BMN_PAM: Boundary Matching Network with Pyramid Attention Module for Temporal Action Proposal Generation

Wu Han, Liang Jiayu

Abstract— **Temporal action proposal generation (TAPG) is a task that aims to generate temporal action proposals, i.e. temporal segments that potentially contain actions, in untrimmed videos. It is crucial for various video analysis and understanding tasks, e.g. temporal action detection, video understanding. However, existing TAPG works generally fail to consider global dependencies of proposals and cannot capture multi-scale features of temporal actions. In this work, a new TAPG method is proposed, termed as BMN_PAM (Boundary Matching Network with Pyramid Attention Module), which can obtain multi-scale feature information and establish long-term/global dependencies between proposals. Specifically, BMN_PAM applies BMN as a baseline method to generate action boundary probabilities. In addition, a new PAM is designed to generate the confidence map of proposals, which exploits multi-scale features and global dependencies of proposals. Then, both the action boundary probabilities and the confidence map are combined to generate accurate action proposals. A benchmark TAPG dataset, i.e. ActivityNet-1.3, is used to evaluate the proposed method. Compared with five updated TAPG methods, BMN_PAM performs best with 75.72 in AR@100 (Average Recall) and 67.38 in AUC (Area Under Curve). In addition, BMN_PAM is generally better than BMN-based methods with other attention mechanisms.**

Index Terms— **Temporal action proposal generation, attention mechanisms, multi-scale action features, global dependencies of proposals.**

1. INTRODUCTION

Temporal action proposal generation (TAPG) refers to the task of generating temporal action proposals or segments in a video [1],[2]. It involves identifying and localizing actions or events of interest within the temporal dimension of a video sequence.

The updated TAPG works [3]-[9] are described as follows. The work[3] generates two types of sliding windows combined to obtain the final proposals, but it does not take into account highly flexible boundary information. The works[4]-[7]adopt the bottom-up approach to generate candidate proposals that closely approximate the ground truth through accurate boundary probabilities. However, these methods[3]-[7] overlook the long-term/global dependencies relationship between proposals. In addition, Khoa et al. address interactions between agents and their environment; while Zhu et al.[8] introduce the CLGNet (Collaborative Local Global Learning Network) to capture dependency relationships of proposals. However, these approaches [8],[9] capture the feature space by convolution address interactions between agents and their environment; while Zhu et al.[8] introduce the CLGNet (Collaborative Local Global Learning Network) to capture dependency relationships of proposals. However, these approaches [8],[9]capture the feature space by convolution kernels with limited sizes, which limits their ability to extract multi-scale features. Therefore, it is necessary to investigate the TAPG methods that can not only establish long-term global dependencies of proposals but also obtain multi-scale feature spaces.

The attention mechanisms have been gradually applied in computer vision tasks. By incorporating attention mechanisms into TAPG methods, they can guide the TAPG models to attend to relevant temporal regions and focus on the most informative temporal segments, which can help develop global dependencies of proposals and improve the efficiency and accuracy of action proposal generation[10]. However, existing TAPG methods with attention mechanisms[10]-[14] rarely consider multi-scale feature spaces of temporal actions, which limits the performance of attention mechanisms for TAPG tasks.

To bridge the gap, this work proposes a new TAPG method, termed as BMN_PAM (Boundary Matching Network with pyramid Attention Module), which can obtain multi-scale feature information and establish global dependencies of proposals. Specifically, BMN_PAM applies BMN as a baseline method to generate action boundary probabilities. Compared with BMN, the PAM(Pyramid Attention Module) is designed and incorporated in BMN_PAM to establish global relationships among proposals and obtain multi-scale feature information, based on which to generate the confidence map of proposals. Then, accurate action proposals can be generated by MN_PAM using the action boundary probabilities and the confidence map of proposals.

The proposed BMN_PAM will be compared with updated TAPG methods and BMN-based methods with various attention mechanisms. Moreover, further analyses of BMN_PAM with different components will also be studied. A benchmark TAPG dataset, i.e. ActivityNet-1.3, will be used to testify the proposed and reference methods. Specifically, we will investigate the following sub-objectives:

1) explore whether BMN_PAM can outperform updated TAPG methods;
2) explore whether the proposed attention mechanism, i.e. PAM, can outperform existing ones for TAPG tasks.

