# Multi-Dimensional Attention With Similarity Constraint for Weakly-Supervised Temporal Action Localization

Zhengyan Chen, Hong Liu, *Member, IEEE*, Linlin Zhang, and Xin Liao, *Senior Member, IEEE*

***Abstract*—Weakly-supervised temporal action localization (WTAL) is a challenging task in understanding untrimmed videos, in which no frame-wise annotation is provided during training, only the video-level category label is available. Current methods mainly adopt temporal attention branches to conduct foreground-background separation with RGB and optical flow features simply concatenated, regardless of the discriminative spacial features and the complementarity between different modalities. In this work, we propose a Multi-Dimensional Attention (MDA) method to explore attention mechanism across three dimensions in weakly supervised action localization, *i.e.*, 1) temporal attention that focuses on segments containing action instances, 2) channel attention that discovers the most relevant cues for action description, and 3) modal attention that fuses RGB and flow information adaptively based on feature magnitudes during background modeling. In addition, we introduce a similarity constraint loss to refine the action segment representation in feature space, which helps the network to detect less discriminative frames of an action to capture the full action boundaries. The proposed MDA with similarity constraints can be easily applied to existing action detection frameworks with few parameters. Extensive experiments on THUMOS'14 and ActivityNet v1.2 datasets show that the proposed method outperforms the current state-of-the-art WTAL approaches, and achieves comparable results with some advanced fully-supervised methods.**

***Index Terms*—Multi-dimensional attention, temporal action localization, video analysis, weakly supervised learning.**

## I. Introduction

With the explosive growth of the video contents from the Internet, action analysis in videos has drawn increasing attention due to its wide applications such as video surveillance, human-computer interaction, video summary, *etc*. Temporal action localization (TAL) aims to figure out both the action category and the accurate temporal location of action instances in untrimmed videos. Most existing action localization methods heavily rely on trimmed videos for model training, requiring frame-level action boundary annotations [1]–[4]. However, it is expensive and time-consuming to acquire a large-scale trimmed video dataset with precise manual annotations. Additionally, the definition of the temporal boundary of an action is often subjective and prone to large variations [5].

To overcome these limitations, several approaches have been proposed in recent years to focus on weakly-supervised temporal action localization (WTAL), using only video-level class labels for training [6]–[8]. Wang *et al.* proposed an end-to-end framework that optimizes the classification module and selection module jointly [6]. Singh *et al.* randomly hide segments to guide the network to detect complete action parts [7]. Paul *et al.* introduced a co-activity similarity loss in consideration of the correlations between videos with similar tags [8].

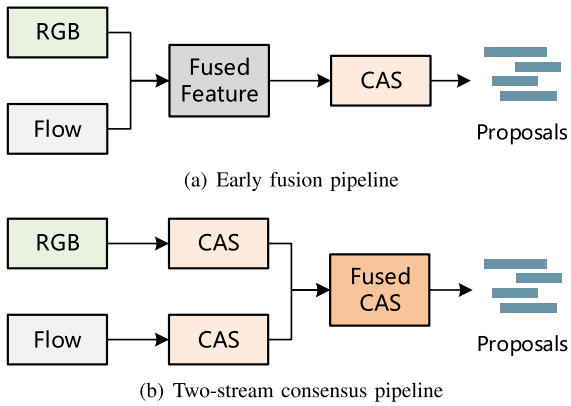
Recently, some works focus on how to distinguish actions from complex backgrounds [9]–[11]. Nguyen *et al.* proposed a similar objective to model the background contents [9]. BaS-Net suggests using a suppression branch to suppress the network activations on the background portions [10]. Lee *et al.* modeled background via uncertainty estimation in consideration of the inconsistency of the background frames [11]. Inspired by some remarkable works that leverage attention mechanism to capture the key motion information in human action analysis task [12]–[15], in order to find out complete action instances from the background, some advanced WTAL methods introduce multiple temporal attention branch to select action instance and suppress background simultaneously, demonstrating the state-of-the-art results [16]. However, we argue that existing methods are focus on designing elaborate temporal attention mechanisms, not able to sufficiently make use of discriminative spatial and temporal information for each video. On the other hand, the temporal action localization task is usually bulit on two feature modalities, *i.e.*, RGB frames and optical flow, which are fused in two possible ways as shown in Fig. 1. Most approaches conduct *early fusion pipline* (Fig. 1(a)) to fuse RGB and optical flow information directly by simply concatenate them as a feature vector after feature extraction [6], [8], [10], [17], which ignore the complementarity localization abilities between different modalities,

Manuscript received 1 September 2021; revised 15 January 2022; accepted 3 May 2022. Date of publication 11 May 2022; date of current version 20 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62073004 and 61972142, and in part by Shenzhen Fundamental Research Program under Grants JCYJ20200109140410340 and GXWD20201231165807007-20200807164903001. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Corresponding author: Hong Liu.)*

Zhengyan Chen, Hong Liu, and Linlin Zhang are with the Key Laboratory of Machine Perception and Shenzhen Graduate School, Peking University, Beijing 100871, China (e-mail: chenzhengyan@pku.edu.cn; hongliu@pku.edu.cn; catherinezll@pku.edu.cn).

Xin Liao is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: xinliao@hnu.edu.cn).

Digital Object Identifier 10.1109/TMM.2022.3174344




(a) Early fusion pipeline

(b) Two-stream consensus pipeline

Fig. 1. Two frameworks of modal fusion in WTAL. **(a) Early fusion pipeline** first performs feature-level aggregation of RGB and optical flow modalities, then generate one Class Activation Sequence (CAS) for localization and classification. **(b) Two-stream consensus pipeline** first predicts CAS of RGB and optical flow streams separately, then generate a fused CAS by weighted average.

*i.e.*, RGB feature provides color and texture information, while optical flow feature focuses on motion information of an action. Another kind of method is based on *two-stream consensus pipeline* (Fig. 1(b)) [18], [19] that computes a weighted sum of their respective output class activation sequences (CAS) before generating action proposals. However, the late-fusion scheme ignores the intrinsic relation between two modalities, and relies on a pre-defined weighting factor for all input videos. This may lead to a limitation when some actions are distinct with action scenes (RGB information) while some actions are distinct with the movements of the target (optical flow information).

To localize complete action instances in untrimmed videos, in this work, we propose a multi-dimensional attention (MDA) mechanism with similarity constraint on action segments. Specifically, we explore attention mechanism cross three dimensions for weakly-supervised temporal action localization, *i.e.*, temporal attention, channel attention and modal attention. Firstly, the temporal attention module with a Temporal Relation Block (TRB) is applied to rescale the temporal weights and catch the temporal correlations among the segments. Then, the channel attention module selects distinguished features for segment-level action classification through a global average pooling. Besides, the modal attention module adaptively fuses RGB and optical flow information to model the background by estimating feature magnitudes from each modality. Moreover, we use a similarity constraint on action features to detect instances more completely. As a result, the proposed MDA attention mechanism can localize the action instances accurately by learning when and what to attend in video segments, as well as make full use of information from different modalities. Experimental results on THUMOS'14 [20] and ActivityNet v1.2 [21] datasets demonstrate that MDA enables the network to generate action proposals more completely, which achieves superior or competitive performance compared to state-of-the-art methods.

In summary, our contributions are as follows:

- We present Multi-Dimensional Attention (MDA) mechanism for weakly-supervised temporal action localization. Different from the previous methods that focus on background-foreground separation by temporal attention and fuse RGB and optical flow information directly, we investigate how to make full use of temporal and spatial information as well as effectively fuse different modalities.
- We introduce a similarity constraint loss on action features from same category in each video, which aggregates the actions features together and helps to detect complete action instances and distinguish the action segments from the background.
- Extensive experiments on two widely used benchmarks, THUMOS'14 and ActivityNet v1.2 show that our approach outperforms previous state-of-the-art methods, and achieve comparable performance with some fully-supervised methods.

## II. Related Work

In this section, we first provide an overview of the recent progress on action recognition, then review existing research about fully-supervised and weakly-supervised temporal action localization in the past few years.

