Real-time Spatio-temporal Action Localization via Learning Motion Representation

Yuanzhong Liu, Zhigang Tu*, Liyu Lin, Xing Xie, and Qianqing Qin

Wuhan University, Wuhan 430079, China
{yzliu.me,tuzhigang,linliyu,Qqqin}@whu.edu.cn,rd_xiex@163.com

Abstract. Most state-of-the-art spatio-temporal (S-T) action localization methods explicitly use optical flow as auxiliary motion information. Although the combination of optical flow and RGB significantly improves the performance, optical flow estimation brings a large amount of computational cost and the whole network is not end-to-end trainable. These shortcomings hinder the interactive fusion between motion information and RGB information, and greatly limit its real-world applications. In this paper, we exploit better ways to use motion information in a unified end-to-end trainable network architecture. First, we use knowledge distillation to enable the 3D-Convolutional branch to learn motion information from RGB inputs. Second, we propose a novel motion cue called short-range-motion (SRM) module to enhance the 2D-Convolutional branch to learn RGB information and dynamic motion information. In this strategy, flow computation at test time is avoided. Finally, we apply our methods to learn powerful RGB-motion representations for action classification and localization. Experimental results show that our method significantly outperforms the state-of-the-arts on dataset benchmarks J-HMDB-21 and UCF101-24 with an impressive improvement of $\sim 8\%$ and $\sim 3\%$.

1 Introduction

Many breakthroughs have been witnessed in spatio-temporal action localization [1–6] mainly due to the progress of deep learning, and the emergence of large datasets [7–9]. Spatio-temporal action localization aims to not only identify the action category, but also localize it in both time and space. Inspired by the success of object detection, most current spatio-temporal action localization methods utilize the popular object detection frameworks [10–17], action bounding boxes are predicted in frame-level, then a dynamic linking strategy is used to generate human action tubes.

The key to accurately recognize human actions in videos is to effectively use both RGB information and motion information [6, 18–20], however, the 2D CNNs can not well model motion information. Many approaches have been proposed to extract motion. Using additional inputs, e.g. RGB difference and optical flow, are the common practices to learn motion information. Although using optical flow has indeed achieved good results, its disadvantages are obvious. Optical

flow needs to be calculated and stored in advance, where the computation of optical flow is very time-consuming. Since both the training and testing phases require optical flow, the previous optical flow based methods cannot be applied in real-time scenarios. The dual-stream network architecture [21, 22] also makes the network unable to be trained in an end-to-end way. Moreover, the motion information cannot be interactively integrated with the RGB information spatio-temporally.

To enable the network to learn motion features from RGB input and avoid the above shortcomings, we use the knowledge distillation method to train the network, which is effective to improve the performance of the network by learning motion features. To better fuse the long distance temporal information in the video, we also added the non-local module at the last layer of the 3D-Convolutional branch. For the 2D-Convolutional branch, which aims to predict the bounding box of the key frame, we use the short-range motion information extraction sub-network to further enhance the motion information near the key frame, so that the final generated action tube is smoother. In summary, we only use optical flow in the training stage of 3D-Convolutional branch, and the pre-trained model which is obtained through knowledge transfer is used in the spatio-temporal positioning task, and in this task optical flow is not required in both the training and the testing stage.

Our method is superior to the existing methods in the frame-mAP and the video-mAP with different thresholds. On the J-HMDB-21 dataset, the frame-mAP and video-mAP under the threshold 0.5 outperforms the second place by 5.8% and 3.9% respectively. On the UCF101-24 data set, compared to the baseline method [4], the accuracy of our method is improved by 2.8% in Frame-mAP (0.5), which is the current best result, and the video-mAP (0.5) is increased by 2%, which is close to the best result. To sum up, our contributions are as follows:

1. We use knowledge distillation to learn motion information from the optical flow stream. Non-local block is also used to help capture accurate long range temporal information.
2. We propose a novel motion cue called short-range-motion (SRM) module to learn short range temporal information, which makes the 2D-Convolutional branch to learn dynamic motion information.
3. The proposed method achieves the state-of-the-art performance by fusing the long and short-term motion information with RGB input.

This paper is organized as follows. In Section 2, we review the related work on S-T action localization. We introduce our method in Section 3. Section 4 presents the experimental results. We conclude our method in Section 5.

