MGN: MULTI-GLIMPSE NETWORK FOR ACTION RECOGNITION

Chaoxu Guo[1,2], *Tingzhao Yu*[1], *Huxiang Gu*[1], *Shiming Xiang*[1], *Chunhong Pan*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
guochaoxu2017@ia.ac.cn, {tingzhao.yu, hxgu, smxiang, chpan}@nlpr.ia.ac.cn

ABSTRACT

Current state-of-the-art action recognition approaches rely on optical flow to extract the local motion information and ignore the importance of global description of the videos. In this paper, we present a novel architecture, named Multi-Glimpse Network (MGN), to boost the performance of action recognition by combining the local and global information of the videos. Specifically, MGN makes predictions through two important modules, Local Glimpse and Global Glimpse. Local Glimpse extracts the local spatiotemporal features of different periods using temporal sampling method. Global Glimpse aggregates the extracted local features to develop global description of the videos. These two modules are complementary and indispensable. Our MGN achieves competitive results on four video action benchmarks of UCF101, HMDB51, ODAR and Penn.

Index Terms— Action Recognition, Multi-Glimpse, Temporal Sampling, Global Description

1. INTRODUCTION

Video-based action recognition has attracted a significant amount of attention from the academic community, owing to its potential applications in many areas such as human-computer interaction and security surveillance. However, it is still a challenging task due to the large variations of scale, viewpoint and camera motion in videos. To address these challenges, many attempts have been made to improve the robustness and performance of recognition systems [1, 2, 3, 4, 5].

Initially, several local descriptors have been proposed to represent the 3D volumes extracted around the interest points, such as Histogram of Gradient (HOG) [6], Histogram of Optical Flow (HOF) [7] and 3D Histogram of Gradient (HOG3D) [8]. Afterwards, a more powerful feature, Improved Dense Trajectories (IDT) [2], which extracts HOG, HOF and Motion Boundary Histograms (MBH) [6] features along the trajectories computed with the interest points in frames of videos, has been proposed and dominated the field of video analysis for years. Despite their superior performance, features such as IDT fail to represent the semantic information of videos. In human action recognition, there are two types of correlated and crucial information: appearance and motion, which are complementary to each other. However, these traditional manually designed features are hard to capture these two types of information simultaneously. In addition, recognition systems using these features are easy to be misled by similar backgrounds and agents of actions.

Recently, Convolutional Neural Networks (CNNs) have witnessed the great success in many domains such as image classification [9, 10, 11, 12], detection [13] and segmentation [14]. To improve the recognition performance, researchers have been trying to design effective neural network structures for action recognition in videos [15, 16, 17, 18]. Karpathy et al. [3] tested four different ConvNets on a large video dataset (Sports-1M) but got the poor performance, which implied that 2D convolutional filters can't capture temporal information directly from stacked frames. To deal with this issue, Simonyan et al. [4] proposed two-stream ConvNets composed of the spatial and temporal network, of which the spatial network captured the features of appearance while the temporal network captured the dynamic features of motions from the pre-computed optical flow. Further, Christoph et al. [19] fused two-stream network spatially in an attempt to align the spatial and temporal information and boosted the performance. Despite the better performance than traditional methods, the dependance on optical flow will lead to heavy computation. To avoid this, Tran et al. [5, 20] designed 3D CNNs to capture the spatiotemporal features from the frames of videos directly. However, the approaches aforementioned aren't capable of dealing with actions of long-range like Rock Climbing. To address this, Wang et al. [18] proposed a temporally segmental network to exploit the information of different segments in videos. And Joe et al. [21] used LSTMs to learn the long-term information from videos by aggregating the frame-level features.

