Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization

Linjiang Huang[1,2]   Liang Wang[3]   Hongsheng Li[1,2] *
[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]Centre for Perceptual and Interactive Intelligence, Hong Kong
[3]Institute of Automation, Chinese Academy of Sciences
ljhuang524@gmail.com, wangliang@nlpr.ia.ac.cn, hsli@ee.cuhk.edu.hk

Abstract

As a challenging task of high-level video understanding, weakly supervised temporal action localization has been attracting increasing attention. With only video annotations, most existing methods seek to handle this task with a localization-by-classification framework, which generally adopts a selector to select snippets of high probabilities of actions or namely the foreground. Nevertheless, the existing foreground selection strategies have a major limitation of only considering the unilateral relation from foreground to actions, which cannot guarantee the foreground-action consistency. In this paper, we present a framework named FAC-Net based on the I3D backbone, on which three branches are appended, named class-wise foreground classification branch, class-agnostic attention branch and multiple instance learning branch. First, our class-wise foreground classification branch regularizes the relation between actions and foreground to maximize the foreground-background separation. Besides, the class-agnostic attention branch and multiple instance learning branch are adopted to regularize the foreground-action consistency and help to learn a meaningful foreground classifier. Within each branch, we introduce a hybrid attention mechanism, which calculates multiple attention scores for each snippet, to focus on both discriminative and less-discriminative snippets to capture the full action boundaries. Experimental results on THUMOS14 and ActivityNet1.3 demonstrate the state-of-the-art performance of our method.

1. Introduction

Temporal action localization in videos has been widely used in various fields [39, 38]. This task aims to localize action instances in untrimmed videos along the temporal dimension. Most existing methods [45, 35, 42, 49, 4, 18, 20] are trained in a fully supervised manner. However, such
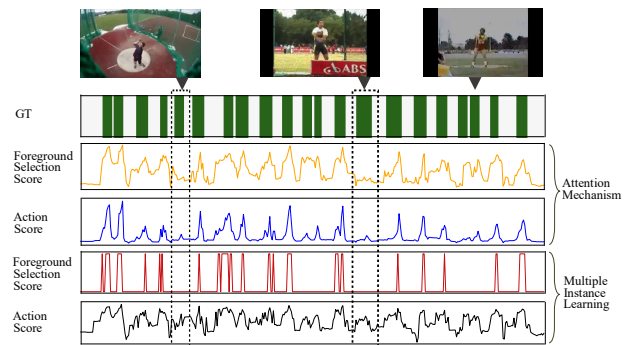
---

*Corresponding author.

Figure 1. An example of the action of "*HammerThrow*". The barcode is the ground-truth (GT). The following line charts are foreground selection scores and frame-wise classification scores. We show the results of two representative methods STPN [31] (attention mechanism) and W-TALC [26] (multiple instance learning). It is obvious that these methods cannot guarantee foreground-action consistency, leading to compromising results.

a requirement of frame-level annotations does not suit real-world applications since densely annotating large-scale videos is expensive and time-consuming. To address this difficulty, weakly supervised methods [14, 1, 41] have been developed with only video-level labels, which are much easier to annotate. Among diverse weak supervisions, video-level category labels are the easiest to collect and are thus most commonly used [41, 26, 31].

Because of the absence of frame-wise annotations, existing works mainly embrace a localization-by-classification pipeline [41, 44], in which an important component is the selector to select snippets with high probabilities of actions, or namely the *foreground*. Existing mechanisms for foreground selection can be categorized into two main strategies, *i.e.*, attention mechanism [26, 19] and multiple instance learning (MIL) [31]. However, both strategies have their intrinsic drawbacks. As shown in Fig. 1, the attention mechanism (either class-agnostic attention or class-wise attention) usually suffers from the discrepancy [37, 50, 19] between classification and detection, *i.e.*, the attention scores being concentrated around most discrim-

inative action snippets or wrongly focus on background snippets. On the other hand, the multiple instance learning should rely on a temporal top-k pooling operation, but there is no guarantee that all the top-k snippets would be foreground, since the number k is generally defined by humans. In summary, existing approaches lack the ability to maintain consistency between foreground and actions, that is, the foreground and actions should be mutually inclusive.

In this work, we propose to tackle the action localization problem by explicitly modeling and regularizing the foreground-action consistency. Given the insight that existing foreground selection strategies only consider the unilateral relation from foreground to actions, we propose a framework to further take bilateral relations into consideration. Based on a common video backbone, our method appends three branches on top of it. The first branch, named as class-wise foreground classification branch (CW branch, Sec.3.2), seeks to model the action-to-foreground relation. Meanwhile, it acts similarly to the noise contrastive estimation (NCE) [6, 28], which actually maximizes a lower bound on mutual information (MI) between the foreground feature and the feature of ground truth action, leading to better foreground-background separation. The second branch (CA branch, Sec.3.3) introduces a class-agnostic attention mechanism that to model the reverse foreground-to-action relation for complementing the first branch, so as to build the foreground-action consistency. Moreover, it enables to learn a semantically meaningful foreground feature. The third branch (MIL branch, Sec.3.4) is an MIL-like pipeline to further improve video classification and facilitate the learning of class-wise attention in the CW branch.

Within each branch, we adopt a hybrid attention mechanism to ease the attention learning and promote precise foreground prediction. In addition to focusing on key frames in the video, the hybrid attention mechanism can learn to accommodate less-discriminative snippets, which benefits capturing accurate action boundaries. To evaluate the effectiveness of our method, we perform experiments on two benchmarks, THUMOS14 [11] and ActivityNet1.3 [2]. Experimental results on the two benchmarks demonstrate the superior performance over state-of-the-art approaches.

