Pattern Recognition 129 (2022) 108725

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Dual relation network for temporal action localization

Kun Xia[a], Le Wang[a,*], Sanping Zhou[a], Gang Hua[b], Wei Tang[c]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] Wormpex AI Research, Bellevue, WA 98004, USA
[c] Department of Computer Science, University of Illinois, Chicago, IL 60607, USA

ARTICLE INFO

Article history:
Received 30 November 2021
Revised 2 April 2022
Accepted 19 April 2022
Available online 22 April 2022

Keywords:
Temporal action localization
Relation reasoning

ABSTRACT

Temporal action localization is a challenging task for video understanding. Most previous methods process each proposal independently and neglect the reasoning of proposal-proposal and proposal-context relations. We argue that the supplementary information obtained by exploiting these relations can enhance the proposal representation and further boost the action localization. To this end, we propose a dual relation network to model both proposal-proposal and proposal-context relations. Concretely, a proposal-proposal relation module is leveraged to learn discriminative supplementary information from relevant proposals, which allows the network to model their interaction based on appearance and geometric similarities. Meanwhile, a proposal-context relation module is employed to mine contextual clues by adaptively learning from the global context outside of region-based proposals. They effectively leverage the inherent correlation between actions and the long-term dependency with videos for high-quality proposal refinement. As a result, the proposed framework enables the model to distinguish similar action instances and locate temporal boundaries more precisely. Extensive experiments on the THUMOS14 dataset and ActivityNet v1.3 dataset demonstrate that the proposed method significantly outperforms recent state-of-the-art methods.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Temporal Action Localization (TAL) aims to localize the temporal starts and ends of some specific action categories in an untrimmed video. It serves as a fundamental tool for several practical applications such as intelligent surveillance [1–3], video summarization [4,5], and action retrieval [6–8]. Therefore, it has received widespread attention from academia and industry in recent years. This task is very challenging due to the complex spatio-temporal backgrounds, ambiguous temporal boundaries, and large variations in person appearance, camera viewpoint and action duration.

Recent TAL methods [9–11] mainly adopt a two-stage pipeline and have significantly pushed forward the state-of-the-art performance. Temporal action proposals are first generated via top-down anchors [9,12] or bottom-up mechanisms [10,13,14]. Then each proposal is classified to an action category and regressed towards more precise action boundaries.

Despite the recent progress, most previous methods are limited in two aspects: (1) they process each action proposal independently of the other proposals, and (2) context in a video is ignored to a large extent. Here the "context" refers to the temporal region outside an action instance. Different from action proposals, context may not correspond to any semantic categories. We argue that relation reasoning based on the interactions between proposals and their dependency on context is critical for precise temporal action localization.

As illustrated in Fig. 1, action proposals are usually imperfect because they may cover much background or incomplete action instances. Therefore, leveraging information beyond each individual action proposal is necessary to improve action localization performance. Modeling the relation between relevant proposals can provide supportive cues to discriminate action categories exhibiting similar appearances or motion patterns because some action categories co-occur more frequently than others. Processing relevant proposals jointly also provides a chance to suppress redundant predictions and false positives. In addition, proposal-context relations are complementary to proposal-proposal relations. On the one hand, the context contains scene information that is useful for action classification but may be omitted in action proposals. For example, frames of a swimming pool will reinforce our belief that the action is more likely to be *diving* than *gymnastics*. On the other hand, the context includes cues, *e.g.*, scene switching, that can help localize ambiguous temporal boundaries.

* Corresponding author.
E-mail address: lewang@xjtu.edu.cn (L. Wang).

https://doi.org/10.1016/j.patcog.2022.108725
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

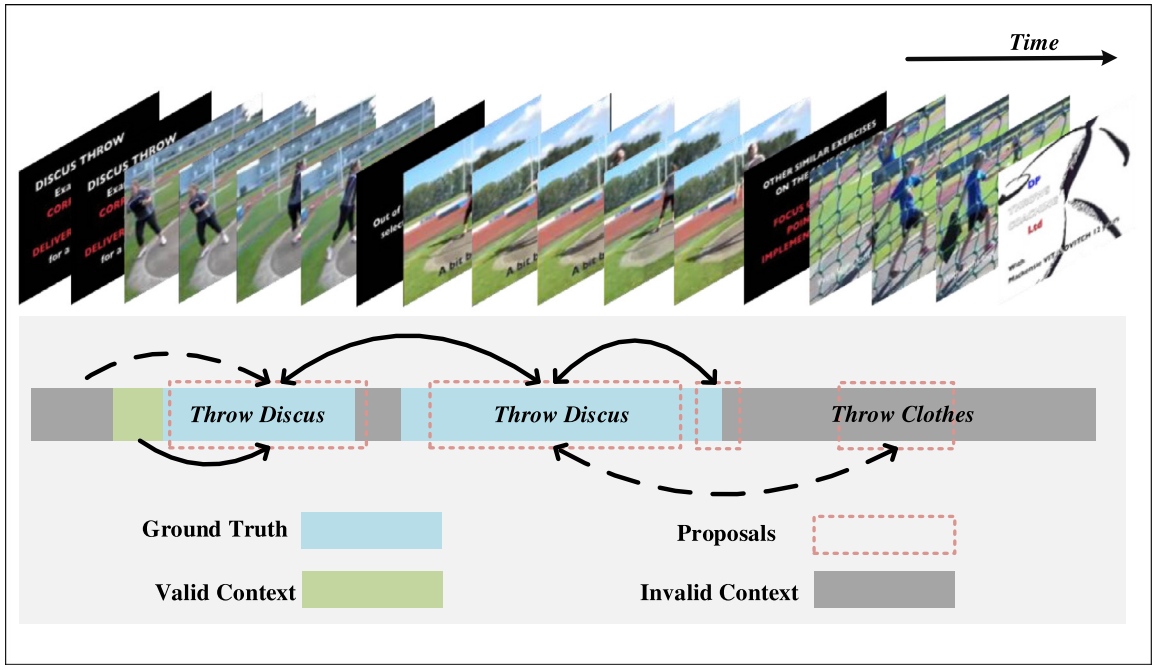

Fig. 1. Illustration of our motivation. The relation between action proposals within a video provides useful clues to discriminate similar action categories and helps suppress redundant action instances and false positives. An untrimmed video contains relevant contextual information for proposal classification or regression and irrelevant context information that adversely affects network learning.

To address the aforementioned issues, we introduce an architecture, termed *dual relation network*, for TAL. It consists of two core components, *i.e.*, a Proposal-Proposal Relation Module (PPRM) and a Proposal-Context Relation Module (PCRM). PPRM performs relation reasoning based on the interactions between proposals. Specifically, it enriches the features of each proposal by taking into account the features of its relevant proposals throughout the whole video. The relatedness between a pair of proposals is measured by both their appearance and geometry similarities. PCRM performs relation reasoning based on the dependency of each proposal on the global context. Instead of simply extending the temporal window of each proposal, PCRM adaptively selects relevant context information from the whole video to capture both short-term and long-term dependency relations between proposals and context. Both modules can be easily incorporated into prior TAL methods and trained end-to-end via standard classification and localization losses.

Experimental results on two benchmark datasets, *i.e.*, THUMOS14 [15] and ActivityNet v1.3 [16], demonstrate that our method outperforms recent state-of-the-art methods. Our quantitative and qualitative ablation studies show that the proposed two relation modules effectively contribute to suppressing false positives and improving temporal localization accuracy.

The main contributions of this paper are summarized as follows.

- We propose a dual relation network for TAL. To the best of our knowledge, this is the first work that unifies proposal-proposal relation reasoning and proposal-context relation reasoning to facilitate TAL. Experimental evaluations show that the two relations are complimentary and critical.
- We propose a proposal-proposal relation module (PPRM), which can effectively perform relation reasoning among relevant proposals based on their appearance and geometry similarities.
- We propose a proposal-context relation module (PCRM), which can adaptively aggregate both short-term and long-term context from the whole video to enrich the proposal features.
- Our method achieves state-of-the-art performance on THUMOS14 and ActivityNet v1.3 datasets.

This paper is organized as follows. In Section 2, we briefly review the related work. Section 3 presents the framework of the proposed method. The experiments are presented in Section 4. Finally, we conclude in Section 5.

2. Related work

2.1. Temporal action localization

The objective of TAL is to identify the temporal boundaries of action instances from an untrimmed video. It is expected that TAL results can cover ground truth action instances under certain temporal intersection-over-union (tIoU) thresholds with high recall as well as high precision. Current mainstream methods [12–14] adopt a two-stage pipeline, where a set of initial action proposals are generated, then classified and refined to more precise temporal locations.