Long-range temporal information is crucial for video-based action recognition. However, only a proportion of frames contain class-specific information. When we see different parts of the videos, we will have a new attention and obtain different informa-

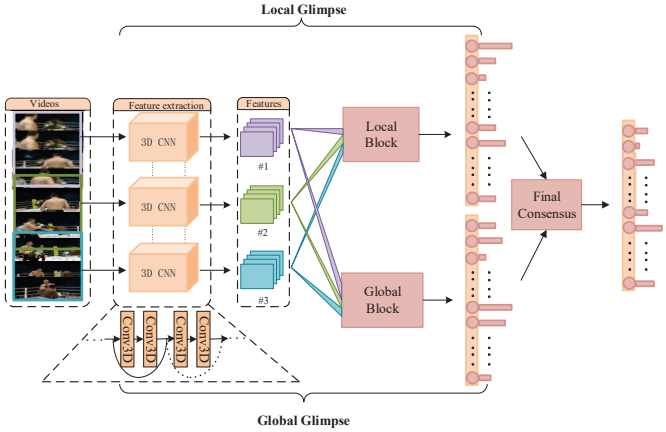

Fig. 1. MGN: Three snippets are chosen from one input video and run through two paths to produce multi-attention features. Local prediction and global prediction are made with corresponding multi-attention features and finally final prediction of MGN is refined from both of them.

This work is supported by the National Natural Science Foundation of China (Grant No. 61773377 and 91646207).

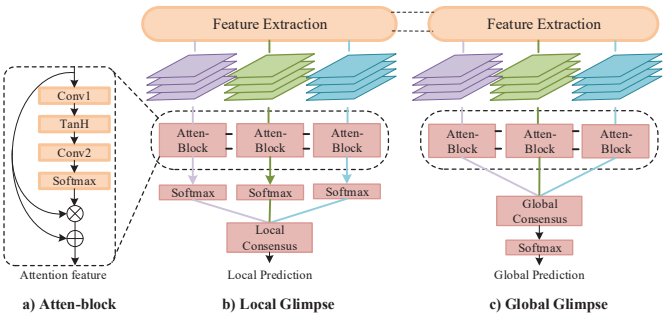

Fig. 2. a) is the attention block in multi-attention block. b) and c) are respectively the Local Glimpse and Global Glimpse in MGN.

tion. Motivated by this, we propose the Multi-Glimpse Network (MGN) as shown in Fig. 1, which makes video-level prediction through Local Glimpse and Global Glimpse. In MGN, feature extraction is implemented with the convolutional part of Res3D [20] and shared by Local Glimpse and Global Glimpse. Specifically, these two modules extract features from three short snippets, which are sampled from videos using temporal sampling method. In [18], Wang et al. divided a video into K segments of equal durations and each snippet of optical flow or single frame was sampled from the corresponding segments. In contrast, we allow the sampled snippets to overlap over each other for a small part to augment the datasets and encourage the correlation of three parts of the video to be exploited by our network. And we only use single network to capture both the appearance and motion information simultaneously with only RGB input. In order to avoid absorbing all information without selection, multi-glimpse attention mechanism is introduced to preserve the most relevant features from different snippets.

Summarily, our contributions are as follows:

- We introduce temporal sampling method into MGN to learn local and global spatiotemporal features and it mitigates the problem of class-inbalance caused by the large variance of length of the videos.
- In order to preserve the most relevant information, we introduce a multi-glimpse attention mechanism into MGN. For each snippet of different periods, the corresponding glimpse of attention will preserve the class-specific features and suppress the disturbance of noise.
- We introduce the modules of Local Glimpse and Global Glimpse, as shown in Fig 2, which are applicable in other clip-level frameworks. Based on the local and global features, these two components make initial predictions from two different perspectives. And the final prediction is refined from both initial predictions.

2. MULTI-GLIMPSE NETWORK

Multiple local class-specific features can provide more evidence for the final prediction. And the global description of videos can be developed by aggregating different local features. Motivated by the mechanism that human can capture distinctive information by few glimpse of the important moments, we show how the Local Glimpse and Global Glimpse can be combined to boost the performance.

