction

Temporal action localization (TAL) aims to detect action intervals in untrimmed videos. As a fundamental task for video understanding [1], [2], [3], it has drawn widespread attention from research, facilitating the rapid and remarkable advance in the fully-supervised settings [4], [5], [6]. Nevertheless, the prohibitively expensive cost of annotations is unacceptable in industrial production. For this reason, weakly-supervised TAL (WS-TAL) [7], [8], [9] with only video-level labels has been advocated to tackle this issue recently.

Owing to label sparsity, a majority of WS-TAL approaches [10], [11], [12], [13], [14], [15] convert the localization into a classification task. In this paradigm, algorithms are designed to detect temporal regions contributing to video-level predictions. Specifically, these approaches first disintegrate untrimmed videos into non-overlapping snippets, on which snippet-wise attention activations and class activation sequence (CAS) are generated. Action regions are localized by thresholding and merging consecutive activations along the time dimension. In particular, there exist two kinds of approaches, *i.e.,* class-aware and class-agnostic. The class-aware mechanism typically employs multiple instance learning (MIL) [16], [17] to generate snippet-wise activations for each class and aggregate them with top-k confidence. MIL-based algorithms focus on extracting salient features that correspond the most to video-level predictions, resulting in them dominating and suppressing the activation values of other areas. However, these suppressed regions with lower confidence are beneficial to guarantee the integrity of actions. On the contrary, the class-agnostic mechanism concentrates on modeling the fine-grained patterns. They independently generate class-agnostic attention scores for each snippet, representing the confidence that it belongs to the foreground, background, and context. Afterwards, a temporal pooling over all snippets with attention scores is utilized to get a compact video-level representation. Despite accessing all snippet features, such class-agnostic attention scores are semantically ambiguous and incapable of detecting accurate temporal boundaries. To address this dilemma, Wang *et. al* [10] proposes a two-branch structure to integrate class-agnostic and class-agnostic paradigms into a unified frame-

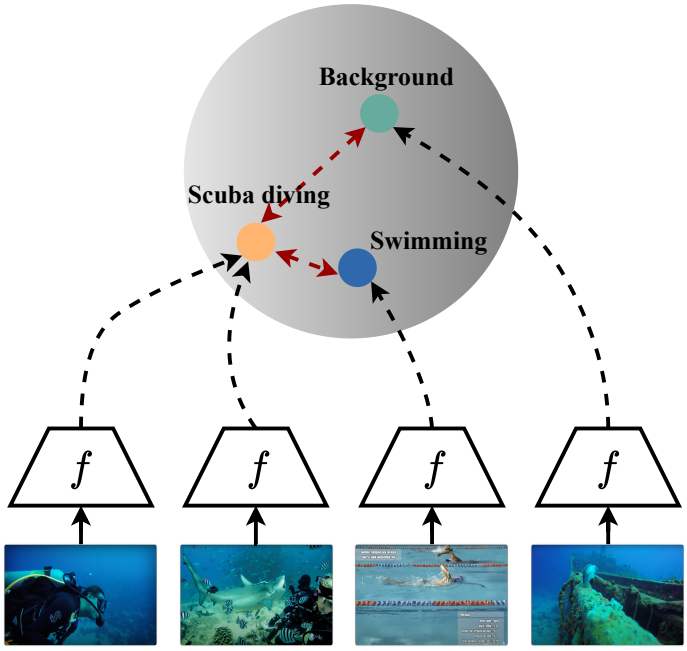


Fig. 1: Given a video containing the action "*Scuba diving*", the normalized representations of frames containing this action should be pulled close, while the context and background should be repelled apart.

Yu Wang is with the School of Software Engineering, Tongji University, Shanghai, China, 201804, and with the Key Laboratory of Embedded System and Service Computing, Ministry of Education.

Shengjie Zhao is with the School of Software Engineering, Tongji University, Shanghai, China, 201804, and with the Engineering Research Center of Key Software Technologies for Smart City Perception and Planning, Ministry of Education, and with the Key Laboratory of Embedded System and Service Computing, Ministry of Education.

Shiwei Chen is with the Department of R&D Data, Microsoft Asia-Pacific Technology CO Ltd, Shanghai, China.

work, where both the fine-grained and salient representations are extracted for more precise localization. Nevertheless, the two branches are independently designed to model different aspects of actions, resulting in potential inconsistencies and conflicts in activation values. For this reason, AICL [18] utilizes contrastive learning with an action consistency constraint to reduce the difference. SMEN [19] introduces a novel slow-motion mining strategy and explicitly induces two branches encoding slow- and normal-motion respectively. In this paper, we follow the two-branch structure component of a class-aware branch and a class-agnostic branch. Since the class-agnostic branch cannot perceive semantic information, it is difficult to search for solutions and is prone to model collapse in the absence of snippet-level annotations. To this end, we further leverage the semantic knowledge of the class-aware branch to instruct the class-agnostic branch's learning process. In this manner, semantic priors from the class-aware branch narrow the solution space of the class-agnostic branch while avoiding prediction conflicts of two branches, which is conducive to learning action-semantic consistency. In fact, the success of semantic distillation in weakly supervised scenarios has also been verified for the action recognition task [20].

Besides, whether the confidence level of each frame is accurate directly affects the quality of subsequent boundary regression. Owing to label sparsity, there are no explicit signals to supervise this procedure. To address this intractable problem, the temporal class activation map (TCAM) [21] is developed to ensure that snippets responding to the video-level classification have high confidence. some attention generation and aggregation strategies [22], [23], [24] are also proposed to alleviate this issue. Wang *et. al* [10] introduce a global dictionary to facilitate similar representations to be considered as the same action class. In this manner, they hope to guarantee semantic consistency between snippets. However, a global dictionary in [10] forces all video actions to have similar representations, which is unreasonable as each video has its specific scenario and context. In this paper, we design a query-based mechanism with contrastive learning to ensure that snippets with similar representations are considered the same actions. Specifically, each query retrieves semantic-specific action and is updated dynamically according to the video context. Then the snippets retrieved by the same dynamic query are clustered together with contrastive learning, as illustrated in Fig. 1. In this fashion, the semantic relationship of snippets is dynamically and explicitly investigated to encourage more reasonable localization in the weakly-supervised setting.

In general, the main contributions and innovations of this work are summarized as follows:

1) We propose a novel **A**ction-**S**emantic **C**onsistency network (ASC-Net) consisting of class-aware and class-agnostic branches to jointly extract salient and fine-grained features of actions for more accurate localization. Besides, the semantic knowledge is distilled from the class-aware branch to the class-agnostic branch to narrow the search space of solutions.
2) Despite the infeasibility of accessing snippet-level annotations, we assume that snippets with similar representations in a specific video should be considered as the same action. Therefore, we propose a novel query-based mechanism with a contrastive loss to dynamically investigate the semantic relationship of actions for more reasonable localization.
3) Extensive experiments on THUMOS-14 and ActivityNet1.3 datasets demonstrate that ASC-Net achieves remarkable advances over existing methods. Also, the detailed ablation studies uncover the effectiveness of the proposed mechanisms.

II. Related Work

A. Action Recognition

As a fundamental task in video understanding, action recognition is mainly dedicated to recognizing categories of actions in trimmed videos. Some investigations even directly utilize off-the-shelf action recognition models to extract video-level features for more complex downstream tasks [25], [26], [27], [28], [20], [29]. Conventional action recognition methods [30], [31] heavily depend on manually well-designed criteria for feature extraction, which is inefficient and time-consuming. With the rise of deep neural networks and their powerful capabilities in tackling computer vision [32], [33], [34], [35], some research has attempted to investigate spatio-temporal dynamics of videos in an end-to-end fashion, *e.g.*, C3D [36], I3D [37], TSM [38], SlowFast [39], and video swin transformer [40]. In this paper, I3D [37] is utilized for preliminary feature extraction from untrimmed videos.

