Learning and Distillating the Internal Relationship of Motion Features in Action Recognition *

Lu Lu[1], Siyuan Li[1], ✉Niannian Chen[1], Lin Gao[1], Yong Fan[1], Yong Jiang[1], and Ling Wu[1]

Southwest University of Science and Technology
✉chenniannian@swust.edu.cn

Abstract. In the field of video-based action recognition, a majority of advanced approaches train a two-stream architecture in which an appearance stream for images and a motion stream for optical flow frames. Due to the considerable computation cost of optical flow and high inference latency of the two-stream method, knowledge distillation is introduced to efficiently capture two-stream representation while only inputting RGB images. Following this technique, this paper proposes a novel distillation learning strategy to sufficiently learn and mimic the representation of the motion stream. Besides, we propose a lightweight attention-based fusion module to uniformly exploit both appearance and motion information. Experiments illustrate that the proposed distillation strategy and fusion module achieve better performance over the baseline technique, and our proposal outperforms the known state-of-art approaches in terms of single-stream and traditional two-stream methods.

Keywords: Action Recognition · Knowledge Distillation · Temporal Modeling · 3D Convolution

1 Introduction

With the advent of convolutional neural networks (CNN) [17], great progress has been made in the research field of activity understanding [21]. Generally, video-based activity understanding is to analyze, recognize, and label the human movements appearing in existing videos. Through training the neural networks with large video datasets [11,16] and the technologies of transfer learning, the CNN-based approaches achieve superior generalization in the task of action recognition. At present, the popular action recognition approaches are mainly divided into two categories: (1) Two-stream Convolutional Networks [24,31]

The first aforementioned architecture, the two-stream convolutional networks, focus on the exploitation of frame-wise information and the corresponding motion between frames, the two types of information are separately processed in

* Supported by organization The Foundation of Sichuan Provincial Education Department(18ZA0501).

two different trained networks to produce classification scores, and the two scores are fused to obtain the final prediction of action classes. [24] is one of the well-known approaches that applied the framework of the two-stream convolutional networks to video-based action recognition. Although classification accuracy can be improved by sampling multiple frames from one video and averaging score in this approach, continuous frames sampling will collect a large number of redundant frames. Therefore, Wang *et al.* [31] proposes the sparse sampling and designs aggregation functions module to utilize the multi-frame sequence in the spatial stream. However, their temporal stream network lacks generalization.

Due to the motion information also exists in the raw RGB frames, the C3D-based approach [28,36] introduces the 3D convolution kernels to learn spatiotemporal features from video clips. Following this fashion, Region Convolutional 3D Network (R-C3D) [34] utilizes the 3D convolution [28] kernel to extract the features and generalizes the region proposal network and regions of interest pooling of Fater R-CNN [22] to the temporal domain. Liu *et al.* [19] also utilizes 3D CNN to extract the features while capturing the correlation between spatial signals and temporal information. Furthermore, Xu *et al.* [35] effectively integrates the two-stream method with3D convolution, which significantly improves the classification accuracy. However, accurately obtaining optical flow require considerable computational cost, and thus increases the latency of action prediction. This prerequisite brings enormous difficulties to practical applications of this work.

To deal with this problem, Motion-Augmented RGB Stream (MARS) [4] trains a standard 3D convolutional neural network that mimics the representation and function of the motion stream. Due to the unitary appearance input (only images input), it can avoid traffic computation of optical flows during the testing phase. However, in this approach, the existing features of the motion stream are not well exploited. It does not learn the features of the middle layer in motion stream, nor does it absorbs the distinct characteristics of flow features in the motion stream and the relationship among these features.