Our main contributions are three-fold. (a) We introduce a class-wise foreground classification pipeline to improve the robustness of foreground prediction. This pipeline models and regularizes the foreground-action consistency that is mostly ignored by existing methods. (b) We propose a hybrid attention mechanism to improve the attention learning and help to capture accurate action boundaries. (c) The proposed class-wise foreground classification pipeline can play a complementary role over existing methods to consistently improve the action localization performance.

2. Related Work

Fully Supervised Temporal Action Localization. Different from action recognition [12, 36, 40, 3], temporal action localization aims to localize the start and end points of action instances, meanwhile recognizing the action category of each action instance. We group fully-supervised methods into two categories. The methods in the first category employ a multi-stage pipeline including proposal generation, classification and proposal refinement. These methods mainly focus on improving the quality of proposals [5, 18] and learning robust and accurate classifiers [35, 49]. In the second category, methods aim to generate action labels at the frame-level granularity [33, 15, 47], which need an additional merging step to obtain the final temporal boundaries. Even though these methods have achieved promising performance, they severely rely on frame-wise annotations.

Weakly Supervised Temporal Action Localization. Recently, many attempts have been made to solve temporal action localization with weak labels. UntrimmedNet [41] proposed to select relevant segments with attention mechanism or multiple instance learning, which is followed by most of the subsequent methods.

Attention-based methods aim to select snippets of high action probabilities by using attention mechanism. For instance, STPN [26] introduced a sparse regularization on the attention sequence to capture key frames of a video. 3C-Net [25] proposed to learn class-wise attention to obtain class-wise features for calculating a center loss. Some methods [27, 24, 9] leveraged the complementary nature of foreground and background to generate both foreground and background attention sequences to explicitly model background. HAM-Net [10] proposed a hybrid attention mechanism that includes temporal soft, semi-soft and hard attentions to capture full action instances. Note that, our method also utilize a hybrid attention mechanism, but totally differs from HAM-Net. First, we utilize different temperature values to generate multiple foreground attention sequences, while HAM-Net generate soft, semi-soft and hard attention sequences mainly by thresholding. Second, we only generate foreground attention sequences, while the hard attention sequence of HAM-Net also contain background snippets.

MIL-based methods [31, 16, 24] can be regarded as a hard selection mechanism based on the principle of multiple instance learning. In contrast to attention-based methods that automatically learn attention weights, MIL-based methods mainly rely on a top-k selection operation to select positive instances in the bag (video). However, as stated above, neither attention-based methods nor MIL-based methods can maintain foreground-action consistency.

There are some methods have noticed the importance of foreground-action consistency. For example, Refine-Loc [29] generated snippet-level hard pseudo labels by expanding previous detection results, TSCN [48] generated pseudo ground truth from the foreground attention sequence, and EM-MIL [21] put the pseudo-label generation into an expectation-maximization framework. There are also some methods [27, 24, 10] attempted to use class

activations as a *top-down* supervision to guide foreground attention generation. In contrast, our method requires no additional supervision to enforce foreground-action consistency, and in theory we can achieve better foreground-background separation.

3. Proposed Method

In this section, we elaborate on the proposed method. The overview of the proposed method is shown in Fig. 2.

Problem definition. Let $V = \{v_t\}_{t=1}^L$ be a video of temporal length L. Assume that we have a set of N training videos $\{V_i\}_{i=1}^N$ which are annotated with their action categories $\{y_i\}_{i=1}^N$, where y_i is a binary vector indicating the presence/absence of each action. During inference, for a video, we predict a set of action instances $\{(c, q, t_s, t_e)\}$, where c denotes the predicted action class, q is the confidence score, t_s and t_e represent the start time and end time.

Relation definition. (1) *Foreground-to-action relation*: the unilateral relation from foreground to actions, *i.e.*, the foreground must be some kind of actions. (2) *Action-to-foreground relation*: the unilateral relation from actions to foreground, *i.e.*, actions must be the foreground.

Overview. There are four modules in our method. Based on the I3D backbone, we utilize a feature embedding module (Sec.3.1) to extract task-oriented features. Then, three branches are appended on the top of it. The first branch, named as class-wise foreground classification branch (CW branch, Sec.3.2), seeks to model the action-to-foreground relation. The second branch (CA branch, Sec.3.3) introduces a class-agnostic attention mechanism to model the reverse foreground-to-action relation for complementing the first branch, so as to build the foreground-action consistency. The third branch (MIL branch, Sec.3.4) is an MIL-like pipeline to improve video classification and facilitate the learning of class-wise attention in the CW branch.

3.1. Feature Embedding Module

To extract task-oriented features, we utilize a feature embedding module that comprises two parts. The first part is a pre-trained network, *i.e.*, I3D [3]. Given a video, we first extract RGB features and optical-flow features respectively by the fixed backbone network. After feature encoding, we employ a two-layer temporal convolutional network [31, 19] to learn task-oriented features $X_e \in \mathbb{R}^{T \times D}$, where T denotes the number of snippets, and D is the dimension.

3.2. Class-Wise Foreground Classification Branch

As mentioned above, foreground and actions should be consistent and mutually inclusive. Nevertheless, most existing methods only consider the foreground-to-action relation, in other words, they only take advantage of the prior that the foreground must be actions. A possible result is that

the obtained foreground scores only focus on discriminative action snippets. Intuitively, a rational relation between foreground and actions should be bilateral, it is essential to further take the action-to-foreground relation into consideration. Inspired by the class-agnostic attention pipeline [26, 19, 16] that exploits the foreground-to-action relation, we propose a symmetric pipeline named class-wise foreground classification pipeline as a branch in our method.