B. Fully-Supervised Temporal Action Localization

Different from action recognition, TAL requires models to predict not only the categories of actions, but also the temporal intervals. Typically, the fully-supervised paradigm has access to frame-wise annotations. The primary approaches are categorized into two groups, *i.e.,* bottom-up and top-down. In detail, the bottom-up methods [41], [42], [43] predict the results of each frame, which are further aggregated by taking some well-designed post-processing strategies. On the contrary, the top-down methods [44], [45], [46], [47], [48], [49], [50] often rely on state-of-the-art object detection techniques in the image domain. They first generate action proposals along the temporal direction. Then the boundaries of intervals are further refined by prior knowledge or regression-based mechanisms. Undoubtedly, the fully-supervised paradigm requires fine-grained annotations, which is fairly labor-intensive and time-consuming.

C. Weakly-Supervised Temporal Action Localization

To alleviate the demand for annotations in a fully-supervised setting, WS-TAL has received surging attention recently. Since only video-level labels are available, the principle of WS-TAL is to discover frames that respond to video-level classification as temporal action intervals. As a result, the mainstream methods are roughly categorized into two groups, *i.e.,* class-aware and class-agnostic. First, the class-aware pipeline utilizes the MIL mechanism to learn category-specific CAS, and then

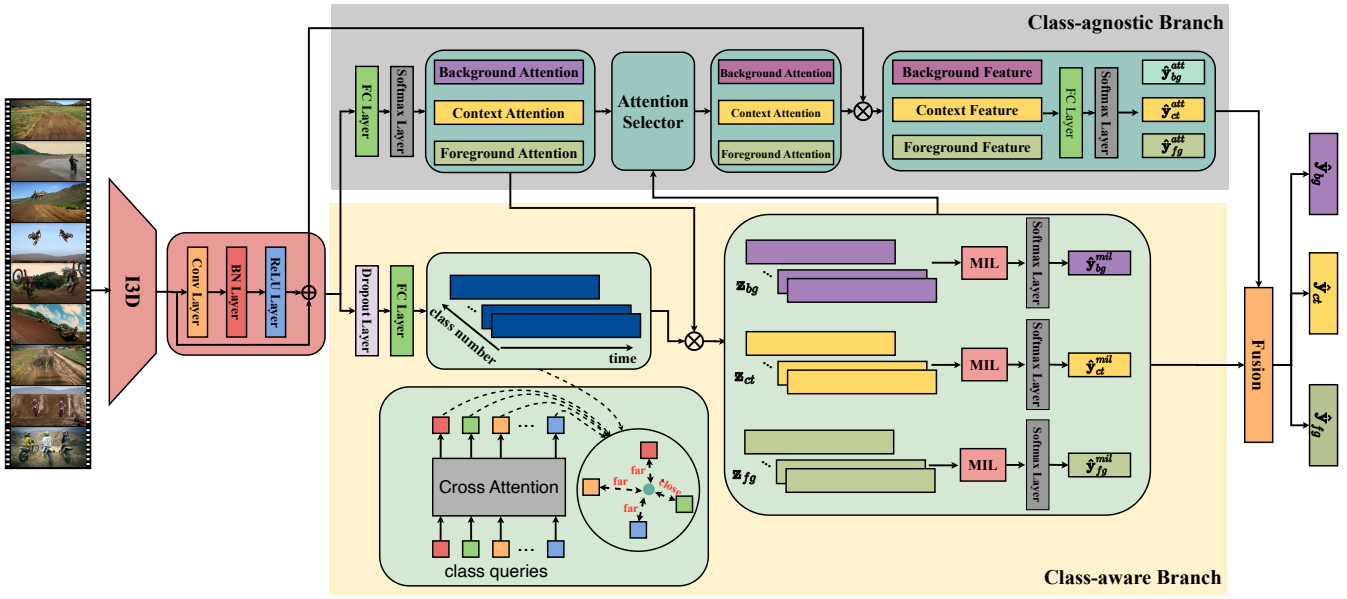


Fig. 2: The overview of ASC-Net. It consists of the class-agnostic and class-aware modules, which jointly extract salient and fine-grained features of actions. Since the class-aware module is sensitive to semantics, it is designed to instruct the class-agnostic module. The final localization results are acquired with a fusion operation on their outputs. Besides, a query-based mechanism is developed to encode the semantic relationship of actions dynamically. $\oplus$ and $\otimes$ represent element-wise addition and tensor multiplication, respectively.

picks up a top-k discriminative subset to construct video-level classification scores. Specifically, W-TALC [17] jointly optimizes MIL loss and co-activity similarity loss to detect activities at a fine granularity. BaS-Net [51] proposes an asymmetrical MIL-based weight-sharing architecture with a filtering module and contrasting objectives to suppress activations from background frames. ACM-Net [11] constructs both a hybrid attention module and a MIL module to distinguish action instances, context, and non-action instances. Ren *et. al* [52] develops proposal-based MIL to inhibit low-quality proposals. CoLA [53] pioneeringly introduces the contrastive representation learning paradigm with an efficient sampling strategy for hard snippet mining. PivoTAL [54] injects prior knowledge into MIL-Based structure from a localization-by-localization perspective to further refine boundaries. DELU [55] proposes a generalized evidential deep learning framework for WS-TAL, where both video- and snippet-level uncertainty are considered. Nevertheless, class-aware methods rely on MIL and have a major limitation in that they only concentrate on the most salient features but ignore fine-grained clues. To address this problem, FC-CRF [56] attempts to erase progressively the most discriminative parts to highlight other less discriminative snippets. In contrast, our method integrates the class-aware mechanism and class-agnostic mechanism to complement each other. Besides, the proposed method forces the class-aware module to instruct the class-agnostic one for effective learning.

Furthermore, the class-agnostic mechanism employs attention-based approaches and focuses on the general actions in the video. In specific, UntrimmedNets [57] designs a soft-attention structure to search for relevant snippets for boosted performance. DGAM [12] finds an action-confusion phenomenon and proposes a conditional Variational Auto-Encoder (VAE) for the effective separation of action and context instances. ASL [8] explores a general independent concept of action by investigating a class-agnostic actionness network. HAM-Net [23] develops a hybrid attention mechanism to model an action in its entirety. LGCA [24] explores an adaptive multi-modal fusion strategy with leaky gated cross-attention. However, these class-agnostic approaches fail to perceive action semantics and stuck in sub-optimal solutions. To overcome this drawback, Wang *et. al* proposes a two-stream network incorporating class-aware and class-agnostic mechanisms to extract both salient and fine-grained features. AICL [18] highlights inconsistency between the class-aware and class-agnostic branches, proposing a consistency constraint to reduce the discrepancy between them. In contrast, we also integrate such two mechanisms into a unified framework, where the class-aware branch's semantic knowledge is distilled to instruct the learning process of the class-agnostic branch, leading to the shrinkage of the search space while ensuring the prediction consistency of two branches.

Besides, an intractable issue in WS-TAL is label sparsity. Due to the absence of frame-wise annotations, it hinders the generation of high-quality CAS. For this reason, some efforts attempt to alleviate it through pseudo-labels or knowledge distillation. Specifically, Xu *et. al* [58] argues that relations at the category and sequence levels are crucial for WS-TAL, and facilitate accurate and complete localization through knowledge distillation. CO_2-Net [14] investigates multi-modal feature re-calibration and modal-wise consistency with pseudo-

labels. RSKP [59] explores a representative snippet knowledge propagation framework, which generates better pseudo labels via representative snippet knowledge propagation. ASM-Loc [60] adopts a multi-step proposal refinement to improve the quality of action proposals with instance-level pseudo-labels progressively. TSCN [61] fuses the attention sequence of RGB and optical-flow modalities to generate segment-level pseudo-labels, while UGCT [62] adopts the RGB and optical-flow modalities to yield pseudo labels for each other by leveraging their complementarity. Nevertheless, these efforts fail to consider the semantics of actions, which is critical for action detection [63]. Intuitively, the same action should have similar representations for a specific video. To this end, we develop a query-based contrastive learning paradigm to ensure action-semantic consistency. This mechanism encourages normalized embeddings from the same class to be pulled closer together, while embeddings from different classes are repelled apart.