### A. Action Recognition

Action recognition is one of the fundamental tasks of video understanding, which has been extensively explored in recent years. Earlier approached such as improved Dense Trajectory (iDT) [22] mainly focused on design hand-crafted features that can represent spatial-temporal features effectively. With the recent availability of Big Data and powerful computational resources, many effective algorithms based on deep learning have emerged [23]–[30]. These methods typically contain three main categories: (i) the two-stream networks that trained on multiple input modalities respectively (e.g., optical flow and warped flow in addition to RGB). The predictions from all modalities are fused to get the final video-level prediction [24], [25]; (ii) CNN+LSTM (Long Short Term Memory), in which recurrent neural networks are built on top of CNN features to capture the long term dynamics for action recognition [26]–[28]; and (iii) 3D CNN based models that extend temporal dimension from 2D convolution operation to capture the spatial and temporal information directly from the raw video frames [13], [29]–[34]. Action recognition networks have achieved significant performance on trimmed video clips, and are usually adopted to extract visual feature sequence from untrimmed videos in action localization task.

### B. Fully-Supervised Temporal Action Localization

Fully-supervised Action Localization relies on frame-level annotations, which aims not only to classify action instances but also to locate the start and end temporal boundary of action instances from long untrimmed videos. Most existing approaches can be summarized in two categories: the top-down (proposal-based) framework and the bottom-up (frame-based) framework. The top-down methods [1], [2], [35]–[39] usually generate action proposals by pre-defined massive anchors,

*e.g.*, fixed-length sliding windows, and then classify them as well as conduct temporal boundary regression. TAL-Net [2] follows Faster R-CNN [40] to perform two-stage action localization, and adopts dilated temporal convolution to control the receptive field. More recently, AFSD [41] conducts the first anchor-free method through learning salient boundary feature. On the contrary, the bottom-up methods [4], [42]–[47] directly predict frame-level action category and location followed by some post-processing techniques. Typically, CDC [43] estimates frame-level actionness score via a Convolutional-De-Convolutional network and then uses the frame-level action confidence to refine action boundaries. G-TAD [46] presents a graph convolutional network model to exploit video context and cast temporal action localization as a sub-graph detection problem. Since the top-down methods can discover most action instances with few omissions while the bottom-up methods are more flexible to cover action instances of various duration, some recent approaches [48], [49] are proposed to utilize the complementarity between these two frameworks, leading to better performances. However, all these fully-supervised methods require precise frame-level annotations during the proposal generation and classification stage, which is time-consuming and not capable of widely employed in real-world scenarios. To alleviate the annotation cost, this work focuses on the same problem in the weakly-supervised manner.

### C. Weakly-Supervised Temporal Action Localization

In recent years, many methods have been proposed to focus on the task of weakly-supervised temporal action localization (WTAL) to reduce the annotation costs with only video-level labels for training. Wang *et al.* firstly introduced WTAL task and proposed UntrimmedNet [6], including a classification module for predicting a classification score for each snippet, and a selection module to select relevant video segments. Later, STPN [18] added a sparsity loss and class-specific proposals. AutoLoc [50] introduced the outer-inner contrastive loss to effectively predict the temporal boundaries. W-TALC [8] considers the co-activity similarity to model inter-video similarities and differences. To obtain reliable and complete proposal from the class activation sequence (CAS), there are some promising works. Specifically, CleanNet [51] introduced an action proposal evaluator that provides pseudo-supervision by leveraging the temporal contrast in snippets. 3C-Net [52] adopted three loss terms to obtain discriminative feature representation. Hide-and-seek [7] and Zeng *et al.* [53] both hide some patches to find out complete action regions. Focused on background-foreground separation, many works regard non-action background as an additional class [9], [10], while [11] consider background as out-of-distribution and propose to learn uncertainty as well as action class scores. Huang *et al.* [54] proposed a GCN-based prototype embedding module to construct action relationships. More recently, some works achieved great performance by designing multiple attention mechanism. HAM-Net [16] introduced a novel hybrid attention mechanism to capture the full temporal boundaries of the actions in the video.

Though these advanced approaches achieved great performance on background-foreground separation and completeness modeling, their response to each temporal point are not discriminative enough, and fail to make full use of the complementary information between different modalities. To address these issues, we propose MDA, which is unique to previous works in three main aspects: **i)** we explore attention mechanism on three dimensions rather than only focus on modeling temporal features; **ii)** we fuse RGB and optical flow information adaptively rather than directly concatenate segment-level features or average the CAS from two streams; **iii)** to detect action instances completely, we apply class-wise similarity constraints on action segments in each video sample.

## III. Methodology

In this section, we first formulate the task of weakly supervised temporal action localization (WTAL). Then we present the proposed framework with multi-dimensional attention (MDA) in detail. The overall pipeline is shown in Fig. 2. The video features are first extracted from video frames by the pre-trained feature extractor. Then the features are embedded by several layers to generate segment-level classification scores. Meanwhile, the background possibilities are measured from the embedded features. The predicted proposals are finally grouped from segments considering both classification scores and background possibilities.

### A. Problem Formulation

Assume an untrimmed video $v$ containing activity instances from $C$ action classes. The corresponding video-level category label is a $C$-dimensional binary vector denoted as $y \in \{0,1\} \in \mathbb{R}^C$, with $y_k = 1$ if there is at least one instance of the $k$-th action class in the video, and $y_k = 0$ if there is no instance of the $k$-th activity. Note that each video may contain multiple action categories and multiple action instances. Different from the fully-supervised temporal action localization that use accurate action instance annotations, the goal of W-TAL task is to detect all action instances with only video-level label for training, *i.e.*, for each test video it predicts a set of action instances $\{(t_s, t_e, \psi, c)\}$, where $t_s$, $t_e$, $\psi$, $c$ represent the start time, the end time, the localization score of the action proposal and the action category, respectively.

### B. Feature Embedding With Temporal Attention

Given an input video $v$, we first divide it into non-overlapping 16-frame segments. As in [10], [18], due to the large variation of video lengths, a fixed number of $T$ segments are sampled from each video. The RGB and optical flow features $\boldsymbol{X}^{rgb} \in \mathbb{R}^{T\times F}$ and $\boldsymbol{X}^{op} \in \mathbb{R}^{T\times F}$ are extracted by pre-trained I3D deep networks [29], respectively. Then the RGB and optical flow features are stacked along feature dimension as input $\boldsymbol{X} \in \mathbb{R}^{T\times 2F}$ of our architecture.

Before embedding the input feature, we first introduce a temporal attention (TA) module that focuses on 'when' is important given an input video, which directly contributes to the accuracy of action localization. The TA module contains a base convolutional process and a **Temporal Relation Block (TRB)**

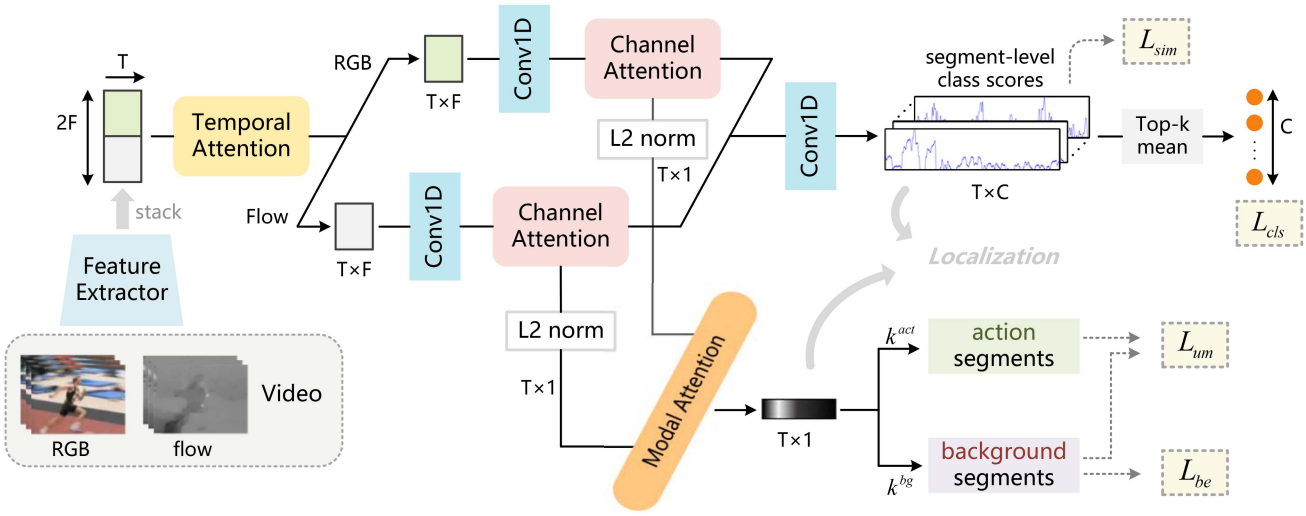



Fig. 2. Framework of the proposed Multi-Dimensional Attention (MDA) method. A pre-trained feature extractor is firstly applied to extract segment-level spatial-temporal features. Then RGB and optical flow features are embedded separately after a temporal attention (TA) module. After that, a channel attention (CA) is employed to emphasis both features before generating the class activation sequences (CAS). We calculate the background probability by estimating magnitudes from RGB and optical flow features through a modal attention (MA). Finally, the localization proposals are generated from both background probability and the CAS.

