Action Coherence Network for Weakly-Supervised Temporal Action Localization

Yuanhao Zhai, *Member, IEEE*, Le Wang, *Senior Member, IEEE*, Wei Tang, *Member, IEEE*, Qilin Zhang, *Member, IEEE*, Nanning Zheng, *Fellow, IEEE*, and Gang Hua, *Fellow, IEEE*

***Abstract*—Weakly-supervised Temporal Action Localization (W-TAL) aims at simultaneously classifying and locating all action instances with only video-level supervision. However, current W-TAL methods have two limitations. First, they ignore the difference in video representations between an action instance and its surrounding background when generating and scoring action proposals. Second, the unique characteristics of the RGB frames and optical flow are largely ignored when fusing these two modalities. To address these problems, an Action Coherence Network (ACN) is proposed in this paper. Its core is a new coherence loss which exploits both classification predictions and video content representations to supervise action boundary regression and thus leads to more accurate action localization results. Besides, the proposed ACN explicitly takes into account the specific characteristics of RGB frames and optical flow by training two separate sub-networks, each of which is able to generate modality-specific action proposals independently. Finally, to take advantage of the complementary action proposals generated by two streams, a novel fusion module is introduced to reconcile them and obtain the final action localization results. Experiments on the THUMOS14 and ActivityNet datasets show that our ACN outperforms the state-of-the-art W-TAL methods, and is even comparable to some recent fully-supervised methods. Particularly, ACN achieves a mean average precision of 26.4% on the THUMOS14 dataset under the IoU threshold 0.5.**

***Index Terms*—Temporal action localization, weakly-supervised learning.**

I. Introduction

TEMPORAL Action Localization (TAL) aims at classifying and locating all action instances in an untrimmed video. It can be applied to many high-level video understanding tasks such as event detection [1]–[3] and video summarization [4], [5]. While fully-supervised TAL methods [6]–[23] have achieved promising performance, they rely on precise annotations of categorical label and the start and end temporal locations of all action instances. This kind of labeling is expensive and time-consuming for large-scale datasets, and could be inconsistent due to ambiguous action transitions [24], [25]. This paper considers a more cost-effective setting: Weakly-supervised Temporal Action Localization (W-TAL), which only requires video-level categorical labels to perform training. These video-level labels are much cheaper to annotate, and could be automatically obtained with textual search terms on video sharing websites. Thus, W-TAL is more practical than its fully-supervised counterpart.

Recent years have witnessed a significant performance improvement on W-TAL [26]–[35]. There are two mainstream methods, *i.e.*, thresholding-based methods and two-stage methods. Thresholding-based methods [29]–[33] first extract RGB and optical flow features via pre-trained models, and then jointly train a classification module that generates Snippet-level Classification Predictions (SCPs) and an attention module that produces Snippet-level Attention Predictions (SAPs). The final TAL results are obtained by directly thresholding the SCPs and SAPs. Two-stage methods [34], [35] first initialize action proposals as anchors of predefined lengths at all temporal locations, and then regress the durations and center locations of these action proposals via additional regression networks.

Despite these recent efforts, two major challenges still persist. On one hand, most methods only exploit the classification scores and attention weights to generate and score action proposals. They do not have an explicit mechanism to model the video content changes between video frames, *e.g.,* caused by action starts or ends, which is critical for accurate action localization. As a result, the thresholding-based methods often generate fragmentary action proposals inside an action instance whose SCPs are fluctuating. Two-stage methods also suffer from this problem because fluctuating SCPs inevitably decrease the confidence score of an action proposal.

On the other hand, how to fuse the predictions from two modalities (*i.e.*, RGB and optical flow) for W-TAL has not been well explored. In action recognition, two-stream Convolutional Neural Networks (CNNs) [36] achieve a significant performance boost by fusing the two-stream classification scores via a weighted summation. This indicates that the two modalities are

Manuscript received March 30, 2020; revised January 15, 2021 and April 4, 2021; accepted April 9, 2021. Date of publication April 14, 2021; date of current version April 6, 2022. This work was supported in part by National Key R and D Program of China under Grant 2018AAA0101400, in part by NSFC under Grants 62088102, 61773312, and 61976171, in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001, and in part by the Natural Science Foundation of Shaanxi under Grant 2020JQ-069. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Concetto Spampinato. *(Corresponding author: Le Wang.)*

Yuanhao Zhai, Le Wang, and Nanning Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: yuanhaozhai@gmail.com; lewang@mail.xjtu.edu.cn; nnzheng@mail.xjtu.edu.cn).

Wei Tang is with the Department of Computer Science, University of Illinois, Chicago, IL 60607 USA (e-mail: tangw@uic.edu).

Qilin Zhang is with ABB Corporate Research Center, Raleigh, NC 27606 USA (e-mail: samqzhang@gmail.com).

Gang Hua is with the Wormpex AI Research, Bellevue, WA 98004 USA (e-mail: ganghua@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3073235.

Digital Object Identifier 10.1109/TMM.2021.3073235

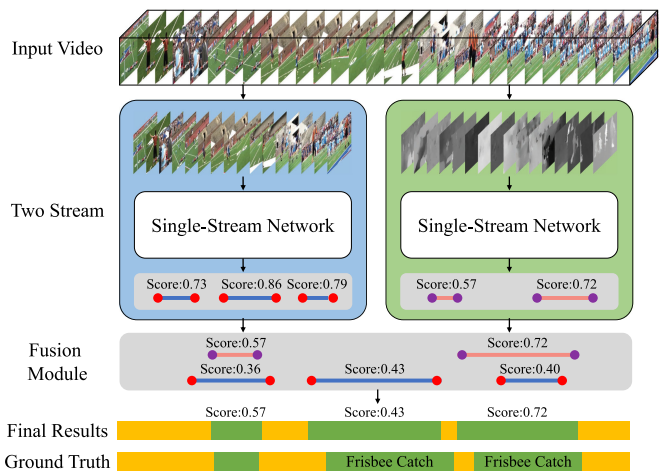


Fig. 1. An overview of the inference process of the proposed ACN. Given an untrimmed video, modality-specific action proposals are generated by two sub-networks, *i.e.*, the RGB stream and the optical flow stream, which are reconciled by a novel fusion module to get the final action localization results.

complementary. However, the unique characteristics of the RGB and optical flow features are usually ignored in the conventional fusion methods. Intuitively, the RGB stream is sensitive to scene transition and large motion displacement but tends to ignore small movements. The flow stream is appearance-invariant [37], and much more sensitive to small movements but may introduce noise during scene transition or camera movement. Therefore, it is desirable to develop an effective way to integrate these two modalities.

This paper presents an Action Coherence Network (ACN) to tackle the aforementioned challenges. For the first problem, inspired by the Outer-Inner-Contrastive (OIC) Loss [34], the authors speculate that a good action proposal should have two qualities: (1) the SCPs within it should be significantly higher than those in the surrounding background, and (2) the video representations within it should differ with those outside the action instance. Therefore, a new coherence loss is introduced to account for both qualities. To address the second challenge, two separate sub-networks are trained by taking RGB frames and optical flow as their respective input to exploit the modality-specific features. The two sub-networks share similar settings and both are trained with the coherence loss. As a result, they are able to generate complementary action proposals by taking advantage of the unique characteristics of each modality.

Fig. 1 presents the overview of ACN, which consists of two sub-networks and a fusion module. Given an input video, features of two modalities are first extracted with pre-trained 2D/3D backbones. Then, action recognition is performed on two modalities respectively to obtain the SCPs and SAPs. After that, initial action proposals with different predefined anchor lengths are generated at all temporal locations. For each stream, a group of regression networks is trained using the coherence loss to regress boundaries of each action proposal to more precise temporal locations. Each stream is able to generate modality-specific action proposals independently. Non-Maximum Suppression (NMS) is then performed to remove duplicated action proposals, and a fusion module is finally proposed to select and combine the two-stream outputs and generate the final TAL results.

In a nutshell, the main contributions are as follows:

- A new coherence loss is proposed to model both SCPs and video representations on action boundaries. It can significantly improve the performance of action localization.
- A novel Action Coherence Network (ACN) is proposed for W-TAL. It has two separate sub-networks taking as input RGB frames and optical flow, respectively. Furthermore, a new fusion module is designed to reconcile the action proposals from the two streams and generate the final TAL results.
- Experiments on two challenging datasets (*i.e.*, THUMOS14 and ActivityNet) demonstrate that the proposed method outperforms state-of-the-art methods. Extensive ablation studies are conducted to validate the contribution of each component.

A short conference version of this paper appeared in [38]. This paper extends the previous version in four aspects. (1) This paper provides more implementation details of ACN. (2) The overfitting problem in the regression network training is addressed. (3) More experiments are conducted to analyze the effectiveness of each component in ACN. (4) The strengths and limitations of our proposed method and future work are discussed.

This paper is organized as follows. Section II briefly reviews the related work. Section III presents the framework of the proposed ACN. The experiments are presented in Section IV. Finally, Section V presents the conclusion.

II. Related Work

Related work on action recognition, fully-supervised temporal action localization, and weakly-supervised temporal action localization are briefly reviewed in this section.