III. Methodology

In this section, an overview and formulation of WS-TAL are first presented, and then the detailed contents of the proposed algorithm are described. Finally, extensive experiments and ablation studies on publicly available THUMOS-14 and ActivityNet1.3 benchmarks are conducted.

A. Overview

Given an arbitrarily long untrimmed video, which embodies a set of action instances $\{A_i = (a_i^s, a_i^e, \mathbf{y})\}_{i=1}^{M}$, where a_i^s and a_i^e represent the start timestamp and the end timestamp for the i-th action, and M is the number of action instances in the video. $\mathbf{y} \in \mathbf{R}^{C+1}$ is the ground truth, where C is the number of action categories and the 0-th dimension denotes the non-action background category. $\mathbf{y}(j) = 1$ if the j-th action presents in the video and $\mathbf{y}(j) = 0$ otherwise. The proposed architecture consists of three sub-modules: the class-agnostic branch, the class-aware branch, and the query-based semantic-aware mechanism. The overview of the proposed framework is displayed in Fig. 2. Specifically, I3D [37] is first adopted to extract spatio-temporal representations of videos. To further enhance the expressiveness, the I3D representations are fed into an extra residual block, leading to the improved features $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_T] \in \mathbf{R}^{T \times D}$, where T is the length of snippets and D is the dimension of features. Afterwards, the class-aware branch utilizes $\mathbf{x}$ to generate class-specific responses, which encode dominant segments but suppress activation values of other regions. So the class-agnostic branch is formulated to capture simultaneously fine-grained clues. Besides, the class-aware branch can perceive action semantics and therefore transfers its knowledge into the class-agnostic branch. Outputs from two branches are late-fused to acquire the resulting predictions. In this fashion, we hope that the model can ensure the action-semantic consistency of the two branches while extracting both salient features and fine-grained features. Furthermore, we assume that frames with similar representations in a given video should be considered the same action. To this end, we design a query-based contrastive learning strategy to ensure action-semantic consistency during training, encouraging normalized embeddings from the same class to be pulled closer together. In contrast, embeddings from different classes are repelled apart. In the following sections, we will elaborate on the details of each part.

B. Class-agnostic Branch

The class-agnostic branch attempts to generate attention for each snippet by optimizing video-level classification. Specifically, our model utilizes $\mathbf{x}$ to yield three attention scores for each frame, *i.e.,* $\mathbf{a}^{fg} = (a_t^{fg})_{t=1}^{T}$, $\mathbf{a}^{ct} = (a_t^{ct})_{t=1}^{T}$, and $\mathbf{a}^{bg} = (a_t^{bg})_{t=1}^{T}$, which stand for the confidence that the t-th frame belongs to the foreground, context, and background. Then these attention scores are adopted to aggregate video-level features through temporal average pooling operations with an attention selector:

$$\begin{aligned} \mathbf{x}_{fg} &= \frac{\sum_{t \in \mathbf{S}_{fg}} a_t^{fg} \mathbf{x}_t}{\sum_{t \in \mathbf{S}_{fg}} a_t^{fg}}, \\ \mathbf{x}_{ct} &= \frac{\sum_{t \in \mathbf{S}_{ct}} a_t^{ct} \mathbf{x}_t}{\sum_{t \in \mathbf{S}_{ct}} a_t^{ct}}, \\ \mathbf{x}_{bg} &= \frac{\sum_{t \in \mathbf{S}_{bg}} a_t^{bg} \mathbf{x}_t}{\sum_{t \in \mathbf{S}_{bg}} a_t^{bg}}, \end{aligned} \tag{1}$$

where $\mathbf{S}_{fg}$, $\mathbf{S}_{ct}$, and $\mathbf{S}_{bg}$ are respectively the foreground set, context set, and the background set hit by the attention selector. Specifically, the selector leverages knowledge from the class-aware branch described in section III-C to filter the attention. Since the class-aware branch is capable of perceiving semantic information, it has extraordinary guiding significance for the generation of class-agnostic attention. Next, a shared fully connected layer following a softmax operation is applied on video-level features $\mathbf{x}_{fg}$, $\mathbf{x}_{ct}$, and $\mathbf{x}_{bg}$ to yield classification predictions $\hat{\mathbf{y}}_{fg}^{ags}$, $\hat{\mathbf{y}}_{ct}^{ags}$, and $\hat{\mathbf{y}}_{bg}^{ags}$ for the foreground, the context, and the background respectively. In fact, the class-agnostic branch employs knowledge from all frames and therefore concentrates on fine-grained clues, which plays a complementary role with the parallel class-aware branch that captures salient features.

C. Class-aware Branch

Since the above class-agnostic mechanism cannot perceive action semantics, the sparsity of action instances is prone to cause the model to fall into suboptimal solutions [21]. For this reason, the class-aware module is dedicated to overcoming this problem. It takes a MIL-based strategy to extract salient features that correspond the most to video-level predictions. In detail, the class-aware module maps the representation $\mathbf{x}$ into the action category space by applying an extra dropout and a fully-connected layer. In this action category space, we get a class activation sequence (CAS) $\mathbf{v} \in \mathbf{R}^{T \times (C+1)}$. The attention $\mathbf{a}^{fg}$, $\mathbf{a}^{ct}$, and $\mathbf{a}^{bg}$ are further used to weight $\mathbf{v}$ and therefore acquire the foreground CAS $\mathbf{z}_{fg}$, the context CAS $\mathbf{z}_{ct}$, and the background CAS $\mathbf{z}_{bg}$:

$$\begin{aligned} \mathbf{z}_{fg} &= \mathbf{a}_{fg} \times \mathbf{v}, \\ \mathbf{z}_{ct} &= \mathbf{a}_{ct} \times \mathbf{v}, \\ \mathbf{z}_{bg} &= \mathbf{a}_{bg} \times \mathbf{v}. \end{aligned} \tag{2}$$

In this manner, the CAS-specific $\mathbf{z}_{fg}$, $\mathbf{z}_{ct}$, $\mathbf{z}_{bg}$ interact with the corresponding class-agnostic attention, which induces the model to learn consistent knowledge despite no explicit supervised signals. Next, we take a MIL strategy which chooses respectively the top-k values of $\mathbf{z}_{fg}$, $\mathbf{z}_{ct}$, and $\mathbf{z}_{bg}$ for the class c along the temporal direction, the mean value of which is utilized to produce video-level classification results as follows:

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

where $\mathbf{O}_{fg}^{c}$, $\mathbf{O}_{ct}^{c}$, and $\mathbf{O}_{bg}^{c}$ are sets that consist of the top-k classification activation values for the class c, and : is a slice operation. The cardinality of sets defined by the hyper-parameter k is proportional to the length of the video and usually assigned as $k = max(\lfloor T/\sigma \rfloor, 1)$, where σ is a hyper-parameter related to datasets. In this fashion, salient features with top-k confidence are encoded, and ω_{fg}^{c}, ω_{ct}^{c}, and ω_{bg}^{c} represent respectively the confidence that the video contains action instances with the class c for the foreground, the context, and the background. Then, a softmax operation is applied on ω_{fg}^{c}, ω_{ct}^{c}, and ω_{bg}^{c} to acquire normalized probability distribution over action categories from a video-level prediction perspective:

$$\begin{aligned}
\hat{\mathbf{y}}_{fg}^{awa}(c) &= \frac{exp(\omega_{fg}^{c})}{\sum_{\tilde{c}=0}^{C} exp(\omega_{fg}^{\tilde{c}})}, \\
\hat{\mathbf{y}}_{ct}^{awa}(c) &= \frac{exp(\omega_{ct}^{c})}{\sum_{\tilde{c}=0}^{C} exp(\omega_{ct}^{\tilde{c}})}, \\
\hat{\mathbf{y}}_{bg}^{awa}(c) &= \frac{exp(\omega_{bg}^{c})}{\sum_{\tilde{c}=0}^{C} exp(\omega_{bg}^{\tilde{c}})}.
\end{aligned} \tag{4}$$

