Feature Weakening, Contextualization, and Discrimination for Weakly Supervised Temporal Action Localization

Md Moniruzzaman, *Graduate Student Member, IEEE,* Zhaozheng Yin, *Senior Member, IEEE*

***Abstract*—Weakly-supervised Temporal Action Localization (W-TAL) aims to train a model to localize all action instances potentially from different classes in an untrimmed video, using a training dataset that has video-level action class labels but has no detailed annotations on the start and end timestamps of action instances. We propose to solve the W-TAL problem from the feature learning aspect, with a new architecture, termed F3-Net, which includes (1) a Feature Weakening (FW) module that can identify and randomly weaken either the most discriminative action or the most discriminative background features over the training iterations to force the network to precisely localize the action instances in both discriminative and ambiguous action-related frames, without spreading to the background intervals; (2) a Feature Contextualization (FC) module that can infer the global contexts among video segments and attentionally fuse them with the local contexts from individual video segments to generate more representative features; and (3) a Feature Discrimination (FD) module that can highlight the most discriminative video segments/classes corresponding to each class/segment, respectively, for localizing multiple action instances from different classes within a video. Experimental results on THUMOS14 and ActivityNet1.3 demonstrate the state-of-the-art performance of our F3-Net, and the FW and FC are also effective plug-in modules to improve other methods. This project will be available at https://moniruzzamanmd.github.io/F3-Net/**

***Index Terms*—Temporal action localization, Feature weakening, Feature contextualization, Feature discrimination**

I. Introduction

TEMPORAL action Localization, which localizes action instances (i.e., time intervals) in untrimmed videos along the temporal dimension, is one of the challenging video understanding tasks. The methods with high performance are under the fully-supervised setting, which requires the video-level action class labels for each training video along with the detailed temporal annotations (start and end time-stamps) of each action instance within the training video [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Since the fully-supervised approach requires a lot of annotation efforts, recently, Weakly-supervised Temporal Action Localization (W-TAL) methods which learn to localize action instances in untrimmed videos based on training datasets with only video-level labels, have been developed [14], [15], [16], [17], [18].

M. Moniruzzaman is with the Department of Computer Science, Stony Brook University, Stony Brook, New York, 11794 (e-mail: mmoniruzzama@cs.stonybrook.edu)

Z. Yin is with the Department of Computer Science and Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, 11794 (e-mail: zyin@cs.stonybrook.edu)

Challenges and motivation: The W-TAL task is challenging from a few aspects:

(1) Actions in ambiguous frames: Many W-TAL methods have achieved good results in coarse localization [14], [19], [20], [21], [22], but they remain unsatisfactory to locate the complete time intervals of action instances. The main reason comes from that, the networks look only at the highly discriminative frames, but apart from the highly discriminative frames, there are some ambiguous frames which are possible to contain action instances and are overlooked by those weakly-supervised networks. The action instances are supposed to be completely localized in both highly discriminative intervals and ambiguous intervals, without spreading to the background intervals which contain unrelated frames, as shown in Fig. 1. Therefore, the motivated research question is: *how to design a feature learning module that can identify and intentionally weaken the most discriminative action and background features so that the network can be enforced to discover the action instances in both discriminative and ambiguous intervals for the complete temporal action localization*?

(2) Local and global contexts: Temporal contextual information is important for the temporal action localization. The most typical approach is to divide a video into short video segments first, and then use a pre-trained network to extract features from each segment independently, which refers to the local contexts. However, this process neglects the global contexts, which provide essential clues for the temporal action localization. For example, as shown in Fig. 2, the "Long Jump" action usually contains the "running" and the "jumping" video segments. Although these two kinds of video segments have their distinct characteristics, they also share common features considering the "Long Jump" action. Therefore, the research question is: *how to design a feature contextualization module that can infer the global contexts between video segments and fuse them with the local contexts from individual video segments, generating more representative features for the temporal action localization*?

(3) Multi-class multi-instance temporal action localization: An untrimmed video may contain multiple action instances with different action class labels. After splitting the video into segments, a video segment may contain multiple action classes, and an action class may have multiple instances in separated video segments. As shown in Fig. 3, the third video segment contains both "Cricket Bowling" and "Cricket Shot" actions, and the "Cricket Bowling" action has two action instances in the first and third video segments, while

Fig. 1. An illustration of background, ambiguous, and highly discriminative frames. The background frames are not related to the action. The ambiguous frames are possible to contain action instances. The highly discriminative frames are highly related to the action class. It is expected to localize the action instances in both highly discriminative and ambiguous action frames for the complete temporal action localization.

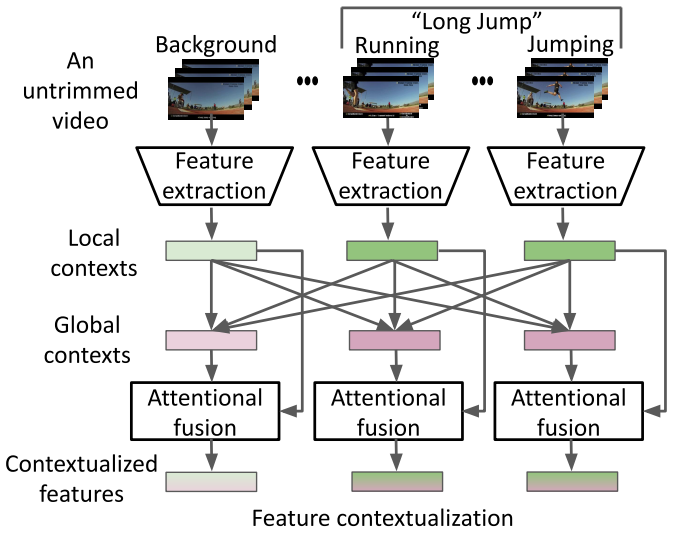

Fig. 2. An illustration of the local and global contexts. The local contexts are extracted from each video segment independently, while the global contexts are computed by exploiting the relationships between different video segments. We introduce a feature contextualization module to attentionally fuse both the local and global contexts into contextualized features for accurate action localization in the temporal domain.

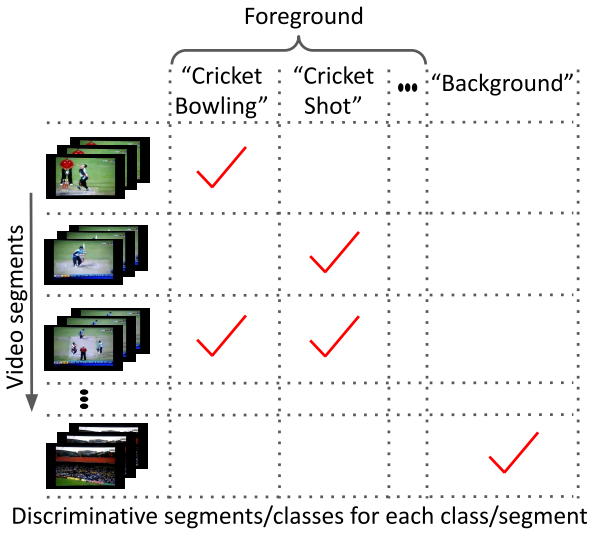

Fig. 3. An illustration of the multi-class multi-instance temporal action localization. Highlighting the discriminative video segments related to action classes and the discriminative action classes within each video segment can be useful for multi-class multi-instance temporal action localization.

the “cricket shot” action has two action instances in the second and third video segments. Therefore, the motivated research question is: *how to design a feature discrimination module that can highlight both the most discriminative video segments corresponding to each action class, and the most discriminative classes within each video segment, eventually improving the performance of temporal action localization*?

Our proposal and contributions: Our three research questions to address the three corresponding challenges are motivated from the feature learning perspective, leading to a novel network architecture, termed F3-Net, which includes Feature Weakening (FW), Feature Contextualization (FC), and Feature Discrimination (FD) modules in a unified network, as shown in Fig. 4. Our main contributions are four-fold:

- We introduce a new FW module that can identify and randomly weaken either the most discriminative action or the most discriminative background features over the training iterations to force the network to precisely localize the action instances in both discriminative and ambiguous action-related frames, without spreading to the background intervals.
- We present a new FC module that first explores the correlative information between different segments in a video to extract the global contexts, and then attentionally fuses both the local and global contexts into more representative contextualized features for accurate temporal action localization.
- We introduce a new FD module that explicitly highlights both the most discriminative video segments corresponding to each action class, and the most discriminative classes within each video segment, for localizing multiple action instances from different classes within a video.
- Our F3-Net based on the three feature learning modules outperforms all the related methods on the THUMOS14 and ActivityNet1.3 datasets, and we also validate that the FW and FC modules are effective plug-in modules to improve the performance of the previous methods.

II. Related Works

Different from action recognition [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], the temporal action localization task aims to localize action instances with temporal boundaries. W-TAL methods address

this problem without temporal annotations. From the aspects of feature learning and pseudo-labeling strategy, we summarize the W-TAL methods into four categories.

