第 51 卷第 10 期
2021 年 10 月

中 国 科 学 技 术 大 学 学 报
JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Vol. 51, No. 10
Oct. 2021

Received: 2021-03-02; Revised: 2021-04-28

doi:10.52396/JUST-2021-0061

Relation aware network for weakly-supervised temporal action localization

ZHAN Yongkun, YANG Wenfei, ZHANG Tianzhu*
Laboratory for Future Networks, University of Science and Technology of China, Hefei 230027, China
* Corresponding author. E-mail: tzzhang@ustc.edu.cn

Abstract: Temporal action localization has become an important and challenging research orientation due to its various applications. Since fully supervised localization requires a lot of manpower expenditure to get frame-level or segment-level fine annotations on untrimmed long videos, weakly supervised methods have received more and more attention in recent years. Weakly-supervised Temporal Action Localization (WS-TAL) aims to predict action temporal boundaries with only video-level labels provided in the training phase. However, the existing methods often only perform classification loss constraints on independent video segments, but ignore the relation within or between these segments. In this paper, we propose a novel framework called Relation Aware Network (RANet), which aims to model the segment relations of intra-video and inter-video. Specifically, the Intra-video Relation Module is designed to generate more complete action predictions, while the Inter-video Relation Module is designed to separate the action from the background. Through this design, our model can learn more robust visual feature representations for action localization. Extensive experiments on three public benchmarks including THUMOS 14 and ActivityNet 1.2/1.3 demonstrate the impressive performance of our proposed method compared with the state-of-the-arts.

Keywords: temporal action localization; weakly-supervised learning; relation modeling
CLC number: TP391.8 **Document code**: A

1 Introduction

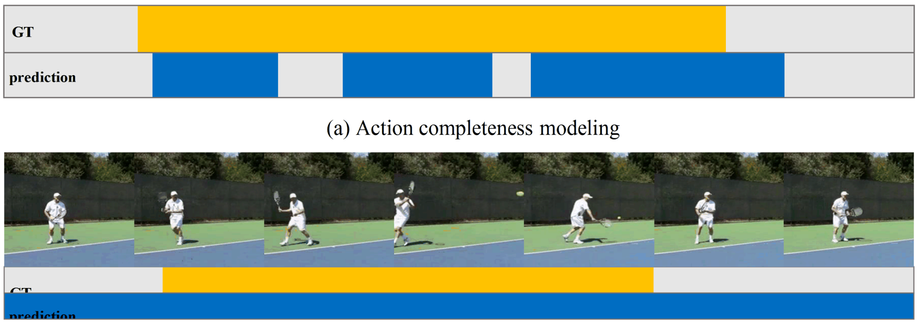
Temporal action localization aims to localize and recognize actions in given long untrimmed videos, which has a wide range of practical applications, e.g., video understanding[1], visual question answering (VQA)[2], video surveillance[3] and video summarization[4].

However, most of existing methods[5-10] tackle this task in a fully supervised way, which relies heavily on expensive and time-consuming manual annotations.

To overcome this issue, researchers have started to study action localization under a weakly-supervised setting recently[11-19]. Various of weak labels have been explored, e.g., action categories[11,13,16], movie scripts[17,20] and sparse spatio-temporal points[18]. Among these methods, action categories based methods have become the main stream. Compared to label precise temporal action boundaries, video-level labels are much easier to collect. Existing methods can be broadly divided into three categories, learning background suppression attention weights[13,21,22], learning discriminative features[12,19] and erasing discriminative segments during training[16,22]. Among these methods, attention based methods have achieved superior performance.

Despite the success of previous approaches, the general framework largely relys on the classification activation, which employs an attention model to identify the action segments and categorizes them into different classes. However, as shown in Figure 1, such method cannot deal with two challenges well: ① action completeness modeling. The first challenge is how to localize each action instance completely. In the weakly-supervised setting, the lack of fine-grained annotations will complicate the complete-ness modeling since only video-level labels are given. As shown in Figure 1(a), the current models tend to divide a complete action into multiple actions. Identifying one fragment of an action is sufficient for video-level classification but not for segment-level localization and hence cause negative influence on the prediction. To solve action discretization issues, the network cannot only focus on the most distinctive individual segments. Therefore, when determining the category of a certain independent video segment, it is important to establish relations

Citation: Zhan Yongkun, Yang Wenfei, Zhang Tianzhu. Relation aware network for weakly-supervised temporal action localization. J. Univ. Sci. Tech. China, 2021, 51(10): 753-765.



(a) Action completeness modeling

(b) Action-context confusion

Figure 1. Two major challenges in WS-TAL: (a) Action completeness modeling, and (b) Action-context confusion. The color yellow indicates the Ground Truth (GT) while blue indicates model predictions. As shown in Figure 1(a), a complete jumping action is divided into multiple segments, which cannot form a whole action. In Figure 1(b), a man is playing tennis. Due to the high similarity of the background, the localization of the prediction often fails to separate the action from the background, results in false predictions.

between the segments within the video. ② action-context confusion. The second challenge is how to distinguish action instances from their context with only video labels. Video-level classifiers learn the correlation between videos with the same label and discover their common contents, which unfortunately include not only the common action but also the closely related context. As shown in Figure 1(b), context segments near to action segments tend to be recognized as action segments, since they are usually surrounded by visually related clips. For this reason, we argue that action-context confusion is inherently difficult with weak supervision, unless employing the prior knowledge about actions. Intuitively, context frame indeed exhibits obvious difference from action frame at the appearance and feature level. An important prior knowledge is that videos with the same action often have different context background frames. To separate context and action, the model should be able to capture the underlying discrepancy between action and context, and therefore, modeling the inter relations between videos is necessary.

Motivated by the above discussions, we propose an end-to-end Relation-Aware Net-work (RANet) for weakly-supervised temporal action localization by jointly modeling the inter and intra video relation of segments. An overview of our algorithm is shown in Figure 2. The Relation Module of RANet is mainly composed of two components:

(Ⅰ) Intra-video relation module. To model action completeness, feature sequences extracted from input videos are fed into the Intra-video Relation Module with a graph convolution neural network (GCN)[23] layer, which builds a bridge of information flow within a video, such that complete actions can be retrieved by aggregating activations from neighbouring multiple segments.

(Ⅱ) Inter-video relation module. As for action-context confusion, we develop a simple and effective strategy by sampling video pairs with the same label to establish relations between videos. we constrain the segments features of the same action between video pairs as close as possible. Therefore our model can pay more attention to the action itself rather than the background during training.

The contributions of this paper are summarized as follows. ①We propose a new end-to-end relation aware network for weakly supervised temporal action localization by jointly modeling inter-video and intra-video relations. ②The proposed relation methods for intra-video and inter-video modeling can learn discriminative relation aware representations for robust weakly supervised action detection and effectively solve two current WS-TAL challenges. ③ Extensive experiments and visualized results on three public benchmarks demonstrate the effectiveness and superiority of RANet over the state-of-the art methods and even compares favorably with some fully-supervised temporal action localization methods.