Current proposal generation methods can be roughly divided into two categories, namely top-down and bottom-up methods. Top-down methods [9,12] generate proposals with predefined anchors, but they lack the ability to precisely determine the temporal boundaries of proposals or generate proposals with flexible duration. Bottom-up methods [13,14] generate action proposals with frame-level probability sequences, *i.e.*, start, actionness, and end probability sequences, then connect start and end points to generate dense proposals. For proposal refinement, top-down methods refine action proposals with a distance loss. Some recent bottom-up methods [14,17,18] train a confidence map for accurate scoring. In addition, several recent methods [11,19] incorporate the graph convolution network (GCN) into TAL to exploit snippet-snippet or proposal-proposal relations.

The work most related to ours is P-GCN [11]. It constructs an action proposal graph to model the interactions between proposals and performs relation reasoning with a GCN. Our dual relation

network differs from P-GCN in two aspects. First, while P-GCN establishes an edge between two proposals only if they have a high tIoU, the proposed PPRM models the relation between each proposal and all other proposals in a video and considers both their geometry (*i.e.*, duration) and appearance similarities as well as feature channel interaction. Second, the proposed PCRM automatically finds relevant snippets for each proposal within the entire video and uses them as global context to update the proposal features, which is neglected by P-GCN. The experimental results show that our proposed method outperforms P-GCN by a large margin on two datasets. In addition, the model size of our network is smaller than that of P-GCN.

Other existing methods [18–20] share similar insight as us and model context for TAL. Concretely, BSN [10] and BMN [17] use the temporal extension operation or the 1D convolution operation to obtain limited local context. Gao et al. [21] directly squeeze the global temporal information through average pooling as the global context, which ignores the long-term dependencies between actions and the context, and may introduce noise. G-TAD [19] formulates video snippets as graph nodes and updates all snippets features via a GCN to classify nodes and score predefined temporal anchors. TCANet [20] updates each snippet of the video by incorporating its local and global temporal relationships. Then, a proposal generator obtains the candidate proposal feature from the encoded video feature to further predict the confidence score and regress boundaries. Both G-TAD and TCANet share the same motivation to integrate the rich context for each snippet by snippet-snippet relations and arrange anchors or proposals based on the encoded video feature.

We argue that they share the following drawbacks. (1) Modeling all snippet-snippet relations might include redundant information into proposals or anchors refinement. (2) Snippet-snippet relations aim to improve the temporal receptive field of each anchor or proposal to contain more temporal information, while they ignore the temporal dependencies between actions and the global context. By contrast, the proposal-context interactions are specific to each proposal and are more flexible to search for supportive clues for proposals of different qualities. BSN++ [18] adopts a nested U-shaped encoder-decoder with a larger temporal receptive field to exploit the rich context for accurate boundary prediction. However, directly upsampling the output of the classification networks cannot recover the degraded temporal information caused by downsampling, which harms precise temporal localization. In addition, all aforementioned methods focus on temporal action proposal generation. They pay more attention to producing reliable confidence scores or achieve accurate boundary prediction using the context.

By contrast, our method aims to improve the representation of imperfect action proposals by incorporating the complementary information from homogeneous relationships between proposals and heterogeneous relations between proposals and the context, driven by the classification task and the localization task, respectively.

2.2. Relation reasoning

Relation reasoning means effectively selecting and integrating information based on the relations between visual entities. It has been widely used in natural language processing [22] and computer vision tasks [23,24]. Su et al. [25] propose a unified framework named as PCG-TAL, which builds upon a two-granularity and two-stream pipeline. It can leverage cross-granularity and cross-stream complementary information obtained by message passing between anchor-based features and frame-based features as well as between RGB stream and flow stream to generate better action segments. Chen et al. [26] introduce a relation attention module to model relations among proposals for temporal action localization. Specifically, the relation attention module could capture the relationship between proposals via a pair-wise relation function, which in spirit is similar to the self-attention mechanism. Sun et al. [27] propose to exploit informative video segments by learning video segment weights for temporal action localization, where the learned weights represent the importance of video segments in recognizing actions and predicting temporal boundaries. Huang et al. [28] introduce a location-aware graph convolutional network (L-GCN) to model the interaction between objects for the video question answering task. It constructs a fully-connected graph where each node is an object and the edges between nodes represent their relationship. Each node also encodes both spatial and temporal object location information. Pan et al. [29] design an Actor-Context-Actor Relation Network (ACAR-Net), which deduces indirect relations between multiple actors and the context for spatio-temporal action localization.

Additionally, the self-attention mechanism, *e.g.*, Transformer [22] and the non-local network [30], has been widely used to model relations for static images or sequence data. It models the bidirectional relations between homogeneous entities. By contrast, our context reasoning block aims to model the unidirectional relation between heterogeneous entities, *i.e.*, from contextual snippets to an action proposal. Zhu et al. [31] propose a cross-layer attention model to aggregate multi-layer features into a single global video representation through weighting global context at different scales for action recognition. In contrast, our proposal-context relation module aims to weight each video snippet in the video using the attention mechanism to obtain a context-aware feature for action proposal refinement. Wu et al. [32] propose a dual attention matching module to better model the whole event for event localization task, where it encodes local temporal information by a global cross-check mechanism. Multiple knowledge representation (MKR) [33] is a new tool to exploit data relations at multiple sources. Our dual relation network has similar spirits to MKR, where we consider and leverage two types of desirable properties of actions, *i.e.*, appearance and geometry features to facilitate the TAL task. Our proposed method is a specific application of MKR.

3. Dual relation network

In this section, we introduce the proposed Dual Relation Network (DRN), which consists of two modules, *i.e.*, a Proposal-Proposal Relation Module (PPRM) and a Proposal-Context Relation Module (PCRM). As shown in Fig. 2, PPRM consists of two blocks, *i.e.*, a Proposal Reasoning Block and a Feature Reasoning Block. They reason the relation between proposals from the temporal and semantic perspectives, respectively. PPRM is designed to obtain discriminative information from relevant proposals for action recognition. PCRM aims to capture supplementary information from the global context for boundary regression.

3.1. Notation and preliminaries

We follow previous action proposal generation methods [10,17] to build our model upon snippet-level features of input videos. As illustrated in Fig. 2, our dual relation network takes an untrimmed video as input, and outputs a category label, a confidence score, and the temporal boundaries of each action instance. Specifically, given an untrimmed video, we encode each successive fixed-length frame with a pre-trained feature extractor (*e.g.*, the I3D network [34]), and denote the output feature sequence as $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, where T is the number of snippets, and $\mathbf{x}_t \in \mathbb{R}^D$ is the feature vector of the t-th snippet, with D representing the channel dimension.

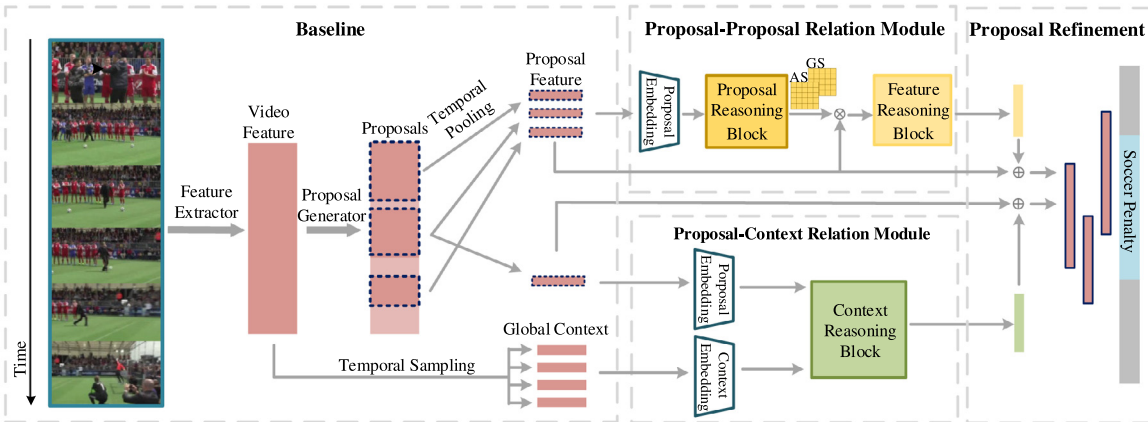

Fig. 2. Architecture Overview. The input is a sequence of snippet-level features. We first generate dense candidate proposals based on the video feature sequence. Then candidate proposals are fed into the Proposal-Proposal Relation Module and the Proposal-Context Relation Module respectively, and both relation modules can automatically integrate the relation features for each proposal. Finally, the refined proposals are post-processed for action classification and localization.

For each video, $\Psi = \{\psi_n | \psi_n = (\mathbf{f}_n, (t_{s,n}, t_{e,n}))\}_{n=1}^{N}$ is the set of action proposals of interest generated by an existing proposal generator (*e.g.*, BSN [10]), where $t_{s,n}$ and $t_{e,n}$ denote the start and end time of the n-th proposal, respectively. N denotes the number of proposals. The feature vector of the n-th proposal, *i.e.*, $\mathbf{f}_n \in \mathbb{R}^D$, is obtained through temporal pooling across the snippet-level features within the start and end time of the proposal.