2 Related Work

Action recognition. Currently, the most successful action recognition methods [21, 23, 22, 24, 25] involve the use of optical flow in a two-stream way, which

typically consists of two branches to learn the appearance and motion information using RGB image and optical flow respectively. Although optical flow has been proven to be an effective motion information, but its estimation is very time-consuming and storage takes up a lot of space. Meanwhile, the contained information between optical flow and RGB images are redundant. How to accurately and efficiently model the motion information in the video to improve the extraction of spatiotemporal features is still far from being solved.

Spatio-temporal action localization. Most spatio-temporal action localization methods are expanded under the popular 2D target detection framework. The extensions mainly include: (1) optical flow is used to capture motion cues; (2) linking algorithms are applied to connect frame-level detections to form action tubes [5]. Although these methods achieved promising results, the temporal property of videos is not fully exploited. The use of pre-calculated optical flow not only hinders the interaction between the motion information and RGB information, but also makes the network unable to be trained end-to-end. Zhang [6] used PWC-Net[26] as a subnet integrated into the action localization network, so flow pre-computation is avoided. But this integration is simply placing the optical flow network before the action positioning network, and its results on the J-HMDB-21 and AVA datasets are not ideal. Köpüklü et al. [4] utilized a single-stage architecture with two branches to extract temporal and spatial information concurrently, and predict bounding boxes and action probabilities directly from video clips in one evaluation. The unified CNN architecture they proposed is novel, but it is unable to make full use of motion information. To fully use temporal information, a long-term feature bank (LFB) [27] is applied to utilize longer clip at inference time. Although the performance is significantly improved, it also brings a lot of extra computation burden and performs poorly on the J-HMDB-21 dataset.

Knowledge distillation. Integrating many models together and then averaging all the results is a simple way to improve model performance. However, the method of model integration causes multiple models to take up a lot of memory and also leads to huge calculations. Hinton et al. [28] put forward the concept of knowledge distillation. First, train a teacher network, and then use the output of the teacher network and the real labels of the data to train the student network. In this way, the small network can learn the knowledge of the large network. When the knowledge of a network is transferred to a network, the similar effect of model integration can be achieved. Nieves et al. [25] used the means of knowledge distillation to pre-train the action recognition stream with optical flow as input, then fix the network parameters, use the optical flow branch and the video action label to jointly train the RGB input branch, so that the network can learn from RGB. The motion information is learned from the input, which effectively improves the accuracy of action classification.

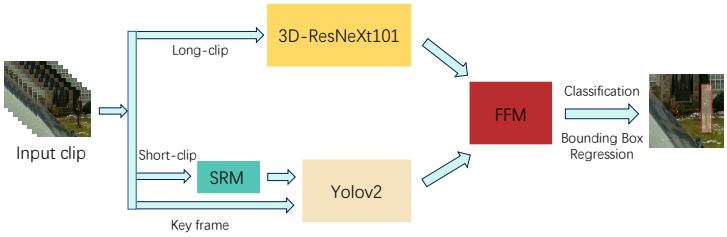

Fig. 1. MENet's architecture.

3 Learning motion representation

Modeling the motion information in the video is the key to realize the analysis and understanding of the video information. Two-stream method is the most popular S-T action localization methods. In the two-stream architecture, RGB images are usually used as an stream to extract RGB appearance information, and optical flow is used as another stream to extract motion information. The two branchs are independent, and the results are averaged at the end. Although the two-stream methods have achieved great success, its shortcomings are also obvious: First, Optical flow estimation is very time-consuming, and pre-saved on the hard disk will also take up a lot of storage space, which limits its real-world application. Second, there is a lot of redundant information between streams, which leads to a lot of unnecessary computing overhead. Third, the two separate branches make it impossible to train the network end-to-end, thus the motion information and RGB information cannot be interactively fused to obtain more robust spatio-temporal features, this leads to the video features obtained may be sub-optimal.

For frame-level video action localization, most algorithms extract key frame RGB information, and then use nearby consecutive n-frames of RGB input to extract temporal information. For the prediction of action categories, motion information at different times may be equally important, but for action positioning tasks, intuitively speaking, the importance of motion at different time distances is obviously different, the closer the time distance is, the more important for generating accurate key frame's action bounding box. Two-stream method ignores this fact, thus performance maybe harmed.