A. Action Recognition

Action recognition has been extensively studied in the past. Traditional methods [39]–[41] extract hand-crafted representations to model spatio-temporal information. Recently, deep learning-based methods show great performance improvement. Among them, there are two mainstream methods: two-stream networks [36], [42]–[44] exploit appearance and motion information from RGB and optical flow respectively; 3D CNNs [8], [45] learn spatio-temporal clues directly from consecutive video frames. Two-Stream Inflated 3D ConvNet (I3D) [46] replaces the 2D CNNs in two-stream networks with 3D CNNs to model the temporal information. Besides, several works [44], [47]–[49] try to model long-term temporal information in action recognition. Several efforts [50], [51] are made to reduce the computational cost in action recognition. There are also several attempts [52]–[55] focusing on directly learning motion clues from RGB frames instead of calculating hand-crafted optical flow.

B. Fully-Supervised Temporal Action Localization

The task of temporal action localization is to temporal localize and classify all action instances in an untrimmed video, and the fully-supervised type requires frame-level annotations of all action instances during training. Some methods [6], [56], [57] exploit a sliding window or a predefined temporal duration to generate action proposals. Inspired by deep learning-based object detection methods, such as Regions with CNN features (R-CNN) [58], many methods [7]–[11], [16], [17], [59] implement a "propose and classify" scheme, where action proposals are first generated and then classified. Some methods [9], [11], [12], [16] apply the Faster R-CNN [60] framework to TAL. Some recent methods [17], [20], [21] focus on generating action proposals with more flexible durations. Zeng *et al.* [23] introduce graph convolutional networks to exploit the relations among action proposals.

Despite the success of fully-supervised TAL, its dependence on temporal annotation, which is expensive and time-consuming, greatly impede its application in real-world scenarios.

C. Weakly-Supervised Temporal Action Localization

Weakly-supervised Temporal Action Localization (W-TAL) aims to achieve temporal action localization with only video-level action categorical labels available during training. Since the frame-level labels are not available during training, several existing methods [27], [28], [30]–[32], [34], [35], [61] adopt a multiple instance learning framework, where a video is treated as a bag of frames/snippets to perform action classification. The trained model generates a per-snippet/frame classification prediction sequence, which is further used to generate the action proposals by thresholding. UntrimmedNet [27] is one of the first W-TAL method, and it uses an attention module to evaluate the relative importance of every snippet and a classification module to perform the snippet-level classification, and generates localization results by thresholding the attention and classification activation sequences. Sparse Temporal Pooling Network (STPN) [28] improves UntrimmedNet by adding a sparsity loss to enforce the sparsity of the segment selection. Hide-and-Seek [26] randomly hides parts of the input video to guide the network to learn the most relevant parts. Paul *et al.* [30] propose a co-activity similarity loss to enforce the learned features to be similar if they belong to the same action category. Liu *et al.* [31] employ multi-attention branches to learn different stages of an action. 3C-Net [33] proposes a classification loss, a count loss and a center loss to improve feature discriminability and delineate proximate action sequences for W-TAL. DGAM [62] learns an attention value with variational auto-encoder to separate actions and context. There are also several works [32], [61] aiming to learn a richer notion of an action with background classification. TSCN [63] proposes an attention normalization loss to replace the background learning, and proposes a pseudo ground truth learning to remove false positives. Some recent works [64], [65] attempt to separate action and context for better action boundary learning.

Apart from these thresholding-based methods, AutoLoc [34] first adopts a two-stage framework in W-TAL. They first traverse all temporal locations with predefined anchors to generate initial action proposals, and then regress the boundaries of them to more precise temporal locations with an Outer-Inner-Contrastive (OIC) loss, which aims at maximizing the activation difference between the action proposal area and its contextual area. CleanNet [35] improves AutoLoc [34] by leveraging the temporal contrast of Snippet-level Classification Prediction (SCP) to regress the action proposals to more precise temporal locations.

The proposed method draws the inspiration from AutoLoc [34], and differs from AutoLoc in three aspects. (1) The proposed coherence loss not only accounts for the prominence of SCPs, but also detects action instances based on their distinctive video representation. (2) Two sub-networks are trained to learn modality-specific action proposals, which are then reconciled by a fusion module. By contrast, AutoLoc only trains one network with concatenated RGB and Flow features as input. (3) The proposed regression networks are designed to make the receptive field the same as the regression field while maintaining a small number of parameters. As discussed in Section IV-D, all these differences contribute to the superiority of the proposed ACN.

III. Action Coherence Network

This section introduces the proposed Action Coherence Network (ACN), which consists of two sub-networks (*i.e.*, an RGB stream and a flow stream), and a fusion module (see Fig. 1). As shown in Fig. 2, each stream is composed of three parts respectively for action recognition, action proposal regression, and Non-Maximum Suppression. The two streams share similar settings and both are trained with the proposed coherence loss, which can supervise action boundary learning and facilitate action proposal regression. Given different input modalities, the two sub-networks can generate action proposals with different characteristics. And these action proposals are finally reconciled via a fusion module. It has been verified from the experiments that the two-stream outputs are complementary and together contribute to a higher performance.

Section III-A first gives the definition of the Weakly-supervised Temporal Action Localization (W-TAL) problem. As the two streams share similar settings, and they only differ in the input modality (*i.e.*, RGB frames or optical flow), Section III-B–Section III-D detail the single-stream structure. Finally, Section III-E introduces how to effectively reconcile the outputs from two streams.

A. Problem Formulation

The task of W-TAL aims to temporally locate and classify all action instances in an untrimmed video. Specifically, given an untrimmed video, only its video-level categorical label $\mathbf{y} \in \mathbb{R}^C$ is available during training, where $\mathbf{y}$ is a normalized multi-hot vector. The k-th element of $\mathbf{y}$ (*i.e.*, y_k) is set to $1/N$ if the video contains action instances of the k-th category, where N is the number of action categories occurred in the video. During testing, the network is expected to output a set of action

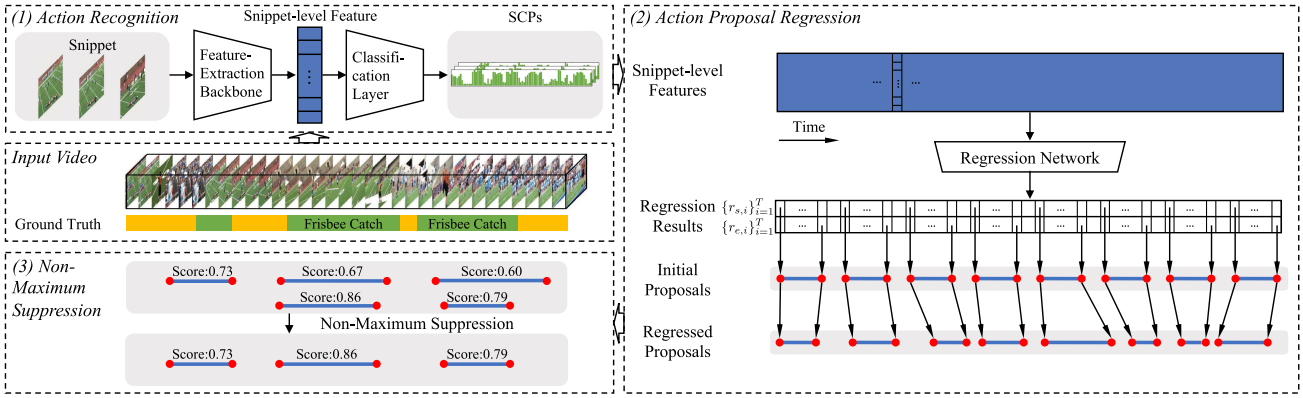


Fig. 2. A single stream overview of the proposed ACN. (1) *Action Recognition*. Pre-trained models are used to extract snippet-level features. The classification model is built upon the snippet-level features, and generates Snippet-level Classification Predictions (SCPs). (2) *Action Proposal Regression*. Initial action proposals with predefined anchor lengths are generated at all temporal locations. The proposals with a certain anchor length are regressed with a corresponding regression network, which is trained with the coherence loss and regresses the boundaries of the proposals to more precise temporal locations. (3) *Non-Maximum Suppression*. Non-Maximum Suppression (NMS) eliminates redundant action proposals.

proposals $\{p_i|p_i = (x_{s,i}, x_{e,i}, c_i, s_i)\}_{i=1}^{M}$ for each testing video, where M is the number of output proposals and p_i is the i-th action proposal, which is a tuple containing a start time $x_{s,i}$, an end time $x_{e,i}$, a category c_i and a confidence score s_i.

B. Action Recognition

The input videos are divided into non-overlapping fixed-length snippets. Then, a pre-trained deep network (*e.g.*, I3D network [46]) is used to extract the RGB frame/optical flow snippet-level features. Specifically, given a video with T snippets, the extracted features are denoted as $\mathbf{F} \in \mathbb{R}^{D\times T}$, where D represents the feature dimension. The extracted features provide a high-level representation of the input video, and are fed to the classification and attention layers of the network.

Following UntrimmedNet [27], the attention weights $\mathbf{a} \in \mathbb{R}^T$ and the classification scores $\mathbf{S} \in \mathbb{R}^{C\times T}$ are obtained by two fully connected (fc) layers with 1 and C output channels, respectively. The Snippet-level Classification Predictions (SCPs) $\hat{\mathbf{S}}$ and action recognition result $\hat{\mathbf{y}} \in \mathbb{R}^C$ are calculated as:

$$\hat{S}(k,t) = \frac{\exp(S(k,t))}{\sum_{i=1}^{C}\exp(S(i,t))}, \tag{1}$$

$$\hat{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\frac{\exp(a(t))}{\sum_{i=1}^{T}\exp(a(i))}\hat{\mathbf{S}}(t), \tag{2}$$

where $S(k,t)$ and $\hat{S}(k,t)$ represent the classification score and SCP at temporal location t and category k, respectively, and $\hat{\mathbf{S}}(t)$ and $a(t)$ represent the SCPs and attention weight at temporal location t, respectively.