We proceed to leverage the proposal-proposal relation module (PPRM) and the proposal-context relation module (PCRM) to enrich the proposal features. Finally, the enriched features of each proposal are employed for action classification and temporal boundary regression. We detail the proposed PPRM and PCRM below.

3.2. Proposal-proposal relation module

As discussed above, reasoning based on the relations among relevant proposals not only helps distinguish action categories with similar appearances or motion patterns, but also provides a chance to suppress redundant action instances and false positives. The proposal-proposal relation module (PPRM) consists of two blocks: the proposal reasoning block and the feature reasoning block. PPRM first augments the features of each proposal by taking into account the features of all relevant proposals across the whole video. Then, the feature reasoning block further promotes discriminative feature channels while suppressing the minor ones.

Proposal Reasoning Block. Given N proposals, we measure their pairwise relatedness by computing an appearance similarity matrix $\mathbf{S}^a \in \mathbb{R}^{N\times N}$ and a geometry similarity matrix $\mathbf{S}^g \in \mathbb{R}^{N\times N}$:

$$S^a_{n,m} = \frac{\mathbf{f}_n^\top \mathbf{f}_m}{\|\mathbf{f}_n\| \|\mathbf{f}_m\|}, \tag{1}$$

$$S^g_{n,m} = \min\left(\frac{t_{e,n} - t_{s,n}}{t_{e,m} - t_{s,m}}, \frac{t_{e,m} - t_{s,m}}{t_{e,n} - t_{s,n}}\right), \tag{2}$$

where $S^a_{n,m}$ and $S^g_{n,m}$ denote the elements at the n-th row and the m-th column of the appearance similarity matrix and the geometry similarity matrix in Eq. (1) and Eq. (2), respectively. $\|\cdot\|$ denotes the L_2 norm. The appearance similarity matrix $\mathbf{S}^a$ contains the cosine similarity scores between each pair of proposal feature vectors. The geometry similarity matrix $\mathbf{S}^g$ comprises the duration similarity scores between proposals based on the assumption that relevant proposals tend to be similar in terms of their temporal scales. Then, a weighted summation of $\mathbf{S}^a$ and $\mathbf{S}^g$ produces the final similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$:

$$\mathbf{S} = \lambda \mathbf{S}^a + (1-\lambda)\mathbf{S}^g, \tag{3}$$

where λ is a hyper-parameter controlling the relative importance between the appearance and geometry similarities. We update the features of the n-th proposal by integrating N proposal features. As a result, the updated features of the n-th proposal, denoted as $\mathbf{f}^s_n$, can be computed by

$$\mathbf{f}^s_n = \frac{1}{N-1}\sum_{m=1}^{N} S_{n,m}\mathbf{W}_S\mathbf{f}_m, \tag{4}$$

where $S_{n,m}$ is an element of $\mathbf{S}$ and represents the relation weight between the n-th proposal and the m-th proposal, and $\mathbf{W}_S \in \mathbb{R}^{D\times D}$ is the weight matrix of a linear projection layer. Note that the bias term is omitted for simplicity.

Feature Reasoning Block. We use the proposal reasoning block to explicitly model the subtle interactions between related proposals. To encourage the network to focus on discriminative features, we subsequently feed the updated proposal features to a feature reasoning block for high-order supportive information. In particular, to model the interdependent relations between channels, our inspiration originates from Hu et al. [35]. The features of the n-th proposal are updated as

$$\mathbf{f}^{pp}_n = \mathbf{W}_r\mathbf{f}_n + \mathbf{f}^s_n \odot \sigma(\mathbf{W}_{ex}\cdot \mathrm{ReLU}(\mathbf{W}_{sq}\mathbf{f}^s_n)), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid activation function and $\odot$ is element-wise multiplication. $\mathbf{W}_r \in \mathbb{R}^{D\times D}$, $\mathbf{W}_{sq} \in \mathbb{R}^{D\times(D/r)}$ and $\mathbf{W}_{ex} \in \mathbb{R}^{(D/r)\times D}$ are weights of three linear projection layers, respectively. r is a predefined integer for the feature reasoning block. This block performs feature reasoning among each relevant proposal to adaptively activate informative features. In other words, it could adaptively evaluate the importance of each semantic feature and assign appropriate weights to it, so as to suppress the noise information within the proposal.

Finally, a softmax-activated fully-connected layer with $C+1$ output channels is used to classify the feature vector of each proposal $\mathbf{f}^{pp}_n$, and output the category prediction result $\hat{\mathbf{y}} \in \mathbb{R}^{C+1}$, where $C+1$ denotes the number of action categories with an additional background category.

Discussion. PPRM takes full advantage of the desirable properties of actions, *i.e.*, appearance, and geometry features to learn relations between all relevant proposals. By contrast, P-GCN [11] only considers proposals with high overlaps and near distance to model their relations via a GCN. In short, PPRM builds proposal-proposal relations from both the temporal and semantic perspectives through the two blocks. Therefore, PPRM can explore and integrate supportive information for action proposal refinement as much as possible.

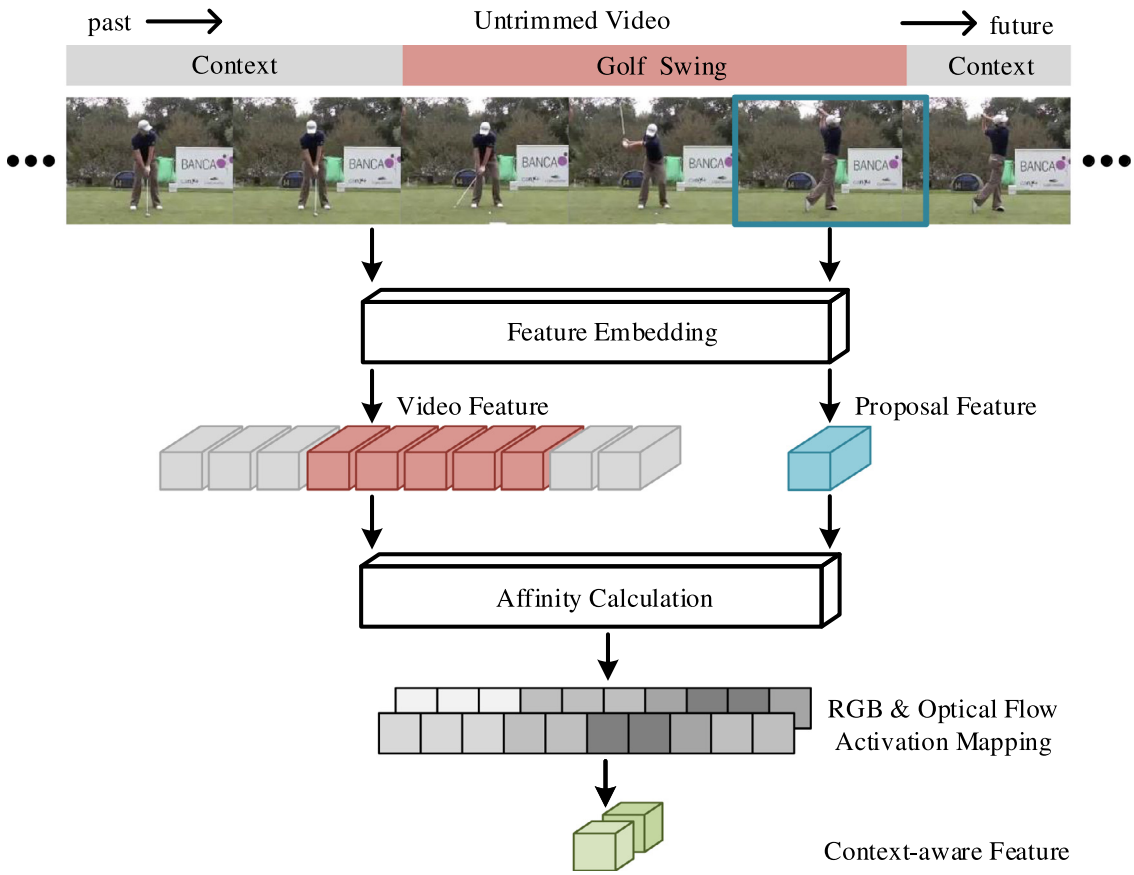

Fig. 3. Context reasoning block is responsible for reasoning the coupling relation between the proposal and global context through the interaction between proposal feature and video feature in two streams. Finally, a context-aware feature is obtained for each proposal.

3.3. Proposal-context relation module

The untrimmed videos contain meaningful contextual information and meaningless noise background. Several previous TAL methods [10,14,36] classify and locate a sparse set of proposals based on region-wise features, but they neglect any available context information. The valid context should be beneficial for both action classification and localization. Even though several existing methods [13,36] use CNNs to capture proposal context, they still suffer from a limited receptive field or limited short-term temporal information. Therefore, we propose the proposal-context relation module (PCRM) to perform reasoning based on the relations between the proposal and the global context. It can automatically retrieve the regions most related to the proposal, and thus can identify and integrate useful contextual information for accurate localization.