Feature weakening-related methods: The feature weakening related methods first identify the most discriminative features and then weaken them to pay more attentions to less discriminative features for the complete temporal action localization. Motivated by the object completeness modeling in object detection task [38], [39], [40], Hide-and-Seek [41] first hid random frame sequences to force the network to discover different action parts for the action completeness modeling. Recently, Step-by-Step Erasion [42] and ACM-BANet [15] trained the network in multiple steps, where during each training step, the network first identified the most discriminative video segments, which were then intentionally weakened to force the network to discover the action instances from the remaining segments. However, it is difficult to set a proper number of training steps, and the attention regions may gradually expand to the background intervals as the number of training steps increases, which significantly decreases the localization performance. *To overcome these problems, we train our network in a single step with a novel feature weakening module that randomly weakens either the most discriminative action or the background features over the training iterations to force the network to precisely localize the action instances in both discriminative and ambiguous action-related frames, without spreading to the background intervals.*

Note, the proposed Feature Weakening (FW) module may seem to be conceptually similar to the Dropout regularization. However, the Dropout regularization randomly zeros out the hidden nodes of the neural network, while our FW identifies and randomly weakens either the most discriminative action or the background features to force the network to further discover more action related frames in the ambiguous intervals.

Feature contextualization-related methods: Most of the existing W-TAL methods [14], [19], [20], [22], [43] neglected the global contexts and only utilized the local contexts for the temporal action localization. The feature contextualization-related methods aim to exploit both the local and global contexts in a video, and then fuse them to generate more representative contextualized features for the temporal action localization. Recently, ContextLoc [44] modeled the local and global contexts in a unified framework for the fully-supervised temporal action localization. More recently, AUMN [45] introduced a weakly-supervised temporal action localization network that computes the relationships between different segments in a video to extract the global contexts, which are then fused with the local contexts by simply performing the element-wise summation. However, the element-wise summation provides a fixed importance to the local and global contexts during the fusion. *Differently, we present a new feature contextualization module that first explores the correlative information between different segments in a video to extract the global contexts, and then learns attention weights on local and global contexts to fuse them accordingly, to generate more representative contextualized features for the accurate temporal action localization.*

Feature discrimination-related methods: The feature discriminative-related methods learn to highlight the most discriminative features by designing W-TAL networks with attention mechanisms and different loss terms. UntrimmedNet [19] proposed attention mechanisms to highlight the most discriminative frames for the W-TAL task. STPN [20] introduced a sparse temporal pooling network to focus on key video segments. AutoLoc [22] introduced a contrastive loss function based on the class activation sequence, and Paul et al. [14] utilized pairwise video similarity constraints to localize the discriminative video segments in the temporal domain corresponding to each action class. 3C-Net [21] utilized category, counting, and center losses to highlight the class-wise discriminative video segments in untrimmed videos. RPN [46] adopted a clustering loss and A2CL-PT [47] employed a triplet loss to highlight the discriminative action frames from the background frames. Recently, some works [18], [48], [49] explicitly modeled the background class to suppress the background frames for the W-TAL task. More recently, FAC-Net [43] introduced a multi-branch architecture to highlight the most discriminative video segments by maintaining the foreground-action consistency. FTCL [50] designed Fine-grained Sequence Distance (FSD) and Longest Common Subsequence (LCS) contrasting objectives for the W-TAL task. Although these methods achieve remarkable progress, they only highlight the most discriminative video segments corresponding to each action class. However, as shown in Fig. 3, an action class may have multiple instances in separated video segments, and a video segment may contain multiple action classes. Therefore, highlighting the most discriminative video segments related to action classes and the most discriminative action classes within each video segment can be very useful for multi-class multi-instance temporal action localization. *To address this challenge, we design our feature discrimination module to highlight both the most discriminative video segments corresponding to each action class, and the most discriminative classes within each video segment.*

Pseudo label-related methods: Pseudo label-related methods iteratively refine a W-TAL network by leveraging segment-wise pseudo labels to distinguish the foreground and background segments. At first, the RefineLoc [51] introduced the pseudo-labeling strategy in W-TAL, where the pseudo labels were generated from the previous detection results to iteratively refine the action localization network. Later, different methods tried to generate high-quality pseudo labels to refine the W-TAL networks. For example, EM-MIL [52] utilized an expectation-maximization framework to generate the pseudo labels, while UGCT [17] generated the pseudo labels from the modality collaborative learning and uncertainty estimation to learn more robust attention weights. More recently, Huang et al. [53] introduced a representative snippet summarization and propagation framework to generate the pseudo labels. However, since the pseudo label-related methods usually generate the initial pseudo labels from the Class Activation Scores (CAS) of an existing W-TAL network, the performances of these methods heavily relied on the quality of the CAS of that W-TAL network.

In this paper, we propose to solve the W-TAL problem from the feature learning aspect. Table I summarizes the in-

TABLE I
INNOVATIONS OF OUR FEATURE WEAKENING, FEATURE CONTEXTUALIZATION, AND FEATURE DISCRIMINATION COMPARED TO RELATED METHODS.

Existing methods related to feature weakening	**Innovation of our feature weakening**
Existing feature weakening-related methods (e.g., [15], [42]) trained the network in multiple steps. During each step, the network first identified the most discriminative features, which were then erased or weakened to force the network to discover the action instances from the remaining features. **Limitation:** It is difficult to set a proper number of training steps to discover different complementary action segments for different action classes, and the attention regions may gradually expand to the background intervals as the training steps increase, which downgrades the localization performance.	We train our network in a single step with a novel feature weakening module that randomly weakens either the most discriminative action or the most discriminative background features over the training iterations to force the network to precisely localize the action instances in both the discriminative and ambiguous action-related segments, without spreading to the background intervals. Please refer to Fig. 5 and Sec. III-B for the details.
Existing methods related to feature contextualization	**Innovation of our feature contextualization**
Existing feature contextualization-related methods (e.g., [45]) first explored the correlative information between different segments in a video to extract the global contexts, which were then fused with the local contexts by simply performing the element-wise summation. **Limitation:** The element-wise summation provides a fixed importance to the local and global contexts during the fusion.	We introduce a new feature contextualization module that first explores the correlative information between different segments in a video to extract the global contexts, and then learns attention weights on local and global contexts to fuse them accordingly, to generate more representative contextualized features for the accurate temporal action localization. Please refer to Fig. 6 and Sec. III-C for the details.
Existing methods related to feature discrimination	**Innovation of our feature discrimination**
Existing feature discrimination-related methods (e.g., [14], [43], [49]) highlighted the most discriminative video segments corresponding to each action class. **Limitation:** These methods did not explicitly highlight the most discriminative action classes within each video segment, which can be useful for multi-class multi-instance temporal action localization.	We design our feature discrimination module to highlight both the most discriminative video segments corresponding to each action class, and the most discriminative classes within each video segment. Please refer to Fig. 7 and Sec. III-D for the details.

novations of our feature weakening, feature contextualization, and feature discrimination modules, compared to the existing methods.

III. METHODOLOGY

The workflow of our W-TAL framework is illustrated in Fig. 4. During the training, given untrimmed videos and their video-level ground-truth labels, we perform a feature embedding (Sec. III-A), a Feature Weakening (FW) module (Sec. III-B), a Feature Contextualization (FC) module (Sec. III-C), and a Feature Discrimination (FD) module (Sec. III-D). Given a testing video, we not only recognize its action classes but also localize the temporal window of each action instance. Note, the FW module is deactivated during the testing.

A. Feature Embedding

Given an untrimmed video $\mathbf{V} = \{s_t\}_{t=1}^{T}$, which is divided into T non-overlapping video segments, we extract the D-dimensional features for each video segment s_t by a pre-trained I3D network [36], generating the feature map of video $\mathbf{V}$ as $\mathbf{X} \in \mathbb{R}^{T\times D}$. Since the extracted features from I3D are learned for the action recognition task originally, we load the feature map $\mathbf{X}$ into a two-layer temporal convolutional network to generate the embedded feature map $\mathbf{X}_e \in \mathbb{R}^{T\times D}$, which is tuned for the W-TAL task.

B. Feature Weakening (FW) Module

Usually, the W-TAL methods follow a localization-by-classification pipeline. Unfortunately, the action classification network tends to focus on the most discriminative features to pursue its classification accuracy, and the discriminative features may be from the most salient portion of the action time interval, which is not sufficient for the temporal action localization task. Ideally, the action instances are supposed to be precisely localized in time intervals with both discriminative and ambiguous action-related features, without spreading to the background intervals. To overcome this challenge, we introduce a Feature Weakening (FW) module to identify and randomly weaken either the discriminative action or the discriminative background features so that the network can be enforced to discover the action instances in both discriminative and ambiguous action intervals for complete localization.