So, how to avoid the use of optical flow while achieving efficient extraction of motion information? How to strengthen motion representation at closer temporal distances?

Now, we propose our long short-term motion enhance method to solve the two problems. By using knowledge distillation and migration, while maintaining the performance of the two-stream method, we avoid using optical flow on the task of S-T action localization. At the same time, in order to enhance the motion information near the video key frames, we use a short-range-motion module(SRM)

to extract short-range motion information. Since the motion information near the key frame is additionally extracted and input into the network together with the RGB information, this is equivalent to explicitly increasing the importance of the motion with a shorter time distance. The experimental results show that our enhancement of long short-term motion effectively improves the accuracy of S-T action localization. We call the network enhanced by long short-term motion as motion-enhanced network (MENet).

3.1 MENet

MENet is based on the recently popular unified network YOWO [4]. YOWO has two branches, a 3D-CNN branch and a 2D-CNN branch, the key frame and the previous n(n=8/16) frames are input into the 3D backbone to extract spatio-temporal video features, key frame is input into 2D backbone to extract RGB features. The features of the two branches are merged together through a channel fusion and attention mechanism(CFAM) module. And then use a convolution layer to predict the action class and bounding box in Yolov2 style. The Yolov2 trained for 2D image target detection is applied to the 2D branch, thus only RGB appearance information can be obtained. Although the 3D-ResNeXt-101 network used in the 3D branch uses 3D convolution to extract video spatio-temporal information and is pre-trained in Kinetics, it does not make good use of motion information, and it has defects in its ability to extract long-range dependencies in videos. YOWO does not use optical flow. Good results are achieved on the J-HMDB-21 and UCF101-24 datasets by YOWO, but there are two obvious shortcomings: one is insufficient extraction of spatio-temporal information in the video, and the other is not using motion information.

The architecture of our MENet is as figure 1, which enhance long range temporal information in 3D-Convolutional branch and short range motion information is strengthened in 2D-Convolutional branch. A feature fusion module(FFM) is used to fusion 3D and 2D features to generate frame-level action detections. FMM is the collective name of CFAM and Yolov2 head. Frame-level outputs are then linked to generate action tubes using the same link strategy as YOWO [4].

3.2 Long range motion information

The typical application of knowledge distillation is to allow small models to learn from large models and obtain knowledge of large models, which is one of the ways of knowledge transfer. To enable the CNNs to learn the motion features provided by optical flow from continuous RGB frames, we use the knowledge distillation method to train the 3D CNN so that the network learns long range optical-flow-like motion features. 3D-ResNeXt-101 are selected as 3D backbone to make full use of long range motion information.As many frames as possible are used to extract long-range motion information, here we put 16 frames of images, which is the limit of YOWO. In video action recognition, Crasto [25] proposed a training method that uses motion to enhance RGB flow, so that the network can avoid calculating optical flow in the test phase. We used the

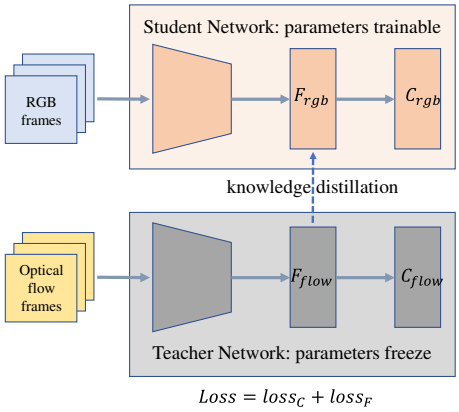

Fig. 2. Use knowledge distillation to train 3D-ResNeXt-101.

same strategy for the human action localization task. First, separately use RGB images and optical flow to train 3D-ResNext101 to classify videos. While the two CNNs are fully trained, the optical flow network parameters are fixed, and then the cross-entropy loss and the mean square error of the output of the global average pooling layer of the RGB branch network and the output of the global average pooling layer of the optical flow branch network are used as the joint loss to train the RGB network, the loss function is:

$$Loss_{K.D.} = CrossEntropy(C_{rgb}, \hat{y}) + \alpha \|F_{flow} - F_{rgb}\|^2 \tag{1}$$

where F_{flow} is the flow feature, F_{rgb} is the RGB feature, $\hat{y}$ is the video ground truth label, C_{rgb} is the predict label, α is a scalar weight modulating the influence of motion features.