As illustrated in Fig. 3, we introduce a context reasoning block as the relation reasoning operator to model the temporal dependencies between the proposal and the whole video and incorporate it into PCRM. The proposed PCRM takes as input proposal features $\{\mathbf{f}_n\}_{n=1}^{N}$ and a sequence of snippet features $\{\mathbf{x}_t\}_{t=1}^{T}$. We first calculate the affinity between the proposal and each time step of the video sequence in the embedding space. We then generate context-aware features through the long-range affinity and use them to augment the original proposal feature.

For each proposal, we first calculate an attention weight a_t based on the relation between the proposal features $\mathbf{f}_n$ and each of the snippet-level features $\mathbf{x}_t$:

$$a_t = \frac{\exp\left(\mathbf{f}_n^\top \mathbf{W}_\mathrm{p}^\top \mathbf{W}_c \mathbf{x}_t\right)}{\sum_{\tau=1}^{T} \exp\left(\mathbf{f}_n^\top \mathbf{W}_\mathrm{p}^\top \mathbf{W}_c \mathbf{x}_\tau\right)}, \tag{6}$$

where $\mathbf{W}_\mathrm{p}, \mathbf{W}_\mathrm{c} \in \mathbb{R}^{D\times D}$ are learnable parameters. To make the model focus on video snippets most relevant to the proposal, the feature of the n-th proposal is updated by linearly aggregating snippet-level features with the attention weights:

$$\mathbf{f}_n^{\mathrm{pc}} = \mathbf{W}_\mathrm{p}\mathbf{f}_n + \mathbf{W}_{\mathrm{pc}} \sum_{t=1}^{T} a_t \mathbf{W}_c \mathbf{x}_t, \tag{7}$$

where $\mathbf{W}_{\mathrm{pc}} \in \mathbb{R}^{D\times D}$ is a learnable parameter. Therefore, the proposed PCRM adaptively captures supportive contextual information from long-range context and filters background noise by generating snippet-level attention weights. Furthermore, PCRM can better refine the region-wise proposal locations by embedding valid contextual information.

Finally, with the context-enhanced features of each proposal $\mathbf{f}_n^{\mathrm{pc}}$, we use a fully-connected (FC) layer with two-dimensional output to predict the start time t_s and end time t_e, and use another FC layer with a sigmoid activation to predict the completeness score c of the proposal, indicating whether the proposal is complete or not.

Discussion. The relation reasoning between proposals and the context aims to supplement missing action evolution information for imperfect action proposals, and it also provides indicative details for boundaries regression, *e.g.*, shot switching. G-TAD [19] updates all snippets features of the video via a GCN and uses them to classify and regress predefined anchors. It aims to improve the temporal receptive field of each anchor through snippet-snippet relations and neglects the temporal dependencies between action instances and the global context. Besides, modeling all snippet-snippet relations might introduce redundant information for proposals refinement. Our PCRM is specific to each proposal and is

more flexible to search for supportive clues for proposals of different qualities.

Summary. Our dual relation network aims to improve the representation of imperfect action proposals by incorporating the complementary information from homogeneous relations between proposals and heterogeneous relations between proposals and the context, driven by the classification task and the localization task, respectively.

3.4. Network optimization

During the training process, the above two relation reasoning modules PPRM and PCRM are jointly trained in an end-to-end manner. The overall loss $\mathcal{L}$ of the proposed dual relation network consists of a classification loss $\mathcal{L}_{\text{cls}}$, a regression loss $\mathcal{L}_{\text{reg}}$, and a completeness loss $\mathcal{L}_{\text{com}}$:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N} \mathcal{L}_{\text{cls}}(\mathbf{y}_n, \hat{\mathbf{y}}_n) \\ &+ \alpha \sum_{n=1}^{N} \mathbf{1}_{\{\hat{c}_n=1\}} \mathcal{L}_{\text{reg}}(t_{s,n}, t_{e,n}, \hat{t}_{s,n}, \hat{t}_{e,n}) \\ &+ \beta \sum_{n=1}^{N} \mathcal{L}_{\text{com}}(c_n, \hat{c}_n),\end{aligned} \tag{8}$$

where $\hat{t}_{s,n}$, $\hat{t}_{e,n}$ are the target start and end time for the nth proposal, respectively. $\hat{\mathbf{y}}_n \in \mathbb{R}^{C+1}$ is a one-hot target category label vector, while α and β are weight hyper-parameters of the regression loss and the completeness loss, respectively. $\mathbf{1}$ is the indicator function. $\hat{c}_n \in \{0, 1\}$ is the completeness label of the n-th proposal. The classification loss $\mathcal{L}_{\text{cls}}$ employs a standard cross entropy loss. The completeness loss $\mathcal{L}_{\text{com}}$ employs the online hard example mining hinge loss [37]. The regression loss $\mathcal{L}_{\text{reg}}$ employs a sum of smooth L_1 losses [38] between the target start/end time and predicted start/end time:

$$\mathcal{L}_{\text{reg}}(t_s, t_e, \hat{t}_s, \hat{t}_e) = S_{L_1}\left(t_s, \hat{t}_s\right) + S_{L_1}\left(t_e, \hat{t}_e\right), \tag{9}$$

where $S_{L_1}(\cdot, \cdot)$ denotes the smooth L_1 loss.

3.5. Inference phase

Proposal Generation. We use the boundary-based method, *e.g.*, BSN [10] or BMN [17], as the proposal generator to produce a dense action proposal set.

Proposal Refinement. Each proposal obtains supplementary information by reasoning the interaction between proposals and the temporal dependencies between proposals and videos. Subsequently, the regression branch predicts offsets and refines the temporal locations of each action proposal. The classification branch predicts the action category of the candidate proposal, and the completeness branch predicts its completeness score. For each proposal, we define its category label as its top-1 action category from its corresponding classification prediction $\hat{\mathbf{y}}$, and its confidence score as the product of its top-1 classification score and completeness score.

Proposal Retrieval. Given candidate proposals with confidence scores, we adopt Soft-NMS (soft non-maximum suppression) to suppress redundant proposals with high overlaps in the post-processing stage.

4. Experiments and discussions

4.1. Datasets and metrics

THUMOS14 [15] includes 413 untrimmed videos over 20 hours from 20 action categories. It is very challenging since each video has more than 15 action instances, and 71% of frames are occupied by background items. Following convention [14], we use the 200 videos in the validation set for training, and evaluate on the 213 videos in the testing set.

ActivityNet v1.3 [16] is another popular benchmark for action localization. We evaluate our method on ActivityNet v1.3. It contains 19,994 untrimmed videos from 200 action categories, which are divided into training, validation, and testing sets by a ratio of 2 : 1 : 1. Following the common practice [14,39], we use the training set for training and the validation set for evaluation. We compare our method with state-of-the-art methods on the THUMOS14 and ActivityNet v1.3 datasets, and we conduct ablation studies on the THUMOS14 dataset.

Evaluation Metrics. To evaluate the proposed method, we use mean average precision (mAP) under different temporal intersection-over-union (tIoU) thresholds. We take the official evaluation code provided by ActivityNet to evaluate the performance of TAL on the two datasets. Specially, the tIoU thresholds are chosen from $[0.1 : 0.1 : 0.7]$ and $\{0.5, 0.75, 0.95\}$ for THUMOS14 and ActivityNet v1.3, respectively. On THUMOS14, we also report average mAP at tIoU thresholds from 0.1 to 0.5 and from 0.3 to 0.7, with a step size of 0.1. On ActivityNet v1.3, we report average mAP over 10 different tIoU thresholds $[0.5 : 0.05 : 0.95]$.

4.2. Implementation details

For the feature extractor, we adopt the I3D network [34] pre-trained on Kinetics [40] to extract RGB features and optical flow features from 16 consecutive frames, respectively. The features are extracted from the global average pooling layer as 1024 dimensional vectors. We use BSN [10] as the proposal generator on THUMOS14. For ActivityNet v1.3, we adopt BSN [10] and BMN [17]. Note that for a fair comparison with previous works, we do not fine-tune the feature extraction backbone or the proposal generator. Particularly, we combine our proposals by BMN [17] with video-level classification results from [41] on ActivityNet v1.3, as in [20]. For each proposal, we obtain its feature through temporal RoI pooling [38] across the snippet-level features within the start and end time of the proposal and map it to the same channels as the video feature. As for the video features, we use linear interpolation to obtain global context features of 100 snippets for THUMOS14 and ActivityNet v1.3. For boundary regression, we predict the offset of the center coordinate and the duration of each proposal instead of directly predicting its start and end time.