The two layers are trained with the cross entropy loss:

$$L_{\text{CE}} = -\sum_{k=1}^{C} y_k \log \hat{y}_k, \tag{3}$$

where $\hat{y}_k$ is the predicted probability of the target video containing action instances in the k-th category. The parameters in the two layers are fixed after the training of action recognition.

C. Action Proposal Regression

This subsection describes the action proposal regression process in a single stream. Without frame-level boundary annotation, it is impossible to directly regress the boundaries of action proposals with an L1-norm distance like in the fully-supervised methods [9], [11], [12], [16]. Instead, this paper introduces a proxy loss–coherence loss for each proposal, and by minimizing the coherence loss, the proposal boundaries are expected to regress to more precise temporal locations.

Action Proposal Initialization: Inspired by Faster R-CNN [60], initial action proposals are generated at all temporal locations with a group of predefined anchor lengths (number of snippets). Formally, given an anchor size P, action proposals are initialized as $\{(x_{s,i}, x_{e,i})\}_{i=1}^{T-P}$, such that $x_{e,i} - x_{s,i} = P$, $x_{s,1} = 0$ and $x_{s,T-P} = T - P - 1$. To account for the contextual information, the inflated start and end boundaries (X_s, X_e) for action proposal (x_s, x_e) are defined as $X_s = x_s - P/4$ and $X_e = x_e + P/4$.

Coherence Loss: An ideal action proposal is expected to have distinctive temporal boundaries, which is modeled via the Outer-Inner-Contrastive (OIC) loss in AutoLoc [34]. Formally, given the SCPs $\hat{\mathbf{S}} \in \mathbb{R}^{C\times T}$ of a video with T snippets and C action categories, the OIC loss for an action proposal (x_s, x_e) of the action category $k \in \{1, \ldots, C\}$ is defined as

$$L_{\text{OIC}} = \frac{\int_{X_s}^{X_e}\hat{S}(k,t)dt - \int_{x_s}^{x_e}\hat{S}(k,t)dt}{(X_e - X_s + 1) - (x_e - x_s + 1)} - \frac{\int_{x_s}^{x_e}\hat{S}(k,t)dt}{x_e - x_s + 1}. \tag{4}$$

The goal of the OIC loss is to regress the start boundary x_s and end boundary x_e of an action proposal so that the temporal regions immediately outside the action proposal have low activation (the first term) while the area within the action proposal has high activation (the second term).

Therefore, the OIC loss only focuses on the snippet-level classification predictions while ignoring the content of the action

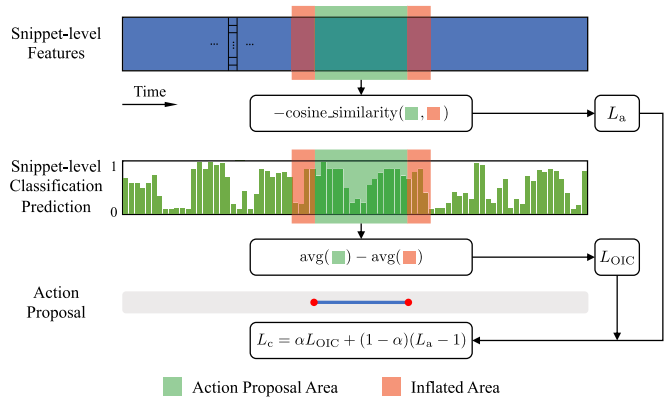


Fig. 3. Illustration of the computation process of the proposed coherence loss L_{c}. It consists of two terms: an appearance term L_{a} that calculates the negative cosine similarity between the action proposal feature and its context feature; an OIC term L_{OIC} that computes the difference between the action proposal classification score and its context classification score.

instance. Intuitively, the video representations of an action instance should also be different from those of snippets immediately preceding and succeeding the action instance. For example, an action may contain a preparation stage, a performing stage and an ending stage, and they should differ not only in the SCPs but also in their feature representations. Following this intuition, the appearance term in the coherence loss is formulated as an arithmetic average of the cosine similarities between the action area and its "start area" and "end area". Formally, given an action proposal (x_s, x_e) and its inflated boundary (X_s, X_e), the features of its start area $\mathbf{R}_s$, the features of its end area $\mathbf{R}_e$ and the features of its action area $\mathbf{R}_a$ are defined respectively as

$$\mathbf{R}_s = \frac{\int_{X_s}^{x_s} \mathbf{F}(t)dt}{x_s - X_s + 1}, \tag{5}$$

$$\mathbf{R}_e = \frac{\int_{x_e}^{X_e} \mathbf{F}(t)dt}{X_e - x_e + 1}, \tag{6}$$

$$\mathbf{R}_a = \frac{\int_{x_s}^{x_e} \mathbf{F}(t)dt}{x_e - x_s + 1}, \tag{7}$$

where $\mathbf{F}(t)$ denotes the feature representation of the t-th snippet, and the feature representation of a temporal area is obtained via average pooling. Here, an appearance term L_{a} accounting for the appearance distinctions between the action area and its start area and end area is introduced as

$$L_{\mathrm{a}} = \frac{1}{2}\frac{\langle \mathbf{R}_a, \mathbf{R}_s \rangle}{\langle \mathbf{R}_a, \mathbf{R}_a \rangle^{\frac{1}{2}} \langle \mathbf{R}_s, \mathbf{R}_s \rangle^{\frac{1}{2}}} + \frac{1}{2}\frac{\langle \mathbf{R}_a, \mathbf{R}_e \rangle}{\langle \mathbf{R}_a, \mathbf{R}_a \rangle^{\frac{1}{2}} \langle \mathbf{R}_e, \mathbf{R}_e \rangle^{\frac{1}{2}}}, \tag{8}$$

where $\langle, \angle$ denotes inner product.

The proposed coherence loss L_{c} (illustrated in Fig. 3) is defined as a linear combination of the new appearance term L_{a} and the OIC term:

$$L_{\mathrm{c}} = \alpha L_{\mathrm{OIC}} + (1-\alpha)(L_{\mathrm{a}} - 1), \tag{9}$$

where $\alpha \in [0, 1]$ is a trade-off hyper-parameter empirically set to 0.6 and fixed thereafter. The coherence loss promotes action proposals having not only high classification activation but also video representations distinctive from its context area.

Action Proposal Regression: The boundaries of initial action proposals are regressed with a set of regression networks. Specifically, each anchor size is assigned an exclusive regression network, and the receptive field size of each regression network is designed to be identical to its corresponding anchor size P to include sufficient (but not excessively redundant) contextual information. The regression network takes as input the video snippet-level features $\mathbf{F}$, and outputs a set of start and end boundary regression results $\{r_{s,i}\}_{i=1}^{T}$ and $\{r_{e,i}\}_{i=1}^{T}$.

To be specific, each regression network consists of three temporal convolutional layers. Denote a temporal convolutional layer with number of kernels n_k, temporal kernel size n_s, dilation rate n_d and activation function Ω as $\mathrm{Conv}(n_k, n_s, n_d, \Omega)$. For the regression network that corresponds to anchor size P, the first two layers are set to $\mathrm{Conv}(256, 3, \lceil P/4 \rceil, \mathrm{ReLU})$. The third output layer is set to $\mathrm{Conv}(2, 1, 1, \mathrm{Sigmoid})$, where the two output channels generate the start and end boundary regression results $\{r_{s,i}\}_{i=1}^{T}$ and $\{r_{e,i}\}_{i=1}^{T}$, respectively. In this way, the receptive field size for the last layer is $(P/4 + P/4 + 2) \times 2 + 1 \approx P$ to include sufficient contextual information. In addition, zero padding is used within the first two layers to retain the output dimensions, and group normalization [66] with group size 32 are used between each two consecutive convolutional layers.

Given the start and end boundary regression predictions $\{r_{s,i}\}_{i=1}^{T}$ and $\{r_{e,i}\}_{i=1}^{T}$, the initial action proposal $(x_{s,i}, x_{e,i})$ with anchor size P are refined as

$$\hat{x}_{s,i} = x_{s,i} + P \cdot \left(r_{s,x_{s,i}+1} - \frac{1}{2} \right), \tag{10}$$

$$\hat{x}_{e,i} = x_{e,i} + P \cdot \left(r_{e,x_{e,i}+1} - \frac{1}{2} \right), \tag{11}$$

where $\hat{x}_{s,i}$ and $\hat{x}_{e,i}$ are the refined start and end boundaries for the i-th initial proposal. In this way, the initial boundaries can be regressed to any temporal locations within its receptive field.

The confidence scores of action proposals are set to the negation of their coherence loss $-L_{\mathrm{c}}$. To balance the positive and negative predictions, only proposals with confidence scores higher than 0.1 are kept. Denote the surviving action proposals as $\{p_i\}_{i=1}^{N_p}$, where N_p is the number of surviving proposals. The total regression loss L_{reg} is defined as the average coherence loss of the remaining proposals:

$$L_{\mathrm{reg}} = \frac{1}{N_p} \sum_{i=1}^{N_p} L_{\mathrm{c}}(p_i), \tag{12}$$

where $L_{\mathrm{c}}(p_i)$ denotes the coherence loss for proposal p_i.