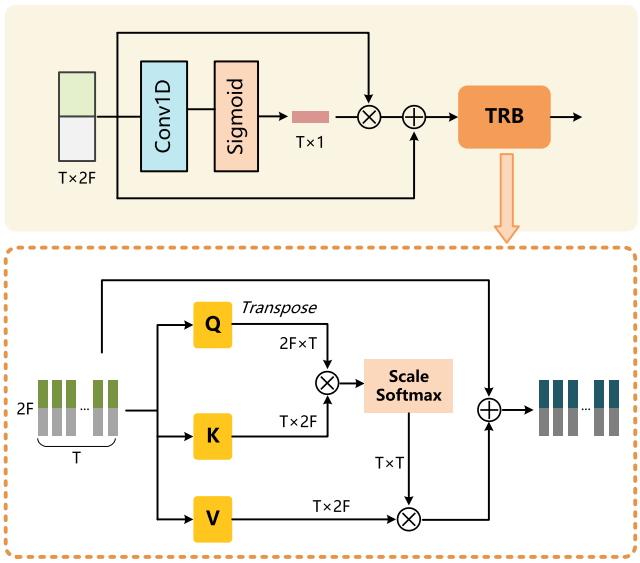



Fig. 3. The illustration of temporal attention (TA) module. $\oplus$ denotes the element-wise summation while $\otimes$ denotes element-wise multiplication. TRB denotes the Temporal Relation Block for modeling temporal relationship. $Q$, $K$ and $V$ are obtained directly from the concatenated segment-level features.

along temporal dimension to model temporal relation of the segments. As illustrated in Fig. 3, an initial temporal attention map $\boldsymbol{M}_{ta} \in \mathbb{R}^{T\times 1}$ is generated by a 1D convolutional layer over the stacked input feature:

$$\boldsymbol{M}_{ta}(\boldsymbol{X}) = \sigma(f_{conv}(\boldsymbol{X}; \phi_{ta})), \tag{1}$$

where $\phi_{ta}$ denotes trainable parameters in the convolutional layer, and $\sigma$ represents the sigmoid function. The output feature is fused with the original $\boldsymbol{X}$ by residual connection:

$$\boldsymbol{X}' = \boldsymbol{M}_{ta}(\boldsymbol{X}) \otimes \boldsymbol{X} + \boldsymbol{X}, \tag{2}$$

where $\otimes$ denotes element-wise multiplication and $\boldsymbol{X}'$ represents the temporal weighted feature. Inspired from the self-attention mechanism, we also introduce a Temporal Relation Block (TRB) in the temporal attention module to catch the relationship among long-term information. Concretely, we treat the emphasized feature $\boldsymbol{X}'$ as *query*, *key* and *value* directly and calculate segment-wise self-attention map by estimate similarities between each segment pair followed by a softmax operation to exploit their relations:

$$\text{Attention}(\boldsymbol{X}') = \text{softmax}\left(\frac{\boldsymbol{X}'\boldsymbol{X}'^{T}}{\sqrt{d_k}}\right)\boldsymbol{X}', \tag{3}$$

where $d_k = 2F$ is the dimension of $\boldsymbol{X}'$ for controlling the scale of dot product. Also, the final emphasised feature is generated by a residual connection. The TRB block can capture relations among different segments regardless of their temporal distance, allowing the network to seek information from the segments in other proposals automatically and boost classification performance. Different from the classical self-attention mechanism proposed in [55], we discard the linear transformation to generate *query*, *key* and *value* which aims to model high-level semantic information, and focus more on emphasizing local information to get reliable class activation scores of each temporal segment. This also makes it more convenient to model different modalities separately in the subsequent operations.

Note that the temporal attention module acts on the temporal dimension, thus the RGB and optical flow features remain relatively independent. To model two modalities separately, we discompose the feature $\boldsymbol{X}'$ into RGB and flow features, *i.e.*, $\boldsymbol{X}'_{rgb} = \boldsymbol{X}'[:, F]$ and $\boldsymbol{X}'_{op} = \boldsymbol{X}'[F, :]$. A temporal 1D convolutional layer followed by LeakyReLU activation is applied to the features of each modality to get the embedded feature maps. Specifically, $\mathcal{A}_m = g_{emb}(\boldsymbol{X}'_m; \phi_{emb})$, where

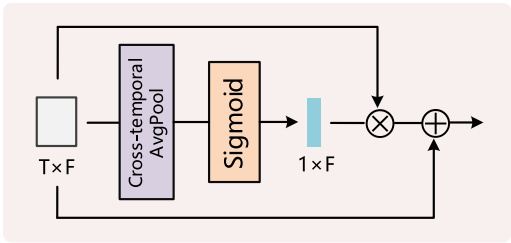



Fig. 4. The illustration of channel attention (CA) module. $\oplus$ denotes the element-wise summation while $\otimes$ denotes element-wise multiplication.

$m \in \{rgb, op\}$ denotes RGB and optical flow respectively, $g_{emb}$ is the convolution operator with the activation function and $\phi_{emb}$ is the corresponding parameters.

### C. Segment-Level Classification With Channel Attention

Convolutional channel features often capture different visual information, which corresponds to different actions. In order to highlight informative features for improving the classification power of the model, a channel attention (CA) module is applied to the embedded feature map for both RGB and optical flow streams. As illustrated in Fig. 4, the self-attention map of the $m$ modality is obtained by performing a global average pooling across the temporal dimension, then the channel attention map $\boldsymbol{M}_{ca}$ is generated from the self-attention map by applying sigmoid activation:

$$\boldsymbol{M}^m_{ca}(\mathcal{A}_m) = \sigma\left(\frac{1}{T}\sum_{t=1}^{T}\mathcal{A}_m(t, 1:F)\right), \tag{4}$$

where $\sigma$ denotes the sigmoid function. Similar to the TA module, the output feature $\mathcal{A}'_m$ is fused with the original feature map $\mathcal{A}_m$ through the use of identity-based skip connection:

$$\mathcal{A}'_m = \boldsymbol{M}^m_{ca}(\mathcal{A}_m) \otimes \mathcal{A}_m + \mathcal{A}_m, \tag{5}$$

where $\otimes$ denotes element-wise multiplication. Note that the CA module does not introduce any trainable parameters. Moreover, the channel attention will also intensify the feature distribution of each segment, which is beneficial to the background modeling through uncertainty estimation described in sec. III-D.

Since segment-level label is not provided in the WTAL task, following previous works [8], [56], [57], we apply the Multiple-Instance Learning (MIL) mechanism to obtain video-level classification scores for training. In this setting, each video is regarded as a bag of segments containing positive (action) and negative (background) instances; segment-level class scores, commonly known as Class Activation Sequences (CAS), are calculated and then temporally pooled to obtain video-level class scores.

For segment-level classification, we first stack the RGB feature $\mathcal{A}'_{rgb}$ and optical flow feature $\mathcal{A}'_{op}$ as $\mathcal{A}' \in \mathbb{R}^{T\times 2F}$ along feature dimension. The segment-level class scores that used for action localization are generated by feeding the feature map into temporal 1D convolutional classification layer:

$$CAS = f_{conv}(\mathcal{A}'; \phi_{cls}), \tag{6}$$

where $\phi_{cls}$ represents the trainable parameters of the classification layer, $CAS \in \mathbb{R}^{C\times T}$ denotes the segment-level action scores, and $C$ is the number of action classes. The concatenation of both enhanced feature helps to model the intrinsic relation between two modalities and get reliable classification scores, especially for the ambiguous action segments.