2 Related work

Action classification is a fundamental task in computer vision area. Traditional methods often rely on manually extracted features, e. g, HOG[24], dense trajectory iDT[25] and motion mode histogram MBH[26]. Subsequently, with the break-through of deep learning, a large number of deep learning methods have been applied to video analysis, such as the classic two-stream network[27], C3D[28], P3D[29], I3D[30], and Temporal segment networks TSN[31], TRN[32], TSM[1]. In our method, I3D is used for feature extraction.

2.1 Fully-supervised action localization

Fully supervision action localization requires frame-level annotations of all action instances during training which has been extensively studied recently. Several large-scale datasets have been collected for this task, such as THUMOS[33], ActivityNet[34], Charades[35] and AVA[36].

Many methods follow the paradigm that has been widely used in the field of object detection[5-8] because they have a common set of problems. Specifically, there are two main directions: two-stage methods and one-stage methods. The two-stage approaches[9-10,37-39] first generate action recommendations, then classify them and further refine time boundaries. The one-stage methods[40-42] replace the direct prediction of action categories and positions from raw data.

2.2 Weakly-supervised action localization

To address the limitation of fully super-vised action detection, weakly supervised action detection has been drawing increasing research attention. Wang et al[11] first proposed a UtrimmedNet network, in which the network first learns the video-level classifier and then selects high category activated frames for action localization. The later work can be broadly divided into three types. The first type is based on attention mechanisms, which aims to highlight foreground segments and suppressing background segments. STPN[13] first added a class-agnostic attention mechanism together with a sparsity loss to encourage the action segments. Followed this framework, a video-level clustering loss was applied in Reference [21] to separate fore-ground and background. Besides, several other methods were proposed by imposing different constraints on attention weights, such as DGAM[22], Bas-Net[43] and TSCN[44]. These methods achieved excellent performance in this field, which shows that fore-ground and background separation is essential. The second type aims at learning more discriminative features by imposing different loss functions. Paul et al[19] proposed a framework named W-TALC consists of Co-Activity Loss is used to encourage class-specific features. Similarly, 3C-Net[12] proposed a Center Loss to force features from the same categories to be as close as possible and features from different categories to be as far as possible. In addition to the above two types, more sophisticated works[45,46] have been proposed in later work. Although attention based methods and learning dis-criminative feature based methods have achieved remarkable progress, a common issue for these methods is that they tend to focus on the most discriminative action segments but ignore trivial action segments, which results in incomplete action localization. To mitigate this issue, the third type works resort to the erasing mechanism to highlight less discriminative segments. For example, Hide-and-Seek[16] proposed to randomly erase input segments during training, which can force the model to discover less discriminative segments. However the erasing strategies cannot be learned in an end-to-end manner, which has many limitations on practical applications, leading to suboptimal performance. In this paper, we focus on the second type and develop relation aware modules within or between videos, which can help learn more robust video features.

2.3 Attention mechanism

Our work is related to the attention mechanism which shines brightly in deep learning. Attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations[47-50]. Similarly, in computer vision, attention mechanisms have been used for image and video recognition, detection, and segmentation[51-54]. To the best of our knowledge, our RANet is the first model using a cross-attention mechanism between videos for the WS-TAL task.

3 Our proposed method

In this section, we introduce the proposed Relation-Aware Network in details. As shown in Figure 2, our RANet mainly consists of four parts: (a) Feature extraction module, (b) Intral-video relation module, (c) Inter-video relation module, (d) Classification & localization module. The details for each module are introduced as follows.

3.1 Feature extraction module

Following recent WS-TAL methods[15,21,22,55], as shown in Figure 2(a), suppose we have a set of training videos and corresponding video-level labels. For each video, we first divide each input video V into 16-frame non-overlapping segments (each segment nearly 0.5s), i.e., $V \in \{V_{\text{RGB}}, V_{\text{flow}}\}$. Then, we feed sampled RGB and Flow segments into the pretrained feature extractor to generate feature vectors $X^{\text{RGB}} \in \mathbb{R}^{d}$ and $X^{\text{flow}} \in \mathbb{R}^{d}$ respectively, where $d = 1024$ is the feature dimension. The video-level label is denoted as $y \in \{0,1,2,\cdots,C\}$ and 0 corresponds to background. The network structure of the RGB and Flow branches is exactly the same and the training is independent of each other. The final prediction result comes from the weighted fusion of two branches. Since the idea of inter-video attention requires us to sample video pairs that contain the same label during the training, we denote the random sampled two videos that have the same label as $X_1 \in \mathbb{R}^{N_1 \times 1024}$ and $X_2 \in \mathbb{R}^{N_2 \times 1024}$.

3.2 Intra-video relation module

Action completeness issue means that a complete action is predicted as multiple sub-actions. In the weakly supervised setting, the lack of fine-grained annotations complicates the completeness modeling. Ignoring the temporal relations of the segments within the video is a major cause for current problem. In our actual observation, a complete action is divided into several parts, but these video parts are usually in a local

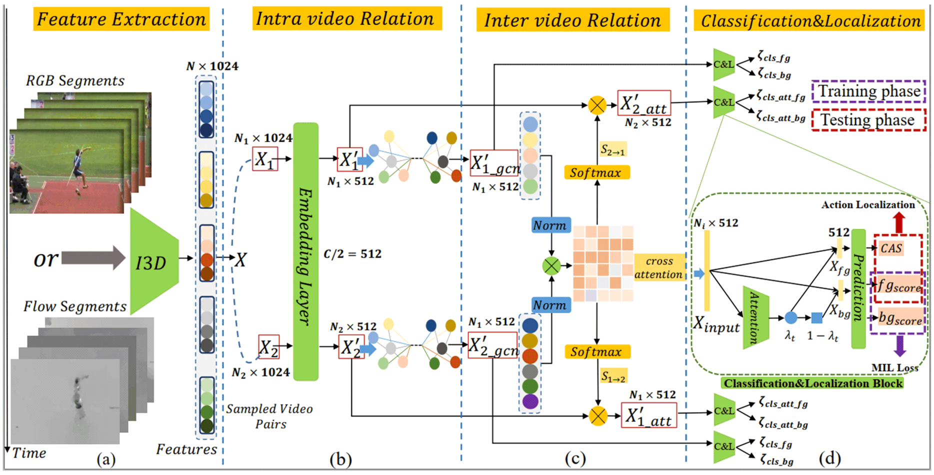


Figure 2. Our RANet consists of 4 components: (a) A feature encoder that extracts video features from input frame/flow sequences. (b) Intra-video relation module for modeling the relation among segments in a video. (c) Inter-video relation module for modeling the relation between segments of two videos. (d) Multiple results are obtained for classification&localization module for video action classification and localization.