In this paper, we also train a dual-action model that only accepts RGB images but generates some feature representation similar to those obtained by a trained motion stream network. Our research denote the dual-action model as Dual-action Stream (DS). Notably, the Dual-action Stream in our research is optimized by mean absolute error of multi-level features and the $L1$ distance between the Gram matrixes of internal feature blocks. This optimization strategy better learns the motion representation among the middle layers of the motion stream, as well as the style and internal relationship of the features. Moreover, similar to the MAR, the student network (Dual-action Stream) not only learns the representation of motion stream network but also is automatically optimized and adjusted by cross-entropy loss of classification and the inputted image. Thus the Dual-action Stream defacto has double functions - the representation of motion feature and the knowledge for RGB input. Because the learned Dual-action Stream does not input the optical flow, it does not need additional computation of optical flow in the inference stage. The evaluation on two well-known action

recognition dataset demonstrates the proposed method outperforms the other known state-of-art approach.

In summary, this paper mainly has the following contributions:

- This paper proposes a novel knowledge distillation method that better learns the characteristics of the representation of motion stream network and the internal relationship of motion features, and it has double functions.
- We propose an efficient, low-memory attention-based fusion module to fuse the classification scores of two different streams, which can be applied to any action recognition approaches that have two more streams.

2 Preliminaries

2.1 C3D

Previous literature [28, 34, 36] have focused on the study of spatiotemporal features, which based on supervised learning of deep 3D convolution network on large-scale video datasets.

The 3D ConvNets (C3D) used in this work has advantages in video processing over 2D ConvNets. Carreira *et al.* [3] not only release the Kinetics video dataset but also expand the 2D-Inception module of Inception-v1 [27] to 3D with a additional time dimension. However, the direct application of 3D ConvNets is time-consuming and thus cannot achieve real-time recognition, and I3D [3] has very low compatibility with long-range temporal video modeling. To deal with these issues, Xie *et al.* [33] extract the spatiotemporal information using 3D and 2D convolution, and balance the trade-off between speed and accuracy. Their work demonstrates that the 3D ConvNets is more suitable to process low-level features than 2D ConvNets. Therefore, the primary network of our research is based on more efficient 3D ConvNets.

2.2 Optical Flow

The optical flow [1] contains the dynamic information of moving objects in video. By utilizing the variation and correlation of pixels in the image domain, optical flow represents the relationship between the current frame and the previous frame. Thus the movement of the object can be discovered in optical flow. Therefore, the optical flow can be used to represent the motion of the target object.

However, the calculation of optical flow takes up considerable resources and consumes a lot of time. In the past, one of the popular real-time methods to extract optical flows is sparse optical flows [2], which can be used for object tracking and so on. The basic principle of sparse optical flows is to calculate feature key points by using the assumption of neighborhood consistency of optical flow, then to track and extract some crucial key points. Rather the dense flow [6] calculates the optical flow for each point, and the calculation means of each point are the same. Therefore, the calculation cost of dense optical flow is much greater. Although the TV-L1 algorithm [20] calculates dense flow, it has a much faster solution. The calculation method of the optical flow used in this paper is TV-L1 [20, 37].

2.3 Knowledge Distillation

In previous research, Hinton *et al.* [13] first propose the concept of knowledge distillation by transferring knowledge of pre-trained teacher networks to the student network. In this work, the teacher network is complex and huge, but the reasoning performance of the teacher network is superior, while the student network is lightweight and has low complexity. In addition, Romero [23] considers feature maps in the middle layer of the teacher network to guide the corresponding layer in the student network. The advantage of this work is that it can efficiently compress the model but does not compromise the performance of the model. However, the inputs of the student network and teacher network in this article have the same modality.

Based on previous work, Garcia *et al.* [8] proposed a new distillation that can be used in multi-mode stream network architecture. One of the recent advances [4] advocates a distillation strategy that transfers the knowledge from the trained motion stream to an RGB stream that only receives appearance input, this cunning methodology avoids heaving network structure and the explicit flow computation in the inference phase of action recognition.