Since the W-TAL networks aim to produce high confidence scores for the target action classes for the action-related segments and low confidence scores for all action classes for the background segments, the networks learn to highlight the action-related segments and suppress the background segments, eventually, produce large embedded feature magnitudes for the action-related segments and small magnitudes for the background segments. Therefore, we identify the discriminative action, ambiguous, and background features based on the

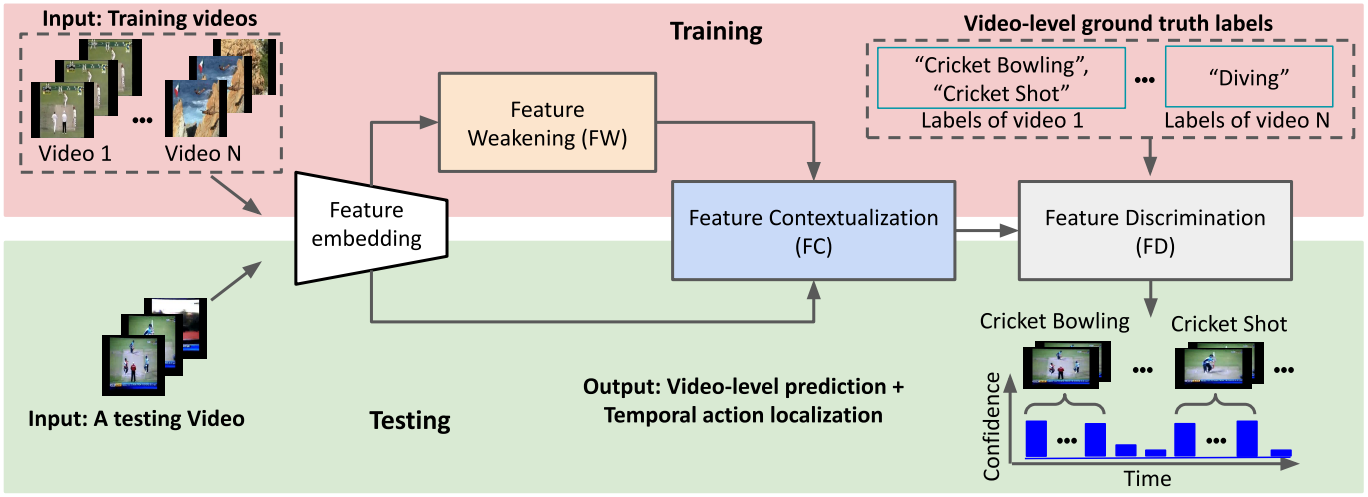

Fig. 4. The workflow of our F3-Net. During the training, given untrimmed videos and their video-level ground-truth labels, the F3-Net first performs the feature embedding. Then, a Feature Weakening (FW) module is designed for action completeness modeling. Thereafter, a Feature Contextualization (FC) module is designed to generate contextualized features. Finally, a Feature Discrimination (FD) module is designed to highlight the most discriminative video segments/classes corresponding to each class/segment, respectively. Given a testing video, the F3-Net not only recognizes its action classes, but also localizes the temporal window of each action instance. Note, the FW module is deactivated during the testing.

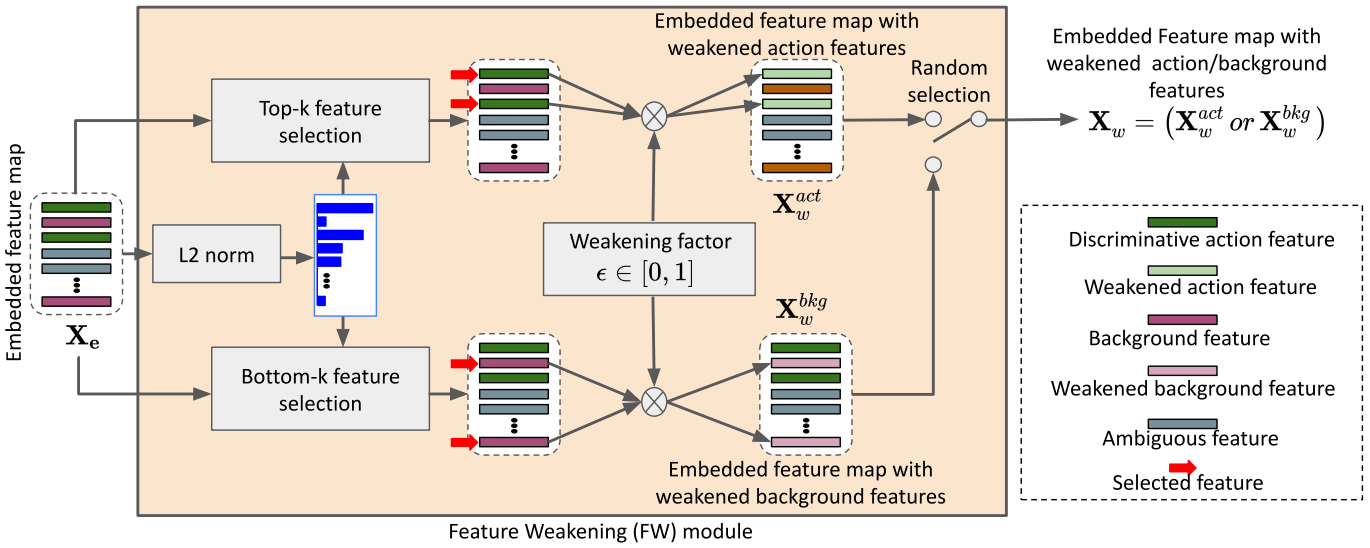

Fig. 5. Illustration of our Feature Weakening (FW) module. During the training, the FW module selects and weakens the most discriminative action and background features based on feature magnitudes. The embedded feature maps with weakened action/background features are randomly selected during the training so that the network is forced to discover the action instances in both discriminative and ambiguous frames for the complete temporal action localization and reduce the distraction from background frames. The FW module is deactivated during the testing phase.

embedded feature magnitude. As shown in Fig. 5, first, we apply the L2 norm on the embedded feature map $\mathbf{X}_e \in \mathbb{R}^{T\times D}$ to compute the segment-wise feature magnitude row-by-row. Then, we treat the top k^{act} and bottom k^{bkg} features in terms of the feature magnitude as the highly discriminative action and background features, respectively. We consider the remaining features in $\mathbf{X}_e$ as ambiguous features. After that, we generate the feature map with weakened action features, $\mathbf{X}_w^{act} \in \mathbb{R}^{T\times D}$, by multiplying the discriminative action features with a weakening factor ϵ ($\epsilon \in [0,1]$), where the ambiguous and background features are unchanged. In this way, we weaken the most discriminative action features and encourage the network to look at ambiguous features. However, if we persistently use the feature map with weakened action features, the discriminative action features are always weakened during the training phase and the network may wrongly shift its focus to unexpected background segments. To remedy this, from the feature map $\mathbf{X}_e$, we generate another feature map named as feature map with weakened background features, $\mathbf{X}_w^{bkg} \in \mathbb{R}^{T\times D}$, by multiplying the background features with a weakening factor ϵ ($\epsilon \in [0,1]$), where the ambiguous and discriminative action features are unchanged. Finally, the output of the FW module, $\mathbf{X}_w$, randomly selects either $\mathbf{X}_w^{act}$ or $\mathbf{X}_w^{bkg}$ with equal chances at every training iteration. With such a FW module, the network is forced to discover action instances in both discriminative and ambiguous intervals, and

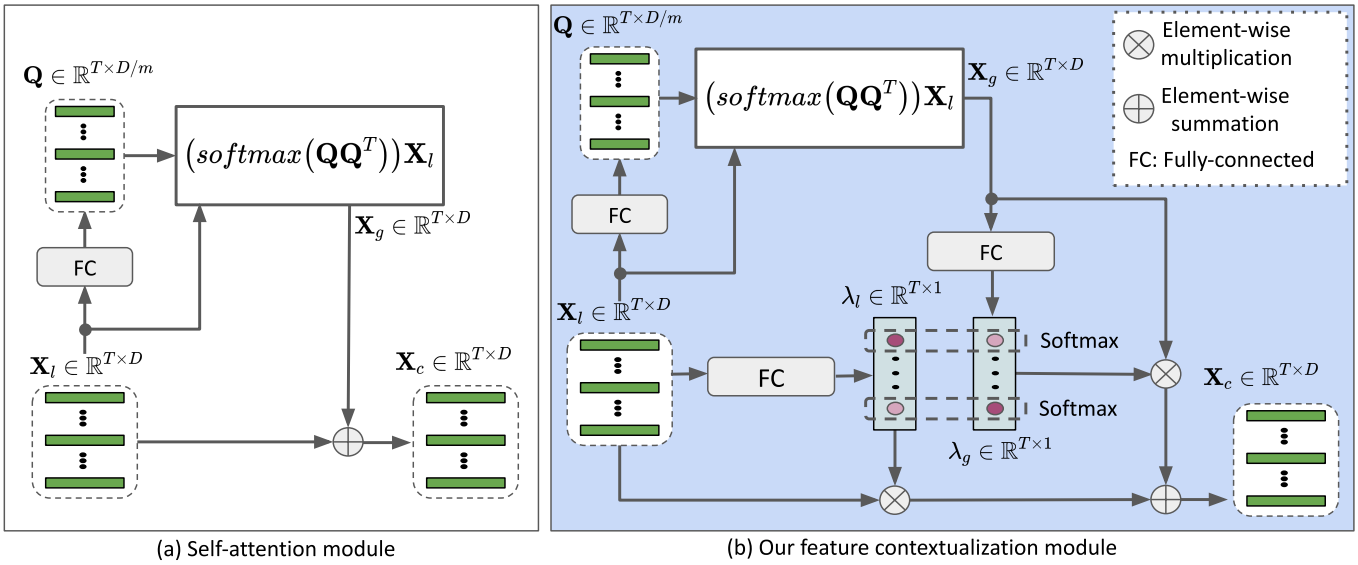

Fig. 6. Typical self-attention module vs. our Feature Contextualization (FC) module.

neglect the distractions brought by the background. It should be noted that the FW module is only used during the training. During the testing phase, the FW module is deactivated, i.e., the embedded feature map is directly fed into the feature contextualization module, as shown in Fig. 4.