2.1. Network Structure of MGN

In MGN, as shown in Fig. 1, one video is split into three snippets and spatiotemporal features are extracted from each snippet with 3D convolutional filters. After that, multi-glimpse attention is used to develop more robust features by preserving the class-specific semantic features and discarding the irrelevant information. Then MGN goes through two paths to make the initial predictions, which separately constitutes Local Glimpse and Global Glimpse. For Local Glimpse, each snippet will generate its own predictions which are combined later as the local prediction while for Global Glimpse global prediction is made after aggregating the local features from each snippet into global description. Finally the final prediction of MGN is the consensus of local and global prediction. The residual attention blocks, Local Glimpse and Global Glimpse are shown in Fig. 2.

2.2. Local Glimpse

Current methods usually only exploit the snippet-level features. However, a single snippet from one video may not contain enough class-specific information and label noise will impede the performance of classification. When we see the videos of HighJump, for example, we can't make the correct judgement if we only see the first few frames before the agent jumps in the final. Motivated by this, we introduce the Local Glimpse into MGN, which extracts the features of snippets from different periods of one video and makes local predictions based on all the local features. To choose different representative snippets, we exploit temporal sampling method to choose the snippets that separately sits in the begin, middle and the end part of the video. Besides, multi-glimpse attention is introduced to preserve the crucial information and suppress the noise like background movement, where each glimpse is for each snippet.

In our case, multi-glimpse attention block consists of three residual attention blocks, each of which is composed of attention block, residual connection and softmax normalization. For the features extracted from each snippet, the attention block scans the features through two consecutive 3D convolutional layers and produces the attention weights. The size of the first convolutional layer is $64 \times 1 \times 7 \times 7$ while that of the second is $1 \times 1 \times 7 \times 7$. After the attention block, the resulting attention weights are normalized by a softmax layer and then combined with the original features. Motivated by [22], we also introduce the residual connection to our attention block. Residual connection sums the original features and the features after attention block. In residual attention block, residual connection allows attention block to keep the good properties of the original features and weaken the selection ability of attention block by bypassing the main branch of the block. Notably, the weights of convolutional layers are shared among all attention blocks. Formally, the residual attention block can be formulated as

$$\mathbf{W_a} = \phi(\mathbf{W_2} \circ \tanh(\mathbf{W_1} \circ \mathbf{F_1})), \tag{1}$$

$$\mathbf{F_1^a} = \mathbf{W_a} \odot \mathbf{F_1} + \mathbf{F_1}. \tag{2}$$

where $\mathbf{W_1}, \mathbf{W_2}$ are the weights of convolutional layers and $\phi(\cdot)$ is the softmax operation. $\mathbf{F_1}, \mathbf{F_1^a}$ are respectively the original features of one snippet and the corresponding attention features after multiplying original features by the attention weights $\mathbf{W_a}$. Attention features of single snippet are not enough to make the correct prediction, so we combine the predictions made with attention features of each snippet. Three snippets come from the different periods of the video so they contain different but complementary information. Formally we name this whole process of feature extraction, multi-glimpse attention and prediction combination as Local Glimpse.

Table 1. Exploration of the modules we proposed on UCF101 (split1).

Structure	UCF	HMDB
Res3D [20]	85.8%	54.9%
Res3D*	87.8%	58.0%
Res3D*+Local	88.6%	58.3%
Res3D*+Global	88.4%	57.9%
MGN	**89.2%**	**59.6%**

2.3. Global Glimpse

As explained above in Sec. 2.2, the Local Glimpse can combine the information from different parts of one video and makes an initial prediction. However, the simple average combination of different predictions from each snippet will make a coarse prediction and can't represent the global description of the video. Therefore, we even introduce the module named Global Glimpse to help make better predictions. Global Glimpse aggregates the local features filtered by multi-glimpse attention to global features and makes a global prediction with the resulting features. Specifically, as shown in Fig. 2, after the multi-glimpse attention, Global Glimpse fuses the local features into global features with the same dimension by a consensus function and then makes a global prediction. Theoretically, the consensus function to aggregate local features can be any type and we use max pooling as the consensus function in our final result.