The class-wise action scores of each video are aggregated by averaging the top $k^{act}$ elements among the CAS, where $k^{act}$ is a hyperparameter to control the number of selected segments. Then, the softmax function is applied to compute the video-level action probability $P_c(v_n)$ for each class $c$ of video $v_n$. The classification loss is defined with binary cross entropy loss:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} -y_{n;c}\log P_c(v_n), \tag{7}$$

where $y_{n;c}$ is the normalized video-level label for the $c$-th class of the $n$-th video.

### D. Background Modeling With Modal Attention

Background segments are usually attributed to an additional action category in many approaches [9], [10], [58]. However, it is undesirable to force all background frames to belong to one specific class, as they do not share any common semantics in most cases. To overcome the limitation that background frames are dynamic and inconsistent, background frames could be formulated as out-of-distribution [11]. Considering the probability for class $c$ of a segment $s_t$, which can be decomposed into the in-distribution action classification and the background identification, is formulated according to the posterior probability:

$$\begin{aligned} P(y_t = c|s_t) &= P(y_t = c, d = 1|s_t) \\ &= P(y_t = c|d = 1, s_t)P(d = 1|s_t), \end{aligned} \tag{8}$$

where $y_t$ is the label of the segment, $d \in \{0, 1\}$ denotes the variable for the background identification, *i.e.*, $d = 1$ when the segment contains any actions while $d = 0$ when the segment belongs to background.

Generally, we note that action segments get high response to a specific action category, while background segments should produce low scores for all actions. Therefore, the features of background segments are prone to have small magnitudes. Since RGB and optical flow streams may focus on different segments due to the modal differences during training, a modal attention (MA) module is proposed to fuse both feature magnitudes:

$$\|f_t\| = \omega_{rgb} * \|a^t_{rgb}\| + \omega_{op} * \|a^t_{op}\|, \tag{9}$$

where $a^t_m = \mathcal{A}'_m[t]$ represents the $t^{th}$ feature vector of the modal $m$, $\|\cdot\|$ is the L2 norm function, $\omega_{rgb}$ and $\omega_{op}$ are trainable attention weights of RGB and optical flow magnitudes, respectively. Afterwards, we define the action probability of $s_t$ via modeling uncertainty with the fused magnitudes:

$$P(d = 1|s_t) = \frac{\min(h, \|f_t\|)}{h}, \tag{10}$$

where $f_t$ is the corresponding feature vector of $s_t$ and $h$ is the pre-defined maximum feature magnitude.

According to the the feature magnitudes, we select top $k^{act}$ and bottom $k^{bg}$ as pseudo action and pseudo background segments, respectively. The uncertainty modeling loss is formulated as:

$$\mathcal{L}_{\text{um}} = \frac{1}{N}\sum_{n=1}^{N}\left(\max(0, h - \|f_n^{\text{act}}\|) + \|f_n^{\text{bg}}\|\right)^2, \tag{11}$$

where $f_n^{act}$ and $f_n^{bg}$ are the mean features of pseudo action and background segments of $v_n$. Meanwhile, to prevent background segments from getting high softmax score for any actions due to the relativeness of softmax function, a background entropy loss is designed to force the background segments to have uniform probability distribution for action classes:

$$\mathcal{L}_{\text{be}} = \frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C} -\log(P_c(s_n^{\text{bg}})), \tag{12}$$

where $P_c(s_n^{\text{bg}})$ is the average action probability for the $c$-th class of the pseudo background segments.

### E. Class-Wise Similarity Constraint

According to the uncertainty modeling loss above, the features of pseudo action segments are prone to have large feature magnitude, which may result in large intra-class variance. This constraint lead to a contradiction between background-foreground separation and segment-level classification. Based on the intuition that features of the same action should be similar, we propose a class-wise similarity constraint loss to push the features from the same action category more similar and closer. Since the frame-level annotations are not available, we use the class activation map from the classification layer to obtain high-activation segments that probably contain action instances. Specifically, for each action classes $c$ in the corresponding ground-truth of a video $v_n$, we select the top $k^{act}$ segment features $\{f_c\}$ from the class activation map to calculate class-wise feature similarity:

$$\text{Sim}(f_i, f_j) = \frac{\sum_{k=1}^{F} f_i^k \times f_j^k}{\sqrt{\sum_{k=1}^{F}\left(f_i^k\right)^2}\cdot\sqrt{\sum_{k=1}^{F}\left(f_j^k\right)^2}}, \tag{13}$$

where $\langle f_i, f_j\rangle$ are feature pairs from the $F$-dimensional features $\{f_c\}$ of the action category $c$. Note that the cosine similarity is bounded between $-1$ and 1, which only considers the angle between the two feature vectors and ignores their magnitudes. Then the class-wise similarity loss is calculated as:

$$\mathcal{L}_{\text{sim}} = 1 - \frac{1}{n_c n_p}\sum_{\forall f_i, f_j \in \{f_c\}} \text{Sim}(f_i, f_j), \tag{14}$$

where $n_c$ is the number of action classes in the ground-truth, and $n_p = \frac{T(T-1)}{2}$ is the number of feature pairs among the selected action features, respectively. By minimizing $\mathcal{L}_{\text{sim}}$, features from the same action category in a video are pushed to be more similar regardless of their high feature magnitude, which is beneficial to balance the background modeling and segment-level action classification.

Finally, we combine the video-level classification loss $\mathcal{L}_{\text{cls}}$, uncertainty modeling loss $\mathcal{L}_{\text{um}}$, background entropy loss $\mathcal{L}_{\text{be}}$ and class-wise similarity constraint loss $\mathcal{L}_{\text{sim}}$ to jointly train the network:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{um}} + \beta\mathcal{L}_{\text{be}} + \gamma\mathcal{L}_{\text{sim}}, \tag{15}$$

where $\alpha$, $\beta$, $\gamma$ are trade-off coefficients to balance the loss components.

### F. Classification and Localization

For multi-label action classification of a test video $V$, we discard categories whose class probability $P_c(V)$ is lower than the threshold $\theta_{act}$. The segment-level posterior probability of the remaining action classes can be calculated as $u_c(t) = P_c(s_t) * P(d = 1|s_t)$ for the $t$-th segment according to (8) and (10). After that, the segments with posterior probabilities larger than $\theta_{seg}$ are selected as candidate segments, the consecutive segments are then grouped into action proposals. Note that we apply a set of $\theta_{seg}$ to generate enough proposals. To obtain action instance score $\psi$ of each proposal $(t_s, t_e, \psi, c)$, we calculate the classification score following the Outer-Inner-Contrastive function proposed in AutoLoc [50]. Specifically, the confidence score is defined as:

$$\begin{aligned}\psi &= \psi_{inner} - \psi_{outer} \\ &= \frac{\sum_{i=t_s}^{t_e} u_c(t)}{t_e - t_s} - \frac{\sum_{i=t_s-t_l}^{t_e} u_c(t) + \sum_{i=t_s}^{t_e+t_l} u_c(t)}{2t_l},\end{aligned} \tag{16}$$

where $t_s$ and $t_e$ are the start time and the end time of the detected instance, $c$ denotes the corresponding action category of the instance, $t_l = \eta(t_e - t_s)$ represents the inflated contrast area, $\eta$ is a hyper-parameter to control the inflated length. Finally, non-maximum suppression (NMS) is employed to generate final proposals.

## IV. Experiments and Discussions

In this section, we introduce the implementation details about our W-TAL method, and perform extensive experiments on THUMOS'14 [20] and ActivityNet v1.2 [21] benchmarks. A set of ablation studies are applied to evaluate the effectiveness of each module component and losses used in our proposed MDA. Also, a comparison is made for the fusion schemes of RGB and optical flow information. We further give qualitative analysis and some visualization results using our proposed approach.

### A. Experimental Setup

*1) Datasets:* We evaluate our model on two popular large-scale action localization benchmark datasets, *i.e.*, THUMOS'14 [20] and ActivityNet v1.2 [21].

**THUMOS'14** [20], which contains 200 validation and 213 test videos from 20 action categories. This dataset is very challenging as the length of an action varies significantly, from less than a second up to 26 minutes, with the mean duration around 3 minutes long. There is also a large variance in the length of an action instance, from less than a second to minutes. There are

15 action instances per video on average, and some videos may contain one or more action instances from other classes.

**ActivityNet v1.2** [21] is a popular large-scale benchmark for temporal action localization task with 100 categories involved in this dataset. Since the frame-level ground-truth of test videos are not available, we train on the training set with 4,819 videos and test on the validation set with 2,383 videos following the common practice. Most videos contain only a single action category, and action instances may last more than half of the duration in most video samples.

