GFNET: A LIGHTWEIGHT GROUP FRAME NETWORK FOR EFFICIENT HUMAN ACTION RECOGNITION

Hong Liu[1], Linlin Zhang[1], Lisi Guan[1], Mengyuan Liu[2]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School
[2]Tencent AI Lab
{hongliu,catherinezll,guanlisi}@pku.edu.cn, nkliuyifang@gmail.com

ABSTRACT

Human action recognition aims at assigning an action label to a well-segmented video. Recent work using two-stream or 3D convolutional neural networks achieves high recognition rates at the cost of huge computation complexity, memory footprint, and parameters. In this paper, we propose a lightweight neural network called Group Frame Network (GFNet) for human action recognition, which imposes intra-frame spatial information sparsity on spatial dimension in a simple yet effective way. Benefit from two core components, namely Group Temporal Module (GTM) and Group Spatial Module (GSM), GFNet decreases irrelevant motion inside frames and duplicate texture features among frames, which can extract the spatial-temporal information of frames at a minuscule cost. Experimental results on NTU RGB+D dataset and Varying-view RGB-D Action dataset show that our method without any pre-training strategy reaches a reasonable trade-off among computation complexity, parameters and performance, which is more cost-efficient than state-of-the-art methods.

Index Terms— Human Action Recognition, Lightweight Network, Convolutional Neural Network

1. INTRODUCTION

Human action recognition (HAR) is an active topic in computer vision due to its wide range of applications in human-robot interaction [1], intelligent video surveillance [2], and video content analysis [3]. Existing work on HAR can be mainly divided into two categories, i.e., image-based methods and video-based methods. Compared with image-based methods, video-based methods take temporal information into account and attract more attention. However, the number of required parameters for video-based neural networks is usually large [4, 5]. Hence, it is inevitable that these networks have high requirements for hardware resources and computing power.

Relation to prior work: Deep learning approaches [1, 6] gain increased attention for their significant improvements in HAR. Two-stream convolutional network [7] takes both RGB frames and optical flow as input to learn spatial-temporal features. Temporal Segment Network [6] combines a sparse temporal sampling strategy with video-level supervision to study temporal relationships. Both [6] and [7] rely on the optical flow that needs a lot of time to pre-compute. Another approach, 3D CNN [4], directly convolves on spatial and temporal dimensions. Nevertheless, 3D CNN has tremendous parameters, limited depth and is hard to train. Compared with 3D CNN, R(2+1)D [8] makes the decomposition of spatiotemporal convolutions to render the optimization easier. Besides, some methods [9, 10, 11] jointly utilize pose, raw depth, and RGB images to achieve higher accuracies.

The aforementioned methods achieve remarkable recognition accuracies but require plenty of parameters and hardware resources. To handle these issues, we propose a lightweight neural network called Group Frame Network (GFNet). Different from recent work [8, 12, 13] which directly processes the concatenated frames in the same pipeline, GFNet adds frame-level decomposition to extract features of each frame at a minuscule cost. In GFNet, frames are convolved separately in parallel. There are two core components: Group Temporal Module (GTM) and Group Spatial Module (GSM). These two modules enable GFNet to obtain temporal-spatial information only from RGB images without any bells and whistles (e.g., optical flow, multi-scale testing). To verify the validity of the model, no pre-training strategy is used in our experiments.

Generally, our contribution is two-fold:

(1) A lightweight Group Frame Network (GFNet) is proposed for human action recognition, which achieves state-of-the-art performance with extreme few parameters.

(2) To implement GFNet, we design two new core components called GTM and GSM to extract spatial-temporal information, where GTM aggregates temporal information across different frames and GSM reduces intra-frame and inter-frame spatial redundancies.

2. GROUP FRAME NETWORK

In this section, we describe the proposed approach for the HAR task as well as discussing two designed modules, GTM and GSM, in detail. These two modules boost the accuracy

This work is supported by National Natural Science Foundation of China (NSFC U1613209, No.61673030), National Key R&D Program from the Ministry of Science and Technology of China (Grant No.2018YFB1308602), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No.ZDSYS201703031405467).

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:04:53 UTC from IEEE Xplore. Restrictions apply.