C. Feature Contextualization (FC) Module

The features of all video segments, $\mathbf{X}_l \in \mathbb{R}^{T\times D}$, which is either the $\mathbf{X}_w$ during training or the $\mathbf{X}_e$ during testing, contains the local context information of individual segments. We design a Feature Contextualization (FC) module to explore the correlation between different video segments to extract the global contexts, and then attentionally fuse the local and global contexts to generate representative contextualized features.

Specifically, our FC module first uses a fully-connected layer to encode the local features $\mathbf{X}_l$ into a set of compact queries $\mathbf{Q} \in \mathbb{R}^{T\times D/m}$, where m is a hyper-parameter to control the memory reading efficiency. Since we are looking for the correlation between video segments, we compute dot products of a query with all queries, followed by a softmax function, to obtain a $T \times T$ matrix, which contains the relevance of each segment to other segments. Then, with this matrix, we obtain the global contexts $\mathbf{X}_g \in \mathbb{R}^{T\times D}$ by aggregating the local contexts, which can be formulated as:

$$\mathbf{X}_g = (\text{softmax}(\mathbf{Q}\mathbf{Q}^T))\mathbf{X}_l \tag{1}$$

The $\mathbf{X}_g$ is finally merged with the $\mathbf{X}_l$ to get the contextualized features $\mathbf{X}_c \in \mathbb{R}^{T\times D}$. One simple way to combine the local and global context features is the element-wise summation, i.e., $\mathbf{X}_c = \mathbf{X}_l + \mathbf{X}_g$, as shown in Fig. 6(a). This fusion gives static equal weights to the local and global contexts.

Differently, we propose to learn the attention weights on local and global contexts first, and then fuse them accordingly, as shown in Fig. 6(b). Specifically, we first use fully-connected layers to compute the attention scores $\lambda_l \in \mathbb{R}^{T\times 1}$ and $\lambda_g \in \mathbb{R}^{T\times 1}$ from $\mathbf{X}_l \in \mathbb{R}^{T\times D}$ and $\mathbf{X}_g \in \mathbb{R}^{T\times D}$, respectively. Then, we apply the softamx across these two attention scores to normalize them and perform attention-weighted merging to fuse the local and global contexts, as follows:

$$\mathbf{X}_c = \lambda_l \otimes \mathbf{X}_l + \lambda_g \otimes \mathbf{X}_g \tag{2}$$

where $\otimes$ represents the element-wise multiplication. We find that the use of the attention-weighted merging leads to improvement in performance when compared to traditional element-wise summation, to be shown in ablation studies.

D. Feature Discrimination (FD) Module

We design a Feature Discrimination (FD) module to highlight both the most discriminative video segments for each action class and the most discriminative classes within each video segment. As shown in Fig. 7, first, we randomly initialize a *class-agnostic foreground classifier* $\mathbf{w}_f \in \mathbb{R}^{1\times D}$ (i.e., the classifier learns weights to classify all action classes (foreground) without depending on any specific action class) and a *class-specific classifier* $\mathbf{W}_{cs} \in \mathbb{R}^{(C+1)\times D}$ (i.e., the classifier learns weights depending on the action classes and the background class), where C represents the number of action classes, and the $(C+1)$th class corresponds to the background. With the help of these two classifiers, we design three branches in the FD module: (1) Action-only classification branch via class-agnostic attention (CA); (2) Action-only classification branch via class-specific attention (CS); and (3) Action and background classification branch (AB).

Complementarity of three branches in the FD module: The CA branch first uses a class-agnostic attention to generate the foreground activation scores and then classifies a video in regard to only the C action classes. The CS branch first uses a class-specific attention to highlight the most discriminative video segments for each class and the most discriminative classes within each video segment, and then classifies a video

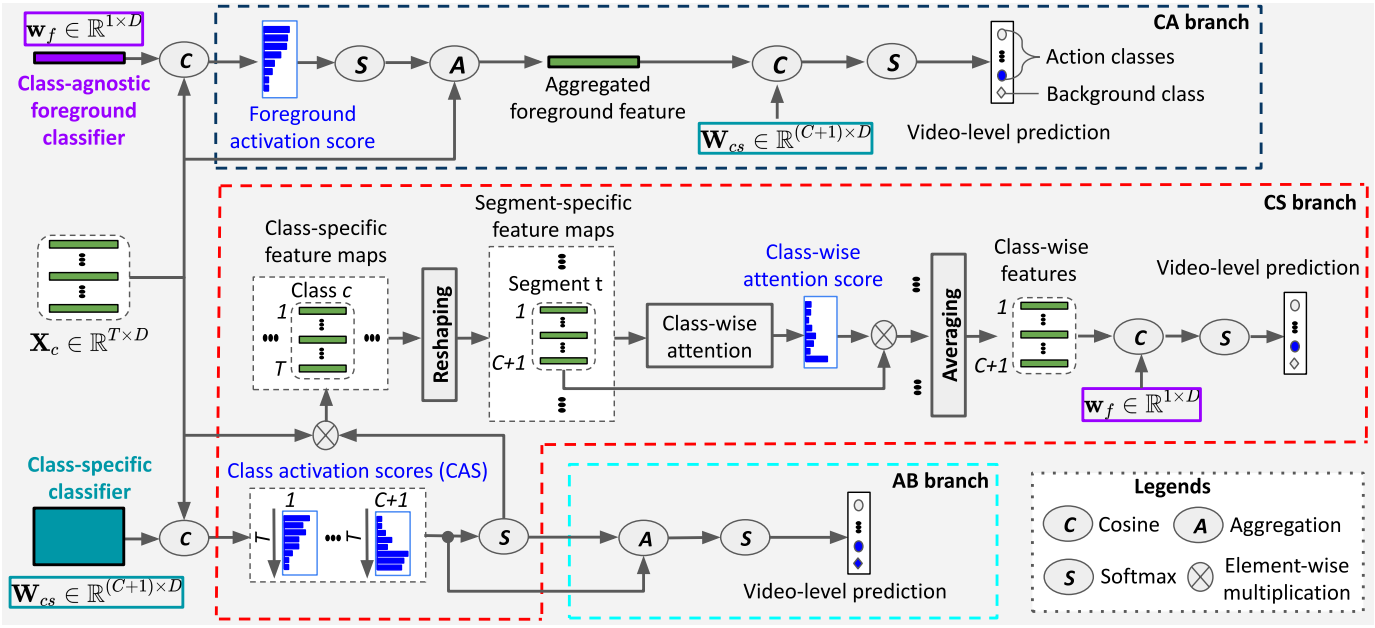

Fig. 7. Illustration of our Feature Discrimination (FD) module. The action-only classification branch via class-agnostic attention (CA) computes the foreground activation scores, while the action-only classification branch via class-specific attention (CS) explicitly highlights the most discriminative video segments for each class and the most discriminative classes within each video segment. Both branches classify the video by focusing on action classes. The action-background classification branch (AB) classifies the input video regarding the action classes and the background class.

in regard to only the C action classes. The training objectives of these two branches are aligned, which regularizes the consistency between the foreground and the C action classes of CS that localizes the action-related frames of various actions in class activation scores (CAS) as foreground. On the other hand, the AB branch classifies an input video in regard to both the C action classes and the $(C+1)$th background class, which helps our FD suppress activations from background segments, leading to effective action-background separation. During training, we employ all three branches and jointly train them to learn the background class as well as the action classes. Without the AB branch, the CS branch cannot learn the background class, nor computes the attention scores between action and background class, thus background segments might be classified wrongly as action classes. Therefore, the joint optimization of these three branches can maintain the consistency between the foreground and the C action classes of CS, and also separate the C action classes from the $(C+1)$th background class.

Action-only classification branch via class-agnostic attention (CA): We first calculate the cosine similarity between the $\mathbf{w}_f$ and $\mathbf{X}_c$ to obtain the *foreground attention score* $\mathbf{a} \in \mathbb{R}^{T\times 1}$, as follows:

$$\mathbf{a}(t) = \cos(\mathbf{X}_c(t,\cdot), \mathbf{w}_f), \qquad t \in [1, T] \tag{3}$$

The foreground attention score vector $\mathbf{a}$ is passed through a softmax layer to get the normalized foreground attention score: $\tilde{\mathbf{a}} \in [0,1]^T$. Then, the *aggregated foreground feature* vector $\mathbf{z} \in \mathbb{R}^{1\times D}$ of the video is computed by:

$$\mathbf{z} = \sum_t \tilde{\mathbf{a}}(t)\mathbf{X}_c(t,\cdot) \tag{4}$$

The video-level classification result $\tilde{\mathbf{y}}_{ca} \in \mathbb{R}^{C+1}$ for the CA branch is finally obtained by computing the cosine similarity between the $\mathbf{z}$ and $\mathbf{W}_{cs}$, and then passing the similarity scores through a softmax layer:

$$\tilde{\mathbf{y}}_{ca}(j) = \frac{\exp(\cos(\mathbf{z}, \mathbf{W}_{cs}(j,\cdot))}{\sum_i \exp(\cos(\mathbf{z}, \mathbf{W}_{cs}(i,\cdot))}, \qquad j \in [1, C+1] \tag{5}$$

The classification loss of the CA branch is defined by:

$$\ell_{ca} = \sum_{j=1}^{C+1} -\mathbf{y}_{ca}(j) \log \tilde{\mathbf{y}}_{ca}(j) \tag{6}$$

where $\mathbf{y}_{ca} = [\mathbf{y}_{ca}(1), ..., \mathbf{y}_{ca}(C), 0]$ is the ground-truth vector, in which $\mathbf{y}_{ca}(c)$ is set to 1 if a video contains action class c. Since this branch aims to classify an input video only on action classes, the label for the background is set to 0.