Manuscript received February 28, 2024

Wu Han School of Software, Tiangong University, TianJin, China, 13840666294

Liang Jiayu School of Computer Science and Technology, Tiangong University, Tianjin, China,15620039978

This work is supported by National Natural Science Foundation of China (**grant number: 61902281**).

The remainder of this paper is organized as follows. Section 2 introduces the background, including TAPG methods and attention mechanisms for TAPG. Section 3 describes the proposed method. In addition, the experiment preparation are presented in section 4. The results of the proposed and reference methods are described and analyzed in section 5. Conclusions are drawn in section 6.

2. BACKGROUND

A. *TAPG methods*

In this section, updated TAPG methods are described, along with their advantages and disadvantages. Gao et al. [3]design a Complementary Temporal Action Proposal (CTAP) method. This method consists of three main modules, each responsible for different tasks in generating high-quality action proposals candidates. The first module generates two sliding windows. The second module combines the candidate segments generated by the two sliding windows to obtain a more comprehensive candidate action proposals. The third module selects the final high-quality action proposals candidate. Through the work of these modules, more accurate and comprehensive candidate action proposals are provided by generating diverse sliding windows, complementing each other, and selecting proposals.

Lin et al. [4] design a boundary-sensitive network (BSN) that aim to locate action boundaries accurately for TAPG. It consists of a proposal evaluation module and a boundary refinement module. The proposal evaluation module enerates initial action proposals based on a set of predefined anchor segments. It employs a two-stream network to capture appearance and motion features for proposal scoring. The boundary refinement module utilizes boundary regression to refine the action boundaries by estimating the temporal offsets for the proposal boundaries. BSN achieves state-of-the-art performance on THUMOS14 and ActivityNet-1.3 datasets in both action proposal recall and localization accuracy.

Lin et al. [5]design a boundary-matching network (BMN)

to generate proposals with both precise temporal boundaries and reliable confidence scores simultaneously. Specifically, a BM (Boundary-Matching) mechanism is proposed to generate confidence scores of densely-distributed proposals, where a proposal is denoted as a matching pair of starting and ending boundaries and all BM pairs are combined into a BM confidence map. Then based on the BM mechanism, BMN is proposed, which is an efficient, effective and end-to-end proposal generation method. BMN outperforms existing TAPG methods on THUMOS14 and ActivityNet-1.3 datasets with remarkable efficiency and generalizability.

Liu et al. [6] design a A Multi-Granularity Generator (MGG). MGG aims to capture actions within different time ranges, thereby improving the ability to recognize and locate actions of varying lengths and sequences. MGG utilizes multiple sliding windows of different time scales or other methods to generate candidate fragments with multiple granularities. Action proposals selects candidate segments with high activity as the final candidate action proposals. The method can generate diverse and highly actionable temporal candidate action proposals, so enhancing the accuracy and effectiveness of temporal action proposals tasks.

Tan et al. [7] propose a simple and effective framework (RTD-Net) for directly generating action proposals. Firstly, the framework proposes a boundary attention module that can remotely capture temporal information in videos. At the same time, adopting a relaxed matching scheme makes the standards between the ground truth more relaxed. Finally, designs a three branch head to obtain the final proposals. This method was tested on THUMOS14 and ActivityNet-1.3, demonstrating its ability to improve the quality of proposals.

Zhu et al.[8] think that existing TAPG methods only handle well the local/short-term dependencies among adjacent frames and generally cannot deal with the global/long-term dependencies among distant frames. They propose CLGNet (Collaborative Local-Global Learning Network), integrating TCN (Temporal Convolution Network) and BLSTM (Bidirectional Long Short-Term Memory), in which TCN handles local dependencies and BLSTM captures the global dependencies. Moreover, an attention mechanism (the background suppression module) is designed to guide the model to focus on actions. Experiments on THUMOS'14 and ActivityNet-1.3 show that CLGNet can outperform the reference methods.

Khoa et al. [9] propose a novel framework called ABN (Agent-aware Boundary Network), leveraging both local and global pathways to capture interactions between agents and their environment. ABN consists of two sub-networks, i.e. the agent-aware representation network and boundary generation network. The agent-aware representation network captures both agent-agent and agent-environment relationships in the video representation. The boundary generation network estimates the confidence score of temporal intervals. By evaluated on the THUMOS-14 and ActivityNet-1.3 datasets, ABN consistently outperforms state-of-the-art methods.

