Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization

Yuanhao Zhai[1], Le Wang[1⋆], Wei Tang[2], Qilin Zhang[3], Junsong Yuan[4], and Gang Hua[5]

[1] Xi'an Jiaotong University, China
[2] University of Illinois, Chicago, USA
[3] HERE Technology, USA
[4] State University of New York at Buffalo, USA
[5] Wormpex AI Research, USA

Abstract. Weakly-supervised Temporal Action Localization (W-TAL) aims to classify and localize all action instances in an untrimmed video under only video-level supervision. However, without frame-level annotations, it is challenging for W-TAL methods to identify false positive action proposals and generate action proposals with precise temporal boundaries. In this paper, we present a Two-Stream Consensus Network (TSCN) to simultaneously address these challenges. The proposed TSCN features an iterative refinement training method, where a frame-level pseudo ground truth is iteratively updated, and used to provide frame-level supervision for improved model training and false positive action proposal elimination. Furthermore, we propose a new attention normalization loss to encourage the predicted attention to act like a binary selection, and promote the precise localization of action instance boundaries. Experiments conducted on the THUMOS14 and ActivityNet datasets show that the proposed TSCN outperforms current state-of-the-art methods, and even achieves comparable results with some recent fully-supervised methods.

Keywords: Temporal Action Localization; Weakly-Supervised Learning

1 Introduction

The task of Weakly-supervised Temporal Action Localization (W-TAL) aims at simultaneously localizing and classifying all action instances in a long untrimmed video given only video-level categorical labels in the learning phase. Compared to its fully-supervised counterpart, which requires frame-level annotations of all action instances during training, W-TAL greatly simplifies the procedure of data collection and avoids annotation bias of human annotators, therefore has been widely studied [18, 41, 34, 27, 30, 1, 23, 46, 24, 28, 26, 43, 20] in recent years.

Several W-TAL methods [41, 30, 27, 23, 28, 26, 20] adopt a Multiple Instance Learning (MIL) framework, where a video is treated as a bag of frames/snippets

⋆ Corresponding author.

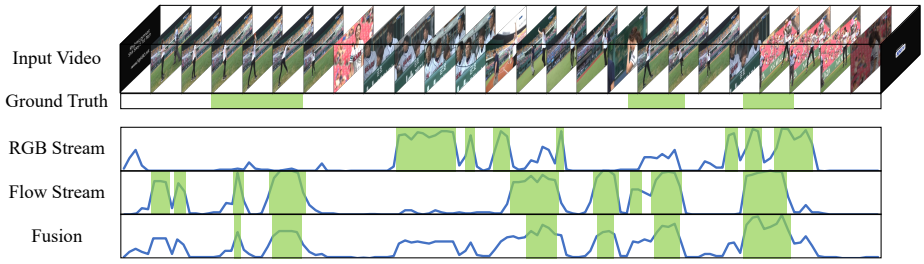

Fig. 1: Visualization of two-stream outputs and their late fusion result. The first two rows are an input video and the ground truth action instances, respectively. The last three rows are attention sequences (scaled from 0 to 1) predicted by the RGB stream, the flow stream and their weighted sum (*i.e.*, the fusion result), respectively, and the horizontal and vertical axes denote the time and the intensity of attention values, respectively. The green boxes denote the localization results generated by thresholding the attention at the value of 0.5. By properly combining the two different attention distributions predicted by the RGB and flow streams, the late fusion result achieves a higher true positive rate and a lower false positive rate, and thus has better localization performance

to perform the video-level action classification. During testing, the trained model slides over time and generates a Temporal-Class Activation Map (T-CAM) [49, 27] (*i.e.*, a sequence of probability distributions over action classes at each time step) and an attention sequence that measures the relative importance of each snippet. The action proposals are generated by thresholding the attention value and/or the T-CAM. This MIL framework is usually built on two feature modalities, *i.e.,* RGB frames and optical flow, which are fused in two possible ways. *Early fusion* methods [30, 34, 1, 23, 24, 20] concatenate the RGB and optical flow features before they are fed into the network, and *late fusion* methods [27, 23, 28, 26] compute a weighted sum of their respective outputs before generating action proposals. An example of late fusion is shown in Fig. 1.

Despite these recent development, two major challenges still persist. One of the most critical problems that prior W-TAL methods suffer from is the lack of ability to rule out false positive action proposals. Without frame-level annotations, they localize action instances that do not necessarily correspond to the video-level labels. For example, a model may falsely localize the action "swimming" by only checking the existence of water in the scene. Therefore, it is necessary to exploit more fine-grained supervision to guide the learning process. Another problem lies in the generation of action proposals. In previous methods, action proposals are generated by thresholding the activation sequence with a fixed threshold, which is preset empirically. It has a significant impact on the quality of action proposals: a high threshold may result in incomplete action proposals while a low threshold can bring more false positives. But how to get out of this dilemma was rarely studied.

In this paper, we introduce a Two-Stream Consensus Network (TSCN) to address the two aforementioned problems. To eliminate false positive action proposals, we design an iterative refinement training scheme, where a frame-level pseudo ground truth is generated from late fusion attention sequence, and serves as a more precise frame-level supervision to iteratively update two-stream models. Our intuition is simple: late fusion is essentially a voting ensemble of the RGB and flow streams, and if a proper fusion parameter (*i.e.*, the hyperparameter to control the relative importance of two streams) is selected, late fusion can provide more accurate result compared with each individual stream. The advantage of combining these two streams has been demonstrated by the Two-Stream Convolutional Networks [37] for action recognition. As shown in Fig. 1, the two streams produce different activation distributions, which lead to different false positives and false negatives. However, when they are combined, the false positive action proposals that only exist in one stream can be largely eliminated, and a high activation value occurs only when both streams are confident that an action instance exists. Since the late fusion result is of higher quality than single stream result, it can in turn serve as a frame-level pseudo ground truth to supervise and refine both streams. To generate high-quality action proposals, we introduce a new attention normalization loss. It pushes the predicted attention to approach extreme values, *i.e.*, 0 and 1, so as to avoid ambiguity. As a result, simply setting the threshold to 0.5 yields high-quality action proposals.