However, we empirically observe that some regression networks tend to be "lazy," and generate regression results with low variance. This problem degrades the performance, as the true action boundaries usually vary for proposals at different locations. To address this problem, a regularization term L_{norm} to maximize the variance of the regression results generated by

each layer is introduced:

$$L_{\text{norm}} = -\frac{1}{T}\sum_{i=1}^{T}\left(r_i^s - \frac{1}{T}\sum_{j=1}^{T} r_j^s\right)^2 - \frac{1}{T}\sum_{i=1}^{T}\left(r_i^e - \frac{1}{T}\sum_{j=1}^{T} r_j^e\right)^2. \quad (13)$$

Denote all anchor sizes as $\{P_i\}_{i=1}^{N_a}$, where N_a is the number of anchor sizes. The overall loss for the regression networks is defined as:

$$L = L_{\text{reg}} + \beta\frac{1}{N_a}\sum_{i=1}^{N_a} L_{\text{norm}}(P_i), \quad (14)$$

where $L_{\text{norm}}(P_i)$ denotes the regularization loss for the regression results corresponding to the anchor size P_i, and β is a trade-off hyper-parameter.

D. Single-Stream Inference

During testing, the authors only keep one action proposal per snippet location which achieves the highest confidence score among all action proposals covering this location of different anchor sizes and discard all others. Subsequently, Non-Maximum Suppression (NMS) is performed with an overlap Intersection-over-Union (IoU) threshold 0.4, which is empirically determined via cross-validation. Since a video may contain action instances of more than one action category, action localization is performed on all action categories whose classification predictions are higher than a predefined threshold 0.1.

Note that the RGB stream and the flow stream are trained separately. They are able to produce reliable modality-specific temporal action localization results independently.

E. Two-Stream Fusion

After obtaining action proposals from both RGB and flow streams, a *filter fusion* strategy is adopted to select and combine them. Empirically, the flow stream typically provides more accurate action proposals thanks to its sensitivity to even subtle motions, which align well with the start and end boundaries of action instances. In contrast, the RGB stream tends to provide longer, coarser, and less accurate action proposals, but corresponds better to scene transition locations. In addition, the two streams also focus on different parts of the video, and generate action proposals at different locations. Based on this observation, the flow stream is used as the primary source and the RGB stream is used as the auxiliary one. Let $\{p_{\text{flow},i}\}_{i=1}^{N_{\text{flow}}}$ and $\{p_{\text{rgb},i}\}_{i=1}^{N_{\text{rgb}}}$ denote the retained action proposals from the flow and RGB steams, respectively, where N_{flow} and N_{rgb} indicate the corresponding action proposal numbers.

The fusion module first retains all flow proposals $\{p_{\text{flow},i}\}_{i=1}^{N_{\text{flow}}}$ and at the same time discounts the confidence scores of all RGB proposals by a factor of 2.[1] Subsequently, for each RGB proposal $p_{\text{rgb},i}$, its IoU with all flow proposals $\{p_{\text{flow},i}\}_{i=1}^{N_{\text{flow}}}$ are calculated to obtain a retention score $I(p_{\text{rgb},i})$ with max-pooling:

$$I(p_{\text{rgb},i}) = \max(\{\text{IoU}(p_{\text{rgb},i}, p_{\text{flow},j})\}_{j=1}^{N_{\text{flow}}}). \quad (15)$$

The reconciled proposals are the union of all flow proposals $\{p_{\text{flow},i}\}_{i=1}^{N_{\text{flow}}}$ and a set of RGB proposals $\{p_{\text{rgb},i}|I(p_{\text{rgb},i}) < 0.4\}_{i=1}^{N_{\text{rgb}}}$.

To validate the superiority of the proposed action proposal fusion module, two alternative fusion methods are also included for comparison. *Union fusion* refers to directly combine the action proposals from RGB and flow streams as the final results. The *early fusion* used in AutoLoc [34] is also included: the early fusion method directly feeds the concatenated RGB and optical flow features to one sub-network and takes the output as the final localization results.

IV. Experiments and Discussions

In this section, the authors present the implementation details of the proposed Action Coherence Network (ACN) and compare it with the state-of-the-art temporal action localization (TAL) methods on two benchmark datasets. Extensive ablation studies are performed to validate the contribution of each component of the ACN. Finally, the authors present discussions and future work.

A. Dataset and Evaluation

Extensive experiments are conducted on two popular large-scale benchmarks, *i.e.*, THUMOS14 [67] and ActivityNet [49]. Note that only video-level action categorical labels are leveraged for training.

THUMOS14 dataset: [67] contains 1010, validation and 1,574 testing videos from 101 action categories. Among them, only 200 validation videos and 213 testing videos within 20 categories have temporal annotations. The authors follow previous methods to use the 200 validation videos to train our model, and use the 213 testing videos to evaluate it. The video length varies significantly from a few seconds to 26 minutes. The duration of an action instance also has a large variance, from shorter than one second to several minutes.

ActivityNet dataset: [49] offers a larger benchmark for TAL task. Two release versions of ActivityNet, *i.e.*, ActivityNet v1.3 and ActivityNet v1.2 are leveraged for experiments. ActivityNet v1.3 covers 200 action categories, with a training set of 10024 videos and a validation set of 4926 videos. ActivityNet v1.2 is a subset of ActivityNet v1.3, and it covers 100 action categories, with 4819 and 2383 videos in the training and validation sets respectively.[2] The training set and the validation set are used for training and testing, respectively. As shown in Fig. 4, the lengths of action instances in ActivityNet are generally longer than those in THUMOS14, and the duration range of ActivityNet is also larger than that of THUMOS14.

[1] To alleviate its overfitting tendency, the factor 2 is empirically determined via cross validation on THUMOS14, and also works well on ActivityNet.

[2] In our experiments, there are 9937 and 4575 videos in the training and validation sets of ActivityNet v1.3 respectively, and 4471 and 2211 videos in the training and validation sets of ActivityNet v1.2 respectively. The other videos are unaccessible from YouTube.

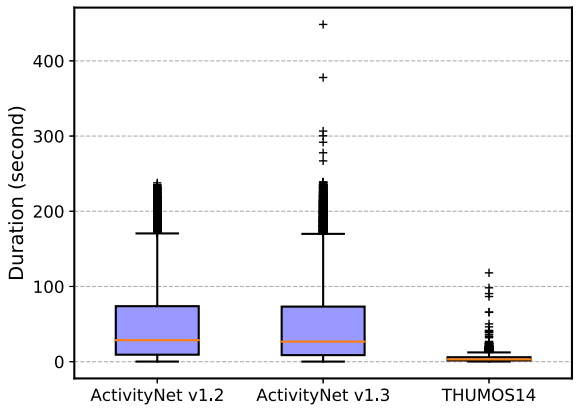


Fig. 4. Box plots of the action instance duration in THUMOS14 [67] and ActivityNet [49] datasets. On average, action instances in both release versions of ActivityNet datasets are significantly longer than those in THUMOS14. Specifically, the median action instance duration in THUMOS14 is 3.0 seconds, while those in ActivityNet v1.2 and ActivityNet v1.3 are 28.5 seconds and 26.6 seconds, respectively.

Evaluation metric: The authors follow the standard evaluation protocol based on mean Average Precision (mAP) values at different Intersection-over-Union (IoU) thresholds. The mAP computes the average per-category Average Precision (AP), which is defined as the area under the Precision-Recall (PR) curve. A proposal is a true positive if its IoU with certain unmatched ground truth is higher than a threshold, otherwise it is a false positive. The recall is defined as the ratio of true positive to ground truth, and the precision is the ratio of true positive to all proposals. The PR curve is drawn by connecting the recall and precision values for increasing set of proposals in a confidence score descending order, starting with the highest-scored proposal to all proposals [68]. The mAP values are calculated by the evaluation codes provided by the corresponding datasets.

B. Implementation Details

The proposed ACN is implemented in PyTorch [69]. The optical flow is estimated with TV-L1 algorithm [70], and only forward optical flow is computed. Two feature-extraction backbones are leveraged, namely UntrimmedNet [27] with a BNInception [71] backbone pre-trained on the ImageNet dataset [72], and I3D [46] pre-trained on the Kinetics dataset [46] for feature extraction, with a snippet length of 15 and 16 frames, respectively. RGB and optical flow features are extracted as 1024-dimensional vectors at the global_pool layer. The feature-extraction backbones are not fine-tuned for fair comparison with previous methods. For hyperparameter selection, we first manually create a validation set by uniformly sampling 20% videos from each class in the original training set, and then conduct grid search on the newly generated validation set. The regression network is trained with stochastic gradient descent (SGD) optimizer for 8 epochs, an initial learning rate of 0.001, and a decay factor of 10 for every 3 epochs. The normalization term in Eq. (14) is only used for the THUMOS14 dataset with β set to 2. The weight decay factor is set to 0.0005. To alleviate the background noise, attention thresholding is employed during testing. All snippets whose attention weights are lower than a threshold are discarded. Specifically, the attention threshold is fixed at 5 for the flow stream and 7 for the RGB stream for UntrimmedNet features, and 0.3 for both streams for I3D features. Following AutoLoc [34], the anchor sizes P (the number of snippets) are set to 1, 2, 4, 8, 16, 32 for THUMOS14 and 16, 32, 64, 128, 256, 512 for ActivityNet. If the length of a video is shorter than the minimal predefined anchor length, the whole video is considered as an action proposal and its confidence score is equal to its classification prediction score.