The above methods either do not take into account the correlation of the features of each layer nor the learning of the intermediate features is not sufficient. Our method differs from the above works as: (a) the distillation strategy in our method can better learn the features of the middle layers, and (b) the strategy can better learn the internal relationship of the features in the middle layers. The distillation method used in this paper is explained in Section 3.2. Our method of learning the internal relationships of features is described in Section 3.3. Section 3.4 is our attention-based fusion module. Our experiments are in Section 4.

3 Simulation of Motion Stream

There are state-of-the-art approaches [18,26] that attempt to retrieve the motion and appearance representation in a single network stream using 3D convolution. They propose distinct modules to exploit motion information better. Nevertheless, these modules lead the architecture cumbersome and only gain modest improvement.

Inspired one of the distillation methods [4], we propose an optimized distillation solution that utilizes not only the explicit privileged knowledge of motion stream but also distillate latent information from the features of the pre-trained motion stream. Different from the MAR [4] approach, the proposed strategy discovers and mines more comprehensive motion information from the intermediate layers of the optical flow stream, which promotes the student network to learn from the teacher network more effectively. Furthermore, this work proposes a lightweight and effective fusion module to fuse the scores produced by the motion streams and the proposed Dual-action Stream, which improve the performance of two-stream architecture that can be applied to the scenario in possession of sufficient computing resources.

3.1 Distilling Motion Information

At the beginning of the training phase, a 3D convolutional network is constructed and trained to classify the category of the inputted flow clips, i.e., the motion stream. Based on the trained motion steam, we can build a similar network to learn the representation and knowledge of the trained motion stream, where the second network acts as the role of the student. The student network inputs and processes RGB clips while mimicking the feature extraction functions of the trained motion stream. According to the past distillation strategies [4], the Motion-Augmented RGB Stream (MAR) adopts Mean Squared Error (MSE) loss to reduce the Euclidean distance of high-level features between the motion stream and targeted stream (i.e., MAR stream). The MSE is formalized as Equation 1.

$$\mathcal{L}_{\mathrm{MSE}} = \left\| f_{\mathrm{MAR}(n-1)} - f_{\mathrm{FLOW}(n-1)} \right\|_2 \tag{1}$$

where the n represents the number of total layers of the network, $f_{\mathrm{MAR}(n-1)}$ refers to the features produced by the layer before the final linear layer of the motion-augmented stream. $f_{\mathrm{FLOW}(n-1)}$ refers to the feature generated by of $n-1$ layer of optical flow stream.

However, in the MAR approach, the significance of low-level features is ignored, and the latent information is not well exploited. To address these issues, we propose a novel distillation stragegy that comprehensively extracts the knowledge of motion stream into the proposed Dual-action Stream meanwhile learning the knowledge of RGB frames. In the proposed distillation strategy, there are three kinds of loss terms to guide network learning.

3.2 Learning Multi-level Knowledge

According to the literature [38], the deeper layers of neural networks produce high-level global representations, while the shallow layers stand for low-level local features. Therefore, the proposed Dual-action Stream adopts the Mean Absolute Error (MAE) loss to simultaneously learn multi-level features of the motion stream. Intuitively, the proposed distillation stragegy can be graphically shown in Figure 1. Denote the $f_{\mathrm{DS}(i)}$ as i'th layer features of Dual-action Stream network. $f_{\mathrm{FLOW}(i)}$ refers to the feature of optical flow stream i'th layer. The multi-level MAE loss term of Dual-action Stream can be expressed as Equation 2.

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left\| f_{\mathrm{DS}(i)} - f_{\mathrm{FLOW}(i)} \right\|_1 \tag{2}$$

This multi-level loss term distills the different level information of the trained motion stream into our proposed Dual-action Stream network that only operates on RGB frames. Besides, to utilize and exploit the existing appearance input in the training phase, the Dual-action Stream also adopts a categorical cross-entropy loss as the second loss term.