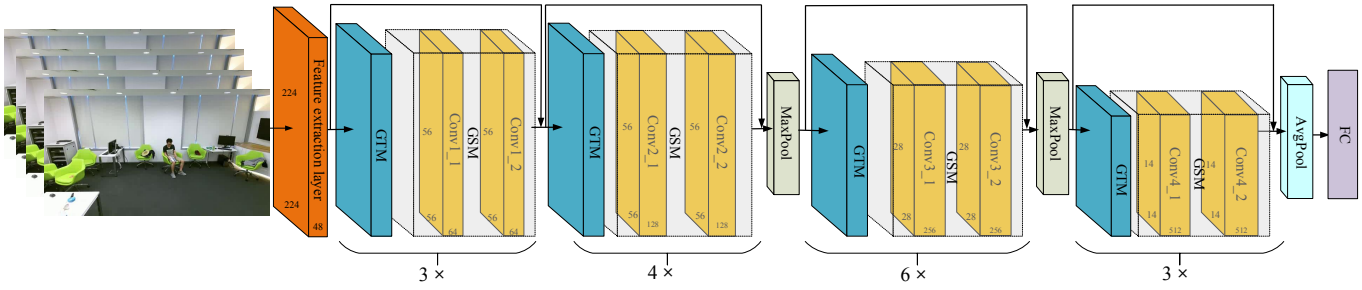


Fig. 1. The overall framework of GFNet, which consists of GTMs and GSMs. $i\times$ denotes the number of blocks is i.

of our baseline while reducing the number of parameters and computation cost.

2.1. Architecture Overview

The overall framework of the proposed GFNet is illustrated in Fig. 1. The entire video with a variable number of frames is provided as the input of the network. Through an average sampling strategy, the video is divided into N equal-length segments and only one frame is selected from each segment. Due to the repeatability of adjacent frames, this sampling strategy can reduce inter-frame redundancy while preserving long-temporal information.

GFNet contains two parts: a feature extraction layer and a series of stacked blocks. The first part is a feature extraction layer consisting of K separated branches. The sampled frames are simultaneously fed into the network to maintain the temporal information among these frames. In the feature extraction layer, each frame is learned independently using a network branch to get its spatial features. Specifically, all the sampled frames are stacked by channel-wise convolution. It means that the input channel of GFNet is $3N$ when using RGB images as input. The second part is a series of stacked blocks for integrating spatial-temporal information. Owing to the impressive performance and strong generalization ability of residual architecture, the block is based on the highly modularized residual unit. GFNet has the same number of stacked blocks as ResNet50 [14]. With the deepening of the network, channels of residual blocks are increased to obtain high-level semantic information. Each residual block is composed of two feed-forward modules, i.e., GTM and GSM. The first GTM follows the feature extraction layer and the others are between two GSMs. The details of the block are shown in Fig. 2. Through a fully-connected (FC) layer, GFNet produces the classification label for the given video.

Considering the extraneous motion and identical texture features in sampled frames, GFNet decomposes frames and reduces the number of channels for each frame to lessen spatial redundancy. To be specific, the number of channels is equally divided among branches. It means that only a small number of channels are used per frame.

2.2. Group Temporal Module

Since each frame is learned separately in the feature extraction layer, GFNet severely hinders the inter-frame information

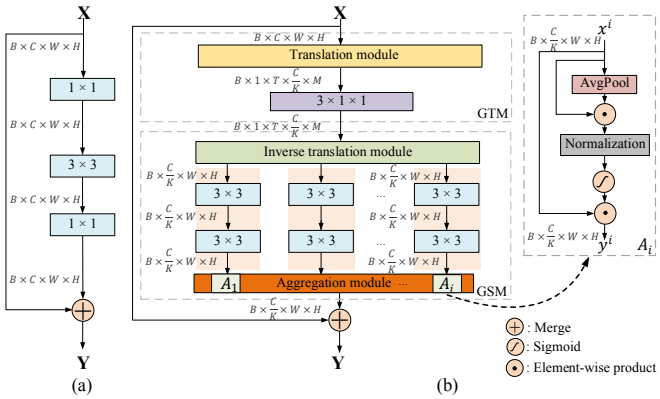


Fig. 2. The details of the proposed block. Here, B, C, T, W, H and K denote the batch size, channel, time dimension, width, height, and branch number, respectively. (a) is the typical bottleneck in ResNet50 [14]. (b) is the proposed block.