Formally, given an input video, RGB and optical flow features are first extracted from pre-trained deep networks. Then two-stream base models are trained with video-level labels on RGB and optical flow features, respectively, where the attention normalization loss is used to learn the attention distribution. After obtaining two-stream attention sequences, a frame-level pseudo ground truth is generated based on their weighted sum (*i.e.*, the late fusion attention sequence), and in turn provides frame-level supervision to improve the two-stream models. We iteratively update the pseudo ground truth and refine the two-stream base models, and the normalization term at the same time forces the predicted attention to approach a binary selection. The final localization result is obtained by thresholding the late fusion attention sequence.

To summarize, our contribution is threefold:

- We introduce a Two-Stream Consensus Network (TSCN) for W-TAL. The proposed TSCN uses an iterative refinement training method, where a pseudo ground truth generated from late fusion attention sequence at previous iteration can provide more precise frame-level supervision to current iteration.
- We propose an attention normalization loss function, which forces the attention to act like a binary selection, and thus improves the quality of action proposals generated by the thresholding method.
- Extensive experiments are conducted on two standard benchmarks (*i.e.*, THUMOS14 and ActivityNet) to demonstrate the effectiveness of the proposed method. Our TSCN significantly outperforms previous state-of-the-art W-TAL methods, and even achieves comparable results to some recent fully-supervised TAL methods.

2 Related Work

Action Recognition. Traditional methods [19, 7, 8, 39] aim to model spatio-temporal information via hand-crafted features. Two-Stream Convolutional Networks [37] use two separate Convolutional Neural Networks (CNNs) to exploit appearance and motion clues from RGB frames and optical flow, respectively, and use a late fusion method to reconcile the two-stream outputs. [10] focuses on studying different ways to fuse the two streams. The Inflated 3D ConvNet (I3D) [3] expands the 2D CNNs in two-stream networks to 3D CNNs. Several recent methods [47, 5, 35, 40, 31] focus on directly learning motion clues from RGB frames instead of calculating optical flow.

Fully-supervised Temporal Action Localization. Fully-supervised TAL requires frame-level annotations of all action instances during training. Several large-scale datasets have been created for this task, such as THUMOS [15, 13], ActivityNet [2], and Charades [36]. Many methods [33, 48, 12, 14, 6, 42, 22, 4] adopt a two-stage pipeline, *i.e.*, action proposal generation followed by action classification. Several methods [42, 6, 11, 4] adopt the Faster R-CNN [32] framework to TAL. Most recently, some methods [22, 25, 21] try to generate action proposals with more flexible durations. Zeng *et al.* [45] apply the Graph Convolutional Networks (GCN) [17, 38] to TAL to exploit proposal-proposal relations.

Weakly-supervised Temporal Action Localization. W-TAL, which only requires video-level supervision during training, greatly relieves the data annotation efforts, and draws more and more attention from the community recently. Hide-and-Seek [18] randomly hides part of the input video to guide the network to discover other relevant parts. UntrimmedNet [41] consists of a selection module to select the important snippets and a classification module to perform per snippet classification. Sparse Temporal Pooling Network (STPN) [27] improves UntrimmedNet by adding a sparse loss to enforce the sparsity of selected segments. W-TALC [30] jointly optimizes a co-activity similarity loss and a multiple instance learning loss to train the network. AutoLoc [34] is one of the first two-stage methods in W-TAL, and it first generates initial action proposals and then regresses the boundaries of the action proposals with an Outer-Inner-Contrastive loss. CleanNet [24] improves AutoLoc by leveraging the temporal contrast in snippet-level action classification predictions. Liu *et al.* [23] propose a multi-branch network to model different stages of action. Besides, several methods [28, 20] focus on modeling the background and achieve state-of-the-art performances.

Recently, RefineLoc [1] uses an iterative refinement method to help the model capture a *complete* action instance. And our method is distinct from RefineLoc in three main aspects. (1) We adopt a late fusion framework, while RefineLoc adopts an early fusion framework. (2) Our pseudo ground truth is generated from two-stream late fusion attention sequences, which provides better localization performance than each single stream, while RefineLoc generates the pseudo ground truth by expanding previous localization results, which might result in coarser and over-complete action proposals. (3) We introduce a new attention normalization loss to explicitly avoid the ambiguity of attention, while RefineLoc has no explicit constraints on attention values.