In CNNs, the size of the receptive field of convolution and pooling operations is limited, so they are processing local block information, and can't well capture the long-range dependence in the video. Non-local module [29] is proven to be a way to effectively enhance long-range motion information in video. We adopted this practice and added a layer of non-local module at the last layer of the network to enhance the network's ability to extract long-range motion information.

3.3 Short range motion information

An optical flow frame is generally calculated from two adjacent frames, it represents the movement of pixels in the time interval of two images. Assuming that the time interval between the two frames is Δt, n optical flow frames represents the movement within $n\Delta t$. The smaller the value of n, the closer the time distance to the key frame. In the two-stream methods, n generally takes the value 16, 32, 64, which does not highlight the importance of a shorter time distance

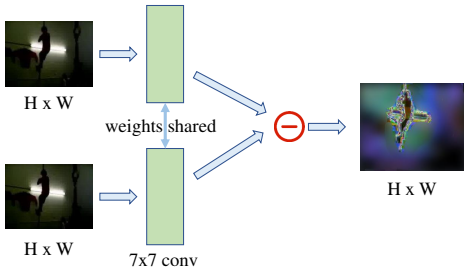

Fig. 3. Short-range-motion module.

for positioning the motion in the key frame. Considering that the motion of shorter time distance can help better locate the motion in the key frame, while extracting the long-range spatiotemporal information, we explicitly extract the motion information in the n neighboring video frames.

In order to verify that short-range motion information as input can indeed enhance the performance of S-T action localization tasks. We first use today's best-performing optical flow estimation method [30] to extract a frame of optical flow near the key frame and input it into the 2D backbone network together with the RGB key frame. The experimental results show that using optical flow as input can indeed greatly enhance the performance of the CNNs. But is optical flow the most appropriate way to express motion characteristics?

Many optical flow estimation algorithms[31, 26, 32, 30] today use parameter-sharing twin networks to extract the features of adjacent RGB video frames, and then use the correlation operation to construct a minimum cost volume to match the pixels in the two frames of images. With many methods such as crose-to-fine, downsampling, warp, and parameter sharing, the speed and accuracy of the optical flow have been greatly improved. But for video tasks that frequently input dozens of consecutive frames, embedding the optical flow estimation network as a sub-network in the video task will cause a large amount of graphics card memory consumption. And the training process will be more difficult and slow. In our practice, it will increase the training time by 2 to 3 times. Research [33] shows that RGB difference can also effectively represent motion information, which can be regarded as a rough representation of motion information. Although RGB difference as one of the motion information representation methods does not improve the performance of video tasks as well as optical flow, it can be obtained simpliy by only performing one subtraction. Based on the above observations, we propose our short-range motion information modeling module(SRM). In our SRM, feature encoding network of parameter sharing is used to extract features from RGB video frames, and then the features are subtracted to obtain motion features.

During our review of the literature, we found that Zhang et al. [34] have made relevant attempts. We implemented our SRM (as fig 3) based on the sug-

gestions given in their experiments, and tested our ideas in the 2D backbone. Experimental results show that with only a small increase in calculations, SRM can quickly and accurately extract motion information and effectively enhance the network's spatiotemporal feature extraction capabilities.

4 Experiments

To evaluate the performance of the algorithm, UCF101-24 and J-HMDB-21 datasets are selected. We strictly follow the official evaluation metrics to measure the performance of the algorithm and compare our results with other state-of-the-art methods.

4.1 Datasets and metrics

JHMDB-21 In order to promote an in-depth understanding of action recognition algorithms, Jhuang et al. [8] used part of the HMDB-51 [35] dataset to annotate human joint points to create the JHMDB-21 dataset. Since the 928 videos they extracted from HMDB-51 belong to 21 action categories, this dataset is also called the JHMDB-21 dataset.The labeled data in the JHMDB-21 dataset includes motion segmentation, dense optical flow, and human joint points. 928 video clips all contain 15 or more video frames.

UCF101-24 UCF101-24 [18] has more data than the JHMDB-21 dataset. Based on the UCF-101 data set, 3194 video action frames are labeled. Since they belong to 24 action categories, they are called UCF101-24 datasets. Although the UCF101-24 dataset has not been officially released, it has been revised by many scholars and its labels is relatively reliable.

metrics We use general indicators for target detection tasks to measure the performance of the algorithm on the task of action positioning. This paper will report the Frame-mAP(F-mAP) index when the IOU threshold is 0.5 and the Video-mAP(V-mAP) index under various IOU thresholds.