exchange, which may lead to dramatic performance degradation. To leverage the inter-frame information effectively and better strengthen temporal relationships, GTM is proposed to efficiently overcome the side effects brought by the separated branch. Focusing on the relationship between frames, GTM consists of a translation layer and a 3D convolution layer. Fig. 2 details GTM. The translation layer makes the replacement of the data dimension. It includes the channel merger and the channel separation, which achieves the conversion of the feature map from four-dimensional data to five-dimensional data. The 3D convolution layer, performing as a 1D convolution layer in GTM, is intended to extract features along the time dimension T. Because frames are processed separately in branches, the number of T is the same as the number of branch K. Thus, the process of GTM contains two steps:

$$X^i_{mid} = \Gamma(X^i_{in}), \tag{1}$$

$$X^i_{out} = f_{conv_{3\times1\times1}}(X^i_{mid}), \tag{2}$$

where $\Gamma(.)$ denotes the translation module, the input feature map of the i-th block is $X^i_{in} \in \mathbb{R}^{B\times C_{in}\times W\times H}$ and the mid feature map is $X^i_{mid} \in \mathbb{R}^{B\times 1\times T\times \frac{C_{out}}{K}\times M}$, $M = W \times H$. The output feature map is $X^i_{out} \in \mathbb{R}^{B\times C_{out}\times W\times H}$. Here, B denotes the batch size and C denotes the channels. W and H denote the width and the height of the feature map, respectively. $f_{conv_{3\times1\times1}}(.)$ is the convolution layer with kernel size $3\times1\times1$.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:04:53 UTC from IEEE Xplore. Restrictions apply.

2.3. Group Spatial Module

For the convolution layer of ResNet50 [14], the computational cost is closely related to the number of channels. Motivated by this, a novel module called GSM is designed to significantly decrease the number of parameters and computational efforts. The details of GSM are illustrated in Fig. 2. Different from recent work [12, 13, 15] that only using channel-wise convolution or group convolution to reduce the parameters, GSM explores the relationship between frames and groups to learn the discriminative features of frames. In GSM, each frame is convolved via the corresponding group. Because of the similarity among frames, the texture information is repetitive. Meanwhile, irrelevant motion inside frames increases the intra-frame redundancy. Aiming at minimizing the interference of redundant information, GSM diminishes the number of channels to extract features per frame.

In GSM, there is an inverse translation module, K branches, and an aggregation module. Each branch contains two stacked $conv_{3\times3}$. To extract features of each frame independently, the branch number K is equal to the sampled frame number N. Motivated by [16], an aggregation module that calculates the similarity between global and local features is added behind branches to enhance the spatial features. Therefore, the process of GSM can be formulated as:

$$(Y^{i,1}_{mid}, ...Y^{i,j}_{mid}, ...Y^{i,K}_{mid}) = \hat{\Gamma}(Y^{i}_{in}), \tag{3}$$

$$Y^{i,j}_{out} = f_{conv_{3\times3}}(f_{conv_{3\times3}}(Y^{i,j}_{mid})), \tag{4}$$

$$Y_{out} = Att(Y^{i,1}_{out}, ..Y^{i,j}_{out}, ...Y^{i,K}_{out}), \tag{5}$$

where the input feature map Y^{i}_{in} is equal to X^{i}_{out} and $\hat{\Gamma}(.)$ is the inverse change formulation of $\Gamma(.)$ in GTM. In the j-th branch, the mid feature map is $Y^{i,j}_{mid} \in \mathbb{R}^{B\times\frac{C_{in}}{K}\times W\times H}$ and the output feature map is $Y^{i,j}_{out} \in \mathbb{R}^{B\times\frac{C_{out}}{K}\times W\times H}$.