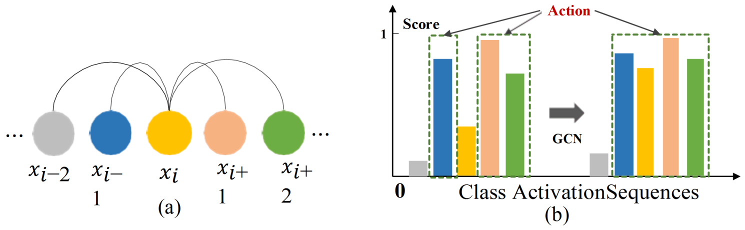


Figure 3. Overall idea of using GCN to solve action completeness issue. We focus on feature segment x_i, which aggregating information from the adjacent segments x_{i+j}, $j \in [-2,+2]$. The action score of nodes x_i becomes higher, and thus the discrete action intervals are connected to form a more complete action during localization stage.

interval. Therefore, a simple and straightforward idea is to use GCN to establish the relation be-tween segments, and update the feature of each segment to aggregate information from the adjacent segments. Figure 3 illustrates our idea in detail, the colored nodes represent action segments while gray nodes represent background segments. Each segment is regarded as a node in the graph, The adjacency matrix A describes the connectivity between nodes, according to the local principle, each node is limited to only connect with its neighborhood of size $P = 2k + 1$. We generally set $k = 2$ for the average action duration of the dataset statistics (average 3s). Furthermore, unlike the self-attention mechanism which learns the correlation between the current segment and the all other segments of the video, using GCN to gather information only from the neighborhood can significantly reduce the computational complexity. The Intra-video Relation Mod-ule is shown in Figure 2 (b), we first feed the extracted video features X_1 and X_2 into an embedding layer to get X'_1 and X'_2.

$$X'_1 = f_{\mathrm{conv}}(X_1;\varphi), X'_2 = f_{\mathrm{conv}}(X_2;\varphi). \qquad (1)$$

The embedding layer is implemented as a convolution layer with a kernel size 1 to reduce the feature dimension from 1024 to 512, whereφ denotes trainable parameters in the convolution layer.

In the intra-video relation module, we view each segment as a node and then utilize graph convolution to model the relation among video segments. In specific. Given an undirected graph with m nodes, a set of edges between nodes, an adjacency matrix $A \in \mathbb{R}^{m\times m}$, and a degree matrix $D_{ij} = \sum_j A_{ij}$. Consider a linear formulation of graph convolution as the multiplication of a graph signal $Y \in \mathbb{R}^{n\times m}$ with a filter with a filter $W \in \mathbb{R}^{n\times c}$:

$$Z = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\,\widetilde{D}^{-\frac{1}{2}}Y^T W). \qquad (2)$$

where $\widetilde{A} = A+I$, I is the identity matrix. $D_{ij} = \sum_j A_{ij}$. σ is a non-linear operation (e. g. , Sigmoid). As a result, the input to a graph convolutional layer is $n \times m$, and the output is $c \times m$. We keep the GCN network's input and output size the same because we only use it for feature updates. Therefore, we set $c = n$ in our experiments. The feature of intra-video updating can be seen in Equation (3):

$$X_{\mathrm{gcn}} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\,\widetilde{D}^{-\frac{1}{2}}X^{*T}W, X^{*} \in \{X'_1, X'_2\}. \quad (3)$$

In this paper, we find that a single GCN layer is enough to model the relation among video segments. We denote the updated intra-video relation feature pairs as $X'_{1\_\mathrm{gcn}} \in \mathbb{R}^{N_1 \times 512}$ and $X'_{2\_\mathrm{gcn}} \in \mathbb{R}^{N_2 \times 512}$ respectively.

3.3 Inter-Video Relation Module

In weakly-supervised setting, only guided by the classification loss, some background frames associated with the action may be activated, which can cause action-context confusion. To solve this problem, the model should be able to capture the underlying discrepancy between action and context. With the observation that the context exhibits notable difference from the action at representation level. In other words, the extracted feature representations for context and action are also different. Inspired by the Image-Text matching task[56], which explored the latent alignments between image regions and words to learn more robust feature representations. Similarly, we can learn the correspondence between videos in the feature space by sampling different videos with the same label during training, and then to update the segment feature so that the network could distinguish action instances from their context well. We adopt such sampling strategy for a potentially important prior knowledge is that the same action contains the same action, but usually contains different context background segments. Specifically, the Inter-Relation Module we proposed embeds the video pairs into the same feature space and requires to reconstruct the segment feature of one video from the other video with the same category label by cross-attention mechanism. As illustrated in Figure 4, by modeling the relation between segments cross video pairs, the network can effectively alleviate the action-context confusion issue in the video. As a result, more accurate video localization results can be obtained.

Figure 2 (c) illustrates the Inter-video Relation Module in detail. The network first normalize the features output from the Intra-video Relation Module because it can stabilize the training and speed up the convergence. And then, in order to establish the relation between all the segments, we use a dot product to measure the similarity matrix S $\in \mathbb{R}^{N_1 \times N_2}$ between videos for convenience.

$$S = \frac{X'_1 X'^{T}_2}{||X'_1|| \, || X'_2 ||}. \quad (4)$$

where N means the feature normalization and T represents the transpose operation. We use the joint similarity matrix S because it can help to update the video segment features cross videos. Unlike using S directly, we use the softmax layer with the parameter β on the similarity matrix to generate the update matrix $S_{1\to2}$ and $S_{2\to1}$:

$$S^{ij}_{1\to2} = \frac{\exp(\beta S^{ij})}{\sum_{j=1}^{N_2}\exp(\beta S^{ij})}, S^{ij}_{2\to1} = \frac{\exp(\beta S^{ij})}{\sum_{i=1}^{N_1}\exp(\beta S^{ij})}. \quad (5)$$

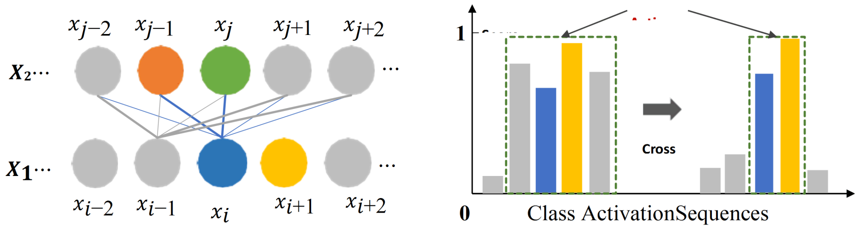
where $S_{1\to2} \in \mathbb{R}^{N_1 \times N_2}$ and $S_{2\epsilon1} \in \mathbb{R}^{N_1 \times N_2}$ and the arrow → indicates the direction of the video relation. We use a softmax layer here because the dot product similarity may not be in the range [0,1], and a softmax operation with the parameter β can enlarge the feature difference for the weight in adjacent matrix. We set $\beta = 10$ in our experiments. Once we get the attention matrix $S_{1\to2}$, $S_{2\to1}$ and the input video segment features X'_1, X'_2, we weight the sum to update the segments features according to Equation (6).

$$X'_{1\_\mathrm{att}} = S_{1\to2}X'_2, X'_{2\_\mathrm{att}} = S_{2\to1}X'_1. \quad (6)$$

where $X'_{1\_\mathrm{att}} \in \mathbb{R}^{N_1 \times 512}$ and $X'_{2\_\mathrm{att}} \in \mathbb{R}^{N_2 \times 512}$. Through this design, the feature update of each segment of the video will utilize the relation information from other videos and the network will pay more attention to the action itself, rather than the high associated background. In addition, we also reuse X_{1_gcn} and X_{2_gcn} as the middle layer of super-vision for better results. All those four updated segment features will been sent into the next module to optimize the entire network or produce predicted results.