*2) Evaluation Metrics:* We follow the standard evaluation protocol by reporting mean Average Precision (mAP) values under different temporal intersection over union (IoU) thresholds. Specifically, the IoU thresholds for THUMOS'14 is set to [0.1:0.1:0.7], and for ActivityNet v1.2 dataset is [0.5:0.05:0.95]. Note that each ground-truth action instance can only match one action proposal. The evaluation on both datasets are conducted using the official evaluation code provided by ActivityNet.[1]

*3) Implementation Details: a) Feature Extraction:* Each video is divided into $T$ 16-frame non-overlapping segments. Due to the memory constraint, the number of input segments $T$ is fixed to 900 for THUMOS'14 and 150 for ActivityNet v1.2. TV-L1 algorithm [59] is applied to generate the optical flow frames from the RGB data. Then, we use the two-stream I3D network [29] pre-trained on Kinetics-400 dataset [29] to extract 1024-dimensional RGB and flow features of each video segment. Note that for fair comparison, the I3D feature extractor is not fine-tuned. Following STPN [18], we use stratified random perturbation during training and uniform sampling during test.

*b) Training and evaluation details:* (1) For THUMOS'14 dataset, we set video snippets length $T = 900$. The number of the pseudo action segments $k^{act}$ and background segments $k^{bg}$ are set to $T/9$ and $T/4$ respectively. We set $h = 100, \alpha = 5 \times 10^{-4}$, $\beta = 0.01$, $\gamma = 1$, $\eta = 0.25$ and $\theta_{act} = 0.2$ by grid search, and use a set of thresholds from 0 to 0.25 with the step 0.025 for $\theta_{seg}$. Non-maximum suppression (NMS) with threshold 0.7 is performed to remove highly overlapped proposals. The batch size is set to 16 during training. (2) For ActivityNet v1.2 dataset, since most video length in this dataset varies from a few seconds to several minutes, which is much shorter than the THUMOS'14 dataset, we set video snippets length $T = 150$ by grid search. Considering the long-term dependence of action classes in ActivityNet v1.2, we set the convolutional kernel size $k = 25$ during feature embedding, and apply additional average pooling to post-process the final CAS. The number of the pseudo action segments $k^{act}$ and background segments $k^{bg}$ are set to $T/5$ and $T/50$ respectively. According to parameter fine-tuning, we set $h = 100, \alpha = 10^{-4}, \beta = 1, \gamma = 1, \eta = 0.3$ and $\theta_{act} = 0.1$, and use a set of thresholds from 0 to 0.10 with the step 0.01 for $\theta_{seg}$. To remove highly overlapped proposals, we perforom NMS with threshold of 0.7. The batch size is set to 512 during training. For both datasets, we use Adam [60] optimizer. All the experiments are based on PyTorch-1.7 [61] on the RTX-2080Ti platform.

[1][Online]. Available: https://github.com/activitynet/ActivityNet/blob/master/Evaluation/

TABLE I
ABLATION STUDY OF THE ATTENTION MODULES IN DIFFERENT DIMENSIONS ON THE THUMOS'14 DATASET

| | TA | CA | MA | mAP@IoU (%) 0.1 | 0.3 | 0.5 | 0.7 | AVG |
|---|---|---|---|---|---|---|---|---|
| Baseline | | | | 67.5 | 50.5 | 29.8 | 10.7 | 39.8 |
| TA-only* | ✓ | | | 40.5 | 51.7 | 30.8 | 10.7 | 40.5 |
| TA-only | ✓ | | | 69.8 | 54.2 | 33.8 | 14.4 | 43.1 |
| CA-only | | ✓ | | 68.4 | 52.2 | 32.1 | 11.6 | 41.3 |
| MA-only | | | ✓ | 67.6 | 51.4 | 31.1 | 11.2 | 40.5 |
| TA+CA | ✓ | ✓ | | **69.9** | 54.3 | 34.1 | **14.4** | 43.5 |
| TA+MA | ✓ | | ✓ | 69.7 | 54.6 | 34.2 | 13.3 | 43.4 |
| CA+MA | | ✓ | ✓ | 69.0 | 52.8 | 31.7 | 11.3 | 41.4 |
| **MDA (Ours)** | ✓ | ✓ | ✓ | 69.7 | **55.2** | **35.6** | **14.4** | **44.2** |

* Indicates the Temporal Attention Module (TA) Without Temporal Relation Block (TRB). AVG Denotes the Averaged mAP Under IoU Thresholds 0.1:0.1:0.7. The bold entities highlight the best result under the certain IoU thresholds.

## B. Ablation Study

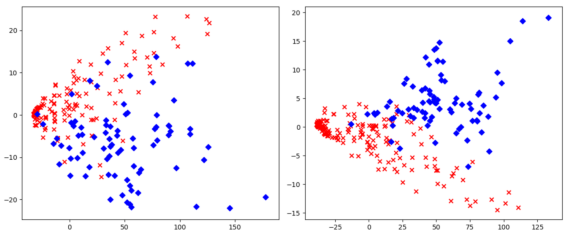
*1) Effectiveness of each attention component:* To investigate the contributions of different components of our MDA method, we conduct ablation study with all possible combinations of the three proposed attention modules. The baseline is the framework without any proposed attention mechanisms. Specifically, the extracted features are directly separated into RGB and flow streams and modeled by the 1D-convolutional layer respectively, then the refined features are stacked and send to the segment-level classifier without channel attention module. We also use the stacked feature to calculate feature magnitude when the modal attention is excluded. Experimental results under various IoU thresholds are shown in Table I. As the table presents, each of the temporal attention (TA), channel attention (CA) and modal attention (MA) module enhances the localization performance, and the combination of them brings about more improvement. The temporal attention module contributes the most, which directly affect the segment-level classification score for final proposal generation. The best average mAP is achieved when all attention modules are combined together. We also evaluate the effectiveness of the proposed Temporal Relation Block (TRB) by excluding it from the temporal attention module. Specifically, we can achieve 3.0% mAP performance gain under IoU threshold 0.5 with our TRB block. This demonstrates that the TRB block can effectively catch the temporal relation by measuring similarities among the action segments in a video, making it easier to find out complete action instances.

*2) Effectiveness of loss components:* In Table II, we evaluate the effectiveness of the loss components including the proposed similarity constraints on THUMOS'14. The baseline is performed only with video-level classification loss $L_{cls}$, which achieves 25.8% mAP on average. Based on this, the uncertainty modeling loss $L_{um}$ trains the model to produce large feature magnitudes for action segments and small ones for background segments, lifting the performance to 43.0%. Furthermore, the background entropy loss $L_{be}$ prevent background segments from generating high softmax score. Since the background frames are dynamic and inconsistent, in some cases where the probability distribution of some atypical background frames are

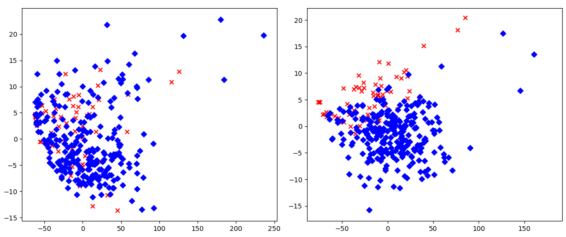
TABLE II
ABLATION STUDY OF LOSS COMPONENTS ON THUMOS'14

| $L_{cls}$ | $L_{um}$ | $L_{be}$ | $L_{sim}$ | mAP@IoU (%) 0.1 | 0.3 | 0.5 | 0.7 | AVG |
|---|---|---|---|---|---|---|---|---|
| ✓ | | | | 50.1 | 32.8 | 16.0 | 4.8 | 25.8 |
| ✓ | ✓ | | | 68.7 | 54.8 | 33.7 | 13.5 | 43.0 |
| ✓ | ✓ | ✓ | | 68.1 | 54.1 | 33.9 | 13.8 | 43.1 |
| ✓ | ✓ | | ✓ | 69.6 | 55.2 | 35.5 | 14.3 | 44.1 |
| ✓ | ✓ | ✓ | ✓ | **69.7** | **55.2** | **35.6** | **14.4** | **44.2** |

AVG Denotes the Averaged mAP Under IoU Thresholds 0.1:0.1:0.7.
The bold entities highlight the best result under the certain IoU thresholds.