During the training stage, we set the mini-batch size to 32 on the THUMOS14 dataset and 64 on the ActivityNet v1.3 dataset. For PPRM, we select 10 proposals with high similarity both in appearance and geometry to reduce computation. We train the model for 60 epochs with the SGD optimizer, set the initial learning rate to 0.01, and reduce it by a factor of 10 for every 15 epochs. All hyper-parameters are determined by empirical grid search, *i.e.*, $\lambda = 0.5$, $r = 2$, $\alpha = 1$, and $\beta = 0.5$. For post-processing, Soft-NMS with tIoU thresholds of 0.2 and 0.4 are used to remove duplicate proposals in experiments on THUMOS14 and ActivityNet v1.3, respectively. We combine the predicted results of RGB and Flow streams by a ratio of 5 : 6 to generate the final predictions, where the Soft-NMS threshold is set to 0.3.

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

ActivityNet v1.3. We compare our method with 13 state-of-the-art methods and their variants, and report the mAP at different tIoU thresholds as well as the average mAP at tIoU thresholds $0.5 : 0.05 : 0.95$ in Table 1. For fair comparison, we report the experimental results with different backbones, I3D [34] and TSN [45],

Table 1
Temporal action localization results on ActivityNet v1.3. The "Avg" column denotes the average mAP at tIoU thresholds from 0.5 to 0.95, with a step size of 0.05.

Method	Year	Backbone	mAP@tIoU (%)			Avg (%)
			0.5	0.75	0.95	0.5:0.05:0.95
BSN [10]	2018	TSN	46.45	29.96	8.02	30.03
BMN [17]	2019	TSN	50.07	34.78	8.29	33.85
G-TAD [19]	2020	TSN	50.36	34.60	**9.02**	34.09
BSN+ [18]	2021	TSN	51.27	35.70	8.33	34.88
Ours [BSN]	-	TSN	51.75	35.97	7.05	34.50
Ours [BMN]	-	TSN	**53.48**	**37.21**	7.54	**35.42**
TAL [36]	2018	I3D	38.23	18.30	1.30	20.22
P-GCN [11]	2019	I3D	42.90	28.14	2.47	26.99
BU-MR [14]	2020	I3D	43.47	33.91	**9.21**	30.12
TCANet [20]	2021	I3D	51.91	34.92	7.46	34.43
AFSD [42]	2021	I3D	52.40	35.30	6.50	34.40
ContextLoc [43]	2021	I3D	56.01	35.19	3.55	34.23
VSGN [44]	2021	I3D	52.38	36.01	8.37	35.07
Ours [BSN]	-	I3D	52.84	36.10	6.32	35.92
Ours [BMN]	-	I3D	**56.10**	**39.92**	6.95	**37.83**

respectively. Particularly, our method achieves the highest average mAP result (as shown in the "Avg" column) compared with previous methods. Although this dataset is huge and complex, our proposed method further improves the mAP at tIoU of 0.75 from 39.13% (TCANet [20]) to 39.92%, and the average mAP from 37.56% to 37.83%. These results clearly indicate the efficacy of the relation reasoning capability of our method in complex scenarios.

Notably, compared to P-GCN [11] and ContextLoc [43], which also belong to the proposal refinement-based method, our method significantly outperforms them by a large margin on average mAP, where we adopt the same proposals generated by BSN [10].

THUMOS14. We report the performances of our method with different backbones and other 16 state-of-the-art methods in Table 2, where the tIoU thresholds are set from 0.1 to 0.7. Clearly, our proposed method achieves significant improvements against all other state-of-the-art methods. At tIoU thresholds from 0.1 to 0.5, the mAP scores of our method are clearly higher than that of the previous best method P-GCN [11]. In addition, we compare the average mAP at tIoU from 0.1 to 0.5 and the average mAP at tIoU from 0.3 to 0.7 with previous state-of-the-art methods. Our method reaches 67.26% and 53.80% respectively, which outperforms the previous best performance (P-GCN [11], AFSD [42]). This demonstrates that our proposed dual relation network has significant advantages.

Model size. The number of parameters of our full network, including PPRM and PCRM, is 4.2M. P-GCN [11], our main competitor, consists of 4.6M parameters. Thus, our superior performance is not caused by additional learnable parameters.

4.4. Ablation studies

To investigate the contribution of each component in our proposed method, we conduct comprehensive ablation studies by comparing it with its variants with certain components changed or removed.

Effectiveness of reasoning modules. To verify the effectiveness of our proposed PPRM and PCRM, we present the experimental results of different combinations of the relation modules on the THUMOS14 dataset in Table 3 and the ActivityNet v1.3 dataset in Table 4. The baseline network is constructed by removing both PPRM and PCRM. By default, we adopt BSN as the proposal generator. Compared with the baseline, the combination of PPRM and PCRM significantly improves the mAP at all tIoU thresholds from 0.1 to 0.7 on THUMOS14. For ActivityNet v1.3, PPRM and PCRM achieve significant performance improvements. Adequate experiments reveal that the relation features obtained by the two reasoning modules can indeed improve the performance. Conecretely, compared with PCRM, PPRM has a clearer contribution to the mAP at tIoU from 0.5 to 0.7 on THUMOS14, while PCRM improves the mAP at tIoU from 0.1 to 0.3 more significantly than PPRM. This phenomenon demonstrates that the two reasoning modules complement the proposal features.

Impact of the reasoning blocks in PPRM. Our proposed PPRM consists of two building blocks, namely the *proposal reasoning block* and the *feature reasoning block*. They reason the relation features between proposals from the temporal and semantic perspectives, respectively. We conduct ablation studies to quantitatively measure their impacts. As shown in Table 5, either individual block improves the mAP at all thresholds, and their combination further boosts the performance. This indicates that these two blocks are complementary, and both are important to our PPRM.

Effects of different proposal generators. We present the TAL results on THUMOS14 by combing our proposed dual relation network with different proposal generation methods in Table 6. The results show that our method significantly improves the performance combined with different state-of-the-art proposal generation methods. Therefore, our proposed dual relation network is flexible and versatile, which can adapt to different TAL frameworks.

Precision-Recall on THUMOS14. We draw the per-category Precision-Recall (PR) curves obtained by different variants of our proposed method in Fig. 4, where we denote the overall framework *dual relation network* as DRN. Clearly, the red categories-wise PR curve indicates that our method can improve precision and recall for most categories (*i.e.*, higher in the y-axis), and thus a larger area is enclosed by the PR curve, x and y axis (*i.e.*, Average Precision, AP).

Moreover, compared with the baseline, PPRM and PCRM have their respective advantages in different action categories, and have their own contributions to the overall framework performance. Specifically, for actions that are not obvious in appearance (*e.g.*, Diving and Cricket Bowling), PPRM improves the representation of each proposal by reasoning the subtle relation between the proposals, and enhances the discrimination of proposals. For actions that rely on contextual information (*e.g.*, Golf Swing and Pole Vault), PCRM enriches the feature of each proposal by reasoning the interaction between the region-wise proposal and the long-range context, and achieves more precise localization. In general, our proposed dual relation network can achieve the best performance in most scenarios.

Duration similarity or IoU similarity. We use the duration similarity between action proposals instead of their IoU as the geometry similarity. Unlike objects in an image, whose sizes depend on their distances to the camera, action instances of the same category in natural videos tend to have similar durations, *e.g.*, Basketball Dunk and Cliff Diving. Thus, we use the duration to measure the proposal's geometric similarity across the entire video. By contrast, IoU only considers overlapping proposals. We conduct an ablation experiment to compare the effects of the two geometric similarities on THUMOS14. Table 7 indicates that duration outperforms IoU.

Analysis of performance improvements. In order to further explore sources of performance improvements, we conduct additional experiments in terms of AR@AN for evaluating the action boundary. As illustrated in Table 8, it can be observed that our method can improve the average recall of action boundaries compared with other state-of-the-art methods. It indicates that the performance improvement comes from more accurate boundary prediction.

Table 2
Temporal action localization results on THUMOS14. The two "Avg" columns denote the average mAP at tIoU thresholds from 0.1 to 0.5 and at tIoU thresholds from 0.3 to 0.7, with a step size of 0.1.