4.2 Training settings

We choose 3D-ResNeXt101 as the 3D backbone and Yolov2 as the 2D backbone. The consecutive 16 frames (including key frames) before the key frame are input into the 3D backbone, the consecutive 4 frames (including key frames) before the key frame are input into the SRM, the key frame will be concated with the output of SRM into the 2D backbone. For the 3D backbone network, a non-local module is added in the last layer, then we use the same training settings as crasto et al. [25] for knowledge distillation training, the pre-trained weights are used to initialize the 3D branch of MENet. For the 2D backbone network, we initialize the network with Yolov2s pretrained weights, the number of input channels of the first layer of convolution is modified to 6 to accommodate the input. We use momentum stochastic gradient descent optimization algorithm for training, the momentum value is set to 0.9, and the weight decay rate is set to 0.0005.

For the JHMDB-21 dataset, set the initial learning rate to 0.0001 and the batch size to 16, and the weight is decayed by half every 10000 batch iterations. For the UCF101-24 dataset, set the initial learning rate to 0.001 and the batch size to 32. When the number of iterations reaches 20000, 30000, 40000, and 50000 nodes, the learning rate will be reduced by half.

4.3 Ablation study

Long-Range-Motion-Enhancement. In common practice, usually 16, 32, 64 frames of continuous images are used as input for video tasks. Limited by the requirements of the network architecture, only 16 frames can be input at most in MENet. We first use knowledge distillation to enhance the motion information, and then add a non-local layer to the last layer of the backbone network to enhance the long-range motion information in the network. In order to verify the validity of the results, we separately use the 3D branch to conduct experiments in the UCF101-24 data set, results are shown in the Table 1.

We compared various metrics, including the classification accuracy and the localization when the IOU threshold is 0.5 on the training set, the F-score and Frame-mAP on the validation set, and the Video-mAP on the test set. It can be clearly seen that the knowledge distillation training significantly improves the network perfermance. The non-local module further enhances the performance of the network in positioning tasks. After using knowledge distillation and non-local modules at the same time, F-mAP increased by 4.31%, and V-mAP increased by 1.9%. This proves that our method is effective and the enhancement of network performance is continuous.

Table 1. 3D backbone performance comparison with different motion enhancement methods. RGB only means to use only RGB training, K.D. refers to knowledge distillation training, K.D.+non-local refers to the use of non-local modules while using knowledge distillation.

Methods	Classif.	local.(0.5)	F-score	F-mAP(0.5)	V-mAP(0.5)
RGB only	96.02	92.70	91.11	77.98	44.30
K.D.	96.10	93.30	91.70	80.28	45.00
K.D.+non-local	**96.10**	**94.0**	**92.40**	**82.29**	**46.20**

Short-Range-Motion-Enhancement. As we all know, 2D networks are good at extracting the apparent features of RGB images. Using optical flow and other motion information as input can also significantly enhance network performance. For action positioning tasks, we enhance the extraction of short-range time distance motion information by inputting the motion information close to the key frame into the network. We use Yolov2 as the 2D backbone network to conduct experiments on the UCF101-24 dataset, and we compare the performance of the network under different input conditions. We first input only the RGB image,

Table 2. 2D backbone performance comparison with different short range motion enhancement methods. RGB only means to use only RGB training, RGB + optical flow(n=1) refers to input RGB and 1 optical flow frame together, RGB+SRM(n=1) refers to input RGB and 1 SRM frame together, RGB+SRM(n=3) refers to input RGB and 3 SRM frame together.

Methods	Classif.	local.(0.5)	F-score	F-mAP(0.5)	V-mAP(0.5)
RGB only	62.4	79.4	51.57	30.11	13.50
RGB+optical flow(n=1)	68.2	82.1	57.59	36.90	14.96
RGB+SRM(n=1)	**75.2**	**87.7**	**68.15**	**50.36**	**29.35**
RGB+SRM(n=3)	**79.9**	**88.9**	**73.09**	**52.91**	**30.61**

and found that the results are poor. Then we input a frame of optical flow together with the RGB image into the 2D network, the F-mAP of the network has increased by 6.79%, and the V-mAP has increased by 1.46%. Then we input a frame of SRM, which greatly improves network performance, compared with optical flow, its F-mAP has increased by 13.46% and V-mAP has increased by 12.9%. Taking into account the noise between the information of the two frames of images, we finally use four frames of images to generate 3 SRM output. With RGB key frames as input, the experiment results demonstrate this further enhances the performance of the network, F-mAP increased by 22.80%, and V-mAP increased by 17.11%.