Notably, the above procedure is aware of action semantics, which is expected to propagate to the previously mentioned class-agnostic module for consistent predictions. Since segments with the top-k confidence for the action instances are supervised by video-level ground truth, it can provide reliable instructions for the class-agnostic module. To this end, we propose to distill their semantic knowledge. Specifically, we let $\mathbf{I}_{fg}$, $\mathbf{I}_{ct}$, and $\mathbf{I}_{bg}$ denote the set of temporal indexes corresponding to the foreground, context, and background, respectively. They are calculated by:

$$\begin{aligned}
\mathbf{I}_{fg} &= \mathop{\arg\max}_{\substack{\mathbf{I}_{fg} \subset \{1,2,\ldots,T\} \\ |\mathbf{I}_{fg}|=k}} \sum_{i \in \mathbf{I}_{fg}} \sum_{j=1}^{C} \mathbf{z}_{fg}[i,j], \\
\mathbf{I}_{ct} &= \mathop{\arg\max}_{\substack{\mathbf{I}_{ct} \subset \{1,2,\ldots,T\} \\ |\mathbf{I}_{ct}|=k}} \sum_{i \in \mathbf{I}_{ct}} \sum_{j=1}^{C} \mathbf{z}_{ct}[i,j], \\
\mathbf{I}_{bg} &= \mathop{\arg\max}_{\substack{\mathbf{I}_{bg} \subset \{1,2,\ldots,T\} \\ |\mathbf{I}_{bg}|=k}} \sum_{i \in \mathbf{I}_{bg}} \sum_{j=1}^{C} \mathbf{z}_{bg}[i,j].
\end{aligned} \tag{5}$$

Here, because the class-agnostic branch ignores class-specific information, the choice of index is based on the sum of all foreground confidences (*i.e.,* all action categories) rather than a specific class c. Then, $\mathbf{I}_{fg}$, $\mathbf{I}_{ct}$, and $\mathbf{I}_{bg}$ are further used to guide attention selection in the class-agnostic module. Specifically, the attention selector hits the foreground set $\mathbf{S}_{fg}$, the context set $\mathbf{S}_{ct}$, and the background set $\mathbf{S}_{bg}$ as follows:

$$\begin{aligned}
\mathbf{S}_{fg} &= \mathbf{I}_{fg}, \\
\mathbf{S}_{bg} &= \begin{cases} \mathbf{I}_{bg} - \mathbf{I}_{fg}, & \mathbf{I}_{bg} - \mathbf{I}_{fg} \neq \varnothing \\ \mathbf{I}_{bg}, & \mathbf{I}_{bg} - \mathbf{I}_{fg} = \varnothing \end{cases}, \\
\mathbf{S}_{ct} &= \begin{cases} \mathbf{I}_{bg} \cap \mathbf{I}_{fg}, & \mathbf{I}_{bg} \cap \mathbf{I}_{fg} \neq \varnothing \\ \mathbf{I}_{ct}, & \mathbf{I}_{bg} - \mathbf{I}_{fg} = \varnothing \end{cases},
\end{aligned} \tag{6}$$

where $-$ and $\cap$ represent difference and intersection operation respectively. $\mathbf{S}_{fg}$, $\mathbf{S}_{bg}$, and $\mathbf{S}_{ct}$ are utilized to guide attention selection in the class-agnostic branch as described in section III-C. This procedure propagates semantic knowledge between two modules and also guarantees consistent predictions to some extent.

Last, the predictions from the class-agnostic and class-aware modules are integrated through a late-fusion operation:

$$\begin{aligned}
\hat{\mathbf{y}}_{fg} &= (\hat{\mathbf{y}}_{fg}^{ags} + \hat{\mathbf{y}}_{fg}^{awa})/2, \\
\hat{\mathbf{y}}_{ct} &= (\hat{\mathbf{y}}_{ct}^{ags} + \hat{\mathbf{y}}_{ct}^{awa})/2, \\
\hat{\mathbf{y}}_{bg} &= (\hat{\mathbf{y}}_{bg}^{ags} + \hat{\mathbf{y}}_{bg}^{awa})/2.
\end{aligned} \tag{7}$$

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

$$\begin{aligned}
\mathcal{L}_{cls}^{fg} &= -\sum_{c=0}^{C} \mathbf{y}_{fg}(c) log \hat{\mathbf{y}}_{fg}(c), \\
\mathcal{L}_{cls}^{ct} &= -\sum_{c=0}^{C} \mathbf{y}_{ct}(c) log \hat{\mathbf{y}}_{ct}(c), \\
\mathcal{L}_{cls}^{bg} &= -\sum_{c=0}^{C} \mathbf{y}_{bg}(c) log \hat{\mathbf{y}}_{bg}(c),
\end{aligned} \tag{8}$$

where $\mathbf{y}_{fg}$, $\mathbf{y}_{ct}$ and $\mathbf{y}_{bg}$ are the corresponding ground truths. Notably, for $\mathbf{y}_{fg}$, we set $\mathbf{y}_{fg}(j) = 1$ if the j-th action presents in the video and $\mathbf{y}_{fg}(j) = 0$ otherwise. For $\mathbf{y}_{bg}$, $\mathbf{y}_{bg}(0)$ is set to 1 and all other class indexes are set to 0. Since the context is action-related but semantically belongs to the background, we set both $\mathbf{y}_{ct}(0)$ and $\mathbf{y}_{ct}(j)$ to 1.

D. Query-based Semantic-aware Mechanism

Since the label sparsity in WS-TAL, learning action consistency is intractable. Nevertheless, we assume that frames with similar representations in a given video should be considered to belong to the same action. For this reason, we develop a query-based contrastive learning paradigm to encourage normalized embeddings with the same class to be pulled closer together while repelling embeddings from different classes apart. Specifically, we formulate a set of learnable queries $\mathbf{g} \in \mathbf{R}^{(C+1)\times D}$, whose cardinality is the same as action categories. Here we hope each query can retrieve action-specific patterns. However, these queries are

Supervision	Method	mAP@t-IoU(%)							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
Fully Supervised	SSN [64]	66.0	59.4	51.9	41.0	29.8	-	-	-
	BSN [65]	-	-	53.5	45.0	36.9	28.4	20.0	-
	BMN [48]	-	-	56.0	47.4	38.8	29.7	20.5	-
	BSN++ [49]	-	-	59.9	49.5	41.3	31.9	22.8	-
	G-TAD [50]	-	-	66.4	60.4	51.6	37.6	22.9	-
Weakly Supervised †	3C-Net [66]	59.1	53.5	44.2	34.1	26.6	-	8.1	-
	PreTrimNet [67]	57.5	50.7	41.4	32.1	23.1	14.2	7.7	23.7
	SF-Net [7]	71.0	63.4	53.2	40.7	29.3	18.4	9.6	40.8
	Ju *et al.* [68]	72.3	64.7	58.2	47.1	35.9	23.0	12.8	44.9
	LACP [69]	75.7	71.4	64.6	56.5	45.3	34.5	21.8	52.8
Weakly Supervised	MAAN [22]	59.8	50.8	41.1	30.6	20.3	12.0	6.9	31.6
	BasNet [51]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	EM-MIL [70]	59.1	52.7	45.5	36.8	30.5	22.7	16.4	37.7
	DGAM [12]	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
	A2CL-PT [71]	61.2	56.1	48.1	39.0	30.1	19.2	10.6	37.8
	CoLA [53]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9
	HAM-Net [23]	65.4	59.0	50.3	41.1	31.0	20.7	11.4	39.8
	ACSNet [72]	-	-	51.4	42.7	32.4	22.0	11.7	-
	ACM-Net [11]	65.3	59.2	49.5	38.4	27.4	16.4	6.9	37.6
	ASL [8]	67.0	-	51.8	-	31.1	-	-	-
	D2-Net [73]	65.7	60.2	52.3	43.4	36.0	-	-	-
	AUMN [13]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	41.5
	UM [74]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	41.9
	FAC-Net [15]	67.6	62.1	52.6	44.3	33.4	22.5	12.7	42.2
	ACG-Net [75]	68.1	62.6	53.1	44.6	34.7	22.6	12.0	42.5
	CO_2-Net [14]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	44.6
	ASM-Loc [60]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	45.1
	DELU [55]	71.5	66.2	56.5	47.7	**40.5**	**27.2**	**15.3**	46.4
	P-MIL [52]	71.8	67.5	58.9	49.0	40.0	27.1	15.1	47.0
	Wang *et al.* [10]	73.0	68.2	60.0	47.9	37.1	24.4	12.7	46.2
	AICL [18]	73.1	67.8	58.2	48.7	36.9	25.3	14.9	46.4
	Xu *et al.* [58]	73.1	66.9	58.3	48.8	36.5	24.4	13.4	45.9
	ASC-Net (ours)	**74.1**	**69.9**	**61.8**	**50.9**	38.3	24.5	12.8	**47.5**