We randomly initialize an action classifier $W_a \in \mathbb{R}^{(C+1) \times D}$ and a foreground classifier $W_f \in \mathbb{R}^D$, where C denotes the number of action categories, and the $(C+1)$-th class corresponds to the background. Given the embedding X_e, we calculate the cosine similarities between X_e and W_a to obtain the class activation scores $S_a \in \mathbb{R}^{T \times (C+1)}$ as

$$S_a(t, j) = \delta \cdot \cos(X_e(t), W_a(j)), \qquad (1)$$

where $X_e(t)$ denotes the embedding of the tth snippet, and δ is a scalar to control the scale of the value.

To build the action-to-foreground relation, we follow the line that action snippets are also foreground snippets and calculate the class-wise attention scores $A_a \in \mathbb{R}^{T \times (C+1)}$, which are used to aggregate the embedding X_e into the video-specific class-wise features $F_a \in \mathbb{R}^{(C+1) \times D}$ as

$$A_a(t, j) = \frac{\exp(\tau \cdot S_a(t, j))}{\sum_k \exp(\tau \cdot S_a(k, j))}, \qquad (2)$$

$$F_a(j) = \sum_t A_a(t, j) X_e(t), \qquad (3)$$

where t denotes the tth snippet, j represents the jth category, and τ is a temperature hyper-parameter controlling the smoothness of the softmax function. It is obvious that the feature $F_a(j)$ should be identified as the foreground if the jth action is performed in the video. Conversely, if the jth action is absent in the video, it should be classified as the background. This observation promotes us to introduce a foreground classification process for the feature F_a. Specifically, given the foreground classifier W_f, we can obtain the class-wise foreground activation scores $R_a \in \mathbb{R}^{C+1}$ and the class-wise foreground confidences $P_a \in \mathbb{R}^{C+1}$ as

$$R_a(j) = \delta \cdot \cos(F_a(j), W_f), \qquad (4)$$

$$P_a(j) = \frac{\exp(R_a(j))}{\sum_i \exp(R_a(i))}. \qquad (5)$$

A normalized cross-entropy loss $\mathcal{L}_{cw}$ is computed as

$$\mathcal{L}_{cw} = -\mathbb{E}[\hat{y}_a^T \log P_a], \qquad (6)$$

where $\hat{y} = y / \sum_{i=1}^{C+1} y(i)$ is the normalized ground-truth vector, and $y(C+1) = 0$. At this time, we actually transform the multi-label classification problem into multiple binary classification problems.

Discussion: Even if the CW branch is simple, it plays an important role in foreground-background separation.
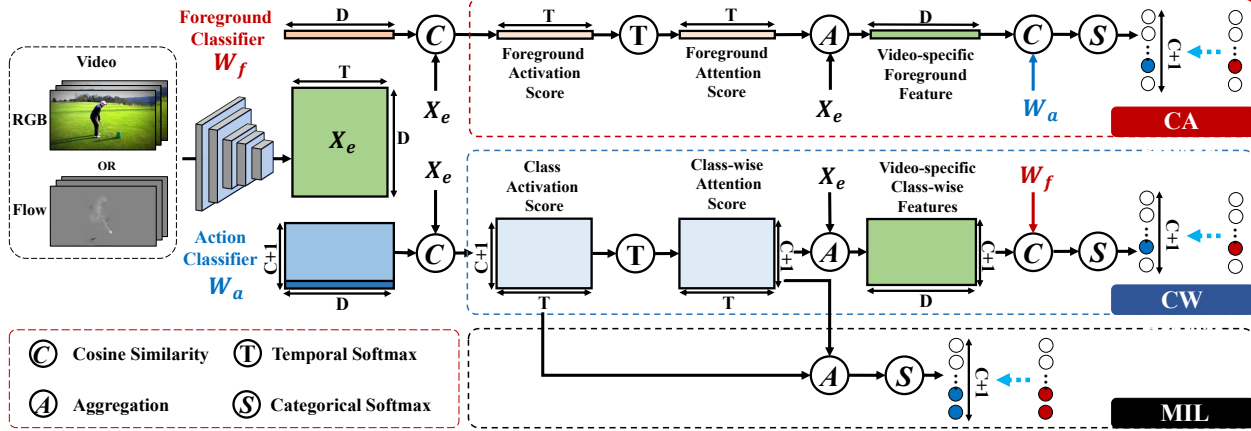
Figure 2. The overview of our method. We have three main branches. The class-wise foreground classification branch (CW branch) seeks to build the relation from actions to foreground, while the class-agnostic attention pipeline (CA branch) complements the reverse relation from foreground to actions for the first branch, so as to build the foreground-action consistency. The multiple instance learning branch (MIL branch) is a MIL-like pipeline to provide a different perspective for video classification and facilitate the learning of class-wise attention. For more details about the hybrid attention strategy, please refer to Fig. 3.