Based on the above analyses, some of the existing works generate proposals using sliding windows or anchors, which limit the boundary flexibility of temporal actions and fail to consider global dependencies of proposals. Moreover, in most existing works, the feature space is obtained only through a single convolution kernel, which cannot capture multi-scale features. Therefore, it is necessary to investigate the TAPG methods that can not only establish long-term global dependencies of proposals but also obtain multi-scale feature spaces.

B. *Attention mechanisms for TAPG*

The attention mechanism has been widely used in the field of natural language processing. Recently, it has been gradually applied in computer vision tasks, e.g. image classification and image segmentation. By incorporating attention mechanisms into TAPG methods, they can guide the TAPG models to attend to relevant temporal regions and focus on the most informative temporal segments, improving the efficiency and accuracy of action proposal generationite[10]-[14].

Hu et al. [10] propose the SE (Squeeze-and-Excitation) attention mechanism. It introduces a gating mechanism that adaptively recalibrates feature maps by assigning importance weights to different channels based on their relevance to the task. In addition, the SE blocks can be stacked together to generate the SENet architectures that won the first in ILSVRC 2017 classification competition.

Li et al.[11] propose a SKNet (Selective Kernel Networks). It is proposed as a deep learning network architecture for

image classification and recognition tasks.The key nnovation is the introduction of a selective convolution module, which enables the extraction of features from different aspects of the feature map and adaptively adjusting the receptive field. This process improves the expressive ability of the network. SKNet has demonstrated impressive results in various tasks, including image classification and object detection. It has achieved leading performance in competitions and challenges, solidifying its effectiveness in the field.

Wang et al.[12] propose an efficient channel attention mechanism called ECA-Net. It addresses the limitations of existing methods (e.g. the SE attention mechanism) that can be computationally expensive and memory-intensive for large-scale models. ECA-Net introduces a lightweight approach that uses a 1D convolutional operation to model channel dependencies. The ECA-Net achieves competitive performance while being computationally efficient and memory-friendly, making it a valuable addition to deep CNNs.

Li et al.[13] design the Hybrid Attention Module (HAM). This module sequentially applies channel attention and spatial attention, with a preference for using the channel attention module first. In the channel attention module, an adaptive mechanism and one-dimensional convolution are employed to capture cross-dimensional connections. Furthermore, based on the weights obtained from channel attention, the spatial attention module incorporates channel separation technology. This involves dividing the features into two groups along the channel and separately extracting spatial features, which are then added together to generate the output. Experimental results demonstrate that HAM, as a versatile module, enhances the performance of image classification tasks.

Deng et al.[14] propose a cross-channel interactive attention mechanism to facilitate information fusion among channels. This attention mechanism computes the interaction between each channel and its neighboring channels within a distance of K, thereby capturing local channel interaction information and enhancing computational efficiency. The value of K is determined using an adaptive approach. Additionally, the computation of channel attention information is performed using one-dimensional convolution. By incorporating the cross-channel interactive attention module at the end of the ShuffleNet-V2 base unit, the model's feature extraction capability is enhanced.

The existing works that aims to capture local relationships using convolutional techniques, while neglecting direct global relationships and multi-scale feature information. This limitation stems from the lack of effective attention mechanisms that can adequately model long-range dependencies without incurring excessive computational costs. Therefore, to address this issue, this paper proposes a pyramid attention mechanism that facilitates multi-scale feature fusion and establishes long-term global dependencies among proposals.

3. METHODOLOGY

In this section, the proposed BMN_PAM is described in details. Fig.1 shows the flowchart of BMN_PAM. It consists of five major parts, i.e. Video pre_processing, Temporal evaluation module in BMN, PAM, Temporal action proposal generation module in BMN and Post_processing. Firstly, video pre_processing generates feature sequences with input videos. In BMN, these features are used to generate boundary probabilities of actions. PAM contains multiple layers of attention mechanisms to capture multi-scale features and long-term dependencies of proposals, thus generating the confidence map of proposals. Then based on the boundary probabilities of actions and the confidence map of proposals, Temporal action proposal generation module in BMN is used to generate the final proposals. Eventually, in post_processing part, the redundant proposals are suppressed by Soft-NMS (Soft Non-Maximum Suppression) technique to generate more accurate proposals.