At the end of the Local Glimpse and Global Glimpse, we fuse the local prediction and global prediction into a final prediction of the video by average pooling. Therefore the final prediction composes of not only local information from different parts of the video but the global description, which are complementary to each other and generate more comprehensive description of the video. Formally, the final prediction can be summarized as

$$P = f(G + L), \tag{3}$$

where

$$G = \phi(g(\mathbf{F_1^a}, \mathbf{F_2^a}, \mathbf{F_3^a})), \tag{4}$$

$$L = l(\phi(\mathbf{F_1^a}), \phi(\mathbf{F_2^a}), \phi(\mathbf{F_3^a})). \tag{5}$$

where L, G and P are respectively the local prediction, global prediction and final prediction. And $l(\cdot), g(\cdot)$ and $f(\cdot)$ are the consensus function of each of those. The softmax function $\phi(\cdot)$ predicts the probability based on the local and global feature. We choose average pooling as $l(\cdot), f(\cdot)$ and max pooling function as $g(\cdot)$ for the final results below. When learning the network, we employ the loss as

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{C} y_j(P_j - \log\sum_{k=1}^{C}\exp(P_k)) + \lambda \parallel \mathbf{W} \parallel_2^2, \tag{6}$$

where y_j is the groundtruth label concerning class j, m is the number of training samples and $\mathbf{W}$ is all the learnable weights of MGN.

3. EXPERIMENT

In this section, we first introduce the evaluation datasets and implementation details of MGN. Then, we demonstrate the effectiveness of Local Glimpse and Global Glimpse. After this, we explore methods of the combination of local prediction and global prediction. Finally we compare the performance of MGN with the state-of-the-art.

Table 2. Exploration study of the consensus function of final prediction on UCF101 (split1).

Consensus Function	UCF	HMDB
Max	88.10%	57.8%
Average	**89.20%**	**59.6%**
Weighted Average	88.53%	59.0%

3.1. Datasets and Implementation Details

We conduct experiments on two commonly used action datasets UCF101 [23] and HMDB51 [24] and another two smaller action datasets ODAR and Penn. UCF101 contains 13K videos and is annotated into 101 action classes while HMDB51 include 6.8K videos of 51 action classes. We follow the 3-fold cross validation policy and report the average accuracy for UCF101 and HMDB51. ODAR is a challenging dataset of open domain action recognition, which consists of multiple publicly available datasets such as IXMAS [25] and KTH [26] with carefully selected action classes that are common across these datasets. And Penn contains 2326 videos of 15 different actions and human joint annotations for each sequence.

We employ the stochastic gradient descent algorithm to train the network, where batch size is set to 8 and the weight decay λ is 0.0005. The initial learning rate is initialized as 0.001. We decrease it every 10K iterations for UCF101 while every 5K iterations for other datasets. We stop after 25K iterations. The techniques of data augmentation we used when training are only horizontal flipping and random cropping because the temporal sampling method provides a natural data augmentation by choosing different parts of videos to train. Unless illustrated specially, MGN is pretrained on Sports1m [3], which is one of the largest video classification benchmarks. MGN is implemented using Caffe [27].

3.2. Evaluation of MGN

In this section, we focus on the study of the Local Glimpse and Global Glimpse and demonstrate their effectiveness on split1 of UCF101. As explained in Sec. 2.2, when learning these models we employ temporal sampling method. We choose the original Res3D [20] as our baseline model and compare it with the models using techniques we proposed. Specifically, we compare 5 settings: (1)Res3D, (2)Res3D with temporal sampling method (Re3D*), (3)Res3D* with Local Glimpse, (4)Res3D* with Global Glimpse, (5)MGN. The results are summarized in Table 1. We observe that the temporal sampling method boosts the performance of original Res3D, which implies that the number of training data and correlation among different parts of one video are crucial. But due to the existence of unnecessary information contained in every snippet, we resort to the multi-glimpse attention in Local Glimpse to filter information like background noise. And it achieves better performance than simple training using temporal sampling method. We also notice that simple combination of local predictions can't produce a good global description so we further utilize the Global Glimpse to make a global prediction. Considering the effectiveness of both components, we combine the results of Local Glimpse and Global Glimpse in MGN and it achieves the best performance. And we employ average pooling as $f(\cdot)$ to combine the global prediction and local prediction in this evaluation experiment.