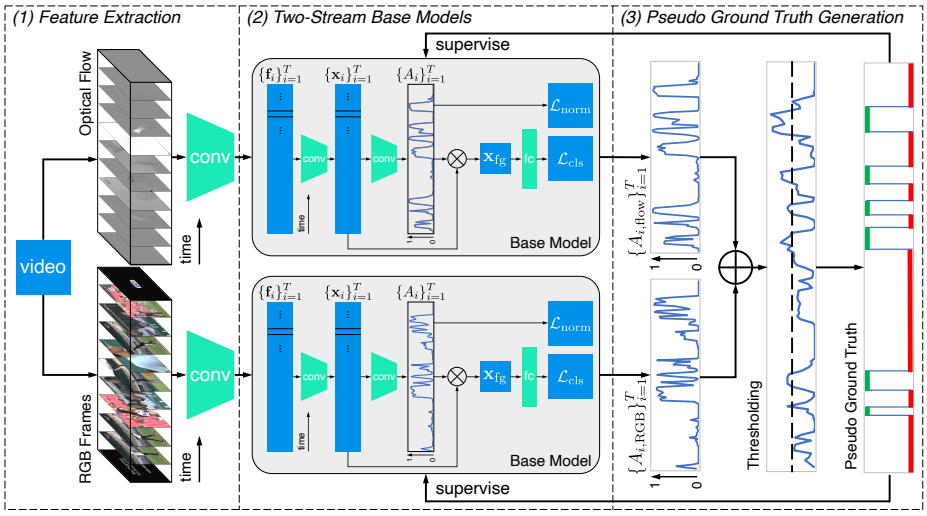

Fig. 2: An overview of the proposed Two-Stream Consensus Network, which consists of three parts: (1) RGB and optical flow snippet-level features are extracted with pre-trained models; (2) two-stream base models are separately trained using these RGB and optical flow features; (3) frame-level pseudo ground truth is generated from the two-stream late fusion attention sequence, and in turn provides frame-level supervision to two-stream base models

3 Two-Stream Consensus Network

In this section, we first formulate the task of Weakly-supervised Temporal Action Localization (W-TAL), and then describe the proposed Two-Stream Consensus Network (TSCN) in detail. The overall architecture is shown in Fig. 2.

3.1 Problem Formulation

Assume we are given a set of training videos. For each video v, we only have its video-level categorical label $\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^C$ is a normalized multi-hot vector, and C is the number of action categories. The goal of temporal action localization is to generate a set of action proposals $\{(t_s, t_e, c, \psi)\}$ for each testing video, where t_s, t_e, c, ψ denote the start time, the end time, the predicted action category and the confidence score of the action proposal, respectively.

3.2 Feature Extraction

Following recent W-TAL methods [34, 27, 30, 23, 24, 28, 26, 43, 20], we construct TSCN upon snippet-level feature sequences extracted from the raw video volume. The RGB and optical flow features are extracted with pre-trained deep networks (*e.g.*, I3D [3]) from non-overlapping fixed-length RGB frame snippets

and optical flow snippets, respectively. They provide high-level appearance and motion information of the corresponding snippets. Formally, given a video with T non-overlapping snippets, we denote the RGB features and optical flow features as $\{\mathbf{f}_{\text{RGB},i}\}_{i=1}^{T}$ and $\{\mathbf{f}_{\text{flow},i}\}_{i=1}^{T}$, respectively, where $\mathbf{f}_{\text{RGB},i}, \mathbf{f}_{\text{flow},i} \in \mathbb{R}^D$ are the feature representations of the i-th RGB frame and optical flow snippet, respectively, and D denotes the channel dimension.

3.3 Two-Stream Base Models

After obtaining the RGB and optical flow features, we first use two-stream base models to perform the video-level action classification, and then iteratively refine the base models with a frame-level pseudo ground truth. The features of two modalities are fed into two separate base models, respectively, and the two base models use the same architecture but do not share parameters. Therefore, in this subsection, we omit the subscript RGB and flow for conciseness.

Since the features are not originally trained for the W-TAL task, we concatenate the T input features $\{\mathbf{f}_i\}_{i=1}^{T}$, and use a set of temporal convolutional layers to generate a set of new features $\{\mathbf{x}_i\}_{i=1}^{T}$, where $\mathbf{x}_i \in \mathbb{R}^{D'}$, and D' denotes the output feature dimension.

As a video may contain background snippets, to perform video-level classification, we need to select snippets that are likely to contain action instances and meanwhile filter out snippets that are likely to contain background. To this end, an attention value $A_i \in (0, 1)$ to measure the likelihood of the i-th snippet containing an action is given by a fully-connected (FC) layer:

$$A_i = \sigma\left(\mathbf{w}_A \cdot \mathbf{x}_i + b_A\right), \tag{1}$$

where $\sigma(\cdot)$, $\mathbf{w}_A$, and b_A are the sigmoid function, weight vector and bias of the attention layer. We then perform attention-weighted pooling over the feature sequence to generate a single foreground feature $\mathbf{x}_{\text{fg}}$, and feed it to an FC softmax layer to get the video-level prediction:

$$\mathbf{x}_{\text{fg}} = \frac{1}{\sum_{i=1}^{T} A_i} \sum_{i=1}^{T} A_i \mathbf{x}_i, \tag{2}$$

$$\hat{y}_c = \frac{e^{\mathbf{w}_c \cdot \mathbf{x}_{\text{fg}} + b_c}}{\sum_{i=1}^{C} e^{\mathbf{w}_i \cdot \mathbf{x}_{\text{fg}} + b_i}}, \tag{3}$$

where $\hat{y}_c$ is the probability that the video contains the c-th action, and $\mathbf{w}_c$ and b_c are the weight and bias of the FC layer for category c. The classification loss function $\mathcal{L}_{\text{cls}}$ is defined as the standard cross entropy loss:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{4}$$

where y_c denotes the value of label vector $\mathbf{y}$ at index c.

Ideally, an attention value is expected to be binary, where 1 indicates the presence of action while 0 indicates background. Recently, several methods [28, 20] introduce a background category, and use the background classification to guide the learning of attention. In this work, instead of using background classification, we introduce an attention normalization term to force the attention to approach extreme values:

$$\mathcal{L}_{\text{att}} = \frac{1}{l} \min_{\substack{A \subset \{A_i\} \\ |A|=l}} \sum_{a \in A} a - \frac{1}{l} \max_{\substack{A \subset \{A_i\} \\ |A|=l}} \sum_{a \in A} a, \tag{5}$$

where $l = \max\left(1, \lfloor \frac{T}{s} \rfloor\right)$ and s is a hyperparameter to control the selected snippets. This normalization loss aims to maximize the difference between the average top-l attention values and the average bottom-l attention values, and force the foreground attention to be 1 and background attention to be 0.