C. Video pre_processing

The target of video pre_processing is to produce the temporal action features of videos. For ActivityNet-1.3 dataset, wo-stream network is widely-used by existing works [15]. Two-stream network refers to the spatial stream and the temporal stream. The information in the input videos is extracted by the streams respectively, and then is combined to form the spatial-temporal features. The settings of feature extraction are the same as those in the work [5].

D. Temporal evaluation module in BMN

The target of this module is to produce the boundary probabilities of temporal actions, i. e. the start probability sequence Ps and the end probability sequence Pe. The BMN contains two major parts, i.e. the temporal evaluation module and the proposal evaluation module. Specifically, the temporal evaluation module is used to generate the boundary probabilities of actions in the proposed method (BMN_PAM). It consists of two one-dimensional convolutions with three convolution kernels.

E. PAM

The target of PAM is to produce the confidence map of boundaries, containing the information, i.e. the starting point of each action and the duration of each action. Fig. 2 shows the flowchart of PAM, which contains three major parts, i.e. the base module, the parallel processing module and the attention module. Firstly, the base module is able to change the receptive fields of the input features. Secondly, the parallel processing module is able to integrate multi-scale features by cluster convolutions with filters of different kernel sizes. The features obtained in the parallel processing module are connected together. Then, these features are processed by the attention module to obtain new feature sequences with attention weights to generate the confidence map.

1) Base module

The base module is used to change the receptive fields of the input feature sequences. It consists of two one-dimensional convolutions with three convolution kernels. Their activation functions are the “relu" function.

2) Parallel processing module

Parallel processing module can process multi-scale and multi-branch feature vectors. The specific procedures of the parallel processing module is as follows. Firstly, the input features are split into four branches. For each branch, the features are processed using the group convolutions. The group convolutions with multi-scale convolution kernels can extract multi-scale features. Secondly, the features generated by the group convolution from each branch are connected to get a new feature vector.

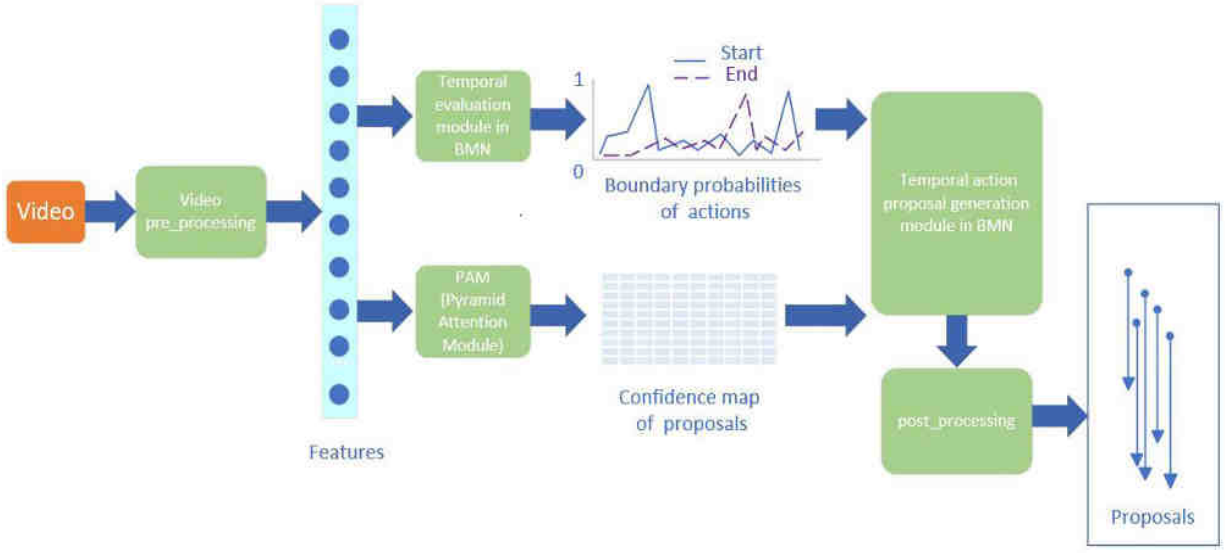

Fig. 1 The flowchart of BMN_PAM.