3.4 Classification and localization

In this part, we introduce how to train our models, and



(a) (b)

Figure 4. Inter-video relation cross videos. The segment features are updated through the cross-attention mechanism, the action and non-action frames are well separated. There-fore, the predicted localization results are more accurate.

then use them to generate local-ization predictions. The overall overview is shown in Figure 2(d).

Classification. In the weakly-supervised setting, only video-level labels are known. Existing methods simply train a foreground model to respond strongly at some locations within the video, but leave the remaining background frames unmodeled. Different from these models, we follow[21], which explicitly accounts for background frames and sub-stantially improves on weakly-supervised action localization. For each video segment feature X_{input} extracted by the pre-trained network, the Attention Module predicts a frame level attention vector $\lambda = ^{[\lambda_1,\lambda_2,\cdots,\lambda_N]}$, which can be used to pool the frame-level features into a single foreground video level feature representation. The attention layer consists of two fully connected layers, with the nonlinear activation functions ReLU[57] and Sigmoid respectively. The Sigmoid activation function ensures that the weight $\lambda_t \in [0, 1]$. We set an extra small constant ϵ = 1e-6 to prevent the denominator from being zero based on our experience.

$$x_{fg} = \frac{\sum_{t=1}^{N} \lambda_t x_t}{\sum_{t=1}^{N} \lambda_t + \epsilon}. \tag{7}$$

The complement of the attention vector 1 -λ, can also be used to pool segments be-longing to the background into a video-level background representation. Similarity

$$x_{bg} = \frac{\sum_{t=1}^{N} (1 - \lambda_t) x_t}{\sum_{t=1}^{N} (1 - \lambda_t) + \epsilon}. \tag{9}$$

Once we get the foreground feature $X_{fg} \in \mathbb{R}^{N_1 \times 512}$ and the background feature $X_{bg} \in \mathbb{R}^{N_1 \times 512}$, we feed the pooled features to the Prediction Layer which produces a video-level prediction. Two fully-connected layers with ReLU are used for mapping the X_{fg} to $C + 1$ categories, a special category 0 means background. Ω means trainable parameters in the prediction layer.

$$fg_{\text{score}} = f_{\text{prediction}}(X_{fg};\Omega), bg_{\text{score}} = f_{\text{prediction}}(X_{bg};\Omega). \tag{8}$$

We encourage high discriminative capability of the foreground feature X_{fg} and simultaneously punish any discriminative capability of the background feature X_{bg}. We use a Cross-Entropy Loss to measure the difference between the ground truth and the prediction label. The classification loss of foreground and background can be generated for constraint. The total loss function consists of foreground and background losses from intra-video or inter-video relation modules respectively.

$$\mathscr{L}_{\text{total}} = (\mathscr{L}_{cls\_fg} + \alpha\epsilon\mathscr{L}_{cls\_bg}) + (\mathscr{L}_{cls\_att\_fg} + \alpha\epsilon\mathscr{L}_{cls\_att\_bg}). \tag{10}$$

where the hyperparameter α represents the weight between the foreground and the background. We set $\alpha = 0.1$ in our experiments. The total loss is the sum of these four terms.

Localization. We have given detailed descriptions about the design of our proposed RANet, and the remaining issue is how to use the trained network for action localization. In this section, we introduce how to use the trained network for action detection. To generate the Class Activation Sequences (CAS) used for action detection, we send the output feature X_{gen} in Intra-video model of each segment to the Prediction layer.

Given the final CAS, following previous methods[12-14], we use a two-stage method to generate action proposals. First, we set the thresholdτ of the video-level prediction fg_{score}, and discard the categories with the confidence lower than the τ. Then, for each remaining action category, we apply a threshold on the corresponding CAS to generate detection proposals. The score of temporal action segment $[t_{\text{start}}, t_{\text{end}}, c]$ can be obtained via Equation (11).

$$\text{score} = \sum_{t=\text{start}}^{t=\text{end}} \frac{\theta\lambda_t^{\text{RGB}}\,\text{CAS}^{\text{RGB}}(t,c) + (1-\theta)\lambda_t^{\text{flow}}\,\text{CAS}^{\text{flow}}(t,c)}{t_{\text{end}} - t_{\text{start}} + 1}. \tag{11}$$

where θ denotes the weight of the RGB and flow branches. According to the score, a certain threshold is set to take the generated continuous segment as the final action localization result. Finally, the Non-Maximal Suppression is used to fuse the “flow” stream detections and the RGB stream detections. We set $\theta = 0.3$ in this work. See experiments for more details.

4 Experiments

In this section, we experimentally evaluate the proposed framework for activity localization from weakly labeled videos. We first discuss the datasets we used, followed by the implementation details, ablation learning and some visualized results.

Table 1. Results on THUMOS14 testing set. Comparison with state-of-the-arts on THU-MOS14, we report mAP values at IoU thresholds 0.1：0.1：0.9. In order to demonstrate the superiority of our proposed method, recent works in both fully-supervised and weakly-supervised setting are reported.

Methods	Supervised	mAP@ IoU								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
S-CNN[38]	Full	47.7	43.5	36.3	28.7	19.0	10.3	5.3	−	−
SSN[40]	Full	66.0	59.4	51.9	41.0	29.8	−	−	−	−
BSN̄[58]	Full	−	−	53.5	45.0	36.9	−13.1	−28.4	−20.0	−
P-GCN[59]	Full	**69.5**	**67.8**	**63.6**	**57.8**	**49.1**	−	−	−	−
CDC̄[60]	Full	−	−	40.1	29.4	23.3	−	13.1	7.9	−
TALNet[37]	Full	−59.8	−57.1	53.2	48.5	42.8	−	−	−	−
TPC[61]	Full	−	−	44.1	37.1	28.2	20.6	−12.7	−	−
R-C3D[39]	Full	−54.5	−51.5	44.8	35.6	28.9	−	−	−	−
UntrimmedNet[11]	Weak	44.4	37.7	28.2	21.1	13.7	−	−	−	−
Hide−and−seek[16]	Weak	36.4	27.8	19.5	12.7	6.8	−	−	−	−
AutoLoc[55]	Weak	−	−	35.8	29.0	21.2	−13.4	5.8	−	−
CleanNet[62]	Weak	−	−	37.0	30.9	23.9	13.9	7.1	−	−
STPN[13]	Weak	−52.0	−44.7	35.5	25.8	16.9	9.9	4.3	−	−
MANN[63]	Weak	59.8	50.8	41.1	30.6	20.3	12.0	6.9	2.6	0.2
W−TALC[19]	Weak	55.2	49.6	40.1	31.1	22.8	−	7.6	−	−
3C−Net[12]	Weak	56.8	49.8	40.9	32.3	24.6	−	7.7	−	−
Liu et al[15]	Weak	57.4	50.8	41.2	32.1	23.1	−15.0	7.0	−	−
Nguyen et al[21]	Weak	60.4	56.0	46.6	37.5	26.8	17.6	9.0	3.3	0.4
DGAM[22]	Weak	60.0	54.2	46.8	38.2	28.8	**19.8**	**11.4**	**3.6**	**0.4**
Ours	Weak	**66.4**	**60.05**	**51.9**	**41.1**	**30.7**	**19.8**	10.8	2.9	0.37