(a) "*HighJump*" video



(b) "*Diving*" video

Fig. 5. Visualizations of feature embeddings $\mathcal{A}'$ before classification layer on two video samples (best viewed in color). **Left:** without similarity constraint; **Right:** with similarity constraint $L_{sim}$. Blue points represent action embeddings and red points denote background embeddings. Our proposed similarity constraint achieves a more separable feature distribution.

not unified by $L_{be}$ successfully, it may lead to imprecise temporal boundary of the generated proposals, especially on the proposals that mostly contain ambiguous background frames with small IoU with the ground-truth. Therefore, the performance at map@0.1 and map@0.3 seem to be inferior due to the increase of the False-Positive samples. As regards the proposed similarity constraint loss $L_{sim}$, since it constraints the action features more closer while maintaining large magnitudes produced by $L_{um}$, we can obtain 1.1% average mAP performance gain both with and without $L_{be}$. While the action features can be more discriminative from the ambiguous background features owing to $L_{sim}$, $L_{be}$ can play a positive role in background-foreground separation and promote the mAP under all IoU thresholds. For better comparison, we also visualize the feature distribution in 2-dimensional space using PCA with and without the similarity constraints, as demonstrated in Fig. 5. It can be observed that the background and action features are more separatable with the proposed similarity loss, while the original features tend to mix together. When combine the loss components together, we can achieve a new stat-of-the-art across all metrics, at 44.2% mAP on average.

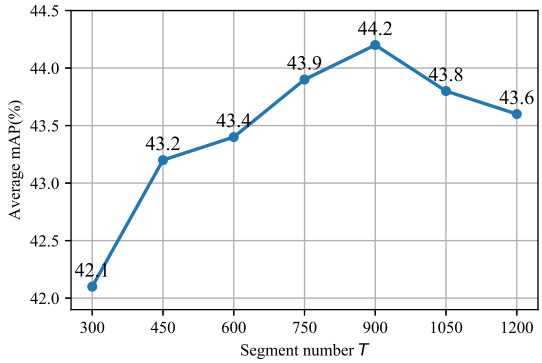



Fig. 6. Effectiveness of the number of segments $T$ on THUMOS'14 dataset. The localization results are reported as the average mAP under IoU thresholds 0.1:0.1:0.7. The best performance is achieved when $T = 900$.

TABLE III
ABLATION STUDY OF THE FUSION SCHEMES BETWEEN RGB AND OPTICAL FLOW ON THUMOS'14

| | Fusion scheme | mAP@IoU (%) 0.1 | 0.3 | 0.5 | 0.7 | AVG |
|---|---|---|---|---|---|---|
| One-stream | feature fusion | 69.6 | 52.8 | 32.2 | 12.5 | 42.2 |
| Two-stream | average CAS | 68.9 | 53.4 | 33.6 | 14.3 | 42.9 |
| | average magnitudes | 68.7 | 53.7 | 33.6 | 14.3 | 43.1 |
| | modal attention (add) | 69.1 | 54.5 | 34.5 | **14.8** | 43.6 |
| | modal attention (concat) | **69.7** | **55.2** | **35.6** | 14.4 | **44.2** |

AVG Denotes the Averaged mAP Under IoU Thresholds 0.1:0.1:0.7.
The bold entities highlight the best result under the certain IoU thresholds.

*3) Comparison of the modal fusion schemes:* As illustrated in Sec. I, there are several fusion strategies to combine complementary RGB and optical flow information. To verify the effectiveness of the fusion scheme, we implement five experiments in Table III, including one-stream and two-stream pipelines. (1): directly concatenate the RGB and optical flow features after feature extraction, as illustrated in Fig. 1(a), hence the modal attention is not used; (2): produce CAS for each modality separately and average them to generate proposals as in Fig. 1(b), similar to STPN [18] and TSCN [19]; (3): average feature magnitudes of RGB and optical flow features for background estimation, while the features are concatenated before classification; (4): use modal attention to fuse RGB and flow features by element-wise weighted addition before generating CAS and fuse both magnitudes for background modeling. (5): use modal attention to fuse the magnitudes during background modeling, while concatenate RGB and flow features directly for segment-level classification. From the table, the two-stream pipelines outperform the one-stream method that conduct feature-level fusion. Comparing (3) to (2), fusing two modalities during background estimation shows better performance than fusing CAS after segment-level classification respectively. The proposed modal attention for background modeling adaptively fuse two modalities and localize action instances more accurately than both average CAS or average feature magnitudes, while concatenating RGB and flow features directly (5) achieves better results than fusing them with model attention by element-wise addition (4).

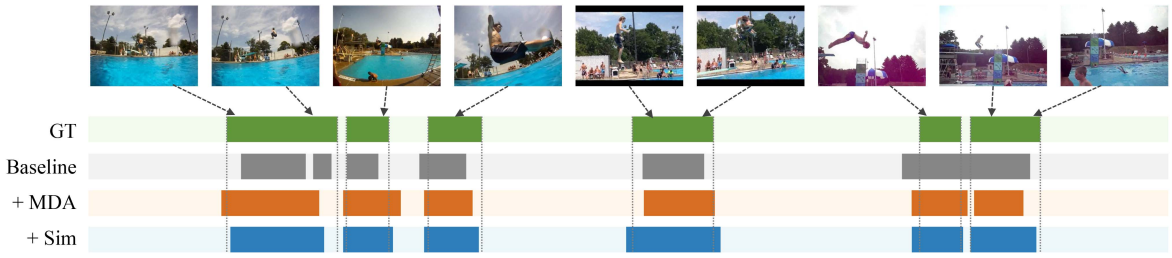



Fig. 7. Qualitative results visualization on THUMOS'14 (best viewed in color). The sample video contains multiple instances of *Diving* activity. The proposed Multi-Dimensional Attention (MDA) and the similarity constraints are added subsequently. GT represents the ground-truth.

We claim that the RGB and flow features have different distribution, and the concatenation retains original information from two modalities and introduce more learnable parameters rather than dimension reduction by addition.

*4) Analysis on the number of video segments:* Under the weakly-supervised setting of the temporal action localization task, we can only use video-level annotations for training. As described in our method, the video-level action scores are aggregated by averaging the top $k^{act}$ elements from the segment-level classification. Therefore, the temporal length of each video is fixed to $T$ to achieve parallel optimization during training. Considering the large variance among the temporal length of video samples, we employ a linear interpolation-based strategy following STPN [18] on the training videos. As shown in Fig. 6, we investigate the effect of the segment number $T$ on THUMOS'14, where $T$ is altered from 300 to 1200. From the experiments, we observe that the localization accuracy is not increase linearly with the temporal length $T$. The average mAP increases with $T$ until $T$ exceeds a certain value. This is because the action instances are not fully sampled when $T$ is not large enough to catch complete motion information, while it tends to be over-sampled when $T$ exceeds a proper length, leading to time intervals within the proposals. The best localization results can be achieved when $T = 900$ on the THUMOS'14 dataset.

## C. Comparison With the State-of-the-art

Table V shows the comparison of our MDA method with current state-of-the-art fully-supervised and weakly-supervised methods on THUMOS'14. "-" denotes the result not available from the corresponding paper. The results indicate that our proposed method outperforms the existing weakly-supervised methods across all metrics. In particular, the proposed MDA method achieves 35.6% mAP at the IoU threshold of 0.5 and obtains a performance gain of more than 1.6% in terms of the average mAP from IoU 0.1 to 0.7 over the recent TS-PCA [69]. Compared to the recent weakly-supervised methods that use extra annotations (count of the instances [52] or single-frame annotations [68]) for training, our model also achieves better performances. Moreover, our MDA even surpasses some fully supervised methods. This proves that our method produces more precise and complete localization.