TABLE I: Quantitative comparisons on THUMOS-14 dataset. The mAP is used as an evaluation criterion at different t-IoU thresholds (from 0.1 to 0.7 in steps of 0.1). The symbol † means extra training data are used.

video-independent, and thus they are improved using a cross-attention mechanism through video-specific representations $\mathbf{x}$. In specific, the representation $\mathbf{x}$ is normalized as $\mathbf{u}$ using a ℓ_2 normalization layer, and then queries are further augmented as follows:

$$\begin{aligned} \mathbf{Q} &= \mathbf{g}\mathbf{W}_q, \\ \mathbf{K} &= \mathbf{u}\mathbf{W}_k, \\ \mathbf{V} &= \mathbf{u}\mathbf{W}_v, \\ \mathbf{g}' &= softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \end{aligned} \tag{9}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable matrices, d_k is the dimension of representations, and $\mathbf{g}'$ is the improved queries. In this manner, $\mathbf{g}'$ is video-related and contains both contextual information and action appearance of video-specific, which is conducive for subsequent semantic learning.

Afterwards, for the representation $\mathbf{u}_t$, suppose it is recognized as the class c when generating the foreground CAS $\mathbf{z}_{fg}$, and then we impose a contrastive constraint on it so that snippets with the same category are clustered. Therefore, we optimize an infoNCE loss [76] as follows:

$$\mathcal{L}_{nce} = -\frac{1}{T}\sum_{t=1}^{T}\log\frac{exp(\mathbf{u}_t \cdot \mathbf{g}'[\arg\max_c \mathbf{v}(t), :])}{\sum_{c=0}^{C} exp(\mathbf{u}_t \cdot \mathbf{g}'[c, :])} \tag{10}$$

In this manner, we explicitly encourage snippets identified as the same category to have similar representations.

E. Overall Loss Function

Following the previous work [10], we also introduce an extra guide loss to mitigate the discrepancy of responses from the class-agnostic and class-aware modules:

$$\mathcal{L}_{gui} = \frac{1}{T}\sum_{t=1}^{T}|1 - a_t^{fg} - \mathbf{z}_{fg}[t, 0]|. \tag{11}$$

Therefore, the overall loss is formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{cls}^{fg} + \lambda_1\mathcal{L}_{cls}^{ct} + \lambda_2\mathcal{L}_{cls}^{bg} + \lambda_3\mathcal{L}_{nce} + \lambda_4\mathcal{L}_{gui} \tag{12}$$

where λ_1, λ_2, λ_3 and λ_4 are weights that control the importance of different terms.

F. Inference

During inference, one can get video-level classification results $\hat{\mathbf{y}}_{fg}$ and attention-weighted CAS $\mathbf{z}_{fg}$. Then, we choose consecutive frames that construct proposals $(\hat{t}^s, \hat{t}^e, \phi(c))$ for class c by imposing a threshold α on $\mathbf{z}_{fg}$ and a threshold β on $\mathbf{a}_{fg}$. Different α are typically adopted to generate proposals

with various scales. $\hat{t}^s$ and $\hat{t}^e$ denote the start and end frames. Following [77], $\phi(c)$ is the refined confidence that action c involves in this proposal. $\phi(c)$ absorbs the confidence scores from adjacent areas:

$$\begin{aligned}\phi_{in}(c) &= \frac{\int_{\hat{t}^s}^{\hat{t}^e} \mathbf{z}_{fg}[t,c]}{\hat{t}^e - \hat{t}^s},\\ \phi_{out}(c) &= \frac{\int_{\hat{t}^s-\hat{t}^v}^{\hat{t}^s} \mathbf{z}_{fg}[t,c] + \int_{\hat{t}^e}^{\hat{t}^e+\hat{t}^v} \mathbf{z}_{fg}[t,c]}{2\times\hat{t}^v},\\ \phi(c) &= \phi_{in}(c) - \phi_{out}(c) + \gamma\hat{\mathbf{y}}_{fg}(c).\end{aligned} \tag{13}$$

Notably, $\mathbf{z}_{fg}$ and $\hat{\mathbf{y}}_{fg}$ are respectively frame-level and video-level responses, and their integration can more comprehensively reflect the confidence of action instances. The hyper-parameter γ controls the importance of them. Besides, $\hat{t}^v = \frac{\hat{t}^e-\hat{t}^s}{5}$ describes the inflated contrast area. The final prediction results are acquired by applying a Non-Maximum Suppression (NMS) on $\phi(c)$.

IV. Experiments

A. Dataset and Setting

In this section, we evaluate the proposed ASC-Net on two commonly used datasets, *i.e.*, THUMOS-14 [80] and ActivityNet-1.3 [81], which cover a large variety of action instances and categories, involving a large range of video lengths and fine-grained discrepancies between actions and background. Next, we will elaborate on the details of datasets and experimental settings.

THUMOS-14 Dataset. THUMOS-14 [80] is a popular action localization dataset, where each video includes 15 action instances on average and the length of videos varies from a few minutes to tens of minutes. The training set and test set consist of 200 untrimmed videos and 213 untrimmed videos, respectively. Besides, 20 action categories occur in this dataset. Since the length of videos varies widely and is suitable for assessing the generalization of algorithms, THUMOS-14 is typically used to evaluate the localization performance of algorithms.

ActivityNet-1.3 Dataset. ActivityNet-1.3 [81] is a larger-scale dataset for temporal action localization. About 35% segments in videos have fine-grained discrepancies between actions and background instances, making it challenging to distinguish them. As a result, we choose this dataset to evaluate the localization performance. The training set, validation set, and test set include 10024 videos, 4926 videos, and 5044 videos, respectively. In this paper, we follow [10] and evaluate the model performance on the validation set.

Evaluation Criteria. In this paper, we follow previous work and employ the mean Average Precision (mAP) with different temporal Intersection over Union (t-IoU) thresholds to evaluate the model performance. Specifically, the t-IoU from 0.1 to 0.7 in steps of 0.1 is adopted for THUMOS-14, and from 0.5 to 0.95 in steps of 0.05 is adopted for ActivityNet-1.3.