Action-only classification branch via class-specific attention (CS): Given the $\mathbf{X}_c$, we compute the cosine similarities between $\mathbf{X}_c$ and $\mathbf{W}_{cs}$ to obtain the segment-level *Class Activation Scores* (CAS) $\mathbf{P} \in \mathbb{R}^{T\times(C+1)}$ as:

$$\mathbf{P}(t,j) = \cos(\mathbf{X}_c(t,\cdot), \mathbf{W}_{cs}(j,\cdot)) \tag{7}$$

where t and j represent the t-th video segment and the j-th class, respectively.

Now, we aim to highlight the discriminative video segments for each class. To achieve this goal, first, we apply the softmax on $\mathbf{P}$ along the temporal dimension, and obtain the normalized class activation scores: $\tilde{\mathbf{P}} \in \mathbb{R}^{T\times(C+1)}$. Then, we perform an element-wise multiplication between the normalized activation scores of each class and the contextualized features to generate

the *class-specific feature maps* $\mathbf{F}_c \in \mathbb{R}^{(C+1)\times T\times D}$, highlighting the most discriminative video segments for the C action classes and the background, as follows:

$$\mathbf{F}_c(j,\cdot,\cdot) = \tilde{\mathbf{P}}(\cdot,j) \otimes \mathbf{X}_c \tag{8}$$

where $\mathbf{F}_c(j,\cdot,\cdot) \in \mathbb{R}^{T\times D}$ represents the feature map that highlights the discriminative video segments for the j-th class.

Later, to highlight the discriminative classes within each video segment, first, we reshape class-specific feature maps $\mathbf{F}_c \in \mathbb{R}^{(C+1)\times T\times D}$ into *segment-specific feature maps* $\mathbf{F}_s \in \mathbb{R}^{T\times(C+1)\times D}$, i.e., for each video segment t, we get a feature map $\mathbf{F}_s(t) \in \mathbb{R}^{(C+1)\times D}$, which has $C+1$ classes and each class corresponds to a D-dimensional feature vector. Then, we apply class-wise attention mechanism on top of the segment-specific feature maps to highlight the most representative classes within each segment. Specifically, the feature map within a segment $\mathbf{F}_s(t) \in \mathbb{R}^{(C+1)\times D}$ is fed into a fully-connected layer followed by a softmax function to compute the *class-wise attention score* $\mathbf{w}_{att}(t) = [0,1]^{C+1}$ for that segment. After that, by performing the element-wise multiplication between the feature map of each segment and the class-wise attention score, we get the segment-specific attention-weighted feature maps $\tilde{\mathbf{F}}_s \in \mathbb{R}^{T\times(C+1)\times D}$, highlighting the discriminative classes for T segments, as follows:

$$\tilde{\mathbf{F}}_s(t,\cdot,\cdot) = \mathbf{F}_s(t,\cdot,\cdot) \otimes \mathbf{w}_{att}(t) \tag{9}$$

where $\tilde{\mathbf{F}}_s(t,\cdot,\cdot) \in \mathbb{R}^{(C+1)\times D}$ represents the feature map that highlights the most discriminative classes for the t-th segment.

Since the weakly-supervised learning only has video-level ground truth rather than the fine-grained segment-level one, we aggregate the $\tilde{\mathbf{F}}_s \in \mathbb{R}^{T\times(C+1)\times D}$, using the average pooling along the temporal dimension, to generate the *class-wise features* $\mathbf{U} \in \mathbb{R}^{(C+1)\times D}$ for the entire video. The classification score $\tilde{\mathbf{y}}_{cs} \in \mathbb{R}^{C+1}$ for the entire video for this CS branch is obtained by computing the cosine similarity between the $\mathbf{w}_f$ and $\mathbf{U}$, and then passing the similarity scores through a softmax layer, as follows:

$$\tilde{\mathbf{y}}_{cs}(j) = \frac{\exp(\cos(\mathbf{U}(j,\cdot),\mathbf{w}_f))}{\sum_i \exp(\cos(\mathbf{U}(i,\cdot),\mathbf{w}_f))} \tag{10}$$

The classification loss of the CS branch is defined by:

$$\ell_{cs} = \sum_{j=1}^{C+1} -\mathbf{y}_{cs}(j)\log\tilde{\mathbf{y}}_{cs}(j) \tag{11}$$

where $\mathbf{y}_{cs} = [\mathbf{y}_{cs}(1),...,\mathbf{y}_{cs}(C),0]$ is the ground-truth vector, in which $\mathbf{y}_{cs}(c)$ is set to 1 if a video contains action class c. Since this branch also aims to classify videos focusing on action classes, the label for the background is set to 0.

Action and background classification branch (AB): We use this branch to classify an input video regarding both the action classes and the background class. In the AB branch, first, we get the classification score $\psi_{ab} \in \mathbb{R}^{C+1}$ for the entire video by aggregating the segment-wise class-activation score with the normalized class-activation score, as follows:

$$\psi_{ab}(j) = \sum_t \mathbf{P}(t,j)\tilde{\mathbf{P}}(t,j) \tag{12}$$

The classification score is then passed through the softmax to get the video-level prediction: $\tilde{\mathbf{y}}_{ab} = \text{softmax}(\psi_{ab}) \in \mathbb{R}^{(C+1)}$. The classification loss of the AB branch is defined by:

$$\ell_{ab} = \sum_{j=1}^{C+1} -\mathbf{y}_{ab}(j)\log\tilde{\mathbf{y}}_{ab}(j) \tag{13}$$

where $\mathbf{y}_{ab} = [\mathbf{y}_{ab}(1),...,\mathbf{y}_{ab}(C),1]$ is the ground-truth vector, in which $\mathbf{y}_{ab}(c)$ is set to 1 if this video contains the action class c. The label for the background class is set to 1, considering that all untrimmed videos in training dataset contain background frames.

E. Network Training and Inference

Training: We compose the three video-level classification losses as follow:

$$\ell_{Total} = \alpha\ell_{ca} + \beta\ell_{cs} + \gamma\ell_{ab} \tag{14}$$

where α, β, γ are the balancing hyper-parameters.

Inference: We take a three-step inference to perform the temporal action localization. First, we threshold on video-level prediction (Eq. 5) of the CA branch and reject the action classes whose prediction score is lower than 0.1. Then, following [15], [43], [49], for each of the remaining action classes, we apply a set of thresholds on the CAS (Eq. 7) of the CS branch to generate action proposals. Finally, we perform the class-wise Non-Maximal Suppression (NMS) to keep the highly overlapped proposals, which are the final proposals.

IV. Experiments

A. Datasets and Metrics

THUMOS14 [62]: THUMOS14 has temporal boundary annotations for 200 validation videos and 213 testing videos, which belong to 20 classes. Following the literature [14], [18], [19], [20], [22], [43], [48], [49], [58], [63], we train our model on the validation set without using the temporal annotations and evaluate it on the test set.

ActivityNet1.3 [64]: ActivityNet1.3 dataset covers 200 action classes, which has temporal boundary annotations for 10,024 videos for training, 4926 videos for validation, and 5044 videos for testing. Since the testing labels are withheld, following the literature [18], [20], [43], [48], [49], [63], we train our model on the training set without using the temporal annotations and evaluate it on the validation set.

Evaluation metrics: We evaluate the temporal action localization performance with the mean Average Precision (mAP) values under different intersection over union (IoU) thresholds.

B. Implementation Details

First, we generate video segments from both RGB and Optic-flow by sliding a non-overlapping temporal window of 16 frames. Then, we use a pre-trained I3D [36] network to extract 1024 dimensional feature vectors for both streams. We separately train our F3-Net for both streams, and collect the generated proposals from both networks during inference. By validation, we set $k^{act} = k^{bkg} = 0.7$, the memory controlling hyper-parameter $m = 2$, weakening factor $\epsilon = 0.8$, the balancing hyper-parameters $\alpha = 0.1$, $\beta = 1.0$, and $\gamma = 0.1$.

TABLE II
RESULTS ON THE THUMOS14 TEST SET. AVG INDICATES THE AVERAGE MAP AT IOU THRESHOLDS 0.1:0.1:0.5.