3) Attention module

The attention module aims to obtain the attention information that are the weights for input features. The specific procedure of the attention module is as follows. Firstly, the initial attention weight is generated by the SE-Net (Squeeze-and-Excitation Network) proposed in the work [10]. Secondly, the Softmax function is used to refine the attention weights, which allows for effective interaction between the local and global attention weights. Eventually, the attention weights are multiplied by the corresponding feature map to obtain the final confidence map.

4) Action proposal generation module in BMN

The target of the proposal generation module in BMN is to obtain the candidate proposals. The specific procedure of this module is as follows. Firstly, the starting and ending positions of actions with high probabilities are selected to generate one set with starting positions and one set with ending positions. Secondly, the starting and ending positions of actions are matched one by one based on the two sets, thus forming candidate action proposals. In this process, if the duration of a proposal is larger than the pre-defined maximum duration D, it will be discarded.

5) Post processing

The target of post processing is to suppress redundant proposals. The specific procedure of this module is as follows. After generating initial candidate proposals, the redundant proposals should be deleted to achieve a higher recall rate. In this work, the Soft-NMS (Soft Non-Maximum Suppression) technique is adopted, which can reduces the confidence scores of redundant proposals. Specifically, if a proposal's overlap degree is higher than the a pre-defined threshold, the proposal's confidence score will be reduced. In this way, more candidate proposals are retained, which avoids the situation of high overlap and improves the recall rate. Eventually, the action proposals with high flexible boundaries are obtained by matching the start and end boundaries flexibly.

4. Experiment Preparation

In this section, the experiment preparations in the dataset, experiment settings and the evaluation measures are introduced.

A. Dataset

A benchmark temporal action dataset, i.e. ActivityNet-1.3 dataset, is selected in this work. It is a large-scale dataset, which contains 19994 temporal videos with 200 action categories.The video samples are divided into training, validation and testing sets with a ratio of 2:1:1. It can be used for TAPG, action recognition, temporal detection and dense captioning tasks[4],[8],[9]

B. Experiment settings

The BMN_PAM model is trained using the Adam optimizer with a learning rate of 0.0001. The batch size is set to 16 and and the training process runs for 9 epochs. In addition, the experiments in this work are run on the paddle platform: the GPU is " Tesla V100"; the version of CUDA is 11.2; the Video Mem is 32GB; the RAM is 32GB.

C. Evaluation measures

Two measures, i.e. AR@AN (Average Recall (AR) under different Average Number (AN) of proposals) and AUC (Area Under the AR vs AN Curve), are selected to evaluate both the proposed and reference methods, since they are widely-used in evaluating TAPG methods. Specifically, AR@AN means the relationship between AR and the AN of candidate actions. The AN value varies from 0 to 100. Moreover, AR is computed under different tIoU (temporal Intersection over Union) thresholds that are set as [0.5:0.05:0.95]. In addition, AUC means the area under the AR and AN curves, which can provide an overall measure of the algorithm's performance.