Table 3. Comparison to state-of-the-art on UCF101 and HMDB51.

Method	UCF	HMDB
MDI+static-rgb [16]	76.9%	42.8%
C3D [5]	82.3%	50.3%
Res3D [20]	85.8%	54.9%
Two Stream [4]	88.0%	59.4%
LSTMs [21]	88.6%	-
TSN [18] (RGB)	85.7%	-
MGN	88.7%	57.9%
TSN [18] (3 modalities)	94.2%	69.4%

Table 4. Comparison to state-of-the-art on Penn and ODAR.

Penn		ODAR	
Actemes [28]	73.4%	two-stream [29]	34.8%
C3D [5]	86.0%	SGM [30]	42.9%
Res3D [20]	86.8%	DTN [29]	57.7%
P-CNN [31]	95.3%	Res3D [20]	58.1%
two-stream C3D [32]	95.3%	C3D [5]	60.9%
MGN	93.5%	MGN	63.0%
MGN*	**95.5%**	MGN*	**65.3%**

*Note that we only report the clip accuracy of C3D and Res3D. Models with * are also finetuned from UCF101.

3.3. Exploration Experiment

As aforementioned, Global Glimpse extracts the global features while Local Glimpse extracts the local information from different snippets of one video. Therefore the performance of our model greatly relies on the consensus function to combine them. In this section we evaluate three candidates:(1) max pooling, (2)weighted pooling, (3)average pooling for $f(\cdot)$. The results are illustrated in Table 2. We can see that the average pooling achieves the best performance while max pooling works worst. This demonstrates that the global information has the same importance as local information and the average combination of them can yield a more robust prediction. Therefore, we present the result of MGN with average pooling as the final result and make comparisons with other methods.

3.4. Comparison with State-of-the-Art

In this section, we compare our MGN with the state-of-the-art approaches on UCF101 and HMDB51 in Table 3, Penn and Odar in Table 4. Models evaluated on HMDB51, ODAR and Penn are finetuned from both UCF and Sports1m.

In Table 3 we compare MGN with several deep architectures on UCF101 and HMDB51. Notably, not all of them are directly comparable due to the difference of basic architectures, training techniques and input modals. Among them the most comparable models are C3D [5], Res3D [20] from which we build our model and other models using only RGB input. As shown in Table 3, we achieve better accuracy than C3D by 6.4% and Res3D by 2.9% on UCF101, which demonstrates the effectiveness of Local Glimpse and Global Glimpse in MGN. We also outperform C3D by 7.6% and Res3D by 3.0% on HMDB51. Notably, when testing Res3D averaged the predictions of 10 clips in one video to get the final prediction while we only exploit the local and global information using 3 snippets from one video. Therefore We achieve better accuracy with less information. When comparing to the two-stream approaches [4, 18], we achieve poorer accuracy for reasons that they use the pre-computed optical flow to extract the temporal features and fused the results of two different networks. [18] even exploits 3 input modalities to boost the accuracy. But optical flow is storage-demanding and slows down the speed of inference due to the heavy computation, which will prohibit the applications in real world. Instead we extract spatiotemporal features from RGB frames directly and implement faster inference when achieving promising accuracy. Therefore less computation is need by MGN. And we don't use the frame-level and IDT features either. As we can know, MGN achieves the best performance of methods using only RGB input and single network.

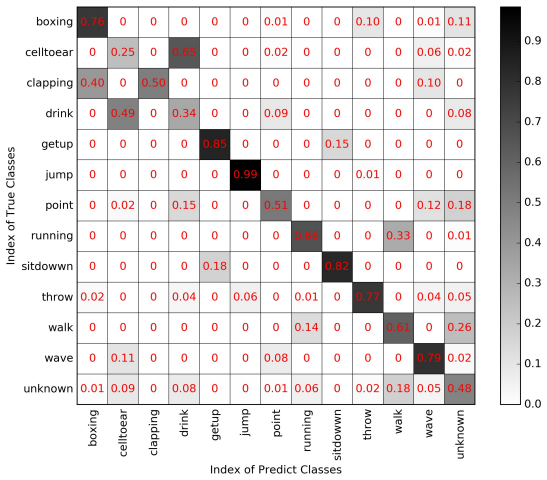

Fig. 3. Confusion matrix of MGN on ODAR dataset.