Category	Method IoU →	0.1	0.2	0.3	0.4	0.5	AVG
(i) Methods related to feature weakening	Step-by-step erasion [42], MM2018	45.8	39.0	31.1	22.5	15.9	30.9
	A2CL-PT (I3D) [47] (ECCV'20)	61.2	56.1	48.1	39.0	30.1	46.9
	ACM-BANet (I3D) [15], MM2020	64.6	57.7	48.9	40.9	32.3	48.9
(ii) Methods related to feature contextualization	AUMN [45], CVPR2021	66.2	61.9	54.9	44.4	33.3	52.1
(iii) Methods related to feature discrimination	STPN (I3D) [20], CVPR2018	52.0	44.7	35.5	25.8	16.9	35.0
	W-TALC (I3D) [14], ECCV2018	55.2	49.6	40.1	31.1	22.8	39.8
	3C-Net (I3D) [21], ICCV2019	56.8	49.8	40.9	32.3	24.6	40.9
	RPN (I3D) [46], AAAI2020	62.3	57.0	48.2	37.2	27.9	46.5
	CoLA (I3D) [54], CVPR2021	66.2	59.5	51.5	41.9	32.2	50.3
	D2-Net (I3D) [16], ICCV2021	65.7	60.2	52.3	43.4	36.0	51.5
	WSAL-BM (I3D) [48], ICCV2019	60.4	56.0	46.6	37.5	26.8	45.5
	BaS-Net (I3D) [49], AAAI2020	58.2	52.3	44.6	36.0	27.0	43.6
	HAM-Net (I3D) [55], AAAI2021	65.4	59.0	50.3	41.1	31.0	49.4
	ACS-Net (I3D) [56], AAAI2021	-	-	51.4	42.7	32.4	-
	UM (I3D) [18], AAAI2021	67.5	61.2	52.3	43.4	33.7	51.6
	FAC-Net (I3D) [43], ICCV2021	67.6	62.1	52.6	44.3	33.4	52.0
	ACM-Net (I3D) [57], arXiv2021	68.9	62.7	55.0	44.6	34.6	53.2
	FTCL (I3D) [50], CVPR2022	**69.6**	63.4	**55.2**	45.2	35.6	53.8
FW+FC+FD	F3-Net (Ours)	69.4	**63.6**	54.2	**46.0**	**36.5**	**53.9**
(iv) Pseudo label-related method	RefineLoc (I3D) [51], ECCV2020	-	-	40.8	-	23.1	-
	DGAM (I3D) [58], CVPR2020	60.0	54.2	46.8	38.2	28.8	45.6
	EM-MIL (I3D) [52], ECCV2020	59.1	52.7	45.5	36.8	30.5	45.0
	TSCN (I3D) [59], ECCV2020	63.4	57.6	47.8	37.7	28.7	47.0
	TSCN [59] + UGCT (I3D) [17], CVPR2021	67.5	62.1	55.3	45.2	33.3	52.7
	WSAL-BM [48] + UGCT (I3D) [17], CVPR2021	69.2	62.9	55.5	46.5	35.9	54.0
	DCC (I3D) [60], CVPR2022	69.0	63.8	55.9	45.9	35.7	54.1
	ASM-Loc (I3D) [61], CVPR2022	71.2	65.5	**57.1**	46.8	36.6	55.4
	FAC-Net [43] + RSKP (I3D) [53], CVPR2022	71.3	65.3	55.8	47.5	38.2	55.6
FW+FC+FD + Pseudo label	F3-Net (Ours) + RSKP (I3D) [53]	**72.0**	**66.1**	56.5	**48.2**	**38.9**	**56.3**

TABLE III
RESULTS ON ACTIVITYNET1.3 VALIDATION SET. AVG INDICATES THE AVERAGE MAP AT IOU THRESHOLDS 0.5:0.05:0.95.

Category	Method IoU →	0.5	0.75	0.95	AVG
(i) Methods related to feature weakening	A2CL-PT (I3D) [47], ECCV2020	36.8	22.0	5.2	22.5
	ACM-BANet (I3D) [15], MM2020	37.6	24.7	6.5	24.4
(ii) Methods related to feature contextualization	AUMN (I3D) [45], CVPR2021	**38.3**	23.5	5.2	23.5
(iii) Methods related to feature discrimination	BaS-Net (I3D) [49], AAAI2020	34.5	22.5	4.9	22.2
	UM (I3D) [18], AAAI2021	37.0	23.9	5.7	23.7
	FAC-Net (I3D) [43], ICCV2021	37.6	24.2	6.0	24.0
FW+FC+FD	F3-Net (Ours)	38.1	**24.9**	**6.6**	**24.6**
(iv) Pseudo label-related methods	TSCN (I3D) [59], ECCV2020	35.3	21.4	5.3	21.7
	TSCN [59] + UGCT (I3D) [17], CVPR2021	38.1	21.2	5.4	22.8
	WSAL-BM [48] + UGCT (I3D) [17], CVPR2021	39.0	21.4	5.1	23.0
	DCC (I3D) [60], CVPR2022	38.8	24.2	5.7	24.3
	FAC-Net [43] + RSKP (I3D) [53], CVPR2022	40.6	24.6	5.9	25.0
	ASM-Loc (I3D) [61], CVPR2022	**41.0**	24.9	6.2	25.1
FW+FC+FD + Pseudo label	F3-Net (Ours) + RSKP (I3D) [53]	39.9	**25.0**	**6.7**	**25.2**

C. Comparisons with the State-of-the-art

Table II and III compare the results of our F3-Net with recent W-TAL methods on the THUMOS14 and ActivityNet1.3, respectively. We separate the W-TAL methods into four categories: (i) Methods related to feature weakening; (ii) Methods related to feature contextualization; (iii) Methods related to feature discrimination; and (iv) Pseudo label-related methods.

Our F3-Net vs. the W-TAL methods designed from the feature learning aspect: Since we propose to solve the W-TAL problem from the feature learning aspect, we mainly compare our F3-Net with other W-TAL methods designed from the feature learning aspect, i.e., the feature weakening, the feature contextualization, and the feature discrimination-related methods, for the fair comparison. As shown in Table II and III, our F3-Net achieves superior performance compared to the latest feature weakening, feature contextualization, and feature discrimination-related W-TAL methods on both THUMOS14 and ActivityNet 1.3, respectively.

Our F3-Net vs. pseudo label-related W-TAL methods: Our F3-Net achieves superior performance compared to many pseudo label-related methods. The performance of our F3-Net is slightly inferior compared to the latest pseudo label-related methods such as ASM-Loc [61] and FAC-Net [43]+RSKP [53]. These pseudo label-related methods improve the performance by refining the W-TAL networks with segment-level pseudo labels over the training iterations, while our F3-Net achieves competitive performance without refining our F3-Net with segment-level pseudo labels. However, the performance of our F3-Net can be further improved by considering the

TABLE IV
ABLATION STUDIES OF DIFFERENT ARCHITECTURES ON THUMOS14 DATASET.

Method	FW	FC	FD			0.1	0.2	0.3	0.4	0.5	AVG(0.1:0.1:0.5)
			AB	CA	CS						
Ours	✓	✓	✓	✓	✓	**69.4**	**63.6**	**54.2**	**46.0**	**36.5**	**53.9**
Case 1	✗	✓	✓	✓	✓	68.7	62.9	53.6	45.6	35.8	53.3
Case 2	✓	✗	✓	✓	✓	68.5	62.7	53.5	45.3	35.1	53.0
Case 3	✓	✓	✗	✓	✓	66.4	61.1	51.8	43.7	33.9	51.4
Case 4	✓	✓	✓	✗	✓	65.1	60.2	50.6	43.1	32.8	50.4
Case 5	✓	✓	✓	✓	✗	62.9	56.4	45.9	36.3	27.6	45.8

TABLE V
EFFECTIVENESS OF OUR FEATURE CONTEXTUALIZATION (FC) OVER SELF-ATTENTION.

Method	0.1	0.2	0.3	0.4	0.5
Our F3-Net with Self-attention	68.6	62.9	53.8	45.6	35.7
Our F3-Net with FC	**69.4**	**63.6**	**54.2**	**46.0**	**36.5**

TABLE VI
FW AND FC AS PLUG-IN'S ON OTHER METHODS.

Method	FW	FC	0.1	0.2	0.3	0.4	0.5
ACM-BANet [15]	✗	✗	64.6	57.7	48.9	40.9	32.3
ACM-BANet [15]	✓	✗	65.1	58.3	49.7	42.1	33.5
ACM-BANet [15]	✗	✓	65.8	58.8	50.1	42.2	33.3
ACM-BANet [15]	✓	✓	66.4	59.6	50.8	42.9	34.1
FAC-Net [43]	✗	✗	67.6	62.1	52.6	44.3	33.4
FAC-Net [43]	✓	✗	68.0	62.5	53.3	45.5	35.4
FAC-Net [43]	✗	✓	68.5	62.6	53.3	45.1	34.9
FAC-Net [43]	✓	✓	69.0	63.2	53.8	45.9	36.0

pseudo-labeling strategy. As shown in the last rows of Table II and Table III, with the pseudo label refinement (e.g., RSKP [53]), we achieve further boost in performances and establish a new state-of-the-art performances on both the THUMOS14 and ActivityNet 1.3 datasets, respectively, on most metrics.

D. Ablation Studies

Ablation studies on different modules and branches: We conduct a series of ablation studies on the THUMOS14 dataset to evaluate the influence of FW, FC, and FD modules (specifically, each branch inside the FD module).

- **Case 1:** We apply our F3-Net without the FW module.
- **Case 2:** We apply our F3-Net without the FC module.
- **Case 3:** We perform the experiment without the AB branch in FD module.
- **Case 4:** We conduct the experiment without the CA branch in FD module.
- **Case 5:** We perform the experiment without the CS branch in FD module.

As shown in Table IV, each algorithm component contributes to our approach, and our F3-Net that combines all modules and branches achieves the best performance.