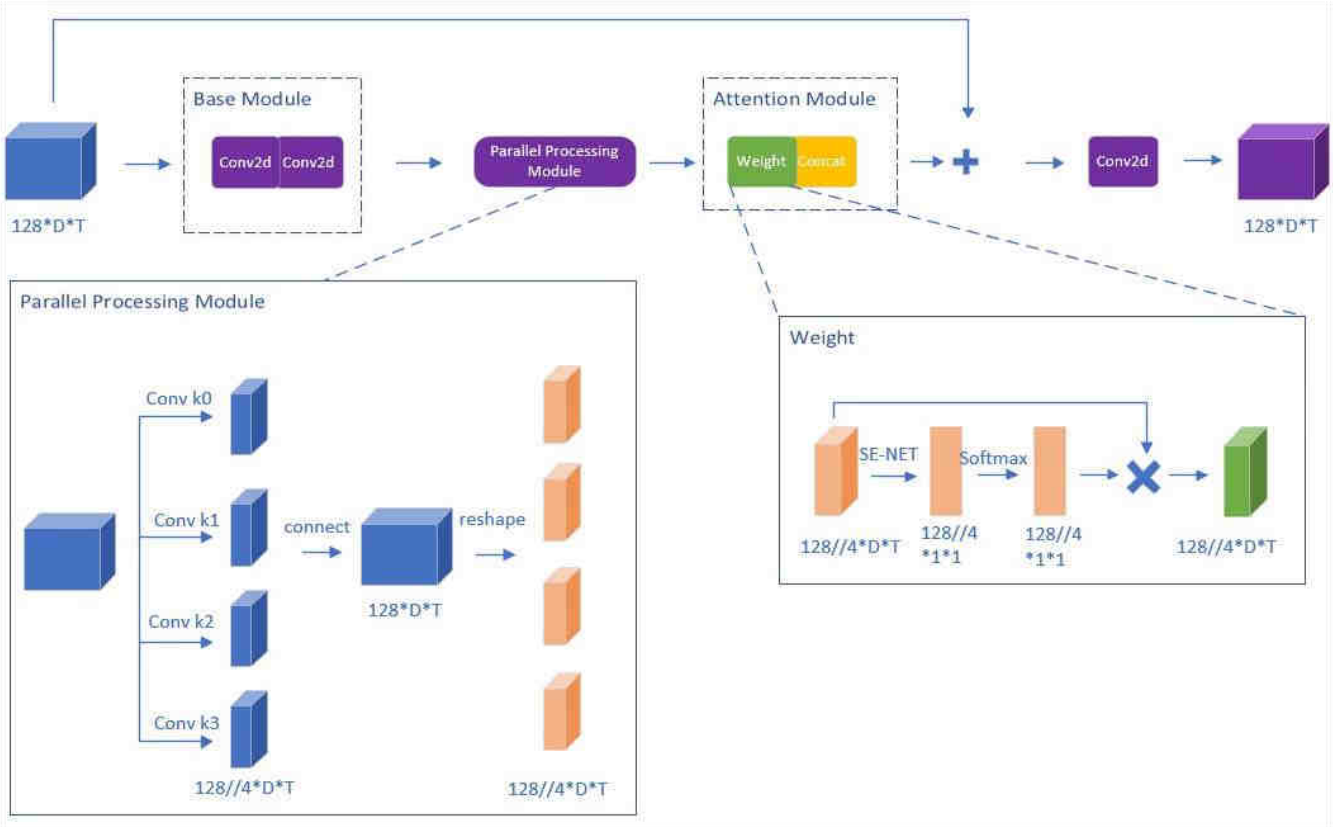

Fig. 2 The flowchart of PAM (T and D represent the length of the input feature sequence and the maximum proposal duration respectively)

5. Result Analyses

A. *Comparison with popular TAPG methods*

In this part, BMN_PAM is compared with five popular TAPG methods, i.e. BMN[5], CATP (Complementary Temporal Action Proposal Generation)[3], BSN (Boundary-Sensitive Network) [4], MGG (Multi-granularity Generator) [6] and RTD-Net (Relaxed Transformer Decoders Network) [7].

Table 1 shows the performance of the six methods in AR@1, AR@100 and AUC. All the evaluation measures are the higher the better. In terms of AR@1, compared with BSN (32.17) and RTD-Net (33.05), BMN_PAM achieves 33.16, which is slightly better. In terms of AR@100, BMN_PAM outperforms all the reference methods, which achieves 75.72. Moreover, MN_PAM also ranks the first in AUC. For example, the AUC value of BMN_PAM is 67.38; while those of others vary from 65.72 (CTAP) to 67.10 (BMN). The results reflect that the proposed method (BMN_PAM) outperforms the five popular reference methods for TAPG tasks.

B. *Comparison with BMN-based methods with different attention mechanisms*

In this part, BMN_PAM is compared with BMN-based methods that apply other attention mechanisms, i.e. SE (Squeeze-and-Excitation) [10], CA (Coordinate Attention) [16] ,SK (Selective Kernel) [11], ECANet (Efficient Channel Attention Network) [12]and SPANet (Spatial Pyramid Attention Network) [17]. In this way, it can be tested that whether the proposed attention mechanism (PAM) in BMN_PAM is better than existing mechanisms.