Therefore, the overall loss for the base model training is the weighted sum of the classification loss and the attention normalization term:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{att}}, \tag{6}$$

where α is a hyperparameter to control the weight of the normalization loss.

In addition, the temporal-class activation map (T-CAM) $\{\mathbf{s}_i\}_{i=1}^{T}$, $\mathbf{s}_i \in \mathbb{R}^C$ is also generated by sliding the classification FC softmax layer over all snippets:

$$s_{i,c} = \frac{e^{\mathbf{w}_c \cdot \mathbf{x}_i + b_c}}{\sum_{j=1}^{C} e^{\mathbf{w}_j \cdot \mathbf{x}_i + b_j}}, \tag{7}$$

where $s_{i,c}$ is the T-CAM value of i-th snippet for category c.

3.4 Pseudo Ground Truth Generation

We iteratively refine the two-stream base models with a frame-level pseudo ground truth. Specifically, we divide the whole training process into several refinement iterations. At refinement iteration 0, only video-level labels are used for training. And at refinement iteration $n+1$, a frame-level pseudo ground truth is generated at refinement iteration n, and provides frame-level supervision for the current refinement iteration. However, without *true* ground truth annotations, we can neither measure the quality of the pseudo ground truth, nor guarantee the pseudo ground truth can help the base models achieve higher performance.

Inspired by two-stream late fusion, we introduce a simple yet effective method to generate the pseudo ground truth. Intuitively, locations at which both streams have high activations are likely to contain ground truth action instances; locations at which only one stream has high activations are likely to be either false positive action proposals or true action instances that only one stream can detect; locations at which both streams both have low activations are likely to be the background.

Following this intuition, we use the fusion attention sequence $\{A^{(n)}_{\text{fuse},i}\}_{i=1}^{T}$ at refinement iteration n to generate pseudo ground truth $\{\mathcal{G}_i^{(n+1)}\}_{i=1}^{T}$ for refinement iteration $n+1$, where $A^{(n)}_{\text{fuse},i} = \beta A^{(n)}_{\text{RGB},i} + (1-\beta) A^{(n)}_{\text{flow},i}$, and $\beta \in [0,1]$ is a

hyperparameter to control the relative importance of RGB and flow attentions. We introduce two pseudo ground truth generation methods.

Soft pseudo ground truth means to directly use the fusion attention values as pseudo labels: $\mathcal{G}_i^{(n+1)} = A_{\text{fuse},i}^{(n)}$. The soft pseudo labels contain the probability of a snippet being the foreground action, but also add uncertainty to the model.

Hard pseudo ground truth thresholds the attention sequence to generate a binary sequence:

$$\mathcal{G}_i^{(n+1)} = \begin{cases} 1, & A_{\text{fuse},i}^{(n)} > \theta; \\ 0, & A_{\text{fuse},i}^{(n)} \leq \theta, \end{cases} \tag{8}$$

where θ is the threshold value. Setting a large value of θ will eliminate the action proposals that only one stream has high activations, and therefore reduces the false positive rate. In contrast, setting a small value of θ will help models to generate more action proposals and achieve a higher recall. Hard pseudo labels remove the uncertainty and provide stronger supervision, but introduce a hyperparameter.

After generating the frame-level pseudo ground truth, we force the attention sequence generated by *each* stream to be similar to the pseudo ground truth with a mean square error (MSE) loss[6]:

$$\mathcal{L}_{\mathcal{G}}^{(n+1)} = \frac{1}{T} \sum_{i=1}^{T} \left(A_i^{(n+1)} - \mathcal{G}_i^{(n+1)} \right)^2. \tag{9}$$

At refinement iteration $n+1$, the total loss for each stream is

$$\mathcal{L}_{\text{total}}^{(n+1)} = \mathcal{L}_{\text{base}} + \gamma \mathcal{L}_{\mathcal{G}}^{(n+1)}, \tag{10}$$

where γ is a hyperparameter to control the relative importance of two losses.

3.5 Action Localization

During testing, following BaS-Net [20], we first temporally upsample the attention sequence and T-CAM by a factor of 8 via linear interpolation. Then, we select top-k action categories from the fusion video-level prediction $\hat{\mathbf{y}}_{\text{fuse}}$ to perform action localization, where $\hat{\mathbf{y}}_{\text{fuse}} = \beta \hat{\mathbf{y}}_{\text{RGB}} + (1-\beta)\hat{\mathbf{y}}_{\text{flow}}$. For each of these categories, following our intention that the attention performs a binary selection, we generate action proposals by directly thresholding the attention value at 0.5 and concatenating consecutive snippets. The action proposals are scored via a variant of the Outer-Inner-Constrastive score [34]: instead of using average T-CAM, we use attention weighted T-CAM to measure the outer and inner temporal contrast. Formally, given action proposal (t_s, t_e, c), fusion attention

[6] Although it is straightforward to use a cross entropy loss for hard pseudo ground truth, we found in practice that the cross entropy loss and the MSE loss achieve similar performance. To simplify training, we use the MSE loss for both kinds of pseudo ground truth.