Specifically, we can transform Eq. (5) into

$$\boldsymbol{P}_a(j) = \frac{\exp(\delta \cdot \cos_j)}{\exp(\delta \cdot \cos_j) + \sum_{i, i \neq j} \exp(\delta \cdot \cos_i)}, \quad (7)$$

where $\cos_j$ is a simplification of $\cos(\boldsymbol{F}_a(j), \boldsymbol{W}_f)$. If the action category i $(i \neq j)$ is absent in the video, it is expected that the feature $\boldsymbol{F}_a(i)$ is a background feature. So there is one positive sample (if there is only one category in the video) from foreground and C negative samples from background. Therefore, Eq. (7) is similar to the noise contrastive estimation (NCE) [6, 28, 7] process, minimizing Eq. (6) actually maximizes the lower bound of mutual information (MI) between the foreground classifier $\boldsymbol{W}_f$ and the feature $\boldsymbol{F}_a(j)$. Besides, the background features are sampled from the same video of the feature $\boldsymbol{F}_a(j)$, they can be viewed as hard negative samples, because action instances are usually surrounded by visually similar clips [19], which further guarantees the foreground-background separation. Therefore, the CW branch does not only introduce the action-to-foreground relation into our method but also enable the learning of robust and discriminative features. However, the above analysis should be based on a meaningful foreground feature, but with the background class, there would be an ambiguity between the feature $\boldsymbol{W}_f$ of the foreground classifier and the background feature $\boldsymbol{W}_a(C + 1)$, leading to inferior performance as shown in our experiments. Therefore, it is essential to enhance the foreground meaning of the $\boldsymbol{W}_f$. Besides, the CW branch only considers the unilateral relation from actions to foreground, which is insufficient to build the foreground-action consistency.

3.3. Class-agnostic Attention Branch

To complement the loss of the relation from foreground to actions, we adopt a class-agnostic attention branch (CA

branch), which also enables to learn a semantically meaningful foreground classifier $\boldsymbol{W}_f$, playing a complementary role with the CW branch. We first calculate the frame-wise foreground activation scores $\boldsymbol{S}_f \in \mathbb{R}^T$ to obtain the foreground attention scores $\boldsymbol{A}_f \in \mathbb{R}^T$ as:

$$\boldsymbol{S}_f(t) = \delta \cdot \cos(\boldsymbol{X}_e(t), \boldsymbol{W}_f), \quad (8)$$

$$\boldsymbol{A}_f(t) = \frac{\exp(\tau \cdot \boldsymbol{S}_f(t))}{\sum_k \exp(\tau \cdot \boldsymbol{S}_f(k))}. \quad (9)$$

Likewise, we can obtain a video-specific foreground feature $\boldsymbol{F}_f \in \mathbb{R}^D$ through a feature aggregation process

$$\boldsymbol{F}_f = \sum_t \boldsymbol{A}_f(t) \boldsymbol{X}_e(t). \quad (10)$$

Then we calculate the cosine similarity between the feature $\boldsymbol{F}_f$ and the action classifier $\boldsymbol{W}_a$ to obtain the video-level class confidence scores $\boldsymbol{P}_f \in \mathbb{R}^{C+1}$

$$\boldsymbol{P}_f(j) = \frac{\exp(\delta \cdot \cos(\boldsymbol{F}_f, \boldsymbol{W}_a(j)))}{\sum_i \exp(\delta \cdot \cos(\boldsymbol{F}_f, \boldsymbol{W}_a(i)))}. \quad (11)$$

A normalized cross-entropy loss $\mathcal{L}_{ca}$ is computed as

$$\mathcal{L}_{ca} = -\mathbb{E}[\hat{\boldsymbol{y}}^T \log \boldsymbol{P}_f], \quad (12)$$

where $\hat{\boldsymbol{y}}$ is the same as the CW branch, *i.e.*, $\boldsymbol{y}(C + 1) = 0$. In this way, the CA branch is exactly a symmetric pipeline with the CW branch, which is also consistent with the inverse relations they introduce.

3.4. Multiple Instance Learning Branch

In addition to the class-agnostic attention pipeline, the multiple instance learning (MIL) pipeline is also a good complement to the CW branch. First, the MIL pipeline also considers the unilateral relation from foreground to actions.
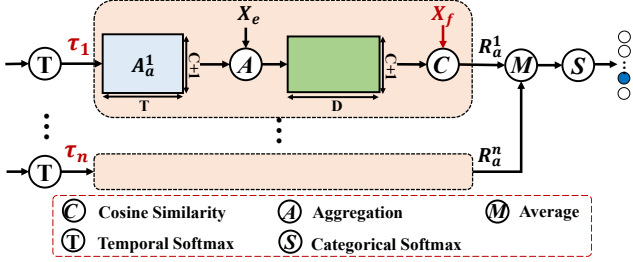
C	Cosine Similarity	A	Aggregation	M	Average
T	Temporal Softmax	S	Categorical Softmax		

Figure 3. Illustration of the hybrid attention strategy of the CW branch. We use N different τ to calculate N class-wise attention scores $\{\boldsymbol{A}_a^i\}_{i=1}^N$, the video-level foreground activation scores $\{\boldsymbol{R}_a^i\}_{i=1}^N$ are averaged to obtain the final foreground activation score. The other two branches use this strategy in the same way.

Second, the temporal top-k average pooling of the MIL is actually a class-wise hard attention operation, which could help to better learn the class-wise attention score in the CW branch. Besides, the MIL is more concerned with whether a class occurs in the whole video, while the class-agnostic attention focuses more on the local (because the aggregation is linear for each frame), the two pipelines provide two different perspectives to classify a video and would be complementary in a way. To better introduce the MIL pipeline into our framework, we change the temporal top-k average pooling into a class-wise soft attention operation, i.e., we share the class-wise attention scores $\boldsymbol{A}_a$ (Eq. (2)) and then aggregate the frame-wise class activation scores $\boldsymbol{S}_a$ (Eq. (1)) into the video-level class activation scores $\boldsymbol{R}_m \in \mathbb{R}^{C+1}$ as

$$\boldsymbol{R}_m(j) = \sum_t \boldsymbol{A}_a(t,j)\boldsymbol{S}_a(t,j). \qquad (13)$$

Similar to the CW branch and the CA branch, we can obtain its corresponding prediction $\boldsymbol{P}_m \in \mathbb{R}^{C+1}$ and a normalized cross-entropy loss $\mathcal{L}_{mil}$. Note that, since the background snippets exist in all videos, the ground truth $\boldsymbol{y}$ of the MIL branch should have the background, i.e., $\boldsymbol{y}(C+1) = 1$.