Implementation Details. In this paper, we first sample consecutively non-overlapping 16 frames of videos as snippets. Based on this, I3D network [37] is employed to extract RGB and optical-flow representations, which are concatenated into 2048-dimensional vectors. These representations encode both the appearance and temporal characteristics of actions. ASC-Net is implemented with PyTorch, trained on GeForce RTX 3090 GPUs, and trained using an Adam optimizer with a learning rate of 1e-4. We train the model for 30 epochs for all datasets. The hyper-parameters λ_1, λ_2, and λ_3 are all set to 0.1, and λ_4 is set to 7e-3. γ is assigned to 0.25 and the dropout rate is 0.5. Besides, the NMS with a t-IoU threshold of 0.4 is employed. For THUMOS-14, σ is set to 8, 8, and 3 for the foreground, context, and background, respectively. The batch size is 8 and the snippet length is set to 750. The threshold α ranges from 0.2 to 0.25 in steps of 0.05, and β ranges from 0.2 to 1.0 in steps of 0.02. For ActivityNet-1.3 σ is assigned to 2, 10, and 10 for the foreground, the context, and the background, respectively. The batch size is 32 and the snippet length is set to 75. The thresholds α and β are range from 0.005 to 0.02 in steps of 0.005.

B. Comparison with State-of-the-Art Methods

In this section, we conduct quantitative experiments on THUMOS-14 and ActivityNet-1.3 benchmarks and compare their results with state-of-the-art methods. In addition, some methods utilizing fully-supervised signals and extra data during training are also displayed for reference. In general, ASC-Net achieves significant progress under the same circumstances. Below we will detail the main results of two datasets.

First, the comparison results with state-of-the-art approaches on the THUMOS-14 dataset are displayed in Table I. We can observe that ASC-Net achieves significant performance improvements over other methods at most t-IoU thresholds. Besides, we also report the average mAP (AVG) of all t-IoU thresholds for comprehensive evaluations, where ASC-Net acquires the highest AVG 47.5. Some algorithms with extra training corpus and supervised signals are also listed in the table so that the boundaries and gaps of performance can be observed. Actually, our approach achieves similar or even better performance. Notably, DELU [55] achieves better localization accuracy over other methods at t-IoU from 0.5 to 0.7. One main reason is that DELU adopts a progressive learning strategy to focus on the entire action instances gradually. Despite achieving good performance, it is inefficient. Nevertheless, a combination of progressive learning and the proposed action-semantic consistency learning may be a future research direction.

Furthermore, we further demonstrate the superiority of ASC-Net on the ActivityNet-1.3 dataset. Results are summarized in Table II. Benefiting from the powerful capabilities of the developed model, we obtain the best localization accuracy at almost all t-IoU thresholds, as well as the AVG. Incredibly, our results are also superior to some algorithms with extra training data and supervised signals, attributed to the well-designed novel framework and action-semantic consistent knowledge learning strategies. Since ActivityNet-1.3 has fine-grained discrepancies among features, the desirable performance achieved by ASC-Net further proves that it has indeed learned some subtle features.

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 [64]	41.3	38.8	35.9	32.9	30.4	27.0	22.2	18.2	13.2	6.1	26.6
	BSN [65]	46.5	-	-	-	-	30.0	-	-	-	8.0	30.0
	G-TAD [50]	50.4	-	-	-	-	34.6	-	-	-	9.0	34.1
Weakly Supervised †	CMCS [78]	36.8	-	-	-	-	22.0	-	-	-	5.6	-
	3C-Net [66]	35.4	-	-	-	-	22.9	-	-	-	8.5	-
	LACP [69]	40.4	-	-	-	-	24.6	-	-	-	5.7	-
Weakly Supervised	UntrimmedNet [57]	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
	AutoLoc [77]	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
	TSM [38]	30.3	-	-	-	-	19.0	-	-	-	4.5	-
	CleanNet [79]	37.1	33.4	29.9	26.7	23.4	20.3	17.2	13.9	9.2	5.0	21.6
	Bas-Net [51]	34.5	-	-	-	-	22.5	-	-	-	5.2	-
	DGAM [12]	40.6	37.0	33.2	29.8	26.6	23.2	19.7	15.1	10.4	5.2	24.1
	EM-MIL [70]	37.4	-	-	-	-	23.1	-	-	-	2.0	-
	TSCN [61]	35.3	-	-	-	-	21.4	-	-	-	5.3	-
	ACM-Net [11]	40.0	36.8	33.9	30.5	27.0	24.0	20.2	15.9	11.0	6.1	24.5
	A2CL-PT [71]	36.8	-	-	-	-	22.0	-	-	-	5.2	-
	AUMN [13]	38.3	-	-	-	-	23.5	-	-	-	5.2	-
	RSKP [52]	40.6	-	-	-	-	24.6	-	-	-	5.9	-
	ASM-Loc [60]	41.0	-	-	-	-	24.9	-	-	-	6.2	-
	Wang *et al.* [10]	41.8	38.5	35.8	32.6	29.2	25.7	22.7	17.5	12.6	6.5	26.3
	P-MIL [52]	41.8	-	-	-	-	25.4	-	-	-	5.2	-
	Xu *et al.* [58]	41.2	-	-	-	-	25.0	-	-	-	**6.5**	-
	ASC-Net (ours)	**43.1**	**40.3**	**37.4**	**34.1**	**31.4**	**28.5**	**25.3**	**21.0**	**15.3**	4.6	**28.1**

TABLE II: Quantitative comparisons on ActivityNet-1.3 dataset. The mAP is used as an evaluation criterion at different t-IoU thresholds (from 0.5 to 0.95 in steps of 0.05). The symbol † means extra training data are used.

$\mathcal{L}_{cls}^{fg}$	$\mathcal{L}_{cls}^{bg}$	$\mathcal{L}_{cls}^{ct}$	$\mathcal{L}_{gui}$	$\mathcal{L}_{nce}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
✓					65.1	58.8	49.0	38.9	27.7	16.4	8.3	37.7
✓	✓				67.1	61.0	50.7	40.5	29.8	18.4	9.4	39.6
✓	✓	✓			68.8	63.4	53.7	43.4	32.8	21.1	10.4	41.9
✓	✓	✓	✓		71.5	65.7	57.7	47.4	35.6	22.7	11.1	44.5
✓	✓	✓	✓	✓	**74.1**	**69.9**	**61.8**	**50.9**	**38.3**	**24.5**	**12.8**	**47.5**

TABLE III: Ablation study on the variants of loss function for THUMOS-14 dataset. The mAP with different t-IoU thresholds is used as evaluation criteria.

$\mathcal{L}_{cls}^{fg}$	$\mathcal{L}_{cls}^{bg}$	$\mathcal{L}_{cls}^{ct}$	$\mathcal{L}_{gui}$	$\mathcal{L}_{nce}$	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	AVG
✓					40.2	37.9	35.2	32.5	29.3	26.8	22.9	19.2	13.6	3.6	26.1
✓	✓				40.7	38.2	35.6	32.6	29.6	27.0	23.2	19.4	13.8	3.5	26.4
✓	✓	✓			41.3	38.5	36.0	32.9	30.0	27.2	23.6	19.7	14.2	3.6	26.7
✓	✓	✓	✓		42.0	39.0	36.5	33.4	30.4	27.6	24.2	20.4	14.5	3.8	27.2
✓	✓	✓	✓	✓	**43.1**	**40.3**	**37.4**	**34.1**	**31.4**	**28.5**	**25.3**	**21.0**	**15.3**	4.6	**28.1**

TABLE IV: Ablation study on the variants of loss function for ActivityNet-1.3 dataset. The mAP with different t-IoU thresholds is used as evaluation criteria.

Method	mAP@t-IoU(%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
ASC-Net/wo	73.2	69.1	60.2	48.7	36.4	22.8	12.4	46.1
ASC-Net/w	**74.1**	**69.9**	**61.8**	**50.9**	**38.3**	**24.5**	**12.8**	**47.5**

TABLE V: Ablation study on model structure for THUMOS-14. ASC-Net/w means the attention selector is adopted for the interactive modeling between the class-agnostic and the class-aware modules, and ASC-Net/wo indicates that there are no interactive behaviors.

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
ASC-Net/wo	41.9	38.7	36.0	33.1	29.4	25.8	22.9	18.5	13.1	4.0	26.3
ASC-Net/w	**43.1**	**40.3**	**37.4**	**34.1**	**31.4**	**28.5**	**25.3**	**21.0**	**15.3**	**4.6**	**28.1**

TABLE VI: Ablation study on model structure for ActivityNet-1.3. ASC-Net/w means the attention selector is adopted for the interactive modeling between the class-agnostic and the class-aware modules, and ASC-Net/wo indicates that there are no interactive behaviors.