$\{A_{\text{fuse},i}\}_{i=1}^{T}$ and T-CAM $\{\mathbf{s}_{\text{fuse},i}\}_{i=1}^{T}$, where $\mathbf{s}_{\text{fuse},i} = \beta\mathbf{s}_{\text{RGB},i} + (1-\beta)\mathbf{s}_{\text{flow},i}$, the score ψ is computed as

$$\psi = \frac{\sum_{i=t_s}^{t_e} A_{\text{fuse},i} s_{\text{fuse},i,c}}{t_e - t_s} - \frac{\sum_{i=T_s}^{T_e} A_{\text{fuse},i} s_{\text{fuse},i,c} - \sum_{i=t_s}^{t_e} A_{\text{fuse},i} s_{\text{fuse},i,c}}{T_e - T_s - (t_e - t_s)}, \quad (11)$$

where $T_s = t_s - \frac{L}{4}$, $T_e = t_e + \frac{L}{4}$, and $L = t_e - t_s$. We discard action proposals with confidence scores lower than 0.

4 Experiments

4.1 Dataset and Evaluation

THUMOS14 dataset [15] contains 200 validation videos and 213 testing videos within 20 categories for the TAL task. We use the 200 validation videos for training, and use the 213 testing videos for evaluation.

ActivityNet dataset [2] has two release versions, *i.e.*, ActivityNet v1.3 and ActivityNet v1.2. ActivityNet v1.3 covers 200 action categories, with a training set of 10,024 videos and a validation set of 4,926 videos. ActivityNet v1.2 is a subset of ActivityNet v1.3, and covers 100 action categories, with 4,819 and 2,383 videos in the training and validation set, respectively.[7] We use the training set and the validation set for training and testing, respectively.

Evaluation Metrics. Following the standard protocol on temporal action localization, we evaluate our method with mean Average Precision (mAP) under different Intersection-over-Union (IoU) thresholds. We use the evaluation code provided by ActivityNet[8] to perform the experiments.

4.2 Implementation Details

Two off-the-shelf feature extraction backbones are used in our experiments, *i.e.*, UntrimmedNet [41] and I3D [3], with snippet lengths of 15 frames and 16 frames, respectively. The two backbones are pre-trained on ImageNet [9] and Kinetics [3], respectively, and are not fine-tuned for fair comparison. The RGB and flow snippet-level features are extracted at the `global_pool` layer as 1024-D vectors.

The networks are implemented in PyTorch [29]. We use the Adam [16] optimizer with a fixed learning rate 0.0001. We train the base models 200 and 80 epochs at refinement iteration 0, and 100 and 40 epochs for later refinement iterations for ActivityNet and THUMOS14, respectively. We set the maximal number of refinement iterations to 4 for the THUMOS14 dataset, and 24 for the ActivityNet datasets, and choose base models that achieve the lowest loss at the

[7] In our experiments, there are 9,937 and 4,575 videos in training and validation set of ActivityNet v1.3, respectively, and 4,471 and 2,211 videos in training and validation set of ActivityNet v1.2, respectively, because the rest of the videos are unaccessible from YouTube.

[8] https://github.com/activitynet/ActivityNet/tree/master/Evaluation

Table 1: Comparison of our method with state-of-the-art TAL methods on the THUMOS14 testing set. UNT and I3D are abbreviations for UntrimmedNet feature and I3D feature, respectively

	Method	mAP@IoU (%)								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Fully-supervised	Yuan *et al.* [44]	51.0	45.2	36.5	27.8	17.8	-	-	-	-
	CDC [33]	-	-	40.1	29.4	23.3	13.1	7.9	-	-
	R-C3D [42]	54.5	51.5	44.8	35.6	28.9	-	-	-	-
	SSN [48]	66.0	59.4	51.9	41.0	29.8	-	-	-	-
	BSN [22]	-	-	53.5	45.0	36.9	28.4	20.0	-	-
	TAL-Net [4]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	-	-
	GTAN [25]	69.1	63.7	57.8	47.2	38.8	-	-	-	-
	BMN [21]	-	-	56.0	47.4	38.8	29.7	20.5	-	-
Weakly-supervised	UntrimmedNet [41]	44.4	37.7	28.2	21.1	13.7	-	-	-	-
	STPN (UNT) [27]	45.3	38.8	31.1	23.5	16.2	9.8	5.1	2.0	0.3
	AutoLoc (UNT) [34]	-	-	35.8	29.0	21.2	13.4	5.8	-	-
	W-TALC (UNT) [30]	49.0	42.8	32.0	26.0	18.8	-	6.2	-	-
	Liu *et al.* (UNT) [23]	53.5	46.8	37.5	29.1	19.9	12.3	6.0	-	-
	RefineLoc (UNT) [1]	-	-	36.1	-	22.6	-	5.8	-	-
	CleanNet (UNT) [24]	-	-	37.0	30.9	23.9	13.9	7.1	-	-
	BaS-Net (UNT) [20]	56.2	50.3	42.8	34.7	25.1	17.1	9.3	3.7	**0.5**
	Ours (UNT)	**58.9**	**52.9**	**45.0**	**36.6**	**27.6**	**18.8**	**10.2**	**4.0**	**0.5**
	STPN (I3D) [27]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
	W-TALC (I3D) [30]	55.2	49.6	40.1	31.1	22.8	-	7.6	-	-
	Liu *et al.* (I3D) [23]	57.4	50.8	41.2	32.1	23.1	15.0	7.0	-	-
	RefineLoc (I3D) [1]	-	-	40.8	-	23.1	-	5.3	-	-
	Nguyen *et al.* (I3D) [28]	60.4	56.0	46.6	37.5	26.8	17.6	9.0	3.3	0.4
	BaS-Net (I3D) [20]	58.2	52.3	44.6	36.0	27.0	18.6	**10.4**	**3.9**	0.5
	Ours (I3D)	**63.4**	**57.6**	**47.8**	**37.7**	**28.7**	**19.4**	10.2	**3.9**	**0.7**

previous refinement iteration to generate the pseudo ground truth. To eliminate fragmentary action proposals, temporal max pooling of kernel size 5 and stride 1 is used on the fusion attention sequence before pseudo ground truth generation on ActivityNet dataset. We use a whole video as a batch. All hyperparameters are determined via grid search: $s = 8$, $\alpha = 0.1$, $\beta = 0.4$, $\gamma = 2$. We set θ to 0.55 and 0.5 for THUMOS14 and ActivityNet, respectively. We choose top-2 action categories and also reject categories whose fusion classification prediction scores are lower than 0.1 to perform action localization.