TABLE I
COMPARISON WITH STATE-OF-THE-ART TAL METHODS ON THE THUMOS14 TESTING SET. "FULLY-SUPERVISED" MEANS THE TEMPORAL ANNOTATIONS ARE USED DURING TRAINING, WHILE "WEAKLY-SUPERVISED" MEANS ONLY VIDEO-LEVEL ACTION LABELS ARE AVAILABLE DURING TRAINING. * INDICATES OUR REPRODUCED VERSION. "BC" DENOTES THE BACKGROUND CLASSIFICATION PROPOSED IN [32]

	Method	mAP@IoU (%)				
		0.3	0.4	0.5	0.6	0.7
Fully-supervised	Karaman *et al.* [73]	0.5	0.3	0.2	0.2	0.1
	Richard and Gall [74]	30.0	23.2	15.2	-	-
	Yeung *et al.* [75]	36.0	26.4	17.1	-	-
	Yuan *et al.* [76]	33.6	26.1	18.8	-	-
	Yuan *et al.* [77]	36.5	27.8	17.8	-	-
	S-CNN [56]	36.3	28.7	19.0	-	5.3
	SST [6]	37.8	-	23.0	-	-
	CDC [8]	40.1	29.4	23.3	13.1	7.9
	Dai *et al.* [11]	-	33.3	25.6	15.9	9.0
	TURN TAP [12]	44.1	34.9	25.6	-	-
	R-C3D [9]	44.8	35.6	28.9	-	-
	Gao *et al.* [59]	50.1	41.3	31.0	19.1	9.9
	SSN [7]	50.6	40.8	29.1	-	-
	BSN [17]	53.5	45.0	36.9	28.4	20.0
	TAL-Net [16]	53.2	48.5	42.8	33.8	20.8
Weakly-supervised	Hide-and-Seek [26]	19.5	12.7	6.8	-	-
	UntrimmedNet [27]	28.2	21.1	13.7	8.3	4.2
	STPN (UNTF) [28]	31.1	23.5	16.2	9,8	5.1
	W-TALC (UNTF) [30]	32.0	26.0	18.8	10.9	6.2
	AutoLoc (UNTF) [34]	35.8	29.0	21.2	13.4	5.8
	CMCS (UNTF) [31]	**37.5**	29.1	19.9	12.3	6.0
	CleanNet (UNTF) [35]	37.0	30.9	23.9	13.9	7.1
	ACN (UNTF)	37.0	**31.1**	**24.9**	**15.7**	**7.5**
	STPN (I3DF) [28]	35.5	25.8	16.9	9.9	4.3
	W-TALC (I3DF) [30]	40.1	31.1	22.8	14.5	7.6
	AutoLoc (I3DF)* [34]	38.1	30.6	23.1	14.2	6.9
	CMCS (I3DF) [31]	41.2	32.1	23.1	15.0	7.0
	3C-Net (I3DF) [33]	40.9	32.3	24.6	-	7.7
	ACN (I3DF)	40.7	34.7	26.4	16.8	8.0
	ACN + BC [32] (I3DF)	**43.4**	**36.3**	**27.3**	**17.6**	**8.7**

C. Comparison With State-of-The-Arts

Experiments on THUMOS14: The results on the THUMOS14 testing set are summarized in Table I, where the UntrimmedNet feature and I3D feature are denoted as UNTF and I3DF, respectively. The proposed ACN outperforms all competing W-TAL methods on the THUMOS14 testing set. Among them, ACN with UNTF outperforms AutoLoc [34] by a large margin at all IoU thresholds. Especially, ACN with UNTF even achieves higher or comparable mAP with previous state-of-the-art methods (*e.g.*, STPN [28], W-TALC [30], AutoLoc [34] and CMCS [31]) with I3DF at high IoU thresholds, which

TABLE II
COMPARISON WITH STATE-OF-THE-ART W-TAL METHODS ON THE ACTIVITYNET V1.2 VALIDATION SET. AVG DENOTES THE MEAN MAP AT IOU THRESHOLDS 0.5:0.05:0.95. $^+$ INDICATES THE REPRODUCED VERSION OF AUTOLOC [34], AND * INDICATES OUR REPRODUCED VERSION

Supervision	Method	mAP@IoU (%) 0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg
Weakly-supervised	UntrimmedNet$^+$ [27]	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
	AutoLoc (UNTF) [34]	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
	ACN (UNTF)	**30.4**	**27.2**	**24.3**	**20.5**	**18.0**	**15.4**	**13.2**	**10.3**	**7.5**	**3.7**	**17.0**
	W-TALC (I3DF) [30]	**37.0**	**33.5**	**30.4**	**25.7**	14.6	12.7	10.0	7.0	4.2	1.5	18.0
	AutoLoc (I3DF)* [34]	31.9	29.3	26.0	22.9	20.0	17.0	13.6	9.7	5.0	1.4	17.7
	ACN (I3DF)	36.0	31.6	28.0	24.2	**21.1**	**17.9**	**14.8**	**11.3**	**7.0**	**3.5**	**19.6**

TABLE III
COMPARISON WITH STATE-OF-THE-ART W-TAL METHODS ON THE ACTIVITYNET V1.3 VALIDATION SET. AVG DENOTES THE MEAN MAP AT IOU THRESHOLDS 0.5:0.05:0.95. * INDICATES OUR REPRODUCED VERSION

Supervision	Method	mAP@IoU (%) 0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg
Weakly-supervised	UntrimmedNet* [27]	7.0	6.1	5.3	4.4	4.0	3.3	2.6	2.1	1.5	0.7	3.7
	AutoLoc (UNTF)* [34]	25.6	22.6	20.0	17.3	14.4	11.6	8.9	6.5	3.7	1.1	13.2
	ACN (UNTF)	**28.8**	**25.9**	**22.9**	**20.6**	**18.2**	**15.4**	**12.9**	**9.6**	**5.5**	**1.3**	**16.1**
	STPN (I3DF) [28]	29.3	-	-	-	-	16.9	-		-	2.6	-
	W-TALC (I3DF)* [30]	33.4	**30.7**	**28.4**	**26.2**	**21.8**	11.6	4.6	2.1	0.8	0.2	16.0
	AutoLoc (I3DF)* [34]	26.1	24.0	21.6	19.0	16.4	14.2	11.6	8.2	4.3	1.3	14.7
	ACN (I3DF)	**33.6**	30.0	26.7	23.4	20.1	**17.4**	**14.0**	**10.8**	**7.2**	**3.9**	**18.7**

demonstrates that the proposed coherence loss is able to detect more precise action boundaries.

When equipped with I3DF, the proposed ACN achieves the state-of-the-art performance compared with all competing W-TAL methods, and is even comparable with some recent fully-supervised methods (*e.g.*, Dai *et al.* [11] and TURN TAP [12]). Besides, adding the background classification [32] as an auxiliary training objective improves the feature representation ability, and further improves the overall performance.

Experiments on ActivityNet: Results on ActivityNet v1.2 and v1.3 validation sets are presented in Table II and Table III, respectively. The proposed ACN outperforms all other methods with the same backbone on the average mAP at IoU thresholds 0.5:0.05:0.95. Furthermore, the proposed ACN with UNTF outperforms AutoLoc [34] at all IoU thresholds on two datasets. Especially, on ActivityNet v1.3, ACN with UNTF even achieves higher mAP than AutoLoc with I3DF at all IoU thresholds.

With I3DF, ACN further improves the performance. Although W-TALC [30] achieves higher mAPs at low IoU thresholds, the performance advantage of ACN becomes more significant as IoU increases, which demonstrates that ACN can locate more precise temporal boundaries and the results have larger overlap with the ground truth. It also should be noted that the performances of almost all the methods degrade on ActivityNet v1.3 compared with ActivityNet v1.2. The reason might be that the durations of action instances on ActivityNet v1.3 vary more than those on ActivityNet v1.2 (see Fig. 4). Particularly, AutoLoc [34] with I3DF drops 3% on average, while ACN with I3DF only drops 0.9% on average, which demonstrates that ACN is able to generate more flexible action proposals.

To summarize, the proposed ACN outperforms all the competing W-TAL methods on both the THUMOS14 and ActivityNet datasets, and even compares favorably with some fully-supervised TAL methods. This clearly demonstrates the efficacy of the proposed ACN.

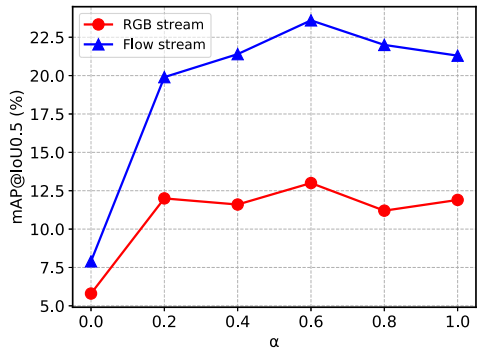


Fig. 5. TAL performance comparison with different α values in (9) during training and evaluation on the THUMOS14 dataset. Both streams achieve the highest performances when $\alpha = 0.6$.

D. Ablation Study

To analyze the contribution of each component of our proposed ACN to the overall performance boost, a set of ablation studies are carried out on the THUMOS14 dataset with UNTF.

Sensitivity Analysis on α: The α value in Eq. (9) is an important trade-off hyper-parameter, which measures the relative importance between SCPs and video representations in the proposed coherence loss. Different α values are evaluated during both training and evaluation phases. The results are measured with mAP at IoU threshold 0.5 and are illustrated in Fig. 5, which justifies our empirical choice of $\alpha = 0.6$.