Method	Year	Backbone	mAP@tIoU (%)							Avg (%)	Avg (%)
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.1:0.5	0.3:0.1:0.7
BSN [10]	2018	TSN	-	-	53.5	45.0	36.9	28.4	20.0	-	36.76
MGG [13]	2019	TSN	-	-	53.9	46.8	37.4	29.5	21.3	-	37.78
BMN [17]	2019	TSN	-	-	56.0	47.4	38.8	29.7	20.5	-	38.48
G-TAD [19]	2020	TSN	-	-	54.5	47.6	40.2	30.8	23.4	-	39.30
ActionDBG [39]	2020	TSN	-	-	57.8	49.4	39.8	30.2	21.7	-	39.78
BSN+ [18]	2021	TSN	-	-	59.9	49.5	41.3	31.9	22.8	-	41.08
Ours	-	TSN	**69.8**	**64.0**	**61.9**	**52.2**	**44.2**	**33.5**	**24.0**	**58.42**	**43.16**
TAL [36]	2018	I3D	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.28	39.82
P-GCN [11]	2019	I3D	69.5	67.8	63.6	57.8	49.1	-	-	61.56	-
BU-MR [14]	2020	I3D	-	-	53.9	50.7	45.4	38.0	28.5	-	43.30
TCANet [20]	2021	I3D	-	-	60.6	53.2	44.6	36.8	26.7	-	44.38
AFSD [42]	2021	I3D	-	-	67.3	62.4	55.5	43.7	**31.1**	-	52.00
ContextLoc [43]	2021	I3D	-	-	68.3	63.8	54.3	41.8	26.2	-	50.88
VSGN [44]	2021	I3D	-	-	66.7	60.4	52.4	41.0	30.4	-	50.18
Ours	-	I3D	**73.0**	**71.9**	**69.2**	**64.7**	**57.5**	**46.9**	30.8	**67.26**	**53.80**

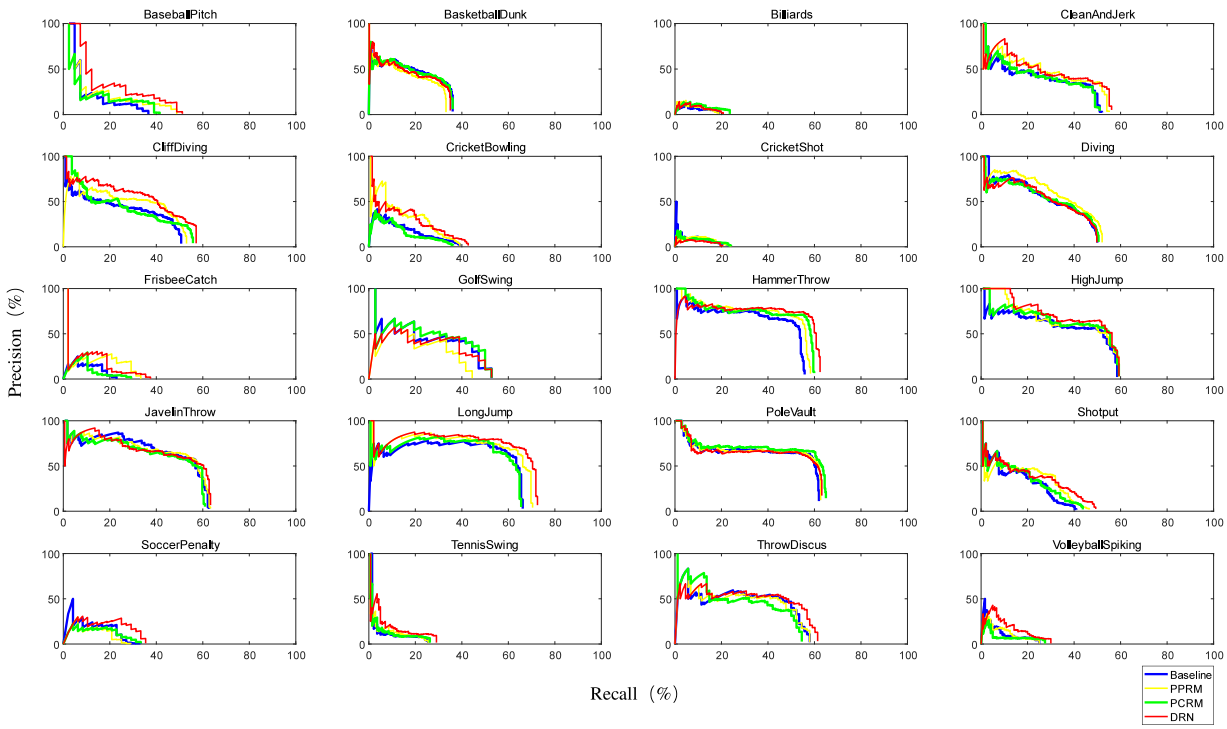

Fig. 4. Per category Precision-Recall (PR) curves on the THUMOS14 testing set. The PR curve is plotted at tIoU threshold 0.7. The area enclosed by the PR curve is Average Precision (AP) of each category. DRN is short for the dual relation network.

Table 3
Ablation study of the proposal-proposal relation module (PPRM) and the proposal-context relation module (PCRM) on THUMOS14.

PPRM	PCRM	mAP@tIoU (%)							Avg (%)
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.1:0.7
		69.7	69.2	65.7	60.9	53.2	41.5	28.0	55.5
✓		70.8	69.9	66.4	62.0	55.8	43.8	29.9	56.9
	✓	72.2	71.0	67.1	61.7	54.9	42.1	28.5	56.8
✓	✓	**73.0**	**71.9**	**69.2**	**64.7**	**57.5**	**46.9**	**30.8**	**59.1**

Table 4
Ablation study of the proposal-proposal relation module (PPRM) and the proposal-context relation module (PCRM) on ActivityNet v1.3.

PPRM	PCRM	mAP@tIoU (%)			Avg (%)
		0.5	0.75	0.95	0.5:0.05:0.95
		49.30	34.65	4.91	33.71
✓		50.15	35.72	5.83	34.90
	✓	51.06	35.26	5.14	34.12
✓	✓	**52.84**	**36.10**	**6.32**	**35.92**

4.5. Visualization

Visualization of the Proposal-Proposal Relation. Our PPRM builds proposal-proposal relations from both the temporal and semantic perspectives through two blocks, a proposal relation reasoning block (PRB) and a feature reasoning block (FRB). PRB first augments the features of each proposal by taking into account features of all relevant proposals across the whole video. Then, FRB further promotes discriminative feature channels while suppressing the minor

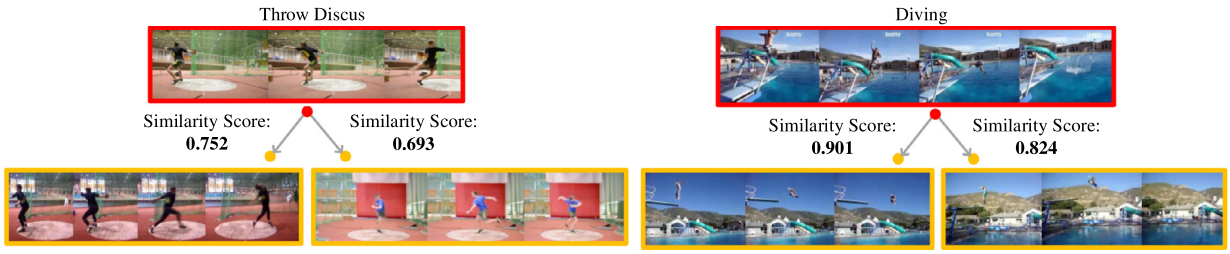

Fig. 5. Visualization of the Proposal-Proposal Relation. PPRM aims to capture discriminative supplementary information from relevant proposals for action recognition through the proposal reasoning block and the feature reasoning block..

Table 5
Ablation study of the proposal reasoning block (PRB) and the feature reasoning block (FRB) on THUMOS14.

PRB	FRB	mAP@tIoU (%)							Avg (%)
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.1:0.7
		72.2	71.0	67.1	61.7	54.9	42.1	28.5	56.8
✓		72.7	71.6	68.7	64.0	57.0	45.4	30.1	58.5
	✓	72.6	71.9	68.2	62.3	55.7	43.0	29.8	57.6
✓	✓	**73.0**	**71.9**	**69.2**	**64.7**	**57.5**	**46.9**	**30.8**	**59.1**

Table 9
Ablation study of the hyper-parameter K on THUMOS14.

K	mAP@tIoU (%)				
	0.3	0.4	0.5	0.6	0.7
1	67.9	62.1	55.4	43.0	29.5
5	68.7	63.5	56.9	45.2	30.0
10	69.2	**64.7**	**57.5**	**46.9**	**30.8**
15	**69.6**	64.5	57.2	46.6	30.4

Table 6
Comparison of different proposal generators on THUMOS14 .

Model	mAP@tIoU (%)					Avg (%)
	0.3	0.4	0.5	0.6	0.7	0.3:0.1:0.7
w/ BSN [10]	69.2	64.7	57.5	46.9	30.8	53.8
w/ ContextLoc [43]	70.3	65.9	58.2	46.5	29.4	54.1
w/ VSGN [44]	69.5	66.2	57.4	48.2	32.5	54.8

Table 7
Comparison of different geometric similarities on THUMOS14.

Method	mAP@tIoU (%)						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
IoU	72.1	70.9	67.8	62.9	55.4	43.4	29.7
Duration	**73.0**	**71.9**	**69.2**	**64.7**	**57.5**	**46.9**	**30.8**

Table 8
Comparison results on THUMOS14 in terms of AR@AN.