4.1 Datasets and evaluation

Datasets. We evaluate our RANet on three popular action localization benchmark datasets, THUMOS14[33] and ActivityNet 1.2/1.3[34]. Both datasets are untrimmed, meaning the videos include frames that contain no target actions, and we do not exploit the temporal annotations for training. THUMOS 14 has video-level annotations of 101 action classes in its training, validation, testing sets, and temporal annotations for a subset of videos the validation and testing sets for 20 classes. The dataset is challenging as some videos are relatively long (up to 26 minutes) and contain multiple action instances. The length of action varies significantly, from less than a second to minutes. ActivityNet 1.2 has 4819 training, 2383 validation and 2480 testing videos from 100 activity categories. Note that the test set annotations for this dataset are withheld. There is an average of 1.5 activity instances per video. As in References [19,38], we use the training set to train and the validation set to test our approach. ActivityNet 1.3 offers a larger benchmark for complex action localization in untrimmed videos which has 10024 videos for training, 4926 for validation, and 5044 for testing with 200 activity classes.

Evaluation. We follow the standard evaluation protocol based on mean average precision (mAP) values at several different levels of intersection over union (IoU) thresholds. The evaluation is conducted using the benchmarking code for the temporal action localization task provided by ActivityNet.

4.2 Implementation Details

We use the Kinetics pretrained two-stream I3D network[30] to extract video features. The inputs to the two-stream are stacks of 16 (RGB or Flow) frame chunks. The output is passed through a 3D average pooling layer to obtain features of dimension 1024 each from two streams. Specifically, we apply the TV-L1[64] algorithm to extract optical flow. Our feature extraction module is fixed during the training time, which will make our model more lightweight, require less training time, and be more friendly to the GPU memory size. Our RANet is trained using the Adam[65] optimizer with 1e-4 learning rate for THUMOUS 14 and 1e-3 learning rate for ActivityNet 1.2/1.3. We set 1e-4 as the weight decay for both datasets. For the hyperparameter a , we find that a needs to be small enough so that the network is driven mostly by the foreground loss. We simply set $\alpha = 0.1$in our experiments. Besides, we generally set $k = 2$for the average action duration of the dataset statistics. All experiments are trained on RTX 2080Ti GPU using Pytorch with version 1.2.

4.3 Comparison with the state-of-the-art methods

We perform a quantitative analysis of our framework by comparing with current state-of-the-art approaches at several IoU thresholds for the task of activity localization. The results on THUMOS14 and ActivityNet are shown in Table 1, Table 2 and Table 3 respectively.

Table 1 shows the quantitative results on THUMOS14. We compare our RANet with existing approaches in both weakly-supervised and fully-supervised action localization. Our method outperforms all other recent weakly-supervised methods, and improves the mAP@0.5 from previous state-of-the-art 28.8% to 30.7%. Even with a much lower level of supervision, our method shows the least gap regarding the latest fully-supervised methods. Furthermore, it can be noticed that our method even outperforms several fully-supervised methods at some IoU thresholds.

We also evaluate our RANet on ActivityNet 1.2 in Table 2. We see that our method outperforms all other weakly supervised approaches. Moreover, similar to THUMOS14, our method significantly outperforms existing weakly-supervised approaches while maintaining competitive with other fully-supervised methods.

Experimental results on ActivityNet 1.3 are shown in Table 3 to compare our method with more baseline methods. Our model outperforms all weakly-supervised methods, following the fully-supervised method with a small gap.

Table 2. Results on ActivityNet 1.2 testing set. Comparison with state-of-the-arts on ActivityNet 1.2, we report mAP values at IoU thresholds 0.5 : 0.05 : 0.95. In order to demon-strate the superiority of our proposed method, recent works in both fully-supervised and weakly-supervised setting are reported.

Methods	Supervised	mAP@ IoU									
		0.5	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
SSN[40]	Full	41.3	**38.8**	**35.9**	**32.9**	**30.4**	**27.0**	**22.2**	**18.2**	**13.2**	**6.1**
AutoLoc[55]	Weak	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3
W-TALC[19]	Weak	37.0	33.5	30.4	25.7	14.6	12.7	10.0	7.0	4.2	1.5
3C-Net[12]	Weak	35.4	-	-	-	22.9	-	-	-	8.5	-
Liu et al[15]	Weak	36.8	-	-	-	22.9	-	-	-	8.5	-
CleanNet[62]	Weak	37.1	-33.4	-29.9	-26.7	23.4	-20.3	-17.2	9.2	5.0	-
DGAM[22]	Weak	41.0	37.5	33.5	30.1	26.9	23.5	19.8	**15.5**	**10.8**	**5.3**
Ours	Weak	**42.3**	**38.7**	**34.5**	**31.2**	**27.3**	**24.0**	**20.1**	15.4	10.1	5.2

Table 3. Comparison of our method with state-of-the-art W-TAL methods on the ActivityNet v 1.3 validation set. The Avg column indicates the average mAP at IoU thresholds 0.5 : 0.05 : 0.95.

Methods	Supervised	mAP@ IoU			
		0.5	0.75	0.95	Avg
TAL-Net	Full	38.2	18.3	1.3	20.2
R-C3D	Full	26.8	-	-	12.7
BSN	Full	**52.5**	**33.5**	**8.9**	**34.3**
Liu et al[15]	Weak	34.0	**20.9**	5.7	21.2
STPN[13]	Weak	29.3	16.9	2.7	-
Nguyen et al[21]	Weak	36.4	19.2	2.9	-21.7
Ours	Weak	**37.5**	20.7	**5.8**	**21.8**

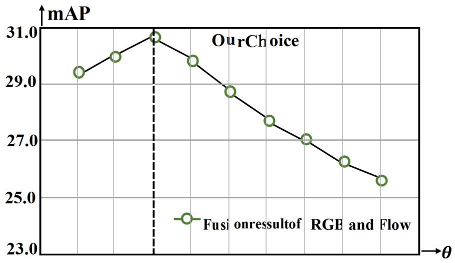


Figure 5. Differentθfor two−stream fusion.

4.4 Ablation learning

We first study the impact of two-stream fusion. We carry out the experimental results with different score fusing weight ($\theta = 0.1, 0.2, \cdots, 0.9$) on THUMOS14 and summarize them in Figure 5. We simply set $\theta = 0.3$ as default in all experiments for its better performance. Then we conduct several ablation experiments on THUMOS14 to investigate the effectiveness of different components in our proposed RANet in Table 5.

Baseline. Our baseline is derived from[21] and contains only foreground and back-ground modeling. No relation modules have been considered.

Intra-video relation module. We add an auxiliary branch for modeling the relation within a video into the baseline, i. e., Intra-video Relation module, which leads to an impressive improvement in a high IoU. Compared to Baseline, our intra-video module gets 1.7% gain in IoU = 0.5. We conjecture that it is because the network is trained to build a bridge for feature update between video segments, and bring out completeness to action especially in IoU 0.5.