Table 2 presents the performance of BMN_PAM and five reference BMN-based methods with other attention mechanisms in AR@100, AUC and parameter number (in millions). In terms of AR@100, BMN_PAM is better than reference methods with a slightly higher value of 75.72; while the reference methods achieve 75.29 (BMN_SE), 75.23(BMN_CA),75.32(BMN_SK),75.44 (BMN_ECANet) and 75.42 (BMN_SPANet) respectively. In terms of AUC, BMN_PAM ranks the second with a value of 67.38, which is slightly lower than BMN_SK with 67.50. Even though BMN_SK is slightly better than BMN_PAM in AUC, it is worse in the AR@100 and the parameter number. Note that a higher parameter number means a higher complexity of the network.

Fig. 3,4,5,6,7,8 show the performance in the average recall under difference tIoU (temporal Intersection over Union) thresholds. Note that in these figures, the area under each curve can be calculated, which represents the AUC value and is the higher the better. In addition, the black curve depicts the average one of the curves with tIoU thresholds within [0.5, 0.95]. It can be seen that under different tIoU thresholds, BMN_PAM performs generally better than the reference methods. For example, when the tIoU is set to 0.7, the AUC of BMN_PAM is 76.34; while those of others are 76.63 (BMN_SK), 75.51 (BMN_SE), 75.39 (BMN_ECANet), 76.26 (BMN_CA) and 75.50 (BMN_SPANet) respectively.Except for BMN_SK that has a slightly higher AUC value, BMN_PAM is better than others in AUC under tIoU of 0.7.

Table 1 Comparison of BMN_PAM and reference TAPG methods (AR@1 means the average recall under .... proposals, and so forth).

Methods	AR@1	AR@100	AUC

BMN	-	75.01	67.10
CTAP	-	73.17	65.72
BSN	32.17	74.16	66.17
MGG	-	74.54	66.43
RTD-Net	33.05	73.21	65.78
BMN_PAM	33.16	75.72	67.38

Table 2 Comparison of BMN_PAM and BMN-based methods with other attention mechanisms (M refers to million).

Methods	AR@100	AUC	Parameter number (M)
BMN_SE	75.29	67.07	9.40
BMN_CA	75.23	67.29	9.42
BMN_SK	75.32	67.50	10.12
BMN_ECANet	75.44	67.06	9.40
BMN_SPANet	75.42	67.05	9.48
BMN_PAM	75.72	67.38	9.81