Method	AR@AN		
	@50	@100	@200
BSN [10]	37.46	46.06	53.23
BMN [17]	39.36	47.72	54.84
MGG [13]	39.93	47.75	54.65
ContextLoc [43]	41.20	49.82	**57.46**
VSGN [44]	40.58	49.23	56.10
Ours	**42.60**	**50.26**	56.74

ones. In order to more intuitively show how they work, we conduct some visualization examples of PRB and FRB through class activation maps (CAM), as shown in Fig. 5. Concretely, PRB models proposal-proposal relation based on their similarity scores calculated by appearance similarity and geometric similarity. We can observe that the relevant proposals often belong to the same category of actions. Thereafter, modeling these relevant proposals with high similarity scores can effectively explore supportive information of an action instance, e.g., jump of "Diving" and run-up of "Javelin Throw". FRB further activates discriminative feature channels and integrates these useful information to help temporal action localization. It can be observed by the heatmaps that our PRB and FRB can capture discriminative semantic features for proposal refinement.

Visualization of the Proposal-Context Relation. To intuitively understand the contribution of PCRM to action localization, we visualize dependencies between action proposals (red) and the global context (blue) on THUMOS14, as shown in Fig. 6. We partition the quality of action proposals by their extent of overlapping with the actual action. Specially, high-quality proposals (left) have a high IoU with action instances and have sufficient information for boundaries regression without additional supplementary information. Medium-quality proposals (*e.g.*, they cover only the continuation and end of actions. (middle)) can leverage their neighborhood context to supplement the missing information for boundary refinement. Low-quality proposals (right), which are frequent and unavoidable, can only glimpse a small part of the action evolution because the missing content is not discriminative. Therefore, they can flexibly capture other action evolution feature from the global context to supplement boundary details and further achieve accurate location.

Analysis of the computation of PPRM. The relevant proposals with high similarity could provide supportive cues to reduce the uncertainty of actions from imperfect predictions. However, it is redundant to consider all relevant proposals in PPRM, since they are often overlapped temporally. As a result, we take K relevant proposals with high similarity both in appearance and geometry for PPRM in the experiment. We evaluate the impact of different numbers of relevant proposals on the performance in Table 9. It can be observed that a small number of action proposals are sufficient to effectively model the relationship between the relevant proposals while reducing the computational burden. Moreover, we also calculate the computation time of the two similarity matrixes S^a and S^g in each iterative process as 0.00157s and 0.00036s, respectively. Therefore, these two similarity matrices do not cost a lot of computation.

Qualitative results. To further demonstrate the effectiveness of our method, we present some example TAL results on both THUMOS14 and ActivityNet v1.3 datasets in Fig. 7. For ease of comparison, the ground truth and the results obtained by our method with and without the two reasoning modules are all presented. These results clearly show that the refined temporal boundaries produced by our full method better correspond to the ground truth, and the confidence is more reliable than the baseline model. It manifests

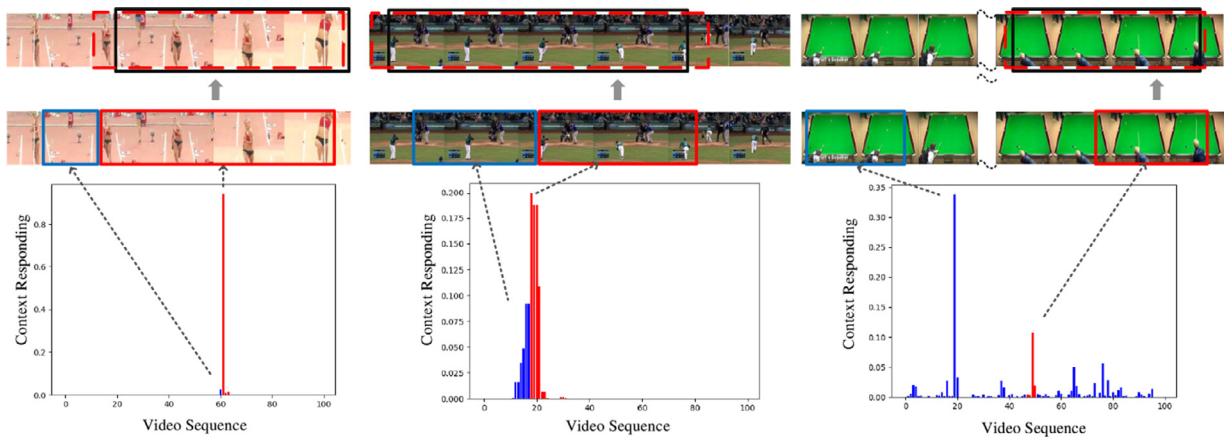

Fig. 6. Visualization of the Proposal-Context Relation. The horizontal axis of the histogram represents the snippet index within a video sequence, and the vertical axis represents the response value of each video snippet. The red solid lines represent temporal locations of action proposals to be updated and their corresponding visual content. The blue solid lines represent the informative context and the corresponding visual content obtained by PCRM. The black solid lines and the red dotted lines represent the ground truth and refined action proposals by PCRM, respectively.

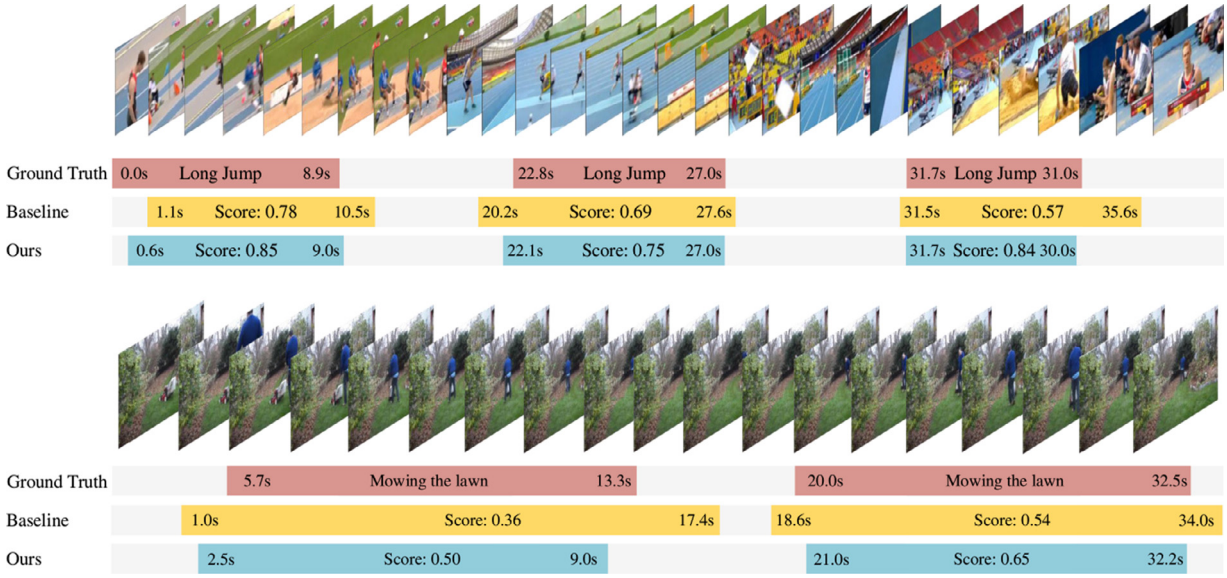

Fig. 7. Qualitative results. We show qualitative detection results on THUMOS14 (top) and ActivityNet v1.3 (bottom).

that our proposed method can recognize and localize the action instances more accurately, benefiting from the relation reasoning.

5. Conclusion

In this paper, we introduce a dual relation network for temporal action localization, which incorporates a proposal-proposal relation module and a proposal-context relation module. The proposal-proposal relation module processes action proposals simultaneously through interaction based on their appearance and geometry similarities, while the proposal-context relation module can efficiently encode temporal inter-dependencies between proposals and the global context. The two relation reasoning modules can jointly learn representative features by adaptively aggregating the proposal relation feature and context relation feature together to facilitate action localization. Our network is lightweight and interpretable, which also verifies the effectiveness of modeling action relations in CNN-based detection. Extensive experimental evaluations demonstrate that our method outperforms the state-of-the-art methods on the THUMOS14 and ActivityNet v1.3 datasets. Especially, it also could be a promising solution for spatio-temporal action localization, and we leave it for our future work.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or orga- nizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Acknowledgment

This work was supported partly by National Key R&D Program of China under Grant 2018AAA0101400, NSFC under Grants 62088102, 61976171, and 62106192, China Postdoctoral Science Foundation under Grant 2020M683490, Natural Science Foundation of Shaanxi Province under Grants 2022JC-41 and 2021JQ-054, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

[1] B. Liu, H. Cai, Z. Ju, H. Liu, Rgb-d sensing based human action and interaction analysis: a survey, Pattern Recognit. 94 (2019) 1–12.