The performance is extremely low when α equals to 0. This reveals that by only using the appearance term accounting for the video content change while ignoring the classification score, it is hard to distinguish the foreground from the background because the background is equally possible to be classified as action instances. When taking the classification score into account, the performance boosts significantly even when $\alpha = 0.2$. When α is

TABLE IV
SINGLE-STREAM TAL PERFORMANCE WITH DIFFERENT TRAINING LOSSES AND SCORING METHODS ON THE THUMOS14 TESTING SET. THE LOSS COLUMN MEANS THE NETWORK IS TRAINED WITH THE CORRESPONDING LOSSES, AND THE SCORE COLUMN MEANS THE PROPOSALS ARE SCORED WITH THE CORRESPONDING METHODS

Loss	Score	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7
L_{a}	$-L_{\mathrm{a}}$	15.3	11.8	7.9	4.2	2.1
	$-L_{\mathrm{OIC}}$	32.8	27.2	20.9	13.0	6.2
	$-L_{\mathrm{c}}$	**33.9**	**28.0**	**21.6**	**14.0**	**7.3**
L_{OIC}	$-L_{\mathrm{a}}$	18.3	13.9	9.8	5.4	2.3
	$-L_{\mathrm{OIC}}$	33.8	27.6	21.3	13.8	6.6
	$-L_{\mathrm{c}}$	**35.5**	**28.9**	**22.0**	**14.1**	**6.9**
L_{c}	$-L_{\mathrm{a}}$	17.3	13.5	9.6	5.4	2.3
	$-L_{\mathrm{OIC}}$	33.7	27.8	22.3	**15.3**	**7.2**
	$-L_{\mathrm{c}}$	**35.8**	**29.8**	**23.6**	14.8	**7.2**

(a) Flow stream-only localization performance.

Loss	Score	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7
L_{a}	$-L_{\mathrm{a}}$	13.6	9.5	5.8	2.6	0.9
	$-L_{\mathrm{OIC}}$	23.7	17.5	11.3	**5.8**	**2.1**
	$-L_{\mathrm{c}}$	**24.2**	**17.6**	**11.6**	5.4	**2.1**
L_{OIC}	$-L_{\mathrm{a}}$	14.2	10.0	6.1	2.7	0.8
	$-L_{\mathrm{OIC}}$	23.7	17.9	11.9	**6.3**	**2.4**
	$-L_{\mathrm{c}}$	**25.0**	**18.6**	**12.4**	6.2	2.2
L_{c}	$-L_{\mathrm{a}}$	14.0	9.8	5.9	2.8	0.8
	$-L_{\mathrm{OIC}}$	24.2	18.1	12.6	6.4	**2.4**
	$-L_{\mathrm{c}}$	**25.3**	**19.2**	**13.0**	**6.7**	2.3

(b) RGB stream-only localization performance.

TABLE V
TAL PERFORMANCE WITH DIFFERENT FUSION METHODS ON THE THUMOS14 TESTING SET. "DISCOUNT" MEANS THE CONFIDENCE SCORES OF RGB PROPOSALS ARE DISCOUNTED BY A FACTOR OF 2

Fusion Method	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7
early fusion	**37.0**	29.3	22.4	14.3	6.2
union fusion	26.5	23.6	19.6	13.8	**7.5**
filter fusion w/o discount	36.1	30.3	24.3	15.1	7.2
filter fusion w/ discount	**37.0**	**31.1**	**24.9**	**15.7**	**7.5**

TABLE VI
PERFORMANCE COMPARISON BETWEEN MODELS TRAINED WITH AND WITHOUT THE PROPOSED REGULARIZATION LOSS L_{norm}. THE VARIANCE COLUMN DENOTES THE AVERAGE VARIANCE OF REGRESSION RESULT FOR DIFFERENT ANCHOR SIZES

L_{norm}	Stream	mAP@IoU (%) 0.3	0.4	0.5	0.6	0.7	Variance
-	RGB	20.9	14.8	9.7	4.5	1.9	0.0173
✓		**25.3**	**19.2**	**13.0**	**6.7**	**2.3**	**0.0189**
-	flow	35.2	29.7	22.8	**15.0**	**7.2**	0.0156
✓		**35.8**	**29.8**	**23.6**	14.8	**7.2**	**0.0201**

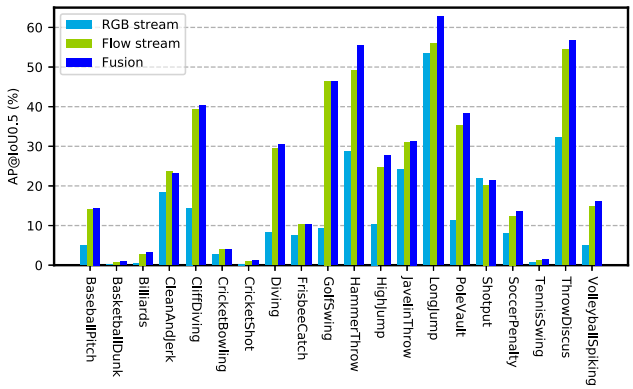


Fig. 6. Per category TAL performance on the THUMOS14 dataset under the IoU threshold 0.5. In all action categories except for *Shotput*, the flow stream outperforms the RGB stream. And the fusion results further improve the performance of most categories.

set to 1, namely only OIC loss is used to train and test our model, the flow stream achieves an mAP of 21.3% at IoU threshould 0.5, which even outperforms the mAP of 21.2% by AutoLoc [34]. This verifies our assumption that the RGB and optical flow modalities are complementary, and the concatenation-based fusion methods such as AutoLoc [34] fail to effectively utilize the two modalities.

Ablation Study on Action Proposal Scoring: As presented in Section III-C, the negation of coherence loss L_c is used to score action proposals. Two additional variants of the scoring method are also included (for evaluation phase only), namely the appearance term $-L_{\mathrm{a}}$ only and the OIC term $-L_{\mathrm{OIC}}$ only. The authors train the network with three different loss functions, namely L_{a}, L_{OIC}, and L_{c}. For each of these models, the authors score the action proposals with three different scoring methods.

The flow stream and RGB stream TAL performances are summarized in Table IV(a) and Table IV(b), respectively. First, the performance advantage of $-L_{\mathrm{c}}$ as the scoring method for all training losses verifies that the OIC term and the appearance term are indispensable, and they jointly contribute to the determination of the relative importance of action proposals. Second, the performance of L_{c} is better than that of L_{OIC}. This verifies that the proposed coherence loss is able to regress the action proposals to more precise action boundaries.

Ablation Study on the Regularization Loss: To help the regression network generate flexible regression results, a regularization term L_{norm} is introduced as in Eq. (13). The comparison results are listed in Table VI. The results reveal that the regularization loss helps improve the localization performance for both streams at all IoU thresholds. Besides, the average testing variance of the regression results with the regularization is also larger than that without the regularization loss, demonstrating that the regularization loss helps to generate more flexible regression predictions. The performance improvement and the variance increase demonstrate the efficacy of the regularization loss.

Ablation Study on Action Proposal Fusion: As discussed in Section III-E, experiments to compare different fusion methods are conducted. The results on the THUMOS14 dataset are presented in Table V. The proposed *filter fusion* which discounts scores of RGB proposals outperforms other methods. Meanwhile, *early fusion* outperforms AutoLoc at all IoU thresholds, which demonstrates the superiority of the proposed coherence loss. However, its performance is even worse than the performance of the flow stream under most of the IoU thresholds, which means the concatenation-based two-modality fusion may introduce some noise to the model and thus lead to performance degradation.