The aggregation module $Att(.)$ that enables features in each branch to autonomously enhance its learned semantic representation, is formulated as:

$$Y^{i,j}_{out} = \psi(\sigma(\varphi_n(\zeta_p(Y^{i,j}_{out}) \odot Y^{i,j}_{out})) \odot Y^{i,j}_{out}), \tag{6}$$

where $\zeta_p(.)$, φ_n, $\sigma(.)$, and $\odot$ denote the average pooling layer, normalization module, sigmoid layer and element-wise product, respectively. ψ is the fusion function of branches.

2.4. Short-Skip Residual Learning

Inspired by the principle of ResNet [14], a shortcut connection is adopted to strengthen gradient back-propagation in the proposed blocks. The residual connection which combines GTM and GSM can be formulated as:

$$X^{i+1}_{in} = X^{i}_{in} + F_{GSM}(F_{GTM}(X^{i}_{in})), \tag{7}$$

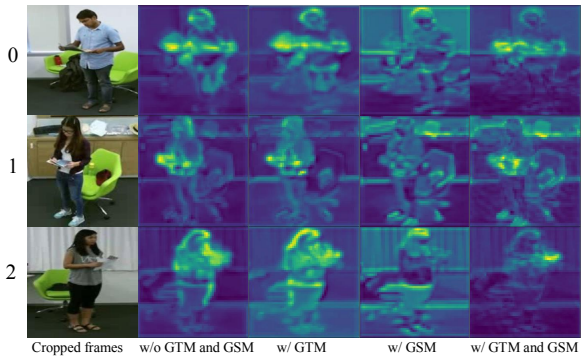
where $F_{GTM}(.)$ and $F_{GSM}(.)$ denote the process of GTM and GSM, respectively.

3. EXPERIMENTS AND DISCUSSIONS

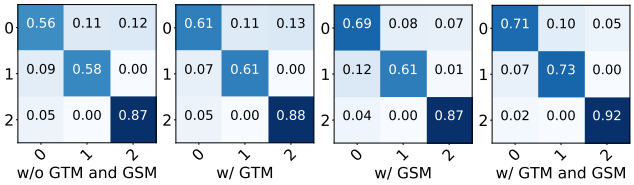
In this section, experiments are conducted on two challenging datasets. We first describe the details of datasets and their evaluation setups and then analyze experimental results.

Table 1. Ablation studies of GTM and GSM on NTU dataset using CV protocol evaluation.

Method	GTM	GSM	FLOPs	♯Params	Acc.
GFNet(Ours)	×	×	697.41G	21.46M	82.31%
	✓	×	698.20G	21.46M	86.25%
	×	✓	**47.88G**	**1.55M**	84.78%
	✓	✓	48.68G	**1.55M**	**89.92%**



(a) The activation maps of three actions.



(b) Confusion matrices of three actions.

Fig. 3. Some representative actions. Here, 0, 1, and 2 are short for acting as "reading", "writing", and "tear up paper", respectively. Best viewed in color.

3.1. Datasets and Protocols

NTU RGB+D (NTU) [17] is a large scale dataset for HAR, which contains 60 daily actions performed by 40 subjects from various views, generating more than 56K videos. There is a considerable variation in viewpoint, intra-class subjects, and sequence lengths, making this dataset challenging. We follow two recommended evaluation protocols [17], i.e., Cross-Subject (CS) and Cross-View (CV) proposed in the dataset.

Varying-view RGB-D Action Dataset (VAD) [18] contains 25K sequences, covering 40 actions performed by 118 subjects. The modalities of data include RGB videos, depth maps, and 3D joints, where only the RGB videos are used for our experiments. To evaluate the recognition performance between neighbor viewpoints, Cross-view recognition II protocol evaluation (CV II) [18] is used in our experiments.

3.2. Implementation Details

The number of sampled frames in each video sequence for all the datasets is set to 16. All frames are resized to 224×224. We train our network using SGD with a Nesterov momentum of 0.9. The initial learning rates are 0.02 for NTU and 0.01 for VAD, respectively. The batch sizes for NTU and VAD are set to 128 and 64. All experiments are conducted on Pytorch with only one NVIDIA GTX 1080Ti GPU.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:04:53 UTC from IEEE Xplore. Restrictions apply.

Table 2. Comparison with state-of-the-arts on NTU dataset. S is short for skeleton; D is short for depth; + is short for the combination. Params: Parameters. * denotes adding an RGB difference stream ($RGB^{diff} = RGB^{t+1} - RGB^{t}$).