4.3 Comparison with the State-of-the-art

Experiments on THUMOS14. Table 1 summarizes the performance comparison between the proposed TSCN and state-of-the-art fully-supervised and weakly-supervised TAL methods on the THUMOS14 testing set. With UntrimmedNet features, TSCN outperforms other W-TAL methods by a large margin, and even

Table 2: Comparison of our method with state-of-the-art W-TAL methods on the ActivityNet v1.2 validation set. The Avg column indicates the average mAP at IoU thresholds 0.5:0.05:0.95

Method	mAP@IoU (%)			Avg
	0.5	0.75	0.95	
UntrimmedNet [41]	7.4	3.2	0.7	3.6
AutoLoc [34]	27.3	15.1	3.3	16.0
W-TALC [30]	37.0	-	-	18.0
Liu *et al.* [23]	36.8	22.0	5.6	22.4
Ours	**37.6**	**23.7**	**5.7**	**23.6**

Table 3: Comparison of our method with state-of-the-art W-TAL methods on the ActivityNet v1.3 validation set. The Avg column indicates the average mAP at IoU thresholds 0.5:0.05:0.95

Method	mAP@IoU (%)			Avg
	0.5	0.75	0.95	
STPN [27]	29.3	16.9	2.7	-
Liu *et al.* [23]	34.0	20.9	**5.7**	21.2
Nguyen *et al.* [28]	**36.4**	19.2	2.9	-
Ours	35.3	**21.4**	5.3	**21.7**

achieves comparable results to some recent W-TAL methods with I3D features (*e.g.*, Nguyen *et al.* [28] and BaS-Net [20]) at high IoU thresholds.

With I3D features, our performance boosts significantly, and outperforms previous W-TAL methods at most IoU thresholds. We note the proposed TSCN can achieve a comparable performance to some recent fully-supervised methods (*e.g.*, R-C3D [42]). TSCN even outperforms TAL-net [4] at IoU thresholds 0.1 and 0.2. However, as the IoU threshold increases, the performance of TSCN drops significantly, because generating more precise action boundaries need true frame-level ground truth supervision.

Experiments on ActivityNet. The performance comparisons on ActivityNet v1.2 and v1.3 are shown in Table 2 and Table 3, respectively, where our models are trained with I3D features. The proposed TSCN outperforms previous W-TAL methods at the average mAP at IoU threshold $0.5 : 0.05 : 0.95$ on both release versions of ActivityNet, verifying the efficacy of our design intuition.

4.4 Ablation Study

In this subsection, a set of ablation studies is conducted on the THUMOS14 testing set with UntrimmedNet feature to analyze the efficacy of each component in the proposed TSCN.

Ablation study on $\mathcal{L}_{\mathbf{att}}$. The goal of $\mathcal{L}_{\text{att}}$ in Equation (5) is to force the attention values to approach extreme values, and therefore generate a clean foreground feature $\mathbf{x}_{\text{fg}}$ and improve action proposal quality. Some recent methods [28, 20] introduce background classification to W-TAL. Particularly, background classification loss $\mathcal{L}_{bg}$ [28] is introduced to classify the background, where a background attention is defined as $1 - A_i$, and a background feature is generated via background attention-weighted pooling over all snippets to perform the background classification. Therefore, $\mathcal{L}_{bg}$ is in essence an implicit attention normalization loss. However, one drawback of such background loss is that assigning background labels to all videos will make the value of the background category in the T-CAM increase. We reproduce $\mathcal{L}_{bg}$ in our model, compare it with our proposed $\mathcal{L}_{\text{att}}$, and

Table 4: Comparison of our method with different attention normalization functions on the THUMOS14 testing set. $\mathcal{L}_{bg}$ is the background classification loss introduced in [28], and $\mathcal{L}_{att}$ is defined in Equation (5). The var column denotes the average attention variance over the whole testing set

$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{bg}$	$\mathcal{L}_{att}$	mAP@IoU (%) 0.3	0.5	0.7	Var
✓	-	-	29.6	16.1	4.1	0.0440
✓	✓	-	34.3	19.3	6.7	0.0599
✓	-	✓	**40.9**	**24.0**	**8.2**	**0.0937**
✓	✓	✓	40.6	23.6	7.8	0.0886

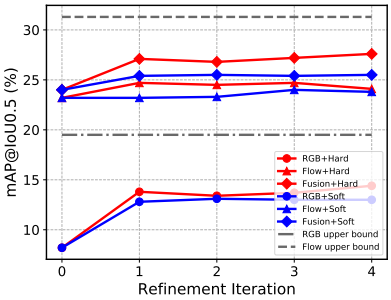

Fig. 3: Comparison between models trained with different pseudo ground truth on the THUMOS14 testing set. The upper bounds denote models trained with ground truth actionness sequence

list the results in Table 4. The results reveal that both $\mathcal{L}_{bg}$ and $\mathcal{L}_{\text{att}}$ help improve the performance. And the proposed $\mathcal{L}_{\text{att}}$ achieves higher attention variance and better localization performance than $\mathcal{L}_{bg}$, demonstrating that the our attention normalization term $\mathcal{L}_{\text{att}}$ can better avoid the ambiguity of attention. Surprisingly, with both $\mathcal{L}_{bg}$ and $\mathcal{L}_{\text{att}}$, the localization performance is still lower than that with only $\mathcal{L}_{\text{att}}$, and we think this is because the noise of background classification reduces the accuracy of action proposal scores.