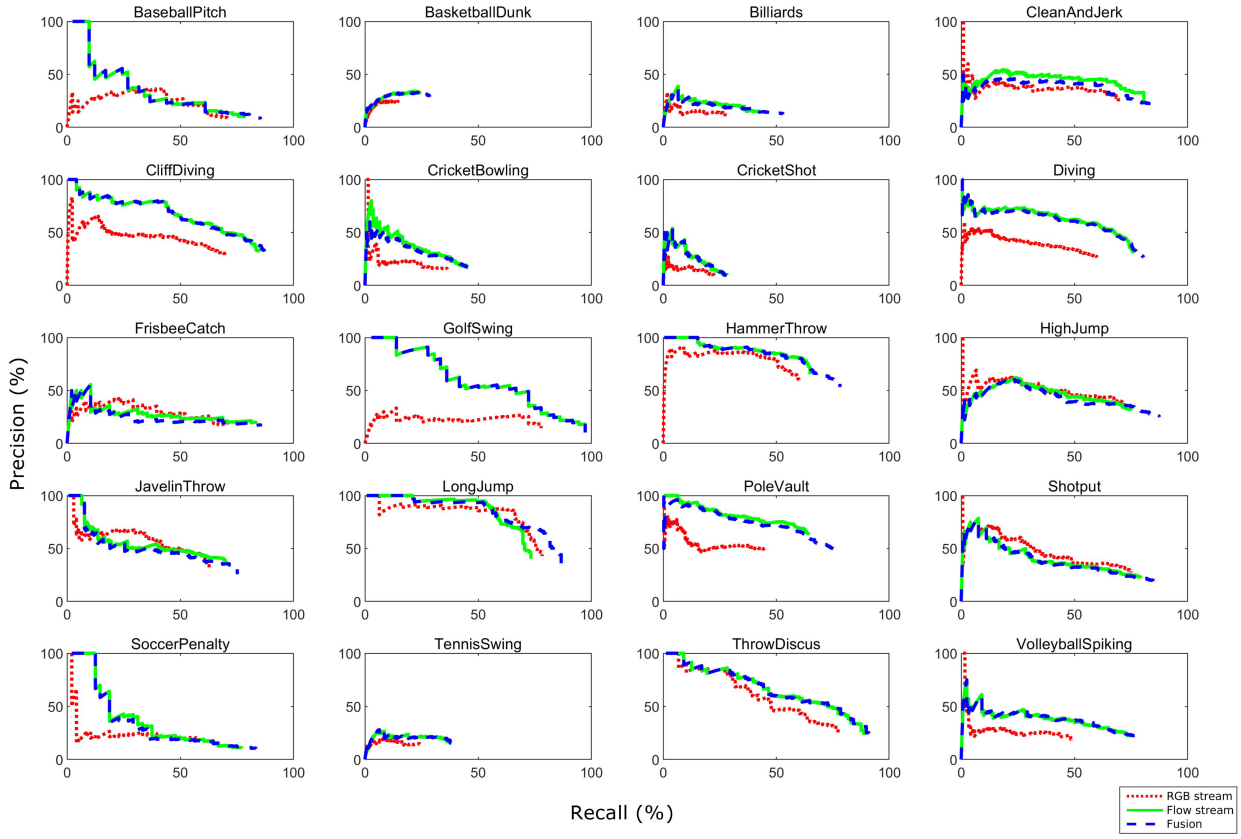


Fig. 7. Per category Precision-Recall (PR) curve under the IoU threshold 0.3 on the THUMOS14 dataset. The x and y axes represent recall and precision, respectively. The area enclosed by the PR curve and x and y axes is the Average Precision (AP).

Fig. 6 compares the TAL performance on all action categories, where the flow stream outperforms the RGB stream on all categories except for *Shotput*. Moreover, the proposed *filter fusion* further helps improve the performance on most of the action categories. For example, on action *Long Jump*, the APs at IoU threshold 0.5 are 53.4% and 56.0% for RGB stream and flow stream respectively, while the fusion result on this category is 62.7%. On action *Hammer Throw*, the APs at the IoU threshold 0.5 are 28.7% and 49.2% for the RGB stream and the flow stream respectively, while the fusion result achieves 55.4%.

Fig. 7 presents the Precision-Recall (PR) curves for all categories under the IoU threshold 0.3. Since the flow stream is chosen as the primary source, the PR curves of the flow stream and the fusion result are largely overlapped on most of the categories. Meanwhile, the *filter fusion* result can achieve higher recall (*i.e.*, longer in the x axis) than the flow stream because the retained RGB proposals contain some true action instances that the flow stream fails to detect. The results also show that the higher recall is the main factor of the performance improvement, because it leads to a larger area enclosed by the PR curve and the x and y axes, and thus a higher AP for each category.

Qualitative Analysis: Several representative examples of TAL results are plotted in Fig. 8 to illustrate the efficacy of the proposed ACN. For the *Frisbee Catch* example, the flow stream and RGB stream both only detect a portion of ground truth results, and thus all two stream proposals are kept in the final fusion results. For the *Billiards* example, the RGB and flow streams are also complementary, but the RGB stream fails to separate two proximate action instances, and these RGB proposals are discarded after fusion because of their large overlap with the flow proposals. For the *Throw Discus* example, both streams provide accurate action proposals, and only flow proposals are retained in the final results under such situation. For the *Javelin Throw* example, the RGB stream fails to detect all true action instances, while the flow stream produces precise action proposals. Although the final fusion results nearly contain all action proposals from two streams, the confidence score discount in the RGB stream helps to alleviate the negative effect of fusing RGB proposals. For the last example of *Clean and Jerk*, both streams cannot provide accurate predictions, and they instead predict fragmentary and low-quality action proposals. In this case, combining these inaccurate predictions leads to even worse prediction results, as the merged predictions will decrease the precision while maintaining the recall.

E. Discussion and Future Work

The proposed coherence loss improves the performance by involving the video content change in the scoring, whose importance has also been verified in the co-activity similarity loss in W-TALC [30]. Specifically, in our work and

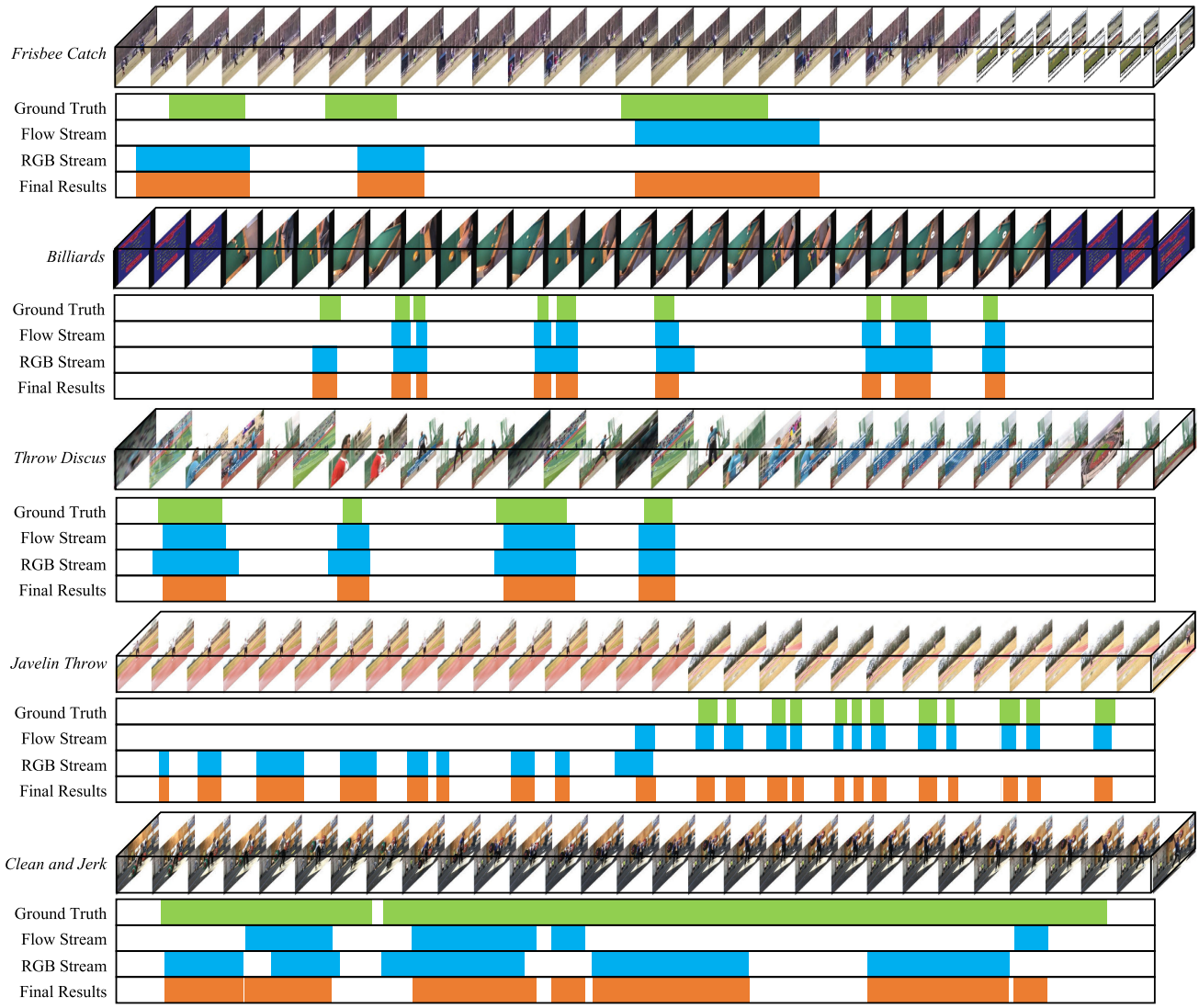


Fig. 8. Qualitative TAL results obtained by ACN on the THUMOS14 dataset. The horizontal axis in the plot is the timestamp. The five rows in each case are 1) input video frames, 2) ground truth of action instances, 3) action proposals from the flow stream, 4) action proposals from the RGB stream, and 5) final fusion results, respectively.

W-TALC [30], the feature similarity is measured via cosine similarity, the values of which will be very approximate for high-dimensional (*e.g.*, 1024-dimensional) features. Moreover, if a video exhibits many scene transitions, the regressed action boundaries tend to be those scene transition locations rather than true action boundaries. This is because the SCPs and features are more distinctive at these locations. Therefore, future work may further exploit effective ways to model the action coherence or similarity, or try to separate scenes in a video before performing temporal action localization.

In addition, the flow stream with only an OIC loss achieves comparable performance with AutoLoc [34]. This means the RGB modality is largely ignored or even not used in the concatenation-based fusion method. Therefore, another future work may continue to focus on properly fusing the two modalities.

Furthermore, the proposed ACN achieves higher performance boost on THUMOS14 than on ActivityNet. The reason is that under large anchor sizes, the dilation can be very large, leading to a very sparse sampling (*e.g.*, the dilation is 128 for anchor size 512). This means the temporal information is not effectively employed. However, on one hand, when increasing the sampling rate, more parameters need to be added to the network, and thus the overfitting problem will become more severe; on the other hand, setting a higher β in Eq. (14) will also lead to performance degradation. Future work may try to solve this dilemma.