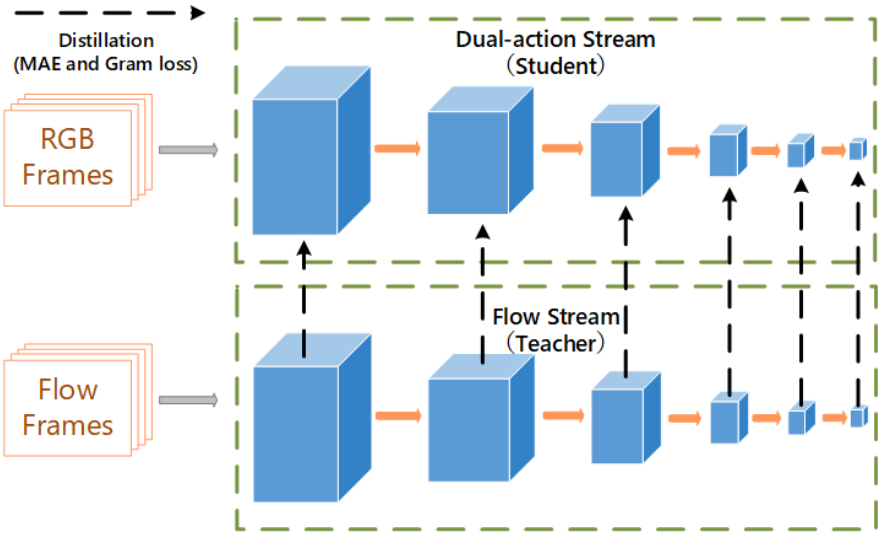

Fig. 1. The distillation strategy operating on multi-level features between the Dual-action Stream and motion stream.

3.3 Learning Internal Relationship of Features

[9, 15] adopt the distance of Gramian matrix of features to learn the texture and style of the input RGB images. The Gramian matrix is calculated by the inner product of flattened vectors of multi-channel features. It represents the characteristics and directionality between features, which can be thought of as texture and style information of RGB images in the case of [9, 15]. Similar to these methodologies, the proposed work applies the Gramian matrix loss to the intermediate features of motion stream and proposed Dual-action Stream, which aims to capture and learn the characteristics and internal relationship of motion features. Let the C_i denote the channel number of features generated by i'th layer, $\overline{f}_{(i)}$ the flattened vectors of i'th features, the G the function of the inner product. The Gramian loss of Dual-action Stream is described as Equation 3.

$$\mathcal{L}_G = \sum_{i}^{n-1} \frac{\left\| G\left(\overline{f}_{\mathrm{DS}(i)}\right) - G\left(\overline{f}_{\mathrm{FLOW}(i)}\right) \right\|_1}{(n-1)C_i T_i H_i W_i} \tag{3}$$

where the T_i, H_i, and W_i respectively refer to the temporal length, height, and width of feature block generated by i'th layer.

To aggregate and exploit the aforementioned useful knowledge, we propose a joint loss to backpropagate the Dual-action Stream network, which yields the network automatically integrate the motion representation and appearance in-

formation, and further improve the classification accuracy. The total loss of Dual-action Stream can be mathematically expressed as Equation 4.

$$\mathcal{L}_{DS} = \text{CrossEntropy}\ (h_{\text{DS}}, y) + \alpha(\mathcal{L}_{\text{MAE}} + \mathcal{L}_G) \tag{4}$$

where h_{DS} refers to the class prediction score of Dual-action Stream network, y is the groud truth label of multi-classification, α is a scalar weight that regulates the influence of all motion information. $\mathcal{L}_{\text{MAE}}$ is multi-level MAE loss, $\mathcal{L}_G$ is the loss of gram function.

3.4 Attention-based Fusion module

Even though Dual-action Stream can identify the action category by itself, the two-stream prediction still can improve the final classification accuracy. For the exploratory studies and the scenarios where computing resources are not strictly required, the two-stream approach can still be used.

In this paper, we employ a linear neural module to replace the averaging fusion to integrate the scores produced by the trained two streams (e.g., Dual-action Stream and motion stream). The architecture of the proposed fusion module is displayed as Figure 2.