4.4 Comparison to the state of the art

We compare with other state-of-the-art S-T action localization algorithms. It should be noted that we did not compare with the VideoCaptureNet [36], because its video-map calculation method is different, it doesn't generate action tubes via linking strategies to calculate Video-mAP. To be fair, we did not compare with YOWO plus LFB. YOWO plus LFB use 8 LFB[27] features which extracted for non-overlapping 8-frame clips using the pretrained 3D ResNeXt-101 backbone, total number of 64 frames are utilized at inference time. Our nerual network architecture is based on YOWO, since we aim to explore how to better use motion information, we did not use the special strategy of LFB at inference time.

As shown in Table 3, on the UCF101-24 dataset, our algorithm has achieved advanced results in various metrics. We have achieved the current best F-mAP 83.5%. Compared with the original YOWO, our MENet has improved significantly in all indicators. MENet without SRM, F-mAP reached 83.5%, the current best, 3.1% ahead of second YOWO; under IOU 0.2 threshold V-mAP reached 75.9%, leading YOWO 0.1%; V-mAP reaches 49.5% under the IOU 0.5 threshold, leading YOWO by 0.7%. MENet with SRM further improved V-mAP at multiple thresholds, especially V-mAP at high threshold, V-mAP reaches 50.8% under the IOU 0.5 threshold, leading YOWO by 2%. There is a slight drop in frame-mAP, which may be due to the fact that the first few frames of the video

Table 3. Perfermance comparison with state-of-the-art algorithms on the UCF101-24 dataset.

Method	F-mAP(0.5)	V-mAP		
		0.1	0.2	0.5
Peng w/o MR [37]	64.8	49.5	41.2	-
Peng w/ MR [37]	65.7	50.4	42.3	-
ROAD [38]	-	-	73.5	46.3
T-CNN [39]	41.4	51.3	47.1	-
ACT [40]	69.5	-	77.2	51.4
MPS [41]	-	82.4	72.9	41.1
STEP [5]	75.0	83.1	76.6	-
YOWO [4]	80.4	82.5	75.8	48.8
MENet w/o SRM (ours)	**83.5**	**82.1**	**75.9**	**49.5**
MENet w/ SRM (ours)	**83.2**	**82.4**	**76.7**	**50.8**

are reused due to too few numbers, and SRM enhances noise. MENet considers motion information and strengthens the extraction of long-range dependent information in the video, various indicators have been significantly improved.

As shown in Table 4, on the J-HMDB-21 dataset, our algorithm has achieved advanced results in various metrics too. We have achieved the current best F-mAP 82.3% and best V-mAP(0.5) 61.7%. MENet without SRM, F-mAP reached 82.3%, the current best, 7.9% ahead of second YOWO; under IOU 0.2 threshold V-mAP reached 92.7%, the current best, leading YOWO 4.9%; under IOU 0.5 threshold V-mAP reached 90.9%, the current best, leading YOWO 5.2%; V-mAP reaches 57.9% under the IOU 0.75 threshold, within an inch of YOWO. MENet with SRM further improved V-mAP at high threshold, V-mAP reaches 61.7% under the IOU 0.75 threshold, leading YOWO by 3.6%. The J-HMDB-21 dataset is much smaller than UCF101-24 dataset, it has a lot of shorter video clips. It is easier to overfit, so the F-mAP and V-mAP under low threshold drop more. But it is clear that our method has achieved the best results currently on this data set, this proves the effectiveness of our work.

5 Conclusion

In this paper, we focus on learning motion representation efficiently without using optical flow.We use the knowledge distillation training method to enable the network to learn motion information from RGB input, and use non-local modules to better improve the network's ability to integrate motion information. To emphasize the motion information closer to the key frame, we use the SRM module to extract the motion information of the neighboring frames of the key frame. Good results prove the effectiveness of our method. In the future, we will explore more effective and more efficient ways to extract motion information to boosting video analysis and understanding.