Method	Year	Modality	Params	CS	CV
C3D [4]	2015	RGB	34M	63.5%	70.3%
MTLN [19]	2017	S	64.45M	79.6%	84.8%
Pose-attention [9]	2017	RGB+S	31.82M	82.5%	88.6%
Rahmani [10]	2017	D+S	-	75.5%	83.1%
DSSCA-SSLM [11]	2018	RGB+S	-	74.9%	-
TSN [6]	2018	RGB	10.37M	-	66.5%
DA-Net [20]	2018	RGB	15.7M	-	75.3%
Res+LSTM [5]	2018	RGB	28M	71.3%	80.2%
TCN + TTN [21]	2019	S	-	77.6%	84.3%
CCP [22]	2019	S	14M	80.1%	86.8%
EleAtt-GRU [23]	2019	RGB	**0.28M**	63.3%	70.6%
DTIs [1]	2019	RGB	60.4M	**85.4%**	-
GFNet(Ours)	2019	RGB	**1.55M**	**82.0%**	**89.9%**
GFNet*(Ours)	2019	RGB	**3.10M**	**84.8%**	**91.8%**

Table 3. Comparison with state-of-the-arts on VAD dataset. Here, Front denotes the average accuracy of testing on viewpoints #front, 2, 4 and 6. Conversely, End denotes the average accuracy of testing on viewpoints #1, 3, 5 and 7.

Method	Year	Modality	Params	Front	End
C3D [4]	2015	D	34M	21%	22%
C3D [4]	2015	RGB	34M	29%	47%
TCN [24]	2017	S	-	33%	53%
SK-CNN [25]	2017	S	139.8M	65%	71%
LRCN(Resnet34) [26]	2017	RGB	155.5M	14%	20%
LRCN(Resnet50) [26]	2017	RGB	157.7M	19%	11%
ST-GCN [27]	2018	S	3.1M	48%	64%
ResNeXt [15]	2018	RGB	44.3M	46%	49%
VS-CNN [18]	2019	S	-	68%	73%
GFNet(Ours)	2019	RGB	**1.54M**[1]	**81%**	**78%**
GFNet*(Ours)	2019	RGB	**3.08M**[1]	**85%**	**84%**

[1] With the usage of different FC layers, parameters vary across different datasets.

3.3. Experimental Results

Evaluation of GTM and GSM. To evaluate the effects of GTM and GSM, the results of the individual module and the combination of both modules are shown in Table 1. In the table, we show the performance of GFNet with three settings. It can be seen that each module contributes to GFNet. GTM provides the opportunity to explore the spatial and temporal correlation and brings 3.94% accuracy improvements compared to the model w/o these two modules. GSM learns spatial information and brings 2.47% accuracy improvements while significantly lessening the network parameters and computation complexity. When combining GTM and GSM, GFNet can capture more abundant spatiotemporal features and attain the best accuracy. The gain over the baseline (stacked frames are processed in one branch) is 7.61%. Comparing the first row with the last row in the table, it can be observed that the parameters saving can be up to 13× and the FLOPs reduction is around 14×. Fig. 3 presents activation maps and confusion matrices of three representative actions. In Fig. 3(a), the activation maps, generated from the first three stacked blocks in GFNet, are based on the cropped person movement areas of 16 frames. Without GTM and GSM, the model studies the spatial-temporal simultaneously, but its areas of concern are too broad. The GFNet with GTM only focuses on the motion area, such as hair, hands, and legs. Furthermore, without GTM, GFNet pays attention to texture information, such as the environment and person's still head. Combining GTM and GSM, GFNet concentrates more on the parts of hands and the papers in hand to obtain discriminative features. Confusion matrices of representative actions are shown in Fig. 3(b). It is illuminated that the error rates of the three similar actions have a reduction relatively by combining these two modules.

Comparison with state-of-the-arts. As we use sole RGB data, our method can be fairly compared with methods using RGB data. The performances of methods using depth data (D), skeleton data (S) or multiple modalities data are listed to show the superior performance of GFNet.