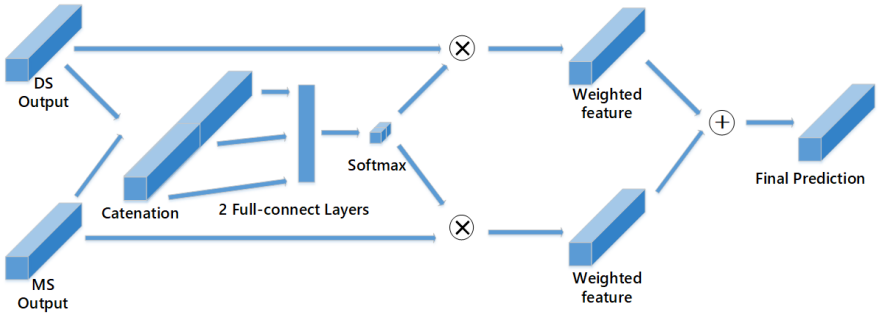

Fig. 2. The architecture of the proposed fusion module, DS output refers to the output of trained Dual-action Stream, the MS means the motion stream.

In the fusion module, the outputs of two streams are concatenated and fed into a twin-layer linear model. After forwarding the two-dimensional output of the linear model (i.e., two full-connect layers) into the softmax layer, the importance weights of each stream are obtained. Finally, the weighted sum of the scores of two streams is considered as the final prediction score. Because the parameters of motion stream and Dual-action Stream are frozen, through the backpropagating the cross-entropy loss, the designed fusion module can obtain the reasonable weights of each stream by updating the learnable parameters of the linear model.

4 Experiment

In this section, this paper describes the experimental details and results of the Dual-action Stream and the two-stream method with the new fusion module. First of all, we introduce the benchmark datasets and details of the implementation and training process. Afterward, we compare our approaches with the state-of-art approach over the two well-known datasets of action recognition. Finally, we explore the effectiveness of applying different components to our proposed Dual-action Stream.

4.1 Datasets

The experiments are conducted and evaluated on two challenging datasets: UCF-101 [25], HMDB-51 [14]. The UCF-101 dataset, one of the most well-known action recognition datasets, consists of 3320 realistic videos from 101 diverse action categories. And the spatial resolution of the original videos in this dataset is 320 $\times$ 240. The UCF-101 dataset provides abundant video resources related to human activity, which aims to help research on realistic action recognition. The Human Motion Database (HMDB-51) is a publicly available database containing 6766 video clips distributed in 51 action categories. These videos are extracted from a variety of sources ranging from digital movies to YouTube videos and manually annotated. For both datasets, this paper adopts the standard evaluation protocol and evaluates the classification accuracy on the split 1 test set of the two datasets.

4.2 The Details of Implementation and Training

In this article, networks are built and trained in the PyTorch framework using four GeForce GTX 1080Ti GPUs with a total of 44G memory. The optical flows that are used for training motion stream and fusion module are generated by the TV-L1 algorithm [20] with the default setting, all the RGB frames are extracted by the raw video at 25 fps, and all these initial features are resized to 256$\times$256.

This paper selects the 3D ResNet-101 network [32] as the backbone of motion stream and Dual-action Stream. Similar to [12], We adopt a mini-batch stochastic gradient descent optimizer to train all models proposed in our approach, and the initial learning rate is set as 0.1, which will reduce by a weight decay of 0.0005 and the momentum of 0.9. In the training process, based on the previous experience [4], this paper set the hyperparameter α as 50 in Equation 4.

Following [4,12], we conduct and evaluate the proposed approach on models with two kinds of input types, 16 consecutive frames clip (16f-clip) and 64 consecutive frames clip (64f-clip). At the training phase, a random clip of the given length (i.e., 16 or 64) is sampled from the video or optical flows, then the clip is cropped into the region of 112$\times$112 and randomly apply horizontal flipping.