Table IV reports the comparison on the AvcivityNet v1.2 dataset. Our method outperforms existing weakly supervised methods by 0.7% under the metric mAP@0.75, and achieves similar high performance on the average mAP to the AC-SNet [64], which mainly considers action-context separation and performs better at mAP@0.95. However, the performance gains on the ActivityNet v1.2 dataset are not as significant as those on the THUMOS'14 dataset. Note that most videos in the ActivityNet v1.2 dataset contains only one action class that composes a big portion of the whole video duration, and is regarded as a single-label classification dataset in some extent compared with the THUMOS'14 dataset which has 10.5% multi-label videos with 71.4% background frames on average. Therefore, the improvement brought by background modeling could be less significant. As discussed in [2], [64], THUMOS'14 is more appropriate for evaluating localization ability with a larger portion of background frames, we also speculate that the action localization performance in the ActivtyNet v1.2 dataset depends more on the classification head, which is not specially designed in the proposed network, while AUMN [65] builds segment-wise classifier that contributes to the classification performance. Additionally, we also find that the annotations of ActivityNet1.2 dataset are coarser than those in THUMOS'14 dataset since the former contains more ambiguous context frames which make our results inferior to [11], [64] at mAP@0.95. Although the performance gain by our method is smaller in ActivityNet v1.2 compared to

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE ACTIVITYNET V1.2 DATASET

| Supervision | Method | mAP@IoU (%) 0.5 | 0.75 | 0.95 | AVG |
|---|---|---|---|---|---|
| Full | CDC (CVPR 2017) [43] | 45.3 | 26.0 | 0.2 | 23.8 |
| | SSN (ICCV 2017) [42] | 41.3 | 27.0 | 6.1 | 26.6 |
| Weak† | 3C-Net (ICCV 2019) [52] | 37.2 | - | - | 21.7 |
| Weak | W-TALC (ECCV 2018) [8] | 37.0 | 12.7 | 4.5 | 18.0 |
| | MAAN (ICLR 2019) [56] | 33.7 | 21.9 | 5.5 | - |
| | Liu *et al.* (CVPR 2019) [62] | 34.0 | 20.9 | 5.7 | 21.2 |
| | TSM (ICCV 2019) [63] | 28.3 | 17.0 | 3.5 | 17.1 |
| | Nguyen *et al.* (ICCV 2019) [9] | 36.4 | 19.2 | 2.9 | - |
| | BaS-Net (AAAI 2020) [10] | 38.5 | 24.2 | 5.6 | 24.3 |
| | TSCN (ECCV 2020) [19] | 37.6 | 23.7 | 5.7 | 23.6 |
| | ACSNet (AAAI 2021) [64] | 40.1 | 26.1 | **6.8** | **26.0** |
| | HAMNet (AAAI 2021) [16] | 41.0 | 24.8 | 5.3 | 25.1 |
| | Lee *et al.* (AAAI 2021) [11] | 41.2 | 25.6 | 6.0 | 25.9 |
| | AUMN (CVPR 2021) [65] | **42.0** | 25.0 | 5.6 | 25.5 |
| | **MDA (Ours)** | 41.4 | **26.8** | 5.8 | **26.0** |

AVG Means the Averaged mAP Under IoU Thresholds 0.5:0.05:0.95. † Denotes the Use of Additional Information. Note That All Methods Employ I3D as the Feature Extractor.
The bold entities highlight the best result under the certain IoU thresholds.

TABLE V
COMPARISON WITH RECENT STATE-OF-THE-ART METHODS ON THE THUMOS'14 DATASET

| Supervision | Method | mAP@IoU (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | AVG[0.1-0.5] | AVG[0.3-0.7] | AVG |
| Full | SSN (ICCV 2017) [42] | 66.0 | 59.4 | 51.9 | 41.0 | 29.8 | 19.6 | 10.7 | 49.6 | 30.6 | 39.8 |
| | TAL-Net (CVPR 2018) [2] | 59.8 | 57.1 | 53.2 | 48.5 | 42.8 | 33.8 | 20.8 | 52.3 | 41.3 | 45.1 |
| | BSN (ECCV 2018) [4] | - | - | 53.5 | 45.0 | 36.9 | 28.4 | 20.0 | - | 36.8 | - |
| | P-GCN (ICCV 2019) [66] | **69.5** | **67.8** | 63.6 | 57.8 | 49.1 | - | - | **61.6** | - | - |
| | G-TAD (CVPR 2020) [46] | 66.1 | 64.2 | 54.5 | 47.6 | 40.2 | 30.8 | 23.4 | 54.5 | 39.3 | **46.7** |
| | BSN++ (AAAI 2021) [45] | - | - | 59.9 | 49.5 | 41.3 | 31.9 | 22.8 | - | 41.1 | - |
| | AFSD (CVPR 2021) [41] | - | - | **67.3** | **62.4** | **55.5** | **43.7** | **31.1** | - | **52.0** | - |
| Weak† | STAR (AAAI 2019) [67] | 68.8 | 60.0 | 48.7 | 34.7 | 23.0 | 11.7 | 6.2 | 47.0 | 24.9 | 36.2 |
| | 3C-Net (ICCV 2019) [52] | 59.1 | 53.5 | 44.2 | 34.1 | 26.6 | 16.7 | 8.1 | 43.5 | 25.9 | 34.6 |
| | SF-Net (ECCV 2020) [68] | **71.0** | **63.4** | **53.2** | **40.7** | **29.3** | **18.4** | **9.6** | **51.5** | **30.2** | **40.8** |
| Weak | Hide-and-Seek (ICCV 2017) [7] | 36.4 | 27.8 | 19.5 | 12.7 | 6.8 | - | - | 20.6 | - | - |
| | UntrimmedNet (CVPR 2017) [6] | - | - | 28.2 | 21.1 | 16.2 | - | 5.1 | 29.0 | - | - |
| | STPN (CVPR 2018) [18] | 52.0 | 44.7 | 35.5 | 25.8 | 16.9 | 9.9 | 4.3 | 35.0 | 18.5 | 26.4 |
| | W-TALC (ECCV 2018) [8] | 55.2 | 49.6 | 40.1 | 31.1 | 22.8 | 14.8 | 7.6 | 39.8 | 23.3 | 31.6 |
| | MAAN (ICLR 2019) [56] | 59.8 | 50.8 | 41.1 | 30.6 | 20.3 | 12.0 | 6.9 | 40.5 | 22.2 | 31.6 |
| | Liu *et al.* (CVPR 2019) [62] | - | - | 41.2 | 32.1 | 23.1 | 15.0 | 7.0 | - | 23.7 | - |
| | TSM (ICCV 2019) [63] | - | - | 39.5 | 31.9 | 24.5 | 13.8 | 7.1 | - | 23.4 | - |
| | Nguyen *et al.* (ICCV 2019) [9] | 64.2 | 59.5 | 49.1 | 38.4 | 27.5 | 17.3 | 8.6 | 47.7 | 28.2 | 37.8 |
| | BaS-Net (AAAI 2020) [10] | 58.2 | 52.3 | 44.6 | 36.0 | 27.0 | 18.6 | 10.4 | 43.6 | 27.3 | 35.3 |
| | TSCN (ECCV 2020) [19] | 63.4 | 57.6 | 47.8 | 37.7 | 28.7 | 19.4 | 10.2 | 47.0 | 28.8 | 37.6 |
| | ACSNet (AAAI 2021) [64] | - | - | 51.4 | 42.7 | 32.4 | 22.0 | 11.7 | - | 32.0 | - |
| | HAMNet (AAAI 2021) [16] | 65.4 | 59.0 | 50.3 | 41.1 | 31.0 | 20.7 | 11.1 | 49.4 | 30.8 | 39.8 |
| | Lee *et al.* (AAAI 2021) [11] | 67.5 | 61.2 | 52.3 | 43.4 | 33.7 | 22.9 | 12.1 | 51.6 | 32.9 | 41.9 |
| | AUMN (CVPR 2021) [65] | 66.2 | 61.9 | 54.9 | 44.4 | 33.3 | 20.5 | 9.0 | 52.1 | 32.4 | 41.5 |
| | TS-PCA (CVPR 2021) [69] | 67.6 | 61.1 | 53.4 | 43.4 | 34.3 | 24.7 | 13.7 | 52.0 | 33.9 | 42.6 |
| | **MDA (Ours)** | **69.7** | **63.1** | **55.2** | **46.6** | **35.6** | **25.0** | **14.4** | **54.0** | **35.4** | **44.2** |

AVG Means the Averaged mAP Under IoU Thresholds 0.1:0.1:0.7. The Algorithms are Divided Into Three Groups According to the Levels of Supervision.
† Denotes the Use of Additional Information. For Fair Comparison, All Methods Employ I3D as the Feature Extractor.
The bold entities highlight the best result under the certain IoU thresholds.

that in THUMOS'14, the improvement performance also verifies the common effectiveness on both datasets.