3.5. Hybrid Attention

As we can see, the attention mechanism plays an important role in our framework. However, even though we strive to build the foreground-action consistency, we find that the attention scores still cannot well cover the ground truth. Just as stated in [19], attention scores are prone to focus on discriminative foreground snippets and visually similar background snippets. In order to address this issue and simultaneously maintain the attention-based structure, we propose a hybrid attention strategy. The motivation comes from a similar observation with [26], i.e., an action can be recognized by identifying a set of key frames. If we can make the attention scores focus on key frames, false positives are expected to be largely reduced. To achieve this, we utilize a simple but effective way that uses a large temperature hyper-parameter τ for attention generation (e.g., Eq. (2) and Eq. (9)). In this way, the attention scores would be concentrated around snippets of high confidences.

Nevertheless, considering only key frames is insufficient, false negatives can also increase. Fortunately, because the key frames are modeled by the new attention scores ($\tau > 1.0$), the original attention scores ($\tau = 1.0$) have to accommodate some less-discriminative action snippets, which provides a fallback mechanism for our method. Therefore, using a hybrid attention strategy would be a reasonable way to improve the performance. As show in Fig. 3, for the CW branch, we first use N different τ to calculate N class-wise attention scores $\{\boldsymbol{A}_a^i\}_{i=1}^N$. With each class-wise attention score $\boldsymbol{A}_a^i$, we can obtain the video-level foreground activation score $\boldsymbol{R}_a^i$ (Eq. (4)). Finally, we average $\{\boldsymbol{R}_a^i\}_{i=1}^N$ and use a softmax operation (Eq. (5)) to obtain the probability confidence scores. The other two branches use the hybrid attention strategy in the same way.

3.6. Training Objectives

Our model is jointly optimized with three video-level classification losses. The overall loss function is as follows:

$$\mathcal{L}_{total} = \lambda_{cw}\mathcal{L}_{cw} + \lambda_{ca}\mathcal{L}_{ca} + \lambda_{mil}\mathcal{L}_{mil} \qquad (14)$$

where λ_{cw}, λ_{ca} and λ_{mil} are balancing hyper-parameters. Our method can also work without the background class. At this time, we use the original ground truth $\boldsymbol{y} \in \mathbb{R}^C$.

4. Experiments

4.1. Datasets

We evaluate our method on two action localization datasets THUMOS14 [11] and ActivityNet1.3 [2]. Note that, we only use video-level category labels for training.
THUMOS14. We use the subset from THUMOS14 that offers frame-wise annotations for 20 classes. We train the model on 200 untrimmed videos in its validation set and evaluate it on 212 untrimmed videos from the test set.
ActivityNet1.3. This dataset covers 200 complex daily activities and provides 10,024 videos for training, 4,926 for validation and 5,044 for testing. We use the training set to train our model and the validation set to evaluate our model.

4.2. Implementation Details

Model details. We use I3D [3] for feature extraction. The network for extracting task-oriented features contains two layers, the output channels are 1024 and 1024, respectively. The scale factor δ of cosine similarity is set as 5.0. We utilize the hybrid attention strategy in a three-head way, with temperature hyper-parameters of 1.0, 2.0 and 5.0. We use ReLU as the activation function in our model, and dropout layers are utilized before all activation functions.
Training details. Our method is implemented with PyTorch [30]. During training, we loop through each video in the mini-batch and accumulate gradients to deal with variable video lengths. We use Adam [13] to optimize our model, the training procedure stops at 100 epochs with the

Table 1. Detection performance comparisons over the THUMOS14 dataset. The column AVG indicates the average mAP at IoU thresholds 0.1:0.1:0.7. UNT and I3D represent UntrimmedNet features and I3D features, respectively. † means the method utilizes additional weak supervisions. *FAC-Net w/o BG* indicates that our method do not use a background class.

| Supervision | Year | Method | mAP @ IoU (%) | | | | | | | AVG (0.1:0.7) |
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	
Full	2017	R-C3D [42]	54.5	51.5	44.8	35.6	28.9	-	-	-
	2018	TAL-Net [4]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1
	2018	BSN [18]	-	-	53.5	45.0	36.9	28.4	20.0	-
	2019	GTAN [20]	69.1	63.7	57.8	47.2	38.8	-	-	-
Weak †	2018	STAR (I3D) [43]	68.8	60.0	48.7	34.7	23.0	-	-	-
	2019	3C-Net (I3D) [25]	59.1	53.5	44.2	34.1	26.6	-	8.1	-
Weak	2017	UntrimmedNet [41]	44.4	37.7	28.2	21.1	13.7	-	-	-
	2018	STPN (I3D) [26]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	27.0
	2018	AutoLoc (UNT) [34]	-	-	35.8	29.0	21.2	13.4	5.8	-
	2018	W-TALC (I3D) [31]	55.2	49.6	40.1	31.1	22.8	-	7.6	-
	2019	MAAN (I3D) [46]	59.8	50.8	41.1	30.6	20.3	12.0	6.9	31.6
	2019	CMCS (I3D) [19]	57.4	50.8	41.2	32.1	23.1	15.0	7.0	32.4
	2019	BM (I3D) [27]	60.4	56.0	46.6	37.5	26.8	17.6	9.0	36.3
	2020	BaS-Net (I3D) [16]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	2020	RPN (I3D) [8]	62.3	57.0	48.2	37.2	27.9	16.7	8.1	36.8
	2020	DGAM (I3D) [32]	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
	2020	TSCN (I3D) [48]	63.4	57.6	47.8	37.7	28.7	19.4	10.2	37.8
	2020	EM-MIL (I3D) [21]	59.1	52.7	45.5	36.8	30.5	22.7	**16.4**	37.7
	2020	A2CL-PT (I3D) [23]	61.2	56.1	48.1	39.0	30.1	19.2	10.6	37.8
	2021	HAM-Net (I3D) [10]	65.4	59.0	50.3	41.1	31.0	20.7	11.1	39.8
	2021	UM (I3D) [17]	67.5	61.2	52.3	43.4	**33.7**	**22.9**	12.1	41.9
	-	FAC-Net w/o BG (I3D)	63.8	57.5	48.1	40.5	31.3	20.0	10.2	38.8
	-	FAC-Net (I3D)	**67.6**	**62.1**	**52.6**	**44.3**	33.4	22.5	12.7	**42.2**