Because the training process of motion stream is exactly the same as [4], we utilize the pre-trained motion model provided by [4], where the last block and the

Table 1. The accuracy of experiment over UCF101 and HMDB-51 dataset

Methods	Pre-train dataset	UCF-101			HMDB-51		
		RGB	Flow	RGB+Flow	RGB	Flow	RGB+Flow
Two-stream Network [24]	ImageNet	73	83.7	88	40.5	54.6	59.4
ConvNet fusion [7]	ImageNet	82.6	86.2.7	90.6	47	55.2	58.2
DTPP [40]	ImageNet	89.7	89.1	94.9	61.5	66.3	75
TLE+Two-stream [5]	ImageNet	-	-	95.6	-	-	71.1
ActionVLAD [10]	ImageNet	-	-	92.7	49.8	59.1	66.9
C3D [28]	sports-1M	82.3	-	-	51.6	-	-
C3D [29]	sports-1M	85.8	-	-	54.9	-	-
R(2+1)D [30]	sports-1M	93.6	93.3	95	66.6	70.1	72.7
TSN [31]	ImageNet	85.7	87.9	93.5	-	-	68.5
I3D [3]	ImageNet	84.5	90.6	93.4	49.8	61.9	66.4
R(2+1)D [30]	ImageNet+Kinetics	96.8	95.5	97.3	74.5	76.4	78.7
TSN [31]	ImageNet+Kinetics	91.1	95.2	97	-	-	-
CCS + TSN [39]	ImageNet+Kinetics	94.2	95	97.4	69.4	71.2	81.9
Distillnation Methods	Pre-train dataset	Mimic	Mimic+RGB	Mimic+Flow	Mimic	Mimic+RGB	Mimic+Flow
MAR(16f) [4]	Kinetics	94.6	95.6	94.9	72.3	73.1	74.5
Dual-action Stream(16f)	Kinetics	95.2	95.6	95.6	73.7	73.5	76.8
MAR(64f) [4]	Kinetics	97.1	95.8	97.5	80.1	80.6	80.9
Dual-action Stream(64f)	Kinetics	97.2	97.6	97.7	80.3	80.8	81.2

last fully-connected layer of motion network are finetuned from the model pre-trained on Kinetics400 [16]. During the phase of training Dual-action Stream, all the parameters of the motion stream are frozen to ensure that the Dual-action Stream properly simulates the optical flow network. In order to train the fusion module, both parameters of motion stream and Dual-action Stream are fixed in favor of convergence.

4.3 Comparison, Ablation study, and Analysis.

So as to compare the proposed training strategy to the state-of-the-art action recognition approaches, we report the performance over the split 1 test dataset of UCF-101 and HMDB-51 in Table 1. Note that "Mimic" in the table refers to Dual-action Stream/MAR stream [4], and the "Mimic+RGB/Flow" column refers to the averaging fusion method used in the traditional two-stream methods [24]. It's obvious that our proposed Dual-action Stream outperforms the other state-of-art approaches on the 64 consecutive frames clip (64f-clip), and our Dual-action Stream even exceeds some of the two-stream methods. On 16f-clip, it shows that our results also have better advantages in the case of less sampling. As for the comparison of the two-stream fusion method, our averaging fusion approach with Dual-action Stream and flow stream is 9.7% better than the original two-stream method [24] on UCF-101 dataset, and 21.8% on HMDB-51 dataset. For more benchmark, Our experimental results are 4.2% higher than TSN [31] on UCF-101 and 12.7% higher on HMDB-51. Experiments show that our method can better learn the features of the middle layers because we achieve state-of-the-art performance.

Compared with 64f-clip, 16f-clip have fewer input frames, which is more demanding for our method. Table 2 shows our ablation study over the UCF-101 dataset. All these studies are conducted on the 16f-clip input. Due to the [4] adopt the original MSE loss between the layers only before the last fully-

Table 2. The ablation study over the UCF101 dataset at split 1(16f-clips), the different modules are gradually added to the baseline.