Table 4. Perfermance comparison with state-of-the-art algorithms on the J-HMDB-21 dataset.

Method	F-mAP(0.5)	V-mAP		
		0.2	0.5	0.75
Peng w/o MR [37]	56.9	71.1	70.6	48.2
Peng w/ MR [37]	58.5	74.3	73.1	-
ROAD [38]	-	73.8	72.0	44.5
T-CNN [39]	61.3	78.4	76.9	-
ACT [40]	65.7	74.2	73.7	52.1
P3D-CTN [42]	71.1	84.0	80.5	-
TPnet [43]	-	74.8	74.1	61.3
YOWO [4]	74.4	87.8	85.7	58.1
MENet w/o SRM (ours)	**82.3**	**92.7**	**90.9**	**57.9**
MENet w/ SRM (ours)	**78.3**	**90.5**	**88.1**	**61.7**

Acknowledgments. The work is supported by the National Key Research and Development Program of China (No.2018YFB1600600)

References

1. Yang, Z., Gao, J., Nevatia, R.: Spatio-temporal action detection with cascade proposal and location anticipation. arXiv preprint arXiv:1708.00042 (2017)
2. He, J., Deng, Z., Ibrahim, M.S., Mori, G.: Generic tubelet proposals for action localization. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2018) 343–351
3. Ye, Y., Yang, X., Tian, Y.: Discovering spatio-temporal action tubes. Journal of Visual Communication and Image Representation **58** (2019) 515–524
4. Köpüklü, O., Wei, X., Rigoll, G.: You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. arXiv preprint arXiv:1911.06644 (2019)
5. Yang, X., Yang, X., Liu, M.Y., Xiao, F., Davis, L.S., Kautz, J.: Step: Spatio-temporal progressive learning for video action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 264–272
6. Zhang, D., He, L., Tu, Z., Zhang, S., Han, F., Yang, B.: Learning motion representation for real-time spatio-temporal action localization. Pattern Recognition **103** (2020) 107312
7. Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision **2** (2012)
8. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: International Conf. on Computer Vision (ICCV). (2013) 3192–3199
9. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6047–6056
10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 779–788

11. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 7263–7271
12. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, Springer (2016) 21–37
14. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2015) 1440–1448
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2017) 2961–2969
16. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 6154–6162
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
18. Singh, G., Saha, S., Sapienza, M., Torr, P.H., Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3637–3646
19. Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Learning to track for spatio-temporal action localization. In: Proceedings of the IEEE international conference on computer vision. (2015) 3164–3172
20. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. (2008)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. (2014) 568–576
22. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6299–6308
23. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2015) 4489–4497
24. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7083–7093
25. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2019) 7882–7891
26. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8934–8943
27. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 284–293
28. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat **1050** (2015) 9
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)

30. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. arXiv preprint arXiv:2003.12039 (2020)
31. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2015) 2758–2766
32. Hui, T.W., Tang, X., Change Loy, C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8981–8989
33. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision, Springer (2016) 20–36
34. Zhang, C., Zou, Y., Chen, G., Gan, L.: Pan: Towards fast action recognition via learning persistence of appearance. arXiv preprint arXiv:2008.03462 (2020)
35. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV). (2011)
36. Duarte, K., Rawat, Y., Shah, M.: Videocapsulenet: A simplified network for action detection. In: Advances in Neural Information Processing Systems. (2018) 7610–7619
37. Peng, X., Schmid, C.: Multi-region two-stream r-cnn for action detection. In: European conference on computer vision, Springer (2016) 744–759
38. Saha, S., Singh, G., Sapienza, M., Torr, P.H., Cuzzolin, F.: Deep learning for detecting multiple space-time action tubes in videos. Pattern Recognition (2015)
39. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (t-cnn) for action detection in videos. In: Proceedings of the IEEE international conference on computer vision. (2017) 5822–5831
40. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 4405–4413
41. Alwando, E.H.P., Chen, Y.T., Fang, W.H.: Cnn-based multiple path search for action tube detection in videos. IEEE Transactions on Circuits and Systems for Video Technology (2018)
42. Wei, J., Wang, H., Yi, Y., Li, Q., Huang, D.: P3d-ctn: Pseudo-3d convolutional tube network for spatio-temporal action detection in videos. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 300–304
43. Singh, G., Saha, S., Cuzzolin, F.: Predicting action tubes. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 0–0