Inter-video relation module. We evaluate a variant with only the Inter-video Re-lation Module branch in order to verify its importance in the network. It can be seen from the experimental result that the Inter branch has achieved significant accuracy improvements from the baseline at all IoU thresholds. The significant improvement in accuracy indicates the utility of the proposed Inter-video Relation module, which aims to separate the action from the context, thus producing more accurate action location results.

RANet. By employing both branches and jointly training them with contrasting objectives, RANet learns the relation within or between video segments and shows the best performance from the others. Compared to the baseline, our RANet gets a remark-able gain in both accuracy and completeness. Under the IoU threshold 0.5, our method improves mAP on THUMOS14 from 25.7% to 30.7%. The excellent performance of action detection indicates that the action predicted by our model is more complete and closer to the Ground Truth.

Table 4. Effect of each component of RANet on the action localization perfor-mance on THUMOS14.

Ours	Intra	Inter	mAP@ IoU		
			0.3	0.5	0.7
RANet	×	×	50.3	25.7	7.3
	√	×	50.6	27.4	8.4
	×	√	51.1	30.1	8.6
	√	√	**51.9**	**30.8**	**9.8**

4.5 Discussion

In this section, we discuss and compare some of our related works, which also use graph convolution and attention mechanisms, or deal with action completeness and action-context confusion issues for action localization task. In addition, we have also added some discussion about the proposed two relation modules.

Graph methods. We observe that two recent explorations[59,66] also refer to graph convolution methods in action localization task. However, they are all used for full supervision, which mainly to model the proposal-proposal interactions and boost the temporal action localization performance. The method in this paper is based on weakly-supervised and can only model relation between segments. Besides, another significant difference is that both of them are related to all proposals to establish the edge of graph convolution, while ours are only based on partial adjacent segments to establish the connection.

Attention methods. Attention mechanism is also widely applied to current video tasks. Chen et al[67] has proposed a relation focus module to enhance expression by extracting useful information from other proposals. However it learns a proposal-wise attention map to capture relative information within a video. In contrast to such approaches, our approach learns a segments-wise attention between different videos, which is inspired by the image-text matching task, using a cross-attention mechanism that embeds video pairs into same feature space, learning the potential alignment of action or background. The goal is to solve the problem of action-context confusion.

Completeness modeling & action-context separation methods. There are two under explored problems in WS-TAL task, namely action completeness modeling and action-context confusion. Some of existing approaches presented have begun to explicitly study these two challenges. As far as we know, these studies are quite different from method proposed in this paper. For example, references[68,69] inspired by the adversarial erasing mechanism to highlight less discriminative segments, thus making the actions

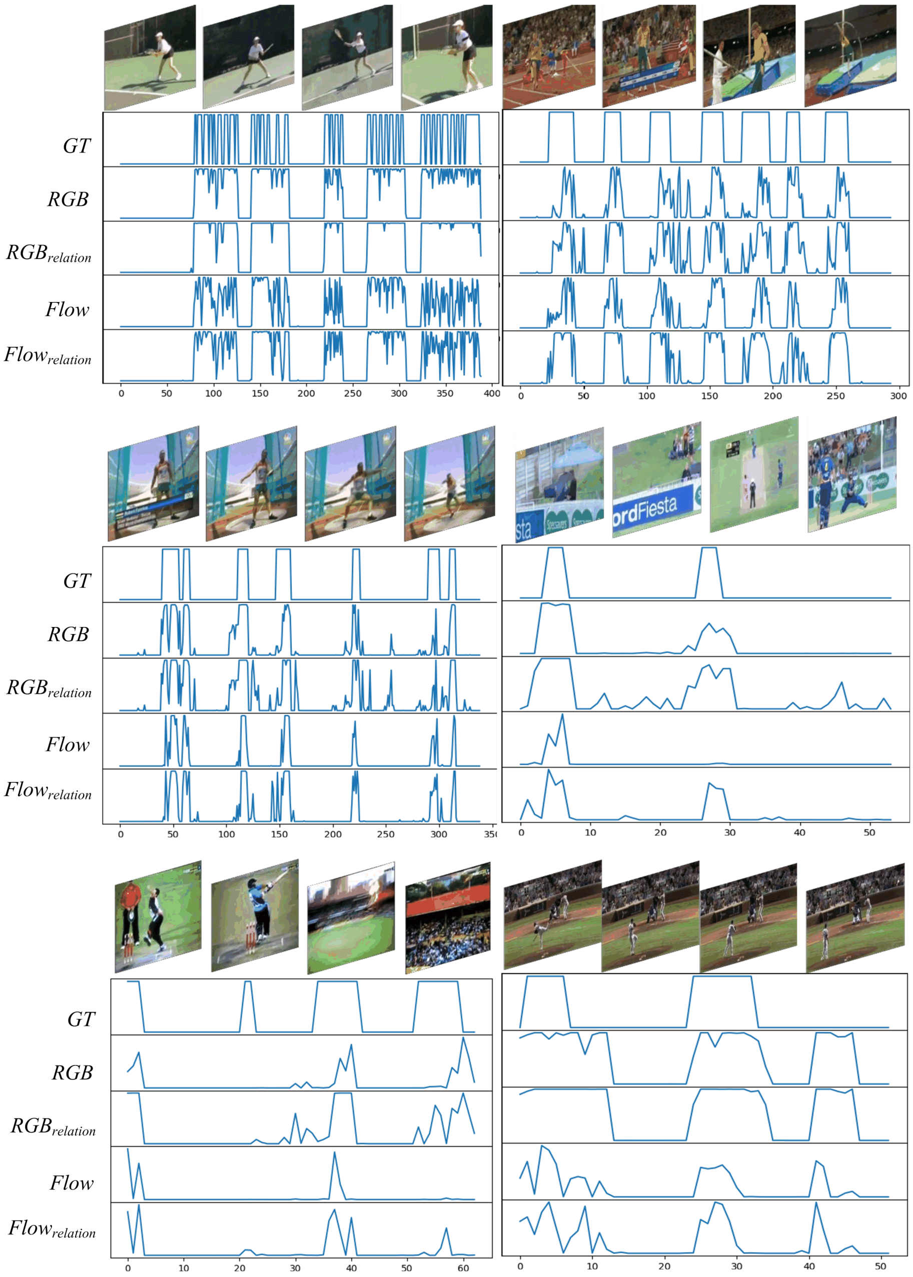


Figure 6. Qualitative results on the THUMOS14 testing set. The six rows in each example are input video, ground truth action instance, RGB stream, $\text{RGB}_{\text{relation}}$, flow stream, and $\text{flow}_{\text{relation}}$ from the model trained with only video level labels.

prediction more complete. However the erasing strategies cannot be learned in an end-to-end manner, which limits its performance. DGAM[22] devotes to solve the action-context confusion issue and devises a conditional variation auto-encoder (CVAE) to construct different feature distributions conditioned on different attentions. However it treats each segment as an individual, which still ignores the relation cross videos.