In Table 4 we compare MGN with state-of-the-art methods on Penn and ODAR. Specifically, we compared it with both traditional methods such as [2],[28],[33] and deep learning approaches such as [29], [31] and [32]. As can be seen, MGN finetuned from UCF101 achieves the best accuracy and outperforms the state-of-the-art on both datasets. It shows that our model has a better ability of cross-domain generalization than other methods such as Res3D. We show the confusion matrix of ODAR in Fig. 3. We can see that MGN recognizes *jump* and *getup* best while confuses *celltoear* and *drink* frequently. There are two reasons for this, where the first reason is the high inter-class similarities caused by the frequent presence of same person or similar scene. And the second reason is the complexity due to the characteristic of cross-domain in ODAR dataset.

4. CONCLUSION

In this paper we have presented a new framework called Multi-Glimpse Network(MGN) to recognize human actions in videos, which achieves competitive results on 4 benchmark datasets. MGN principally consists of two important components, Local Glimpse and Global Glimpse. Local Glimpse and Global Glimpse develop features from two different perspectives, which complete each other and produce more robust descriptions of the videos.

In the future work, we will explore a more effective technique to encode local features to global features and a better method of combining both of them. We anticipate that more distinctive features and a better consensus function will improve the overall performance.

5. REFERENCES

[1] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, pp. 60–79, 2013.

[2] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*. IEEE, 2013, pp. 3551–3558.

[3] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*. IEEE, June 2014.

[4] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*. IEEE, 2015, pp. 4489–4497.

[6] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[7] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*. IEEE, 2008, pp. 1–8.

[8] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*. British Machine Vision Association, 2008, pp. 275–1.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*. IEEE, 2015, pp. 1–9.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016, pp. 770–778.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*. IEEE, 2015, pp. 3431–3440.

[15] Tingzhao Yu, Huxiang Gu, Lingfeng Wang, Shiming Xiang, and Chunhong Pan, "Cascade temporal spatial features for video action recognition," in *ICIP*. IEEE, 2017.

[16] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould, "Dynamic image networks for action recognition," in *CVPR*. IEEE, 2016, pp. 3034–3042.

[17] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes, "Spatiotemporal residual networks for video action recognition," in *NIPS*, 2016, pp. 3468–3476.

[18] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*. Springer, 2016, pp. 20–36.

[19] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*. IEEE, 2016, pp. 1933–1941.

[20] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[21] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*. IEEE, 2015, pp. 4694–4702.

[22] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," *arXiv preprint arXiv:1704.06904*, 2017.

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*. IEEE, 2011, pp. 2556–2563.

[25] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[26] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *ICPR*. IEEE, 2004, vol. 3, pp. 32–36.

[27] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[28] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *ICCV*. IEEE, 2013, pp. 2248–2255.

[29] Yamin Han, Peng Zhang, Tao Zhuo, Wei Huang, and Yanning Zhang, "Video action recognition based on deeper convolution networks with pair-wise frame motion concatenation," .

[30] Tingzhao Yu, Huxiang Gu, Lingfeng Wang, Shiming Xiang, and Chunhong Pan, "Semantic guided network for open domain action recognition," `https://github.com/Tsingzao/SemanticGuidedBlock`.

[31] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid, "P-cnn: Pose-based cnn features for action recognition," in *ICCV*. IEEE, 2015, pp. 3218–3226.

[32] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu, "Body joint guided 3d deep convolutional descriptors for action recognition," *arXiv preprint arXiv:1704.07160*, 2017.

[33] Umar Iqbal, Martin Garbade, and Juergen Gall, "Pose for action-action for pose," in *FG*. IEEE, 2017, pp. 438–445.