learning rate 0.0001. The balancing hyper-parameters λ_{cw}, λ_{ca} and λ_{mil} are 1.0, 0.1 and 0.1, respectively.

Testing details. We take the whole sequence of a video as input for testing. When localizing action instances, the class activation sequence is upsampled to the original frame rate. We reject the category whose class probability $P_f(j)$ (Eq. (11)) is lower than 0.1. Following [16], we use a set of thresholds to obtain the predicted action instances, then we perform non-maximum suppression to remove overlapping segments among rgb stream and optical-flow stream.

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

As shown in Tab. 1, on THUMOS14, even though we do not adopt a background class (*i.e.*, *FAC-Net w/o BG*), our method still outperforms the existing background modeling approaches [27, 16], indicating the effectiveness of building foreground-action consistency. Besides, with the background class, our method obtains a new state-of-the-art performance, achieving gains in terms of mAPs at most IoU thresholds and average mAP. Notably, our method outperforms some fully-supervised methods at IoU 0.1 and 0.2, manifesting the potential of weakly-supervised method.

Tab. 2 demonstrates the results on the ActivityNet1.3 dataset. As we can see, despite the simple architecture, our method obtains comparable performance with state-of-the-

Table 2. Results on ActivityNet1.3 validation set. AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95.

| Method | mAP @ IoU | | | |
	0.5	0.75	0.95	AVG
R-C3D [42]	26.8	-	-	12.7
TAL-Net [4]	38.2	18.3	1.3	20.2
CMCS (I3D) [19]	34.0	20.9	5.7	21.2
MAAN (I3D) [46]	33.7	21.9	5.5	-
BaS-Net (I3D) [16]	34.5	22.5	4.9	22.2
A2CL-PT (I3D) [23]	36.8	22.0	5.2	22.5
TSCN (I3D) [48]	35.3	21.4	5.3	21.7
UM (I3D) [17]	37.0	23.9	5.7	23.7
ACM-BANet (I3D) [24]	**37.6**	24.7	**6.5**	**24.4**
FAC-Net (I3D)	**37.6**	24.2	6.0	24.0

art approaches, and surpasses the fully-supervised methods R-C3D [42] and TAL-Net [4] by large margins of 11.3% and 3.8% in terms of average mAP, respectively.

4.4. Ablation Studies

We conduct a set of ablation studies on the THUMOS14 dataset to analyze the contribution of each component.

Branch analysis. To figure out the contribution of each branch, we should consider two questions: What is the performance of the individual branch? What is the relationship
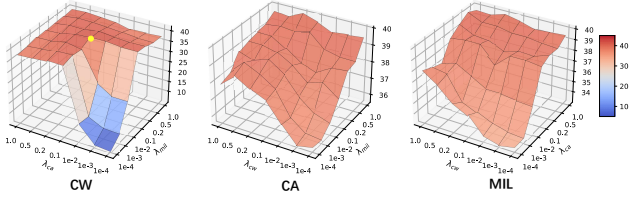
Figure 4. Ablation study on the balancing hyper-parameters λ_{cw}, λ_{ca} and λ_{mil}. We use a background class in these experiments.

between different branches? In Tab. 3, even if each of the three branches only considers unilateral relation between foreground and actions, the CW branch obtains better performance (31.2%), indicating that the relation from actions to foreground is more important, which can also enforce the foreground-background separation. Moreover, combining any two branches can consistently improve the performance, demonstrating the complementary relations among the three branches. Especially, the CW branch can significantly boost the performance of the CA branch and the MIL branch by 5.4% and 4.8% on average mAP, respectively.

In Fig. 4, we further explore the relations among branches by controlling the three balancing hyper-parameters. For example, when we evaluate the CW branch, we fix the λ_{cw} as 1.0, and adjust the balancing hyper-parameters of other two branches. Note that, we use a background class in these experiments. We can find that, when λ_{ca} and λ_{mil} are small, the performance drops to a very low level (about 6%). The reason is that the foreground classifier W_f along with the background class arouse an ambiguity of foreground and background, leading to inferior performance. When λ_{ca} and λ_{mil} are large, the performance is insensitive to the λ_{cw}, and remains at a promising level (above 40.0%). Besides, taking the CW branch as the main branch obtains the optimal result, even if it is used as an auxiliary branch, it also improves the performance.