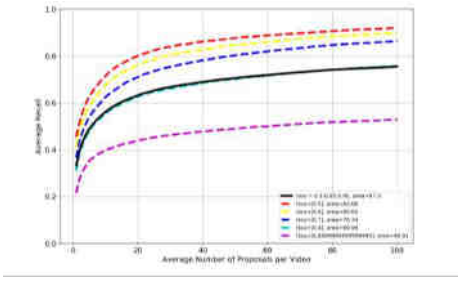

Fig. 3 The performance of BMN_PAM

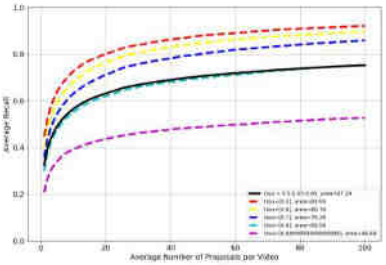

Fig. 4 The performance of BMN_CA

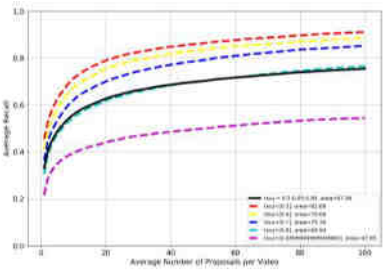

Fig. 5 The performance of BMN_ECANet

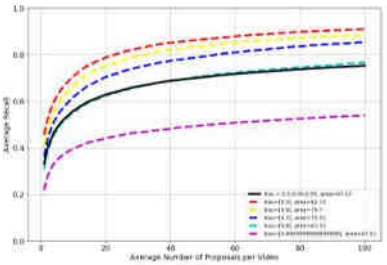

Fig. 6 The performance of BMN_SE

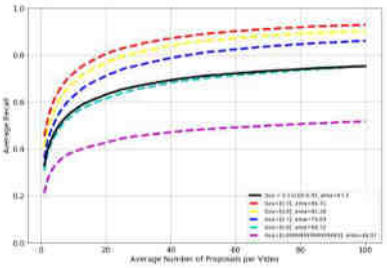

Fig. 7 The performance of BMN_SK

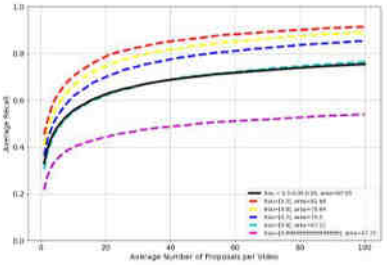

Fig. 8 The performance of BMN_SPANet

CONCLUSIONS

In this work, a new TAPG method is proposed, termed as BMN_PAM (Boundary Matching Network with Pyramid Attention Module). BMN_PAM can obtain multi-scale feature information and capture global dependencies between proposals. Specifically, BMN_PAM applies BMN as a baseline method to generate action boundary probabilities. In BMN_PAM, a new attention mechanism, PAM, is designed to generate the confidence map of proposals. PAM can exploit multi-scale features and global dependencies of proposals. Eventually, both the action boundary probabilities and the confidence map are combined to generate accurate action proposals. Compared with updated TAPG methods on ActivityNet-1.3, BMN_PAM performs best with 75.72 in AR@100 (Average Recall) and 67.38 in AUC (Area Under Curve).

In addition, BMN_PAM is generally better than BMN-based methods with other attention mechanisms. Attention mechanisms have achieved significant success in diverse fields, e.g. natural language processing, image processing and object detection. There are various attention mechanisms that can be combined together for better performance in TAPG tasks. Therefore, the optimal fusion of attention mechanisms deserves investigation in the future.

REFERENCES

[1] Lin C , Li J , Wang Y , et al. Fast learning of temporal action proposal via dense boundary generator. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(07): 11499–11506
[2] Yang H , Wu W , Wang L ,et al. Temporal action proposal generation with background constraint. Proceedings of the AAAI conference on artificial intelligence, 2022, 36(3): 3054–3062
[3] Gao J, Chen K, Nevatia R. Ctap: Complementary temporal action proposal generation. Proceedings of the European Conference on Computer Vision (ECCV), September 2018: 68-83
[4] 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 (ECCV), 2018: 3-19
[5] Lin T, Liu X, Li X, et al. Bmn: Boundary-matching network for temporal action proposal generation. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 3889-3898
[6] Liu Y, Ma L, Zhang Y, et al. Multi-granularity generator for temporal action proposal. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 3604-3613
[7] Tan J, Tang J, Wang L, et al. Relaxed transformer decoders for direct action proposal generation. Proceedings of the IEEE/CVF international conference on computer vision, 2021: 13526-13535
[8] Zhu Y, Han H, Liu G, et al. Collaborative Local-Global Learning for Temporal Action Proposal. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-14
[9] Vo K, Yamazaki K, Truong S, et al. Abn: Agent-aware boundary networks for temporal action proposal generation. IEEE Access, 2021, 9: 126431-126445
[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7132-7141
[11] Li X, Wang W, Hu X, et al. Selective kernel networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 510-519
[12] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 11534-11542
[13] Li G, Fang Q, Zha L, et al. HAM: Hybrid attention module in deep convolutional neural networks for image classification. Pattern Recognition, 2022, 129: 108785
[14] Deng Y, Zhang Y, Zhou Z, et al. A lightweight fetal distress-assisted diagnosis model based on a cross-channel interactive attention mechanism. Frontiers in Physiology, 2023, 14: 1090937
[15] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 2014, 27
[16] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021: 13713-13722
[17] Guo J, Ma X, Sansom A, et al. Spanet: Spatial pyramid attention network for enhanced image recognition. 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020: 1-6

Wu Han

She is a graduate student in Software Engineering at Tianjin University of Technology. And she focuses on temporal action detection and temporal action nomination generation.

Liang Jiayu

She is an associate professor at Tianjin Polytechnic University, focusing on artificial intelligence (mainly intelligent computing, computer vision/image processing, and data analysis).