[2] D. Zhang, L. He, Z. Tu, S. Zhang, F. Han, B. Yang, Learning motion representation for real-time spatio-temporal action localization, Pattern Recognit. 103 (2020) 107312.
[3] Y.H. Kim, S. Nam, S.J. Kim, Temporally smooth online action detection using cycle-consistent future anticipation, Pattern Recognit. 116 (2021) 107954.
[4] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, G. Hua, Two-stream consensus network for weakly-supervised temporal action localization, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 37–54.
[5] Y. Ge, X. Qin, D. Yang, M. Jagersand, Deep snippet selective network for weakly supervised temporal action localization, Pattern Recognit. 110 (2021) 107686.
[6] J. Zhang, H. Hu, Domain learning joint with semantic adaptation for human action recognition, Pattern Recognit. 90 (2019) 196–209.
[7] H. Eun, J. Moon, J. Park, C. Jung, C. Kim, Temporal filtering networks for online action detection, Pattern Recognit. 111 (2021) 107695.
[8] X.-Y. Zhang, H. Shi, C. Li, P. Li, Z. Li, P. Ren, Weakly-supervised action localization via embedding-modeling iterative optimization, Pattern Recognit. 113 (2021) 107831.
[9] J. Gao, Z. Yang, K. Chen, C. Sun, R. Nevatia, Turn tap: Temporal unit regression network for temporal action proposals, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 3628–3636.
[10] T. Lin, X. Zhao, H. Su, C. Wang, M. Yang, Bsn: Boundary sensitive network for temporal action proposal generation, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.
[11] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, C. Gan, Graph convolutional networks for temporal action localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 7094–7103.
[12] Z. Shou, D. Wang, S.-F. Chang, Temporal action localization in untrimmed videos via multi-stage cnns, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1049–1058.
[13] Y. Liu, L. Ma, Y. Zhang, W. Liu, S.-F. Chang, Multi-granularity generator for temporal action proposal, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 3604–3613.
[14] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, Q. Tian, Bottom-up temporal action localization with mutual regularization, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 539–555.
[15] Y.G. Jiang, J. Liu, A.R. Zamir, G. Toderici, I. Laptev, M. Shah, R. Sukthankar, Thumos challenge: action recognition with a large number of classes, 2014,
[16] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 961–970.
[17] T. Lin, X. Liu, X. Li, E. Ding, S. Wen, Bmn: Boundary-matching network for temporal action proposal generation, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 3889–3898.
[18] H. Su, W. Gan, W. Wu, Y. Qiao, J. Yan, Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation, in: Proc. AAAI. Conf. Artif. Intell. (AAAI), 2021, pp. 2602–2610.
[19] M. Xu, C. Zhao, D.S. Rojas, A. Thabet, B. Ghanem, G-TAD: Sub-graph localization for temporal action detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 10156–10165.
[20] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, N. Sang, Temporal context aggregation network for temporal action proposal refinement, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 485–494.
[21] J. Gao, Z. Shi, G. Wang, J. Li, Y. Yuan, S. Ge, X. Zhou, Accurate temporal action proposal generation with relation-aware pyramid network, in: Proc. AAAI. Conf. Artif. Intell. (AAAI), 2020, pp. 10810–10817.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of Neural Information Processing Systems, 2017, pp. 5998–6008.
[23] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, R. Girshick, Long-term feature banks for detailed video understanding, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 284–293.
[24] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, T. Yao, Exploring object relation in mean teacher for cross-domain detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 11457–11466.
[25] R. Su, D. Xu, L. Sheng, W. Ouyang, Pcg-tal: progressive cross-granularity cooperation for temporal action localization, IEEE Trans. Image Process. 30 (2020) 2103–2113.
[26] P. Chen, C. Gan, G. Shen, W. Huang, R. Zeng, M. Tan, Relation attention for temporal action localization, IEEE Trans. Multimed. 22 (10) (2019) 2723–2733.
[27] C. Sun, H. Song, X. Wu, Y. Jia, J. Luo, Exploiting informative video segments for temporal action localization, IEEE Trans. Multimed. (2021).
[28] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, C. Gan, Location-aware graph convolutional networks for video question answering, in: Proc. AAAI. Conf. Artif. Intell. (AAAI), volume 34, 2020, pp. 11021–11028.
[29] J. Pan, S. Chen, M.Z. Shou, Y. Liu, J. Shao, H. Li, Actor-context-actor relation network for spatio-temporal action localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 464–474.
[30] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7794–7803.
[31] L. Zhu, H. Fan, Y. Luo, M. Xu, Y. Yang, Temporal cross-layer correlation mining for action recognition, IEEE Trans. Multimed. (2021).
[32] Y. Wu, L. Zhu, Y. Yan, Y. Yang, Dual attention matching for audio-visual event localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 6292–6300.
[33] Y. Yang, Y. Zhuang, Y. Pan, Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies, Front. Inf. Technol. Electron. Eng. 22 (12) (2022) 1551–1558.
[34] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 6299–6308.
[35] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.
[36] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D.A. Ross, J. Deng, R. Sukthankar, Rethinking the faster R-CNN architecture for temporal action localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 1130–1139.
[37] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 761–769.
[38] R. Girshick, Fast R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1440–1448.
[39] C. Lin, J. Li, Y. Wang, Y. Tai, D. Luo, Z. Cui, C. Wang, J. Li, F. Huang, R. Ji, Fast learning of temporal action proposal via dense boundary generator, in: Proc. AAAI. Conf. Artif. Intell. (AAAI), 2020, pp. 11499–11506.
[40] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
[41] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, X. Tang, Cuhk & ethz & siat submission to activitynet challenge 2016. arXiv preprint arXiv:1608.00797.
[42] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Learning salient boundary feature for anchor-free temporal action localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 3320–3329.
[43] Z. Zhu, W. Tang, L. Wang, N. Zheng, G. Hua, Enriching local and global contexts for temporal action localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 13658–13667.
[44] C. Zhao, A.K. Thabet, B. Ghanem, Video self-stitching graph network for temporal action localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 13658–13667.
[45] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of Neural Information Processing Systems, 2014, pp. 568–576.

Kun Xia received the B.E. degree in Automation from Shenyang University of Technology, China, in 2017, and the M.E. degree in Control Science and Engineering from Northeastern University, China, in 2020. He is currently pursuing the Ph.D. degree at the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University. His research interests include computer vision and machine learning.

Le Wang 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 Stevens Institute of Technology, Hoboken, New Jersey, USA. From 2016 to 2017, he is a visiting scholar with Northwestern University, Evanston, Illinois, USA. He is currently a Professor with the Institute of Artificial Intelligence and Robotics of Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, and machine learning. He is an area chair of CVPR'2022, and a senior program committee member of AAAI'2022. He holds 7 China patents and has 16 more China patents pending. He is the author of more than 60 peer reviewed publications in prestigious international journals and conferences. He is a senior member of the IEEE.

Sanping Zhou received the Ph.D. degree in control science and engineering from Xian Jiaotong University, Xian, China, in 2020. From 2018 to 2019, he was a Visiting Ph.D. Student with the Robotics Institute, Carnegie Mellon University. He is currently an Assistant Professor with the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University. His research interests include machine learning, deep learning, and computer vision, with a focus on person re-identification, salient object detection, medical image segmentation, image classification, and visual tracking.

Gang Hua was enrolled in the Special Class for the Gifted Young of Xi'an Jiaotong University (XJTU), Xi'an, China, in 1994 and received the B.S. degree in Automatic Control Engineering from XJTU in 1999. He received the M.S. degree in Control Science and Engineering in 2002 from XJTU, and the Ph.D. degree in Electrical Engineering and Computer Science at Northwestern University, Evanston, Illinois, USA, in 2006. He is currently the Vice President and Chief Scientist of Wormpex AI Research. Before that, he served in various roles at Microsoft (2015–18) as the Science/Technical Adviser to the CVP of the Computer Vision Group, Director of Computer Vision Science Team in Redmond and Taipei ATL, and Principal Researcher/Research Manager at Microsoft Research. He was an Associate Professor at Stevens Institute of Technology (2011–15). During 2014-15, he took an on leave and worked on the Amazon-Go project. He was an Visiting Researcher (2011–14) and a Research Staff Member (2010-11) at IBM Research T. J. Watson Center, a Senior Researcher (2009-10) at Nokia Research Center Hollywood, and a Scientist (2006-09) at Microsoft Live labs Research. He is an associate editor of TIP, TCSVT, CVIU, IEEE Multimedia, TVCJ and MVA. He also served as the Lead Guest Editor on two special issues in TPAMI and IJCV, respectively. He is a general chair of ICCV'2025. He is a program chair of CVPR'2019&2022. He is an area chair of CVPR'2015&2017, ICCV'2011&2017, ICIP'2012&2013&2016, ICASSP' 2012&2013, and

ACM MM 2011&2012&2015&2017. He is the author of more than 200 peer reviewed publications in prestigious international journals and conferences. He holds 19 US patents and has 15 more US patents pending. He is the recipient of the 2015 IAPR Young Biometrics Investigator Award for his contribution on Unconstrained Face Recognition from Images and Videos, and a recipient of the 2013 Google Research Faculty Award. He is an IEEE Fellow, an IAPR Fellow, and an ACM Distinguished Scientist.

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