To attain an intuitive insight into the three branches, in Fig. 5, we visualize the foreground activation scores (Eq. (8)) and action activation scores of ground truth under different combinations of branches. After adding the CW branch, foreground activation scores can better cover the ground truth, leading to more accurate detection results.

Effect of hybrid attention. From Tab. 3, we can see that the hybrid attention can consistently improve the performance, especially for the single branch. Besides, Tab. 4 shows the ablation studies on the number of attentions and temperature hyper-parameters τ. We can find that the number of attentions is not the more the better, too many attentions would degenerate the performance. Likewise, a large τ, e.g., 10.0, also works against the model, making it to focus too much on discriminative snippets. In Fig. 6, we also visualize the foreground activation scores and action activation scores under different hybrid attention settings. It is obvious that the introduction of the hybrid attention enables to obtain more accurate foreground prediction.

Table 3. Ablation studies on the THUMOS14 dataset. The column AVG indicates the average mAP at IoU thresholds 0.1:0.7.

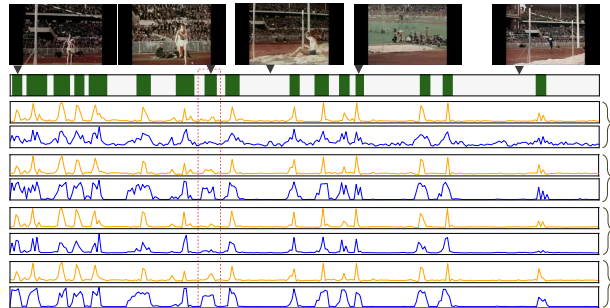| Method					AVG (0.1:0.7)
CW Branch	CA Branch	MIL Branch	Hybrid Attention	Background Class	
✓					31.2
	✓				30.5
		✓			29.6
✓			✓		34.6
	✓		✓		32.5
		✓	✓		32.5
✓	✓				35.9
✓	✓		✓		36.5
✓		✓			34.4
✓		✓	✓		35.7
	✓	✓			35.7
	✓	✓	✓		36.3
✓	✓	✓			37.5
✓	✓	✓	✓		38.8
✓		✓		✓	38.3
✓		✓	✓	✓	39.8
	✓	✓		✓	37.6
	✓	✓	✓	✓	38.4
✓	✓	✓		✓	40.8
✓	✓	✓	✓	✓	**42.2**

Figure 5. Visualization of foreground activation scores and action activation scores under different model setting. We show an example of "*Pole Vault*" in THUMOS14. For fair comparison, we do not use a background class.
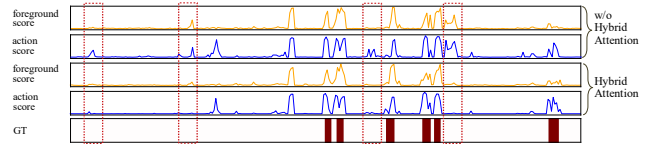

Figure 6. Visualization of foreground activation scores and action activation scores without and with the hybrid attention strategy. We show an example of "*Cliff Diving*" in THUMOS14.

Complementary role of the CW branch. In light of the observation that existing methods ignore the action-to-foreground relation, intuitively, the CW branch may play a complementary role to existing methods. In Tab. 5, we plug the CW branch into four weakly-supervised methods. We can find that the CW branch can significantly boost the per-
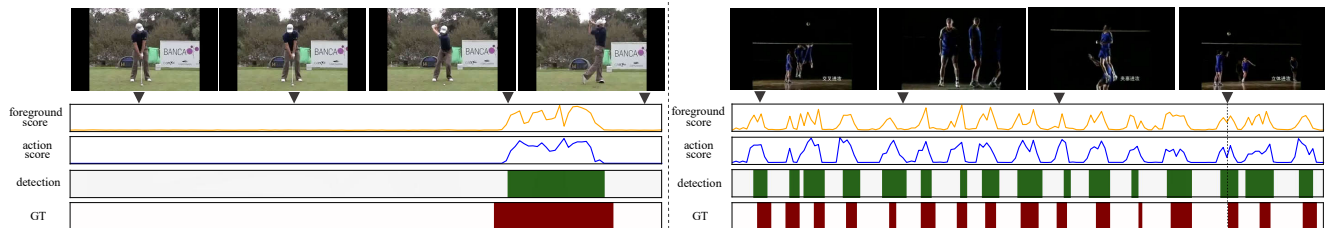
Figure 7. Qualitative results on THUMOS14 [11]. We show: 1) foreground activation scores, 2) activation scores of ground truth action, 3) the detected action instances and 4) ground truth. **Left:** An example of *Golf Swing*. **Right:** An example of *Volleyball Spiking*.
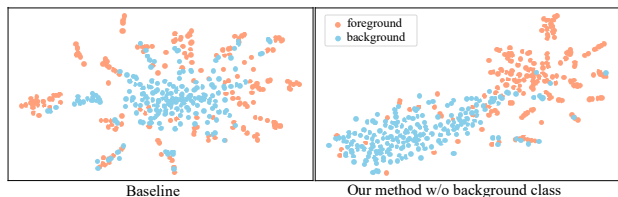

Figure 8. Visualization of foreground-background separation of embedding features of (a) only the CA branch and (b) our method via t-SNE [22] on the THUMOS14 test set. Note that, for a fair comparison, we do not use a background class in both models.

Table 4. Evaluation of the hybrid attention on THUMOS14. Each line shows a hybrid attention setting, *e.g.*, "(2) $\sim$ 1.0, 2.0" represents we use two attention scores with temperature hyper-parameter of 1.0 and 2.0, respectively.