Methods	Accuracy
Baseline [MAR] [4]	94.6
Dual-action Stream (Multi-level MAE Loss)	94.7
Dual-action Stream (Gram Loss)	94.8
Dual-action Stream (Multi-level MAE Loss+Gram Loss)	95.2
Dual-action Stream+Flow	95.6
Dual-action Stream+Flow+Attention-based Fusion	**95.7**

connected layer, we consider it as our baseline. Based on the baseline, our ablation experiment evaluate the performance by adding multi-level MAE loss and Gram loss step by step. Finally, we evaluate the attention-based fusion module by integrating Dual-action Stream with the optical flow stream. Although the MAR [4] achieve amazing high performance, this table shows that the proposed Dual-action Stream and the attention-based fusion module can efficiently improve the performance.

5 Conclusion

In this paper, a novel distillation strategy is proposed to learn from the motion stream comprehensively, therefore it only receives RGB clips but hiddenly utilizes both appearance and motion information. The proposed gram loss and multi-level feature loss are proved to be able to learn motion information more effectively. The evaluation and comparison showed that our Dual-action Stream outperforms most of the two-stream approaches over the UCF101 and HMDB51 dataset.

Acknowledgment

This research is supported by The Foundation of Sichuan Provincial Education Department (NO. 18ZA0501).

References

1. Beauchemin, S.S., Barron, J.L.: The computation of optical flow. ACM computing surveys (CSUR) **27**(3), 433–466 (1995)
2. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. International journal of computer vision **61**(3), 211–231 (2005)
3. 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. pp. 6299–6308 (2017)

4. 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. pp. 7882–7891 (2019)
5. Diba, A., Sharma, V., Van Gool, L.: Deep temporal linear encoding networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2329–2338 (2017)
6. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. pp. 363–370. Springer (2003)
7. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
8. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 103–118 (2018)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
10. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 971–980 (2017)
11. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV. vol. 1, p. 3 (2017)
12. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
14. Jhuang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T.: A large video database for human motion recognition. In: Proc. of IEEE International Conference on Computer Vision. vol. 4, p. 6 (2011)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
16. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
17. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
18. Lee, M., Lee, S., Son, S., Park, G., Kwak, N.: Motion feature network: Fixed motion filter for action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 387–403 (2018)
19. Liu, H., Tu, J., Liu, M.: Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)
20. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: Tv-l1 optical flow estimation. Image Processing On Line **2013**, 137–150 (2013)
21. Poppe, R.: A survey on vision-based human action recognition. Image and vision computing **28**(6), 976–990 (2010)
22. 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. pp. 91–99 (2015)

23. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
24. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
25. 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)
26. Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W.: Optical flow guided feature: A fast and robust motion representation for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1390–1399 (2018)
27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
28. 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. pp. 4489–4497 (2015)
29. Tran, D., Ray, J., Shou, Z., Chang, S.F., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038 (2017)
30. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
31. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence (2018)
32. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
33. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 305–321 (2018)
34. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision. pp. 5783–5792 (2017)
35. Xu, H., Das, A., Saenko, K.: Two-stream region convolutional 3d network for temporal activity detection. arXiv preprint arXiv:1906.02182 (2019)
36. Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., Maybank, S.J.: Asymmetric 3d convolutional neural networks for action recognition. Pattern Recognition **85**, 1–12 (2019)
37. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint pattern recognition symposium. pp. 214–223. Springer (2007)
38. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
39. Zhang, J., Shen, F., Xu, X., Shen, H.T.: Cooperative cross-stream network for discriminative action representation. arXiv preprint arXiv:1908.10136 (2019)
40. Zhu, J., Zhu, Z., Zou, W.: End-to-end video-level representation learning for action recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 645–650. IEEE (2018)