Table 2 shows the detailed results on NTU dataset compared with state-of-the-art methods. For a fair comparison, we only report the results from methods without using optical flow. As the benchmark codes are not available, we cannot get the precise number of parameters of some models. The parameters are estimated to be similar to the baseline used by these models. GFNet gains slightly lower accuracy than DTIs [1], but gives a 38× reduction in the number of parameters. Besides, the combination of RGB images and RGB differences (GFNet*) boosts recognition performance. Notably, the comparison with other reported methods is not entirely fair to GFNet since other methods use different modalities or pre-training strategies. The results indicate that our approach achieves the best trade-off between the number of parameters and performance.

Table 3 shows the recognition accuracies of state-of-the-art methods and our method on VAD dataset. GFNet gains a notable improvement compared with all state-of-the-art methods. However, since the recognition is between neighbor viewpoints, it is very challenging to distinguish some categories, such as "forward lunging" and "left lunging". By integrating an RGB difference stream to complement the motion information, the overall performance of GFNet* on recognizing human actions is significantly enhanced compared to GFNet.

4. CONCLUSIONS

In this paper, we present a lightweight Group Fame Network (GFNet) for HAR. Specifically, our method consists of two core components: Group Temporal Module (GTM) and Group Spatial Module (GSM). GTM is designed to provide temporal information and boost inter-frame correlation. GSM obtains diverse intra-frame information while exploiting image-space redundancy to reduce the number of parameters. With these two modules, GFNet can lessen the irrelevant information of frames. Experimental results demonstrate that the proposed GFNet not only achieves precise recognition results but also relieves the burden of resource utilization, which makes it of great potentiality for deployment on resource-constrained devices.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:04:53 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] M. Liu, F. Meng, C. Chen, and S. Wu, "Joint dynamic pose image and space time reversal for human action recognition from videos," in *AAAI*, 2019, pp. 8762–8769.

[2] H. Liu, J. Tu, M. Liu, and R. Ding, "Learning explicit shape and motion evolution maps for skeleton-based human action recognition," in *ICASSP*, 2018, pp. 1333–1337.

[3] M. Liu, H. Liu, and C. Chen, "3D action recognition using multiscale energy-based global ternary image," *TCSVT*, vol. 28, no. 8, pp. 1824–1838, 2018.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[5] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *CVPR*, 2018, pp. 469–478.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *TPAMI*, vol. 41, no. 11, pp. 2740–2755, 2018.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014, pp. 568–576.

[8] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[9] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *ICCV*, 2017, pp. 604–613.

[10] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *ICCV*, 2017, pp. 5833–5842.

[11] A. Shahroudy, T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in RGB+D videos," *TPAMI*, vol. 40, no. 5, pp. 1045–1058, 2018.

[12] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *ICCV*, 2019.

[13] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *ICCV*, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[15] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet," in *CVPR*, 2018, pp. 6546–6555.

[16] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Enhancing semantic feature learning in convolutional networks," *arXiv preprint arXiv:1905.09646*, 2019.

[17] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[18] Y. Ji, F. Xu, Y. Yang, F. Shen, H. Shen, and W. Zheng, "A large-scale varying-view RGB-D action dataset for arbitrary-view human action recognition," *arXiv preprint arXiv: 1904.10681*, 2019.

[19] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *CVPR*, 2017, pp. 3288–3297.

[20] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *ECCV*, 2018, pp. 451–467.

[21] S. Lohit, Q. Wang, and P. Turaga, "Temporal transformer networks: Joint learning of invariant and discriminative time warping," in *CVPR*, 2019, pp. 12426–12435.

[22] A. Porrello, D. Abati, S. Calderara, and R. Cucchiara, "Classifying signals on irregular domains via convolutional cluster pooling," in *AISTATS*, 2019, pp. 1388–1397.

[23] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks," *TIP*, vol. 99, pp. 1–1, 2019.

[24] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *CVPR*, 2017, pp. 1003–1012.

[25] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *PR*, vol. 68, no. 68, pp. 346–362, 2017.

[26] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *TPAMI*, vol. 39, no. 4, pp. 677–691, 2017.

[27] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018, pp. 7444–7452.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on May 30,2021 at 16:04:53 UTC from IEEE Xplore. Restrictions apply.