Hybrid Attention	AVG			
	CW	CA	MIL	Full
(1) $\sim$ 1.0	31.2	30.5	29.6	40.8
(2) $\sim$ 1.0, 2.0	32.0	32.1	31.7	41.6
(2) $\sim$ 1.0, 3.0	33.6	32.3	31.6	41.8
(2) $\sim$ 1.0, 5.0	**34.7**	31.7	32.0	41.9
(2) $\sim$ 1.0, 10.0	32.9	31.2	26.4	33.8
(3) $\sim$ 1.0, 2.0, 3.0	34.4	32.2	31.4	41.9
(3) $\sim$ 1.0, 2.0, 5.0	34.6	**32.5**	**32.5**	**42.2**
(3) $\sim$ 1.0, 5.0, 10.0	33.4	32.0	26.7	33.2
(4) $\sim$ 1.0, 2.0, 3.0, 5.0	33.6	31.6	29.1	42.0

formance of two classic methods STPN [26] and W-TALC [31] by 1.7% and 1.9%, respectively. Even though the BM [27] explicitly models the background, the CW branch can further improve its performance. Besides, the performance of the recent method UM [17] can be also improved.

4.5. Qualitative Results

We visualize some examples of detected action instances in Fig. 7. In the first example of *Golf Swing*, our method pinpoints the only one action instance. In the second example of *Volleyball Spiking*. Even though this action is frequently performed in the video, our method successful detects all the action instances, which shows the ability to handle dense action occurrence. As we can see, our method significantly suppresses the responses of background. Besides, the foreground scores and action scores are consistent and well cover the ground truth. Fig. 8 shows the visualiza-

Table 5. Evaluation of the complementary role of the CW branch. Note that, our method needs to learn task-oriented features, which is impossible in STPN, so we show the result of "STPN + Embedding", which represents STPN plus a feature embedding module.

Method		mAP @ IoU		
		0.3	0.7	AVG
Attention based	STPN [26] (reproduced)	35.2	4.2	26.8
	STPN + Embedding	38.4	4.7	28.9
	STPN + Embedding + CW	40.3	5.6	30.6
	BM [27] (reproduced)	46.5	9.1	36.0
	BM + CW	47.9	10.0	37.6
MIL based	W-TALC [31] (reproduced)	40.4	7.2	31.6
	W-TALC + CW	42.0	8.7	33.5
	UM [17] (reproduced)	51.0	10.9	40.4
	UM + CW	51.6	11.1	40.8

tion of the features X_e about their foreground-background separation. As we can see, our method can better separate foreground from background than the baseline model.

5. Conclusion

We proposed a weakly-supervised action localization approach, named FAC-Net that consists of three branches. Different from the existing methods that only considers unilateral relation from foreground to actions, our method takes bilateral relations between actions and foreground into consideration. The proposed class-wise foreground classification branch introduces the action-to-foreground relation to maximize the foreground-background separation. Besides, the class-agnostic attention branch and the multiple instance learning branch are adopted to regularize the foreground-action consistency and learn a meaningful foreground feature. According to our experiments, the class-wise foreground classification branch can play a complementary role to existing methods to improve their performance.

References

[1] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

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

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

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

[5] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *IEEE International Conference on Computer Vision*, pages 5793–5802. IEEE, 2017.

[6] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738. IEEE, 2020.

[8] Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Relational prototypical network for weakly supervised temporal action localization. In *AAAI Conference on Artificial Intelligence*, 2020.

[9] Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Modeling sub-actions for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 2021.

[10] Ashraful Islam, Chengjiang Long, and Richard J Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *AAAI Conference on Artificial Intelligence*, 2021.

[11] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, 2014.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[15] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165. IEEE, 2017.

[16] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI Conference on Artificial Intelligence*, 2020.

[17] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. *arXiv preprint arXiv:2006.07006*, 2020.

[18] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*, pages 3–19. Springer, 2018.

[19] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1307. IEEE, 2019.

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

[21] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. *arXiv preprint arXiv:2004.00163*, 2020.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[23] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *European Conference on Computer Vision*, pages 283–299. Springer, 2020.

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

[25] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *IEEE International Conference on Computer Vision*, pages 8679–8687. IEEE, 2019.

[26] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761. IEEE, 2018.

[27] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. *arXiv preprint arXiv:1908.06552*, 2019.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[29] Alejandro Pardo, Humam Alwassel, Fabian Caba, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In *IEEE Winter Conference on Applications of Computer Vision*, pages 3319–3328, 2021.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *European Conference on Computer Vision*, pages 563–579. Springer, 2018.

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

[33] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5734–5743. IEEE, 2017.

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

[35] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058. IEEE, 2016.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[37] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE International Conference on Computer Vision*, pages 3544–3553. IEEE, 2017.

[38] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488. IEEE, 2018.

[39] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM International Conference on Multimedia*, pages 371–380. ACM, 2015.

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, pages 4489–4497. IEEE, 2015.

[41] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4325–4334. IEEE, 2017.

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

[43] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. *arXiv preprint arXiv:1811.07460*, 2018.

[44] Le Yang, Dingwen Zhang, Tao Zhao, and Junwei Han. Equivalent classification mapping for weakly supervised temporal action localization. *arXiv preprint arXiv:2008.07728*, 2020.

[45] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687. IEEE, 2016.

[46] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. *arXiv preprint arXiv:1905.08586*, 2019.

[47] Zehuan Yuan, Jonathan C Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692. IEEE, 2017.

[48] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 37–54. Springer, 2020.

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

[50] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: A weakly supervised temporal action detector. In *ACM International Conference on Multimedia*, pages 35–44. ACM, 2018.