Ablation study on Pseudo Ground Truth. Fig. 3 plots performance comparison between different pseudo ground truth methods at different refinement iterations. Both soft and hard pseudo ground truth help improve the localization performance. The hard pseudo ground truth removes uncertainty to the model, and thus achieves higher performance improvement. However, with the same frame-level supervision, the flow stream outperforms the RGB stream by a large margin. We think this is because of the nature of two modalities: the RGB modality is less sensitive to actions than the optical flow modality. To demonstrate this, we generate a *true* frame-level ground truth actionness sequence (action categories are not used), train our model in the same way as the pseudo ground truth. The results are plotted in Fig. 3 as an upper bound. The results verify our hypothesis and demonstrate that the optical flow modality is more suitable for the action localization task than the RGB modality.

Table 5 lists the detailed performance comparison between the model trained with only video-level labels and that trained with the hard pseudo ground truth. The results show that pseudo ground truth improves the localization performance for both modalities at all IoU thresholds, and thus improves the performance of the fusion result. Also, the pseudo ground truth greatly improves the precision and recall for the RGB stream and the fusion result, and improves the precision for the flow stream with a minor loss of recall (the overall F-measure improves significantly), which demonstrates that the pseudo ground truth can help eliminate false positive action proposals.

Table 5: Comparison between the model trained with only video-level labels and the model trained with hard pseudo ground truth on the THUMOS14 testing set. The label column denotes the supervision used in training, where "video" indicates only video-level labels are leveraged, and "frame" indicates the hard pseudo ground truth is also leveraged during training. Precision, recall and F-measure are calculated under IoU threshold 0.5

Modality	Label	mAP@IoU (%)					Precision (%)	Recall (%)	F-measure
		0.3	0.4	0.5	0.6	0.7			
RGB	video	19.8	13.2	8.2	4.5	1.9	10.2	20.9	0.1371
RGB	frame	31.4	22.1	14.4	8.9	5.2	20.9	30.8	0.2489
Flow	video	40.2	32.0	23.2	15.4	7.2	25.5	43.3	0.3207
Flow	frame	40.8	32.7	24.1	16.8	8.7	30.9	42.4	0.3573
Fusion	video	40.9	32.4	24.0	15.9	8.2	23.6	44.4	0.3078
Fusion	frame	45.0	36.5	27.6	18.8	10.2	31.3	44.6	0.3680

Qualitative Analysis. Three representative examples of TAL results are plotted in Fig. 4 to illustrate the efficacy of the proposed pseudo supervision. In the first example of diving and cliff diving, with only video-level labels, the RGB stream provides worse localization result than the flow stream, and thus leads to a noisy fusion attention sequence. The pseudo ground truth guides the RGB stream to identify false positive action proposals and discover true action instances, and further leads to a cleaner fusion attention sequence, where high activations correspond better to the ground truth. In the second example of cricket shot, with only video-level supervision, the RGB stream can only distinguish certain scenes, and fails to separate proximate action instances. In contrast, the flow stream can precisely detect action instances. Therefore, the pseudo ground truth helps the RGB stream to separate consecutive action instances. In the last example of soccer penalty, both streams have high activations on certain false positive temporal locations. Under this circumstance, the false positive action proposals will have higher activations under frame-level pseudo supervision. To eliminate such false positive action proposals, however, need true ground truth supervision. To summarize, the two modalities have their own strengths and limitations: the RGB stream is sensitive to appearance, thus it fails in scenes shot from unusual angles or separating proximate action instances; the flow stream is sensitive to motion, and provides more accurate results, but it fails in slow or occluded motion. Qualitative results reveal that the pseudo ground truth helps two streams reach a consensus at most temporal locations. Therefore, the fusion attention sequence becomes cleaner and helps generate more precise action proposals and more reliable confidence scores.

5 Conclusions

In this paper, we propose a Two-Stream Consensus Network (TSCN) for W-TAL, which benefits from an iterative refinement training method and a new attention