Intra-video VS inter-video. In this work, the proposed two relation modules correspond to solving two different challenges respectively, making the predicted action localization more close to the ground truth without the frame-level or segment-level fine annotations. But they' re going in different directions. The intra-video relation module pays more attention to forming the discrete action parts into a complete action, which is a process from discrete to complete, while the inter-video relation module aims to separate the action from the highly relevant context, which is a process from redundant to compact. Therefore, they are complementary to each other. Experiments also show that the excellent performance can be obtained by combining the two modules together. In addition, unlike most existing methods that only solve one of the issues in WS-TAL, the RANet proposed in this paper is an end-to-end network which could solve both problems simultaneously.

4.6 Visualized results

In Figure 6, we present some visualization results on THUMOS14 testing set which show that our RANet can effectively learn to localize temporal action instances without any direct temporal boundary information during training. The first row of each sub-figure shows video frames uniformed sampled frames from a video. The second rows show Ground Truth. Rows three to six show different action duration predictions, respectively. Comparison with ground-truth, we can find that each instance's temporal boundary is close to the ground-truth annotation. Compared to a separate RGB or Flow branch. When considering the relations, the network is more accurate in the prediction results, whether in the time span or confidence. The $\mathrm{RGB}_{\mathrm{relation}}$ or $\mathrm{Flow}_{\mathrm{relation}}$ visualization results indicate the robustness of our model.

5 Conclusions

In this work, we first introduce the two existing problems in the current weakly-supervised action localization methods, which are called action completeness modeling and action-context confusion. To tackle the two issues, we proposed a RANet consisting of an Intra-video Relation Module and an Inter-video Relation Module. Our experiments on three challenging datasets demonstrate that the proposed method achieves state-of-the-art results in the WS-TAL task. We hope that this work will foster further research in video localization.

Acknowledgments

This work was partially supported by National Defense Basic Scientific Research Program (JCKY2020903B002), Strategic Priority Research Program of Chinese Academy of Sciences (No. XDC02050500), National Nature Science Foundation of China (Grant 62022078, 62121002), and Youth Innovation Promotion Association CAS 2018166.

Conflict of interest

The authors declare no conflict of interest.

Author information

Zhan Yongkun is currently a Master student in the Laboratory for Future Networks, Department of of Information Science and Technology under the supervision of Prof. Zhang Tianzhu at University of Science and Technology of China (USTC). His research mainly focuses on Action Recognition and Localization.

Yang Wenfei is currently a PhD student in the Laboratory for Future Networks, Department of Information Science and Technology under the tutelage of Prof. Zhang Tianzhu at University of Science and Technology of China (USTC). His research interests focus on Weakly-supervised Learning.

Zhang Tianzhu is currently a Professor at University of Science and Technology of China (USTC). His research interest includes pattern recognition, computer vision, multimedia analysis, and machine learning. He has authored or co-authored over 70 journal and conference papers in these areas, including over 40 IEEE/ACM Transactions papers (TPAMI/TIP/IJCV) and over 30 top-tier computer vision conference papers (ICCV/CVPR/ECCV). He has served as the Area Chair for ECCV 2020, ACM Multimedia 2020, CVPR 2020, ICCV 2019, and ACM Multimedia 2019, the Associate Editor for IEEE T-CSVT and Neurocomputing, and the publicity chair for ICIMCS 2015 and ACM Multimedia Asia 2019.

References

[1] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding. Proceedings of the International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 7083–7093.

[2] Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering. Proceedings of the International Conference on Computer Vision. Long Beach, USA: IEEE, 2015: 2425–2433.

[3] Akti S, Tataroglu G A, Ekenel H K. Vision-based fight detection from surveillance cameras. Ninth International Conference on Image Processing Theory, Tools and Applications. Vancouver, Canada: IEEE, 2019: 1–6.

[4] Lee Y J, Ghosh J, Grauman K. Discovering important people and objects for egocentric video summarization. IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012: 1346–1353.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 91–99.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014: 580–587.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 779–788.

[8] Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1440–1448.

[9] Dai X, Singh B, Zhang G, et al. Temporal context network for activity localization in videos. Proceedings of the International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 5793–5802.

[10] Gao J, Yang Z, Nevatia R. Cascaded boundary regression for temporal action detection. 2017, arXiv:1705.01180 .

[11] Wang L, Xiong Y, Lin D, et al. Untrimmednets for weakly supervised action recognition and detection. Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 4325–4334.

[12] Narayan S, Cholakkal H, Khan F S, et al. 3C-Net: Category count and center loss for weakly-supervised action localization. Proceedings of the International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 8679–8687.

[13] Nguyen P, Liu T, Prasad G, et al. Weakly supervised action localization by sparse temporal pooling network. Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 6752–6761.

[14] Singh G, Saha S, Sapienza M, et al. Online real-time multiple spatio temporal action localisation and prediction. Proceedings of the International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 3637–3646.

[15] Liu D，Jiang T，Wang Y. Completeness modeling and context separation for weakly supervised temporal action localization. Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE，2019：1298- 1307.

[16] Singh K K，Lee Y J. Hide-and-seek：Forcing a network to be meticulous for weakly-supervised object and action localization. International Conference on Computer Vision. Venice，Italy：IEEE，2017：3544-3553.

[17] Laptev I，Marszalek M，Schmid C，et al. Learning realistic human actions from movies. Proceedings of the Conference on Computer Vision and Pattern Recognition. Anchorage，USA：IEEE，2008：1-8.

[18] Cholakkal H，Sun G，Khan F S，et al. Object counting and instance segmentation with image-level supervision. Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach，USA，2019：12397 - 12405.

[19] Paul S，Roy S，Roy-Chowdhury A K. W-TALC：Weakly-supervised temporal activity localization and classification. Proceedings of the European Con-ference on Computer Vision. 2018：563-579.

[20] Bojanowski P，Bach F，Laptev I，et al. Finding actors and actions in movies. Proceedings of the International Conference on Computer Vision. Sydney，Australia：IEEE，2013：2280-2287.

[21] Nguyen P X，Ramanan D，Fowlkes C C. Weakly-supervised action localization with background modeling. Proceedings of the International Conference on Computer Vision. Seoul，South Korea：IEEE，2019：5502-5511.

[22] Shi B，Dai Q，Mu Y，et al. Weakly-supervised action localization by generative attention modeling. Proceedings of the Conference on Computer Vision and Pattern Recognition. New York：IEEE，2020：1009-1019.

[23] Hamilton W，Ying Z，Leskovec J. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems. 2017：1025-1035.

[24] Kläser A，Marszałek M，Schmid C. A spatio-temporal descriptor based on 3D-gradients. Proceedings of the 2008 British Machine Vision Conference. [2021-03-26]，http：//citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 167. 974&rep=rep1&type=pdf.

[25] Wang H，Schmid C. Action recognition with improved trajectories. Proceedings of the International Conference on Computer Vision. Sydney，Australia：IEEE，2013：3551-3558.

[26] Dalal N，Triggs B，Schmid C. Human detection using oriented histograms of flow and appearance. European Conference on Computer Vision. Springer，2006：428 - 441.

[27] Simonyan K，Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems. 2014：568-576.

[28] Tran D，Bourdev L，Fergus R，et al. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the International Conference on Computer Vision. Santiago，Chile：IEEE，2015：4489-4497.

[29] Qiu Z，Yao T，Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. Proceedings of the International Conference on Computer Vision. Venice，Italy：IEEE，2017：5533-5541.