C. Ablation Study

To verify the effectiveness of model components, we make sufficient ablation studies in terms of network structure and loss terms on two datasets. Especially, we first explore the influence of different loss terms on localization accuracy, prediction results from the corresponding variants are shown in Table III and IV. Results from THUSMOS-14 and ActivityNet-1.3 consistently demonstrate that the comprehensive loss function can yield the best performance at all t-IoU thresholds. Also,

Method	mAP@t-IoU(%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
Class-agnostic	58.7	53.1	42.5	32.8	22.0	12.3	4.6	32.3
Class-aware	65.1	59.8	49.7	40.6	29.2	18.3	8.9	38.8
ASC-Net	**68.8**	**63.4**	**53.7**	**43.4**	**32.8**	**21.1**	**10.4**	**41.9**

TABLE VII: Ablation study on model structure for THUMOS-14. The integration of class-agnostic and class-aware modules is investigated.

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
Class-agnostic	27.4	25.1	23.5	21.8	19.1	17.3	14.9	11.2	8.8	2.5	17.1
Class-aware	38.9	35.2	33.8	30.1	26.7	23.3	20.2	15.7	11.9	3.1	23.9
ASC-Net	**41.3**	**38.5**	**36.0**	**32.9**	**30.0**	**27.2**	**23.6**	**19.7**	**14.2**	**3.6**	**26.7**

TABLE VIII: Ablation study on model structure for ActivityNet-1.3. The integration of class-agnostic and class-aware modules is investigated.

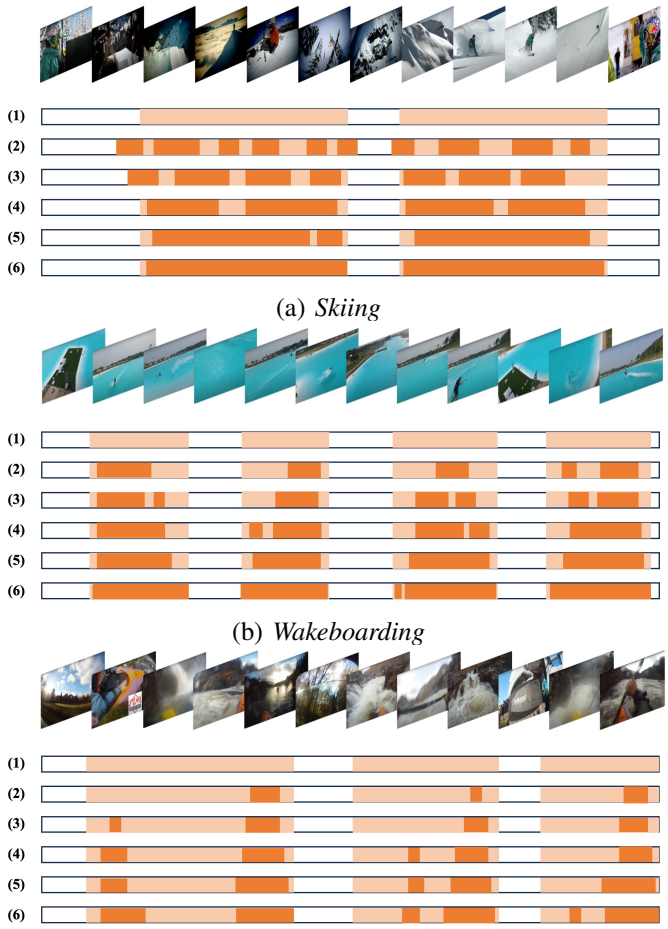
employing any loss term can achieve better performance than removing them, which proves their utility. Notably, introducing the proposed semantic queries with a $\mathcal{L}_{nce}$ loss brings a more significant performance boost than other loss terms. Compared with the variant without semantic queries, the average mAP increases from 44.5 to 47.5 for THUMOS-14 and from 27.2 to 28.1 for ActivityNet-1.3. This phenomenon also reveals the importance of action-semantic information for WS-TAL in the absence of fully-supervised signals.

Based on the above observation, we observe that the performance of some castrated variants is superior to the previous approaches. The reason for this is the advanced model structure. To this end, we further investigate the impact of network structure. First, the relationship between the class-agnostic and the class-aware modules is investigated. The attention selector of the class-agnostic module is discarded, and therefore there are no interactions with the class-aware module. Experimental results are demonstrated in Table V and VI. Undoubtedly, the version with interactions between two modules achieves better performance, where the class-aware module can propagate semantic knowledge into the class-agnostic module and guarantee action-semantic consistency. On the other hand, the integration of class-agnostic and class-aware modules is critical for prediction. For fair comparison, only $\mathcal{L}_{cls}^{fg}$, $\mathcal{L}_{cls}^{ct}$, and $\mathcal{L}_{cls}^{bg}$ are used because $\mathcal{L}_{nce}$ and $\mathcal{L}_{gui}$ are only involved in the class-aware module. Then we can localize action instances by post-processing $\mathbf{a}_{fg}$ and $\mathbf{z}_{fg}$ for class-agnostic and class-aware counterparts, respectively. The experimental results on THUMOS-14 and ActivityNet-1.3 are displayed in Table VII and Table VIII. Unsurprisingly, the complete model acquires the best results compared to its castrated variants. We also discover that the performance of the class-aware module is better than that of the class-agnostic module, which is reasonable since the class-aware module is sensitive to semantic information, but the class-agnostic module is not.

In addition, we also analyze the computational expenses of different structures during inference, and report the average time for each sample on Nvidia GeForce RTX 3090 GPUs, as shown in Table IX. Since the independent class-agnostic branch does not need to tackle semantic information of action instances, it spends minimal inference time. The complete ASC-Net consumes slightly more time than the class-aware branch, but almost the same, due to the parallel structure of the two branches. Also, the video length of ActivityNet-1.3 is longer than THUMOS-14, so the inference cost is greater. In a nutshell, the proposed two-branch structure achieves better localization performance at a tolerable time cost.

Dataset	ASC-Net	Class-aware Branch	Class-agnostic Branch
THUMOS-14	0.083s	0.074s	0.041s
ActivityNet-1.3	0.131s	0.127s	0.092s

TABLE IX: The average time cost of different model structures during inference. Results are acquired by testing samples on Nvidia GeForce RTX 3090 GPUs.



(a) *Skiing*

(b) *Wakeboarding*

(c) *Kayaking*

Fig. 3: We show the qualitative results of three examples with different loss terms. (a) A sample contains the action of "*Skiing*". (b) A sample contains the action of "*Wakeboarding*". (c) A failed sample contains the action of '*Kayaking*". (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) $\mathcal{L}_{cls}^{fg}+\mathcal{L}_{cls}^{bg}+\mathcal{L}_{cls}^{ct}+\mathcal{L}_{gui}+\mathcal{L}_{nce}$

D. Qualitative Visualization

In this section, we visualize some examples and qualitatively investigate the localization performance of ASC-Net with different loss terms. In these three examples, ASC-Net with a complete loss detects more accurate action intervals compared

with other castrated variants. Notably, for the first example of *Skiing*, the background changes are more obvious so that action features are more prominent. Therefore, it is easier to learn. However, background transitions are smoother in the second example, and thus it is more challenging and the algorithm needs to focus on semantic information. When ASC-Net is equipped with $\mathcal{L}_{nce}$, the localization accuracy is significantly improved. This phenomenon also indicates the utility of the proposed query-based semantic mechanism. In addition, we also give a failed example. Since the third example is shot from a first-person perspective and some content is incomplete, ASC-Net fails to localize action intervals.