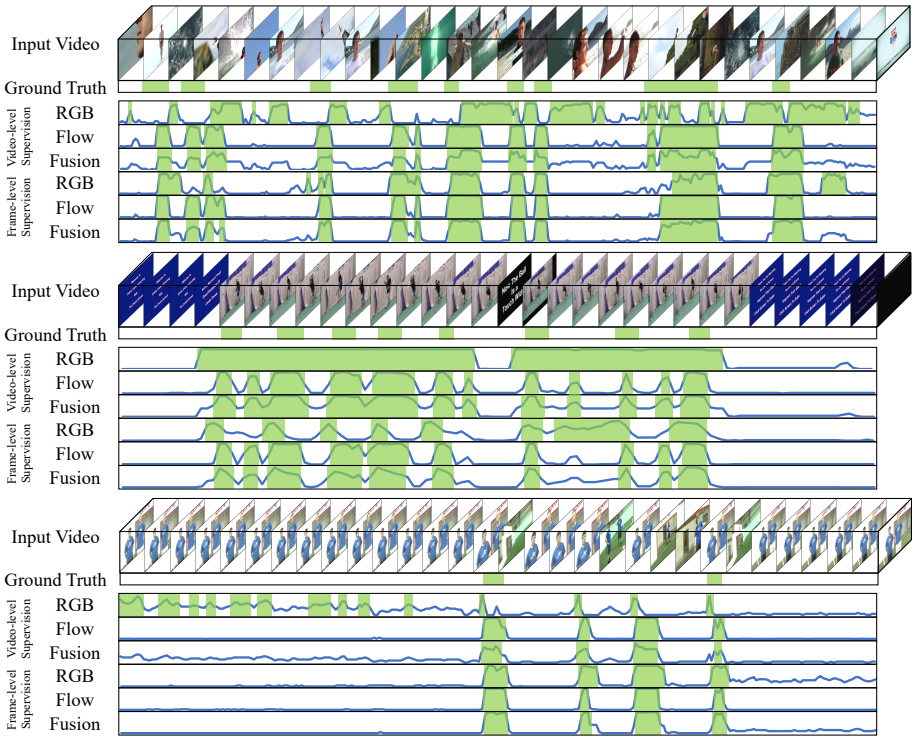

Fig. 4: Qualitative results on the THUMOS14 testing set. The eight rows in each example are input video, ground truth action instance, RGB stream, flow stream, and fusion attention sequences from the model trained with only video-level labels and frame-level pseudo ground truth, respectively. Action proposals are represented by green boxes. The horizontal and vertical axes are time and intensity of attention, respectively

normalization loss. The iterative refinement training uses a novel frame-level pseudo ground truth as fine-grained supervision, and iteratively improves the two-stream base models. The attention normalization loss function reduces the ambiguity of attention values, and thus leads to more precise action proposals. Experiments on two benchmarks demonstrate the proposed TSCN outperforms current state-of-the-art methods, and verify our design intuition.

6 Acknowledgement

This work was supported partly by National Key R&D Program of China Grant 2018AAA0101400, NSFC Grants 61629301, 61773312, and 61976171, China Postdoctoral Science Foundation Grant 2019M653642, Young Elite Scientists Sponsorship Program by CAST Grant 2018QNRC001, and Natural Science Foundation of Shaanxi Grant 2020JQ-069.

References

1. Alwassel, H., Pardo, A., Heilbron, F.C., Thabet, A., Ghanem, B.: Refineloc: Iterative refinement for weakly-supervised action localization. arXiv preprint arXiv:1904.00227 (2019)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1130–1139 (2018)
5. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7882–7891 (2019)
6. Dai, X., Singh, B., Zhang, G., Davis, L.S., Qiu Chen, Y.: Temporal context network for activity localization in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5793–5802 (2017)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 886–893 (2005)
8. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision. pp. 428–441 (2006)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 248–255 (2009)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1933–1941 (2016)
11. Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3628–3636 (2017)
12. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. arXiv preprint arXiv:1705.01180 (2017)
13. Gorban, A., Idrees, H., Jiang, Y.G., Zamir, A.R., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes (2015)
14. Heilbron, F.C., Barrios, W., Escorcia, V., Ghanem, B.: Scc: Semantic context cascade for efficient action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3175–3184 (2017)
15. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes (2014)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

18. Kumar Singh, K., Jae Lee, Y.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3524–3533 (2017)
19. Laptev, I.: On space-time interest points. International Journal of Computer Vision pp. 107–123 (2005)
20. Lee, P., Uh, Y., Byun, H.: Background suppression network for weakly-supervised temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
21. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3889–3898 (2019)
22. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018)
23. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019)
24. Liu, Z., Wang, L., Zhang, Q., Gao, Z., Niu, Z., Zheng, N., Hua, G.: Weakly supervised temporal action localization through contrast based evaluation networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3899–3908 (2019)
25. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 344–353 (2019)
26. Narayan, S., Cholakkal, H., Khan, F.S., Shao, L.: 3c-net: Category count and center loss for weakly-supervised action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8679–8687 (2019)
27. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6752–6761 (2018)
28. Nguyen, P.X., Ramanan, D., Fowlkes, C.C.: Weakly-supervised action localization with background modeling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5502–5511 (2019)
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems, pp. 8024–8035 (2019)
30. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European Conference on Computer Vision. pp. 563–579 (2018)
31. Piergiovanni, A., Ryoo, M.S.: Representation flow for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of Neural Information Processing Systems. pp. 91–99 (2015)
33. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5734–5743 (2017)

34. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision. pp. 154–171 (2018)
35. Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.F., Yan, Z.: Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1268–1277 (2019)
36. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Proceedings of the European Conference on Computer Vision. pp. 510–526 (2016)
37. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of Neural Information Processing Systems. pp. 568–576 (2014)
38. Tan, M., Shi, Q., van den Hengel, A., Shen, C., Gao, J., Hu, F., Zhang, Z.: Learning graph structure for multi-label image classification via clique generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4100–4109 (2015)
39. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3169–3176 (2011)
40. Wang, L., Koniusz, P., Huynh, D.Q.: Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8698–8708 (2019)
41. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
42. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5783–5792 (2017)
43. Yu, T., Ren, Z., Li, Y., Yan, E., Xu, N., Yuan, J.: Temporal structure mining for weakly supervised action detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5522–5531 (2019)
44. Yuan, Z., Stroud, J.C., Lu, T., Deng, J.: Temporal action localization by structured maximal sums. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3684–3692 (2017)
45. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7094–7103 (2019)
46. Zhai, Y., Wang, L., Liu, Z., Zhang, Q., Hua, G., Zheng, N.: Action coherence network for weakly supervised temporal action localization. In: Proceedings of the IEEE International Conference on Image Processing. pp. 3696–3700 (2019)
47. Zhao, Y., Xiong, Y., Lin, D.: Recognize actions by disentangling components of dynamics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6566–6575 (2018)
48. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923 (2017)
49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)