[30] Carreira J，Zisserman A. Quo vadis，action recognition? a new model and the kinetics dataset. Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE，2017：6299-6308.

[31] Wang L，Xiong Y，Wang Z，et al. Temporal segment networks：Towards good practices for deep action recognition. European Conference on Computer Vision. Springer，2016：20-36.

[32] Zhou B，Andonian A，Oliva A，et al. Temporal relational reasoning in videos. Proceedings of the European Conference on Computer Vision. Springer，2018：803 - 818.

[33] Idrees H，Zamir A R，Jiang Y G，et al. The THUMOS challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding，2017，155：1-23.

[34] Caba Heilbron F，Escorcia V，Ghanem B，et al. Activitynet：A large-scale video benchmark for human activity understanding. Proceedings of the Conference on Computer Vision And Pattern Recognition. Boston，USA：IEEE，2015：961-970.

[35] Sigurdsson G A，Varol G，Wang X，et al. Hollywood in homes：Crowd-sourcing data collection for activity understanding. European Conference on Computer Vision. Springer，2016：510-526.

[36] Gu C，Sun C，Ross D A，et al. AVA：A video dataset of spatio-temporally localized atomic visual actions. Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE，2018：6047-6056.

[37] Chao Y W，Vijayanarasimhan S，Seybold B，et al. Rethinking the faster R-CNN architecture for temporal action localization. Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE，2018：1130-1139.

[38] Shou Z，Wang D，Chang S F. Temporal action localization in untrimmed videos via multi-stage CNNs. Proceedings of the Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE，2016：1049-1058.

[39] Xu H，Das A，Saenko K. R-C3D：Region convolutional 3D network for temporal activity detection. Proceedings of the International Conference On Computer Vision. Venice，Italy：IEEE，2017：5783-5792.

[40] Zhao Y，Xiong Y，Wang L，et al. Temporal action detection with structured segment networks. Proceedings of the IEEE International Conference on Computer Vision. Springer，2017：2914-2923.

[41] Lin T，Zhao X，Shou Z. Single shot temporal action detection. Proceedings of the 25th International Conference on Multimedia. Bucharest，Romania：ACM，2017：988 - 996.

[42] Zhang D，Dai X，Wang X，et al. S3D：Single shot multi-span detector via fully 3D convolutional networks. 2018，arXiv：1807.08069，.

[43] Lee P，Uh Y，Byun H. Background suppression network for weakly-supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence. 2020，34：11320-11327.

[44] Zhai Y，Wang L，Tang W，et al. Two-stream consensus network for weakly-supervised temporal action localization. European Conference on Computer Vision. Springer，2020：37-54.

[45] Huang L，Huang Y，Ouyang W，et al. Relational prototypical network for weakly supervised temporal action localization. Proceedings of the AAAI Conference on Artificial Intelligence. 2020，34：11053-11060.

[46] Gong G，Wang X，Mu Y，et al. Learning temporal co-attention models for unsupervised video action localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：9819-9828.

[47] Cheng J，Dong L，Lapata M. Long short-term memory-networks for machine reading. 2016，arXiv：1601.06733 .

[48] Parikh A P，Täckström O，Das D，et al. A decomposable

attention model for natural language inference. 2016, arXiv:1606.01933 .

[49] Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization. 2017, arXiv:1705.04304 .

[50] Davenport T H, Beck J C. The Attention Economy. Harvard Bus. SC, 2001.

[51] Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7794-7803.

[52] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.

[53] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision. Munich, Germany: ACM, 2018: 3-19.

[54] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: ACM, 2019: 3146-3154.

[55] Shou Z, Gao H, Zhang L, et al. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. Proceedings of the European Conference on Computer Vision. Munich, Germany: ACM, 2018: 154-171.

[56] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching. Proceedings of the European Conference on Computer Vision. Munich, Germany: ACM, 2018: 201-216.

[57] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Lauderdale, USA: IEEE, 2011: 315-323.

[58] Lin T, Zhao X, Su H, et al. BSN: Boundary sensitive network for temporal action proposal generation. Proceedings of the European Conference on Computer Vision. Springer, 2018: 3-19.

[59] Zeng R, Huang W, Tan M, et al. Graph convolutional networks for temporal action localization. Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea, 2019: 7094-7103.

[60] Shou Z, Chan J, Zareian A, et al. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 5734-5743.

[61] Yang K, Qiao P, Li D, et al. Exploring temporal preservation networks for precise temporal action localization. 2017, arXiv:1708.03280.

[62] Liu Z, Wang L, Zhang Q, et al. Weakly supervised temporal action localization through contrast based evaluation networks. Proceedings of the International Conference on Computer Vision. Seoul, South, Korea: IEEE, 2019: 3899-3908.

[63] Yuan Y, Lyu Y, Shen X, et al. Marginalized average attentional network for weakly-supervised learning. 2019, arXiv:1905.08586.

[64] Wedel A, Pock T, Zach C, et al. An Improved Algorithm for TV-L^1 optical flow. Statistical and geometrical approaches to visual motion analysis. Springer, 2009: 23-45.

[65] Kingma D P, Ba J. Adam: A method for stochastic optimization. 2014, arXiv:1412.6980.

[66] Xu M, Zhao C, Rojas D S, et al. G-TAD: Sub-graph localization for temporal action detection. Proceedings of the Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10156-10165.

[67] Chen P, Gan C, Shen G, et al. Relation attention for temporal action localization. IEEE Transactions on Multimedia, 2019, 22(10): 2723-2733.

[68] Zhong J X, Li N, Kong W, et al. Step-by-step erasion, one-by-one collection: A weakly supervised temporal action detector. Proceedings of the 26th International Conference on Multimedia. ACM, 2018: 35-44.

[69] Zeng R, Gan C, Chen P, et al. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. IEEE Transactions on Image Processing, 2019, 28(12):5797-5808.

基于关系建模的弱监督时序动作定位

占永昆,杨文飞,张天柱*

中国科学技术大学未来网络实验室,安徽合肥 230027

*通讯作者. E-mail:tzzhang@ustc.edu.cn

摘要: 时序动作定位因其广泛的实际应用成为重要且具有挑战性的方向. 由于全监督定位方法需要大量的人力对长视频进行视频帧或视频片段级别的细腻标注,近些年来,弱监督学习受到了越来越多的关注. 弱监督动作定位在训练阶段只需提供视频级别类别标签,即可定位出视频中动作的区间位置. 然而,大多数现存的方法往往只对独立的视频片段进行分类损失约束,而忽略了这些视频片段之间的关系. 本文提出一种新的关系感知网络实现了基于弱监督的行为时序定位. 通过考虑对视频内和视频间的片段进行关系建模,从而学习出更加鲁棒的视频动作定位特征表示. 具体来说,视频内关系模块的目的是使得网络预测出更加完整的动作,而视频间关系模块的目是将动作从高度依赖的背景中分离出来. 通过在 THUOUS14, ActivityNet1.2/1.3 等三个公共基准定位数据集上进行实验,与最新的方法比,我们提出的方法取得了更好的结果.

关键词: 时序动作定位;弱监督学习;关系建模