V. Conclusion

This paper proposes an Action Coherence Network (ACN) for weakly-supervised temporal action localization, which benefits from a new coherence loss and a novel fusion module. The

coherence loss helps action proposals regress more precise temporal locations, which have high classification activation and clear appearance boundaries. The novel fusion module is capable of reconciling modality-specific action proposals generated by the RGB and flow streams. Experimental results on the THUMOS14 and ActivityNet datasets demonstrate the superiority of our ACN over previous states-of-the-art. Extensive ablation studies are also conducted to validate our design intuitions.

REFERENCES

[1] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. G. Hauptmann, "Multimedia event detection using a classifier-specific intermediate representation," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1628–1637, Nov. 2013.

[2] H. Zhang and C. Ngo, "A fine granularity object-level representation for event detection and recounting," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1450–1463, Jun. 2019.

[3] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[4] K. Kumar and D. D. Shrimankar, "F-DES: Fast and deep event summarization," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 323–334, Feb. 2018.

[5] Yu-Fei Ma, Xian-Sheng Hua, L. Lu, and Hong-Jiang Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[6] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2911–2920.

[7] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2914–2923.

[8] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5734–5743.

[9] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3d network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5783–5792.

[10] F. C. Heilbron, W. Barrios, V. Escorcia, and B. Ghanem, "SCC: Semantic context cascade for efficient action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3175–3184.

[11] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Qiu Chen, "Temporal context network for activity localization in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5793–5802.

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

[13] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos." in *Proc. Brit. Mach. Vis. Conf.*, May 2019, pp. 93.1–93.12.

[14] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 988–996.

[15] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, pp. 7477–7484.

[16] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1130–1139.

[17] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[18] Y. Xu *et al.*, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 9070–9078, 2019.

[19] H. Alwassel, F. Caba Heilbron, and B. Ghanem, "Action search: Spotting actions in videos and its application to temporal action localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 251–266.

[20] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 344–353.

[21] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3889–3898.

[22] H. Song, X. Wu, B. Zhu, Y. Wu, M. Chen, and Y. Jia, "Temporal action localization in untrimmed videos using action pattern trees," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 717–730, Mar. 2019.

[23] R. Zeng *et al.*, "Graph convolutional networks for temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7094–7103.

[24] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[25] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 536–548.

[26] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3524–3533.

[27] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4325–4334.

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

[29] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2018, pp. 35–44.

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

[31] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1298–1307.

[32] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5502–5511.

[33] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3C-NET: Category count and center loss for weakly-supervised action localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8679–8687.

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

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

[36] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[37] L. Sevilla-Lara *et al.*, "On the integration of optical flow and action recognition," in *German Conf., Pattern Recognit.*, Oct. 2018, pp. 281–297.

[38] Y. Zhai *et al.*, "Action coherence network for weakly supervised temporal action localization," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3696–3700.

[39] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*,vol. 64, no.2–3, pp. 107–123, Sep. 2005.

[40] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1817–1824.

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

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

[43] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," 2015, *arXiv:1507.02159*.

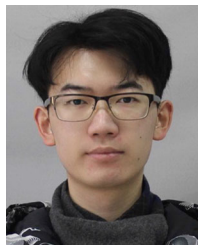
[44] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 20–36.

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

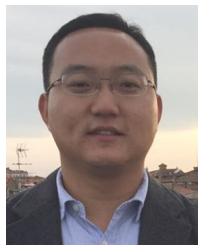
[46] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[47] J. Yue-Hei *et al.*, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4694–4702.

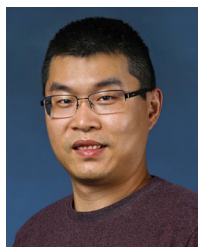
[48] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2017.
[49] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.
[50] B. Korbar, D. Tran, and L. Torresani, "SCSampler: Sampling salient clips from video for efficient action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6232–6242.
[51] C. Luo and A. L. Yuille, "Grouped spatial-temporal aggregation for efficient action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5512–5521.
[52] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7882–7891.
[53] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9945–9953.
[54] Z. Shou *et al.*, "DMC-Net: Generating discriminative motion cues for fast compressed video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1268–1277.
[55] L. Wang, P. Koniusz, and D. Q. Huynh, "Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8698–8708.
[56] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1049–1058.
[57] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "DAPS: Deep action proposals for action understanding," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 768–784.
[58] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
[59] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *Proc. Br Mach. Vis. Conf.*, May 2017, pp. 52.1–52.11.
[60] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.
[61] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11320–11327, Apr. 2020.
[62] B. Shi, Q. Dai, Y. Mu, and J. Wang, "Weakly-supervised action localization by generative attention modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1009–1019.
[63] Y. Zhai *et al.*, "Two-stream consensus network for weakly-supervised temporal action localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 37–54.
[64] Z. Liu *et al.*, "ACSNet: Action-context separation network for weakly supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2021.
[65] Z. Liu, L. Wang, W. Tang, J. Yuan, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through learning explicit subspaces for action and context," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2021.
[66] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[67] Y.-G. Jiang *et al.*, "Thumos challenge: Action recognition with a large number of classes," vol. 1, no. 2, p. 2, Sep. 2014.
[68] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2016, pp. 198–213.
[69] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 8024–8035.
[70] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Pattern Recognit.*, 2007, pp. 214–223.
[71] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[72] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, Jun. 2009.
[73] S. Karaman, L. Seidenari, and A. Del Bimbo, "Fast saliency based pooling of fisher encoded dense trajectories," in *Proc. Eur. Conf. Comput. Vis. THUMOS Workshop*, vol. 1, no. 2, p. 5, Sep. 2014.
[74] A. Richard and J. Gall, "Temporal action detection using a statistical language model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3131–3140.
[75] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2678–2687.
[76] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3093–3102.
[77] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3684–3692.



Yuanhao Zhai (Member, IEEE) received the B.Eng. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2020. He is currently a Research Intern with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision and machine learning.



Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a Visiting Ph.D. Student with the Stevens Institute of Technology, Hoboken, NJ, USA. From 2016 to 2017, he was a Visiting Scholar with Northwestern University, Evanston, IL, USA. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. He is the author of more than 50 peer reviewed publications in prestigious international journals and conferences. His research interests include computer vision, pattern recognition, and machine learning.



Wei Tang (Member, IEEE) received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 2019. He is currently an Assistant Professor with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA. His research interests include computer vision, pattern recognition, and machine learning.

Qilin Zhang (Member, IEEE) received the B.E. degree in electrical information engineering from the University of Science and Technology of China, Hefei, China, in 2009, the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2011, and the Ph.D. degree in computer science from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2016. He is currently a Senior Research Scientist with ABB Corporate Research Center, Raleigh, NC, USA. From 2016 to 2018, he was a Senior Research Engineer and from 2018 to 2020, a Lead Research Engineer with HERE Technologies, Chicago, IL, USA. His research interests include computer vision and signal processing.

Nanning Zheng (Fellow, IEEE) Graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975. He received the M.E. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1981 and the Ph.D. degree in electrical engineering from Keio University, Keio, Japan, in 1985. He is currently a Professor and the Director with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, computational intelligence, and hardware implementation of intelligent systems. Since 2000, he has been the Chinese Representative on the Governing Board of the International Association for Pattern Recognition. In 199, he became a member of the Chinese Academy Engineering.

Gang Hua (Fellow, IEEE) was enrolled in the Special Class for the Gifted Young of Xi'an Jiaotong University (XJTU), Xi'an, China, in 1994. He received the B.S. degree in automatic control engineering and the M.S. degree in control science and engineering from XJTU in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, IL, USA, in 2006. He is currently the Vice President and Chief Scientist of Wormpex AI Research. Before that, he was in various roles with Microsoft, from 2015 to 2018, as the Science or Technical Adviser to the CVP of the Computer Vision Group, the Director of Computer Vision Science Team, Redmond and Taipei ATL, and the Principal Researcher or Research Manager with Microsoft Research. From 2011 to 2015, he was an Associate Professor with the Stevens Institute of Technology. During 2014–2015, he took an on leave and worked on the Amazon-Go project. From 2011 to 2014, he was an Visiting Researcher and from 2010 to 2011, a Research Staff Member with IBM Research T. J. Watson Center, from 2009 to 2011, a Senior Researcher with Nokia Research Center Hollywood, and from 2006 to 2009, a Scientist with Microsoft Live labs Research. He is the author of more than 190 peer reviewed publications in prestigious international journals and conferences. He holds 19 U.S. patents and has 15 more U.S. patents pending. He is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Computer Vision and Image Understanding*, IEEE MULTIMEDIA, *The Visual Computer Journal,* and *Journal of Machine Vision and Applications*. He was also the Lead Guest Editor on two special issues in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *International Journal of Computer Vision*. He is the General Chair of ICCV'2025. He is the Program Chair of CVPR'2019&2022. He is the Area Chair of the CVPR'2015&2017, ICCV'2011&2017, ICIP'2012&2013&2016, ICASSP'2012&2013, and ACM MM 2011&2012&2015&2017. He was the recipient of the 2015 IAPR Young Biometrics Investigator Award for his contribution on Unconstrained Face Recognition from Images and Videos, and was also the recipient of the 2013 Google Research Faculty Award. He is an IAPR Fellow and an ACM Distinguished Scientist.