### D. Qualitative Results

To demonstrate the superiority of the proposed framework, we further visualize some localization results from a test video containing several *Diving* action instances. As illustrated in Fig. 7, the localization results of our method are relatively complete and precise. More specifically, the baseline may result in temporal interval interruption (the first instance) or attribute ambiguous segments as actions (the last instance), and fail to cover complete actions in some cases. With our proposed MDA, the network is more effective to differentiate between the foreground and background by extracting distinguished spatial and temporal features, while the modal attention makes full use of the complementary information from RGB and optical flow through background estimation. On this basis, when further introduce similarity constraints that group the action segments in the feature space, more confusing parts can be selected out. The accurate qualitative results prove the effectiveness of our proposed method under weak supervision.

## V. CONCLUSION

In this paper, we peopose a multi-dimensional attention (MDA) mechanism to explore temporal, channel and modal attention for weakly-supervised temporal action localization. The temporal attention helps capture most discriminative segments containing action instances. Meanwhile, a channel attention module highlights the action-related features to do segment-level action classification. Furthermore, we introduce a modal attention to fuse RGB and optical flow modalities during background estimation. Another similarity constraint helps to refine action instances in feature space and generate complete and precise action proposals. Each of the components is proved to be effective through ablation studies. Experiments on THUMOS'14 and ActivityNet v1.2 dataset demonstrate that the proposed method outperforms current state-of-the-art methods for weakly-supervised temporal action localization.

Our future works include three directions. First, more feature extractors will be explored to extract discriminative video features. Second, the proposed MDA can be applied to other video analysis tasks like spatial-temporal action detection. Third, we could further adopt audio information to conduct audio-visual action localization.

## REFERENCES

[1] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1049–1058.
[2] Y.-W. Chao *et al.*, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1130–1139.
[3] F. Long *et al.*, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 344–353.
[4] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[5] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 536–548.
[6] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4325–4334.

[7] K. K. Singh and Y. J. Lee, "Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3544–3553.

[8] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-TALC: Weakly-supervised temporal activity localization and classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 563–579.

[9] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5501–5510.

[10] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11320–11327.

[11] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1854–1862.

[12] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, Jun. 2022.

[13] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-C3D: Temporal convolutional 3D network for real-time action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7138–7145.

[14] X. Shu, L. Zhang, Y. Sun, and J. Tang, "Host-parasite: Graph LSTM-in-LSTM for group activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 663–674, Feb. 2021.

[15] P. Chen *et al.*, "Relation attention for temporal action localization," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2723–2733, Oct. 2020.

[16] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1637–1645.

[17] H. Su, X. Zhao, T. Lin, S. Liu, and Z. Hu, "Transferable knowledge-based multi-granularity fusion network for weakly supervised temporal action detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1503–1515, 2020.

[18] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6752–6761.

[19] Y. Zhai *et al.*, "Two-stream consensus network for weakly-supervised temporal action localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 37–54.

[20] H. Idrees *et al.*, "The THUMOS challenge on action recognition for videos "in the wild"," *Comput. Vis. Image Understanding*, vol. 155, pp. 1–23, 2017.

[21] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.

[22] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.

[23] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4041–4049.

[24] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Informat. Process. Syst. (NeurIPS)*, vol. 1, 2014, pp. 568–576.

[26] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[27] X. Shu, J. Tang, G. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1110–1118, Mar. 2021.

[28] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 636–647, Feb. 2022.

[29] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[30] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[32] D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[33] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imageNet?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6546–6555.

[34] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "carreira2017quo,xie2018rethinking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5552–5561.

[35] D. Guo, W. Li, and X. Fang, "Fully convolutional network for multi-scale temporal action proposals," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3428–3438, Dec. 2018.

[36] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3628–3636.

[37] C. Lin *et al.*, "Fast learning of temporal action proposal via dense boundary generator," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11499–11506.

[38] H. Liu, S. Wang, W. Wang, and J. Cheng, "Multi-scale based context-aware net for action detection," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 337–348, Feb. 2020.

[39] Y. Chen *et al.*, "Refinement of boundary regression using uncertainty in temporal action localization," in *Proc. Brit. Mach. Vis. Conf.*, 2020.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[41] C. Lin *et al.*, "Learning salient boundary feature for anchor-free temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3320–3329.

[42] Y. Zhao *et al.*, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2914–2923.

[43] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5734–5743.

[44] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3888–3897.

[45] 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 *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2602–2610.

[46] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10156–10165.

[47] F. Long *et al.*, "Coarse-to-fine localization of temporal action proposals," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1577–1590, Jun. 2020.

[48] J. Gao, K. Chen, and R. Nevatia, "Ctap: Complementary temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 68–83.

[49] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3604–3613.

[50] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "AutoLoc: Weakly-supervised temporal action localization in untrimmed videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 154–171.

[51] Z. Liu *et al.*, "Weakly supervised temporal action localization through contrast based evaluation networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3899–3908.

[52] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3C-Net: Category count and center loss for weakly-supervised action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8678–8686.

[53] R. Zeng *et al.*, "Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5797–5808, Dec. 2019.

[54] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Relational prototypical network for weakly supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11053–11060.

[55] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[56] Y. Yuan *et al.*, "Marginalized average attentional network for weakly-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2019.

[57] Z. Luo *et al.*, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 729–745.
[58] 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 *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2166–2174.
[59] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time tv-l 1 optical flow," in *Proc. Joint Pattern Recognit. Symp.*, 2007, pp. 214–223.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[61] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[62] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1298–1307.
[63] T. Yu *et al.*, "Temporal structure mining for weakly supervised action detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5521–5530.
[64] Z. Liu *et al.*, "ACSNet: Action-context separation network for weakly supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2233–2241.
[65] W. Luo *et al.*, "Action unit memory network for weakly supervised temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9969–9979.
[66] R. Zeng *et al.*, "Graph convolutional networks for temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7093–7102.
[67] Y. Xu *et al.*, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9070–9078.
[68] F. Ma *et al.*, "SF-Net: Single-frame supervision for temporal action localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 420–437.
[69] Y. Liu *et al.*, "The blessings of unlabeled background in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6176–6185.

**Zhengyan Chen** received the B.S. degree in information security from Hunan University, Changsha, China, in 2019. She is currently working toward the master's degree with the Key Laboratory of Machine Perception, School of Electronic and Computer Engineering, Peking University, Beijing, China, under the supervision of Prof. Hong Liu. Her research interests include computer vision, human action recognition, and video analysis and understanding.

**Hong Liu** (Member, IEEE) received the Ph.D. degree in mechanical electronics and automation in 1996. He is currently a Full Professor with the School of Electrical Engineering and Computer Science, Peking University (PKU), Beijing, China. Since 2013, he has been selected as Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan. He is also the Director of Open Lab on Human Robot Interaction, PKU. His research interests include computer vision and robotics, image processing, and pattern recognition. He has authored or coauthored more than 200 papers. Recently, he publishes many papers on international journals and conferences, including TMM, TCSVT, TCYB, TALSP, TRO, PR, IJCAI, ICCV, CVPR, ICRA, and IROS. Prof. Liu was the recipient of the Chinese National Aerospace Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is the Vice President of the Chinese Association for Artificial Intelligent (CAAI), and Vice-Chair of the Intelligent Robotics Society of CAAI. He was a Keynote Speaker, the Co-Chair, Session Chair, and PC Member of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC, and IIHMSP.

**Linlin Zhang** received the B.S. degree in automation from the Minzu University of China, Beijing, China, in 2018, and the M.S. degree from the Key Laboratory of Machine Perception, School of Electronic and Computer Engineering, Peking University, Beijing, China, in 2021, under the supervision of Prof. Hong Liu. Her research interests include computer vision, human action recognition, video analysis and understanding, and industrial defect detection.

**Xin Liao** (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2012, respectively. He is currently an Associate Professor and a Doctoral Supervisor with Hunan University, Changsha, China.

He was a Postdoctoral Fellow with the Institute of Software, Chinese Academy of Sciences, Beijing, China, and also a Research Associate with The University of Hong Kong, Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. His research interests include multimedia forensics, computer vision, and image/video analysis. He is an Associate Editor for the *IEEE Signal Processing Magazine*. He is also a member of Technical Committee (TC) on Multimedia Security and Forensics of Asia Pacific Signal and Information Processing Association, TC on Computer Forensics of Chinese Institute of Electronics, and TC on Digital Forensics and Security of China Society of Image and Graphics.