V. Conclusion

In this paper, a novel ASC-Net consisting of class-aware and class-agnostic modules is developed to jointly extract salient and fine-grained features. The class-aware module is also utilized to guide the class-agnostic module for action-semantic consistency. Furthermore, we assume that frames with similar representations in a given video should be considered as the same action. So we design a query-based contrastive learning paradigm to ensure action-semantic consistency. ASC-Net is verified on publicly available THUMOS-14 and ActivityNet-1.3 datasets, and extensive experiments and ablation studies reveal its effectiveness.

References

[1] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3d spatio-temporal u-net," *IEEE Transactions on Image Processing*, vol. 31, pp. 1573–1586, 2022.

[2] S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, and G. Yu, "X-pool: Cross-modal language-video attention for text-video retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5006–5015.

[3] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 020–20 029.

[4] Y. Zhao, H. Zhang, Z. Gao, W. Guan, J. Nie, A. Liu, M. Wang, and S. Chen, "A temporal-aware relation and attention network for temporal action localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 4746–4760, 2022.

[5] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua, "Enriching local and global contexts for temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 13 516–13 525.

[6] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, "Learning to refactor action and co-occurrence features for temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 884–13 893.

[7] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou, "Sf-net: Single-frame supervision for temporal action localization," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 420–437.

[8] J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, "Weakly supervised action selection learning in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7587–7596.

[9] M. Chen, J. Gao, S. Yang, and C. Xu, "Dual-evidential learning for weakly-supervised temporal action localization," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 192–208.

[10] Y. Wang, Y. Li, and H. Wang, "Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 878–18 887.

[11] S. Qu, G. Chen, Z. Li, L. Zhang, F. Lu, and A. Knoll, "Acm-net: Action context modeling network for weakly-supervised temporal action localization," *arXiv preprint arXiv:2104.02967*, 2021.

[12] B. Shi, Q. Dai, Y. Mu, and J. Wang, "Weakly-supervised action localization by generative attention modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1009–1019.

[13] W. Luo, T. Zhang, W. Yang, J. Liu, T. Mei, F. Wu, and Y. Zhang, "Action unit memory network for weakly supervised temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9969–9979.

[14] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng, "Cross-modal consensus network for weakly supervised temporal action localization," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 1591–1599.

[15] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 8002–8011.

[16] M. Moniruzzaman, Z. Yin, Z. He, R. Qin, and M. C. Leu, "Action completeness modeling with background aware networks for weakly-supervised temporal action localization," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2166–2174.

[17] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 563–579.

[18] Z. Li, Z. Wang, and Q. Liu, "Actionness inconsistency-guided contrastive learning for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 1513–1521.

[19] W. Sun, R. Su, Q. Yu, and D. Xu, "Slow motion matters: A slow motion enhanced network for weakly supervised temporal action localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 354–366, 2023.

[20] J. Wu, W. Sun, T. Gan, N. Ding, F. Jiang, J. Shen, and L. Nie, "Neighbor-guided consistent and contrastive learning for semi-supervised action recognition," *IEEE Transactions on Image Processing*, vol. 32, pp. 2215–2227, 2023.

[21] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6752–6761.

[22] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," in *Proceedings of the International Conference on Learning Representations*, 2019.

[23] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1637–1645.

[24] J.-T. Lee, S. Yun, and M. Jain, "Leaky gated cross-attention for weakly supervised multi-modal temporal action localization," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 3213–3222.

[25] D. Li, T. Yao, Z. Qiu, H. Li, and T. Mei, "Long short-term relation networks for video action detection," in *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 629–637.

[26] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 426–12 433.

[27] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[28] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141–154, 2015.

[29] Y. Shi, Z. Wei, H. Ling, Z. Wang, P. Zhu, J. Shen, and P. Li, "Adaptive and robust partition learning for person retrieval with policy gradient," *IEEE Transactions on Multimedia*, vol. 23, pp. 3264–3277, 2020.

[30] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[31] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1817–1824.

[32] Y. Wang and S. Chen, "Multi-agent trajectory prediction with spatio-temporal sequence fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 13–23, 2021.

[33] Y. Wang and S.-J. Zhao, "Consistent representation learning across modalities for zero-shot image recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.

[34] Y. Wang, S. Zhao, R. Zhang, X. Cheng, and L. Yang, "Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 236–248, 2020.

[35] Y. Wang and S. Zhao, "Cross-modal representation reconstruction for zero-shot classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 2820–2824.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[37] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[38] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7083–7093.

[39] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6202–6211.

[40] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3211.

[41] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 492–510.

[42] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proceedings of the ACM International Conference on Multimedia*, 2017, pp. 988–996.

[43] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022.

[44] S. Yan, X. Xiong, A. Nagrani, A. Arnab, Z. Wang, W. Ge, D. Ross, and C. Schmid, "Unloc: A unified framework for video localization tasks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 13 623–13 633.

[45] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, "Tridet: Temporal action detection with relative boundary modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 857–18 866.

[46] C. Lin, J. Li, Y. Wang, Y. Tai, D. Luo, Z. Cui, C. Wang, J. Li, F. Huang, and R. Ji, "Fast learning of temporal action proposal via dense boundary generator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 499–11 506.

[47] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Learning salient boundary feature for anchor-free temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3320–3329.

[48] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3889–3898.

[49] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, "Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2602–2610.

[50] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 156–10 165.

[51] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 320–11 327.

[52] H. Ren, W. Yang, T. Zhang, and Y. Zhang, "Proposal-based multiple instance learning for weakly-supervised temporal action localization," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2394–2404.

[53] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "Cola: Weakly-supervised temporal action localization with snippet contrastive learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 010–16 019.

[54] M. N. Rizve, G. Mittal, Y. Yu, M. Hall, S. Sajeev, M. Shah, and M. Chen, "Pivotal: Prior-driven supervision for weakly-supervised temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 992–23 002.

[55] M. Chen, J. Gao, S. Yang, and C. Xu, "Dual-evidential learning for weakly-supervised temporal action localization," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 192–208.

[56] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 35–44.

[57] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4325–4334.

[58] Z. Xu, K. Wei, E. Yang, C. Deng, and W. Liu, "Bilateral relation distillation for weakly supervised temporal action localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 458–11 471, 2023.

[59] H. Linjiang, W. Liang, and L. Hongsheng, "Weakly supervised temporal action localization via representative snippet knowledge propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3272–3281.

[60] B. He, X. Yang, L. Kang, Z. Cheng, X. Zhou, and A. Shrivastava, "Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 915–13 925.

[61] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, and G. Hua, "Two-stream consensus network for weakly-supervised temporal action localization," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 37–54.

[62] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 53–63.

[63] J. Shen, D. Tao, and X. Li, "Modality mixture projections for semantic video event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1587–1596, 2008.

[64] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2914–2923.

[65] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[66] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8679–8687.

[67] X.-Y. Zhang, H. Shi, C. Li, and P. Li, "Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 886–12 893.

[68] C. Ju, P. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Point-level temporal action localization: Bridging fully-supervised proposals to weakly-supervised losses," *arXiv preprint arXiv:2012.08236*, 2020.

[69] P. Lee and H. Byun, "Learning action completeness from points for weakly-supervised temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 13 648–13 657.

[70] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 729–745.

[71] K. Min and J. J. Corso, "Adversarial background-aware loss for weakly-supervised temporal activity localization," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 283–299.

[72] Z. Liu, L. Wang, Q. Zhang, W. Tang, J. Yuan, N. Zheng, and G. Hua, "Acsnet: Action-context separation network for weakly supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2233–2241.

[73] S. Narayan, H. Cholakkal, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 13 608–13 617.

[74] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1854–1862.

[75] Z. Yang, J. Qin, and D. Huang, "Acgnet: Action complement graph network for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 3090–3098.
[76] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
[77] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 154–171.
[78] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1298–1307.
[79] Z. Liu, L. Wang, Q. Zhang, Z. Gao, Z. Niu, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through contrast based evaluation networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3899–3908.
[80] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.
[81] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.