Effectiveness of FC over self-attention: We first plug the self-attention module Fig. (6(a)) into our approach, and then replace it with our FC module (Fig. 6(b)). We find that our FC improves the performance when compared with the self-attention module on the THUMOS14, as shown in Table V.

Effectiveness of our FW and FC modules as plug-ins on existing W-TAL methods: To test the effectiveness of our FW and FC modules on existing methods, we plug them into two latest W-TAL methods [15], [43]. As shown in Table VI, we find that both the FW and FC modules can boost the performance of ACM-BANet [15] and FAC-Net [43].

E. Qualitative Analysis

We visualize some qualitative results in Fig. 8, where we show activation scores of the predicted classes according to different branches in our FD and different modules:

- **FD (CS):** The CS branch in our FD module localizes all the action instances roughly and has some false positives.
- **FD (CS+CA):** The CA branch collaborates with the CS branch to improve the performance on the localized instances by localizing more actual-action-related frames.
- **FD (CS+CA+AB):** The AB branch explicitly models the background class and helps the FD suppress activations from background frames, removing false positives.
- **FD (CS+CA+AB) + FC:** The FC module helps us get smooth localization by exploiting the local and global contexts.
- **FD (CS+CA+AB) + FC + FW:** Finally, the FW module further improves the performance by localizing the action instances in ambiguous frames.

In the FW module, we assume that the features of action segments generally have larger magnitudes than those of background segments. As shown in Fig. 9, we validate our assumption by performing the qualitative analysis on the feature magnitudes of different segments for different videos. Fig. 9 shows that features of action segments have larger magnitudes compared to ones from background segments.

F. Discussions

In our proposed F3-Net, we first introduce a new FW module for action completeness modeling. The existing feature weakening-related methods [15], [42] trained the W-TAL network in multiple steps. During each training step, the W-TAL network first identifies discriminative features, which are then erased and fed into the network of the next training step. However, one key problem is that the attention regions may gradually expand to the background intervals mistakenly as the training steps increase, which downgrades the localization performance. Differently, we train our network in a single step with a novel FW module that first identifies and then randomly weakens either the discriminative action or the discriminative background features over the training iterations to force the network to precisely localize the action instances in both the discriminative and ambiguous action-related frames, without

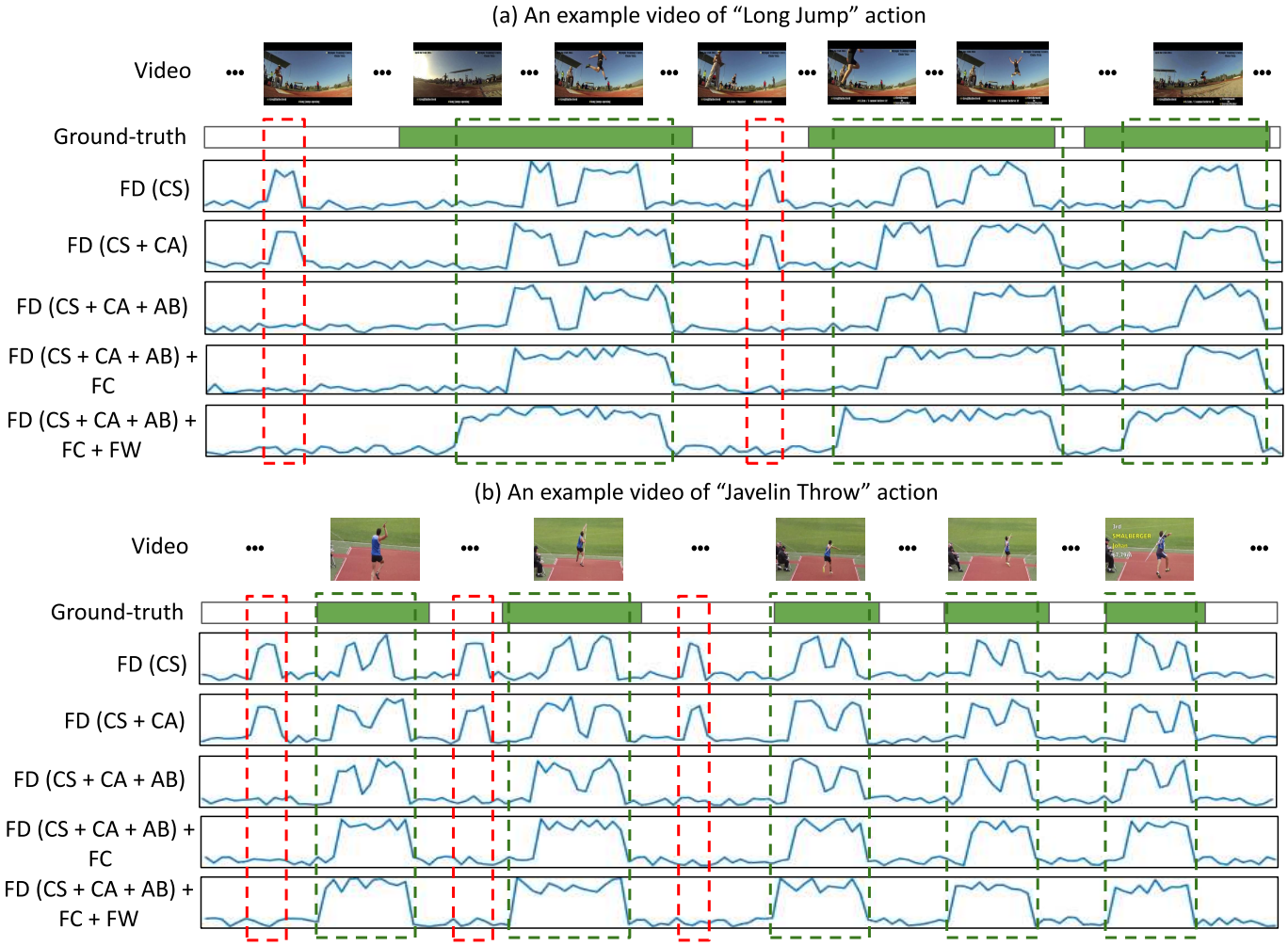

Fig. 8. Qualitative results. With many background frames, the video (a) and (b) contain multiple instances of “Long Jump” and “Javelin Throw” actions, respectively. We show the activation scores of the predicted action classes under five settings: three cases on different branches in the FD module, and the effects of FC and FW in addition to the FD with all branches.

spreading to the background intervals. Since our FW module prohibits attentions from spreading to background intervals, we achieve better performance compared to the existing feature weakening-related methods.

On the other hand, we also introduce a new FC module to generate more representative features for temporal action localization. The existing feature contextualization-related methods (e.g., [45]) first explored the correlative information between different segments in a video to extract the global contexts, which were then fused with the local contexts by simply performing the element-wise summation. In contrast, our FC module learns attention weights on local and global contexts to fuse them accordingly. Intuitively, the conventional element-wise summation is predefined, which provides fixed importance to the local and global contexts during the fusion. On the other hand, the attention-weighted fusion in our FC module automatically learns the attention weights to attentionally fuse the local and global contexts. Since the attention-weighted fusion automatically learns the attention weights, it provides varying importance to the local and global contexts during the fusion, which generates more representative contextualized features and achieves better performance compared to the element-wise summation that provides fixed importance.

Furthermore, we design a novel FD module to highlight the most discriminative video segments/classes corresponding to each class/segment, respectively. The existing feature discrimination-related methods (e.g., [14], [43], [49]) only highlighted the most discriminative video segments corresponding to each action class. But, an action class may have multiple instances in separated video segments and a video segment may contain multiple action classes. Since our FD module highlights both the most discriminative video segments corresponding to each action class and the most discriminative classes within each video segment, we achieve better performance compared to the previous feature discrimination-related methods that only highlight the most discriminative video segments corresponding to each action class.

Although we achieve state-of-the-art performances, the localization performance is not 100% correct yet. Most of the failure cases are related to false positives, i.e., the network wrongly localizes the background frames as foreground, particularly the background frames which are visually similar

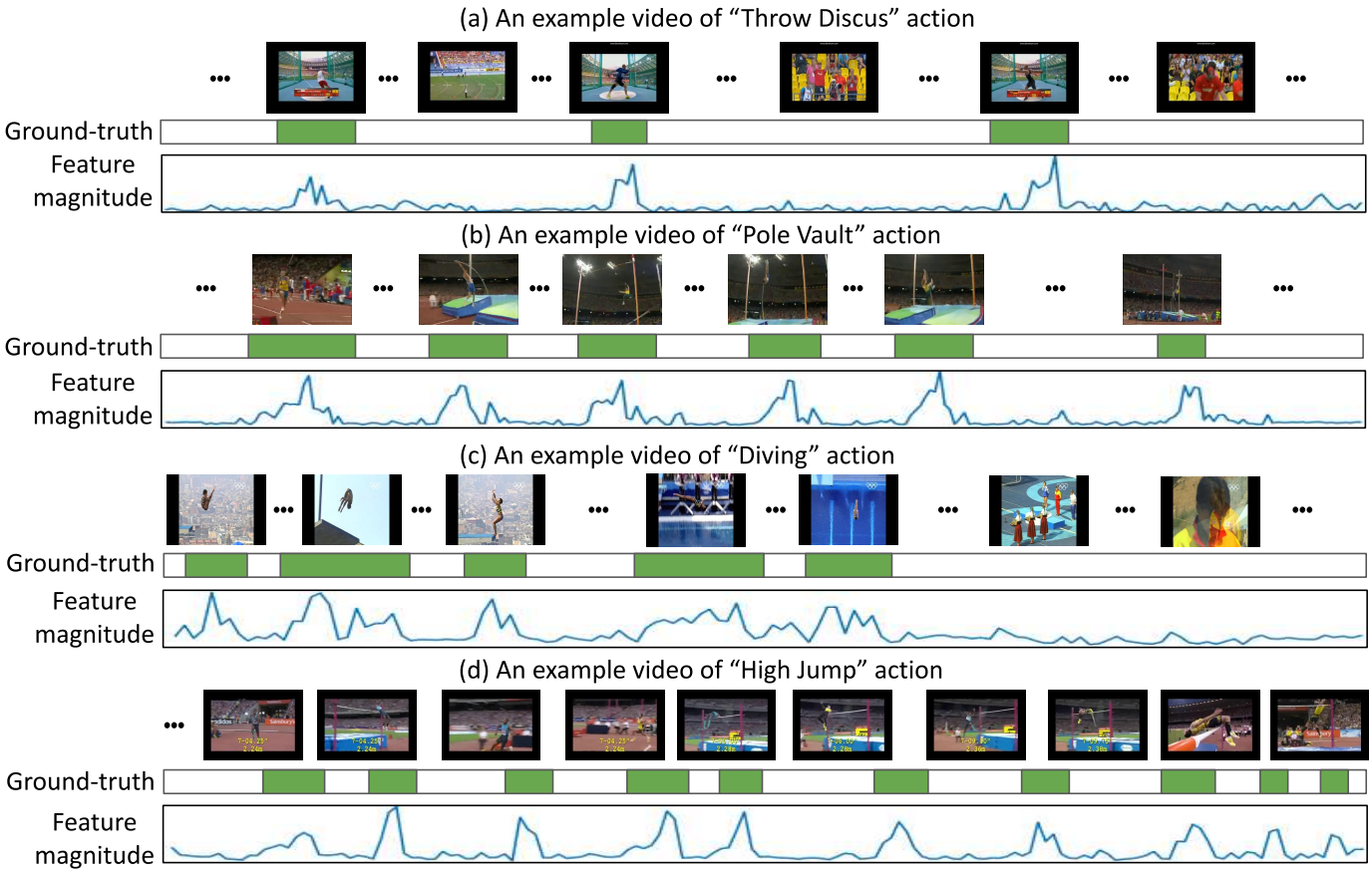

Fig. 9. Visualization of feature magnitudes of different video segments. The video (a) contains multiple instances of the “Throw Discus” action, video (b) contains multiple instances of the “Pole Vault” action, video (c) contains multiple instances of the “Diving” action, and video (d) contains multiple instances of the “High Jump” action. Generally, features of action segments have larger magnitudes compared to ones from background segments.

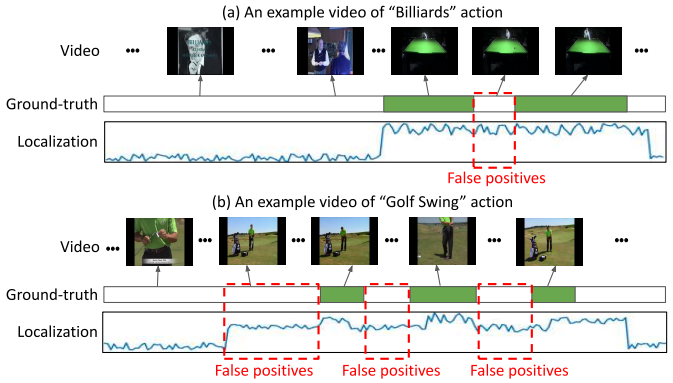

Fig. 10. Visualization of false positive cases. The video (a) contains the “Billiards” action and video (b) contains the “Golf Swing” action, where some background frames are wrongly localized as foreground since they are visually similar and frequently co-occur with the action-related frames, but do not belong to the actual actions.

and frequently co-occur with the action-related frames, but do not belong to the actual actions, as shown in Fig. 10. These background frames can be viewed as hard negative samples, which are inherently difficult to suppress with weak supervision. In the future, we will investigate how to mingle the fully and weakly-supervised localization so that accurate localization can be achieved with less annotation efforts.

V. Conclusion

We propose an F3-Net for W-TAL, including three modules: (1) The FW module to discover the action instances in both discriminative and ambiguous frames for localizing the complete action interval; (2) The FC module to utilize the local and global contexts, generating more representative contextualized features; and (3) The FD module to highlight the most discriminative video segments/classes corresponding to each class/segment, respectively. Our F3-Net outperforms related W-TAL methods on THUMOS14 and ActivityNet1.3. Besides, our FW and FC modules are effective plug-in’s to improve other methods.

Acknowledgment

This project was supported by National Science Foundation grants CMMI-1954548 and ECCS-2025929.

References

[1] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage cnns,” in *CVPR*, 2016.
[2] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, “Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos,” in *CVPR*, 2017.
[3] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, “Temporal action detection with structured segment networks,” in *ICCV*, 2017.
[4] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, “Exploring temporal preservation networks for precise temporal action localization,” in *AAAI*, 2018.

[5] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *CVPR*, 2016.

[6] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *ACMMM*, 2017.

[7] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *ICCV*, 2017.

[8] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *CVPR*, 2018.

[9] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Qiu Chen, "Temporal context network for activity localization in videos," in *ICCV*, 2017.

[10] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *ECCV*, 2018.

[11] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *CVPR*, 2019.

[12] P. Chen, C. Gan, G. Shen, W. Huang, R. Zeng, and M. Tan, "Relation attention for temporal action localization," *IEEE-TMM*, 2019.

[13] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE-TIP*, 2022.

[14] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *ECCV*, 2018.

[15] 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 *ACMMM*, 2020.

[16] 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 *ICCV*, 2021.

[17] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *CVPR*, 2021.

[18] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *AAAI*, 2021.

[19] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *CVPR*, 2017.

[20] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *CVPR*, 2018.

[21] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *ICCV*, 2019.

[22] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *ECCV*, 2018.

[23] M. Moniruzzaman, Z. Yin, Z. H. He, R. Qin, and M. Leu, "Human action recognition by discriminative feature pooling and video segmentation attention model," *IEEE-TMM*, 2021.

[24] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE-TMM*, 2017.

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[26] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE-TMM*, 2017.

[27] H. Song, X. Wu, B. Zhu, Y. Wu, M. Chen, and Y. Jia, "Temporal action localization in untrimmed videos using action pattern trees," *IEEE-TMM*, 2018.

[28] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *IEEE-TMM*, 2018.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[30] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE-TMM*, 2019.

[31] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE-TMM*, 2018.

[32] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019.

[33] T. Yu, L. Wang, C. Da, H. Gu, S. Xiang, and C. Pan, "Weakly semantic guided action recognition," *IEEE-TMM*, 2019.

[34] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3d action recognition," *IEEE-TMM*, 2016.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[36] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[37] M. Hasan and A. K. Roy-Chowdhury, "A continuous learning framework for activity recognition using deep hybrid feature models," *IEEE-TMM*, 2015.

[38] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *NeurIPS*, 2018.

[39] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *CVPR*, 2017.

[40] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *CVPR*, 2018.

[41] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*, 2017.

[42] 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 *ACMMM*, 2018.

[43] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *ICCV*, 2021.

[44] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua, "Enriching local and global contexts for temporal action localization," in *CVPR*, 2021.

[45] 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 *CVPR*, 2021.

[46] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Relational prototypical network for weakly supervised temporal action localization," in *AAAI*, 2020.

[47] K. Min and J. J. Corso, "Adversarial background-aware loss for weakly-supervised temporal activity localization," in *ECCV*, 2020.

[48] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *ICCV*, 2019.

[49] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization." in *AAAI*, 2020.

[50] J. Gao, M. Chen, and C. Xu, "Fine-grained temporal contrastive learning for weakly-supervised temporal action localization," in *CVPR*, 2022.

[51] A. Pardo, H. Alwassel, F. Caba, A. Thabet, and B. Ghanem, "Refineloc: Iterative refinement for weakly-supervised action localization," in *WACV*, 2021.

[52] 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 *ECCV*, 2020.

[53] L. Huang, L. Wang, and H. Li, "Weakly supervised temporal action localization via representative snippet knowledge propagation," in *CVPR*, 2022.

[54] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "Cola: Weakly-supervised temporal action localization with snippet contrastive learning," in *CVPR*, 2021.

[55] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *AAAI*, 2021.

[56] 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 *AAAI*, 2021.

[57] 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.

[58] B. Shi, Q. Dai, Y. Mu, and J. Wang, "Weakly-supervised action localization by generative attention modeling," in *CVPR*, 2020.

[59] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, and G. Hua, "Two-stream consensus network for weakly-supervised temporal action localization," in *ECCV*, 2020.

[60] J. Li, T. Yang, W. Ji, J. Wang, and L. Cheng, "Exploring denoised cross-video contrast for weakly-supervised temporal action localization," in *CVPR*, 2022.

[61] 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 *CVPR*, 2022.

[62] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "Thumos challenge: Action recognition with a large number of classes," 2014.

[63] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *CVPR*, 2019.

[64] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.