The 15th International Conference on
Computer Science & Education (ICCSE 2020)
August 18-20, 2020. Online

Multi-Scale Adaptive Graph Convolutional Network for Skeleton-Based Action Recognition

1st Yiqi Fan
Beijing University of Posts and Telecommunications
Beijing, China
fanyiqi@bupt.edu.cn

2nd Xiaojuan Wang*
Beijing University of Posts and Telecommunications
Beijing, China
wj2718@163.com

3rd Tianqi Lv
Beijing University of Posts and Telecommunications
Beijing, China
lvtianqi@bupt.edu.cn

4th Lingrui Wu
Beijing University of Posts and Telecommunications
Beijing, China
wlr@bupt.edu.cn

***Abstract*—Skeleton-based action recognition is a branch of action recognition which uses dynamic skeletons as input. Recent research based on graph convolutional networks (GCN) has achieved remarkable performance in this area. However, feature extraction and fusion at different physical scales have not been well studied. To solve these issues, we propose a novel Multi-Scale Adaptive Graph Convolutional Network (MSGCN) which contains a Multi-Scale Graph Convolutional Module and a Multi-Scale Selective Fusion Module. Extensive experiments on NTU-RGBD dataset demonstrate the effectiveness of our method, our method achieved competitive performance on NTU-RGBD dataset.**

***Index Terms*—multi-scale, graph convolutional networks, attention mechanism, skeleton-based action recognition**

I. Introduction

Action recognition is a fundamental yet challenging task in computer vision. It has a wide range of potential applications for sports, security, and entertainment. There is a branch of action recognition which is called skeleton-based action recognition uses dynamic human skeletons as input, since the dynamic skeletons convey significant action information and have the robustness to occlusion, complex scenarios and changes of viewpoints.

There are three types of methods in the task of skeleton-based action recognition in general, RNN-based methods [1]–[8], CNN-based methods [9]–[14] and GCN-based methods [15], [16]. With the development of graph convolutional networks [17]–[20], GCN-based methods have made significant progress. These methods represent the skeleton connection relationship as a graph and apply graph convolutional networks to extract spatial features.

However, the GCN-based methods still have certain shortcomings. First, current GCN-based methods often extract spatial information on only one scale, without considering feature extraction on multiple scales. Second, although some GCN-based methods [16] introduce a spatial attention mechanism which allows the network to aggregate information adaptively in the spatial dimension, attention mechanism in the channel dimension are rarely used.

To solve the above problems, we proposed a novel Multi-Scale Adaptive Graph Convolutional Network for skeleton-based action recognition. We design a Multi-Scale Graph Convolutional Module to extract spatial features of different scales by downsampling and upsampling the spatial-temporal graph. For effectively fusing features of different scales, we design a Multi-Scale Selective Fusion Module under the attention mechanism. Then, we evaluate our proposed method on the NTU-RGBD dataset, and ablation studies demonstrate the effectiveness of the Multi-Scale Graph Convolutional Module and Multi-Scale Selective Fusion Module we proposed. Our method exceeds previous methods and achieves competitive performance.

In summary, there are three contributions in our paper:

- We propose a Multi-Scale Graph Convolutional Module which extracts spatial features of different scales.
- We propose a Multi-Scale Selective Fusion Module which selectively fuses features of different scales through scale-wise attention mechanisms.
- Our proposed method achieves competitive performance on the NTU-RGBD dataset.

II. Related work

A. Skeleton-based Action Recognition

Action recognition is divided into several categories based on the type of input data. Since dynamic skeletons convey significant action information and are robust to occlusion, complex scenarios and changes of viewpoints, it became an alternative data type replacing RGB and depth data in the action recognition task. Generally, there are two types of methods for skeleton-based action recognition: handcrafted-based methods and deep learning methods. Handcrafted-based methods design features like covariance matrices [21], relative positions of joints [22] and points in a lie group [23] to model human actions. With the development of deep learning, deep learning methods exceed handcrafted-based methods. RNN-based [1]–[8] methods usually represent skeleton data as a sequence of joints while CNN-based [9]–[14] methods represent skeleton data as a spatial-temporal image. However, these representations lose important structural information when transforming skeleton data into grid data. In [15], Yan first proposes a spatial-temporal graph convolutional networks(ST-GCN) which shows that graph convolutional networks(GCN) can construct spatial information better since it directly represents the human skeleton data as a graph and uses adjacency

∗ is the corresponding author.

978-1-7281-7267-5/20/$31.00 ©2020 IEEE

matrix to maintain the structural information. 2s-AGCN [16] uses two-stream GCN architecture to fuse joint score and bone score, and it generates adaptive adjacency matrices according to the input data.

B. Attention Mechanism

The attention mechanism is inspired by the human visual attention mechanism, and has achieved great success in many vision tasks, such as classification, object detection, and image segmentation. SENet [24] proposes a Squeeze-and-Excitation block that adaptively recalibrates channel-wise feature responses by element-wise multiplication. CBAM [25] uses max pooling and average pooling with channel attention module and proposes a spatial attention module. The spatial attention module is similar to the channel attention module except the pooling operations are conducted at the channel dimension. Another approach of self-attention uses matrix multiplication instead of element-wise multiplication. In Non-local Networks [26], Wang designs a non-local block that computes the response at a position as a weighted sum of the features at all positions capturing long-range dependencies. DANet [27] proposes a Dual Attention Network with a position attention module and channel attention module, both of them perform matrix multiplication to adaptively aggregate information. Formally, the attention mechanism based on matrix multiplication can be regarded as a kind of GCN. From this perspective, each element in the feature map is a node in the graph, and the attention matrix can be regarded as an adaptive adjacency matrix.

III. Method

In this section, we introduce the specific details of our proposed Multi-Scale Adaptive Graph Convolutional Network.

A. Revisiting Spatial-Temporal Graph Convolutional Networks

Skeleton sequence data can be represented as a $C \times T \times N$ tensor, where N is the number of human joints, T is the temporal length and C is the number of channels. ST-GCN first represents the skeleton sequence data as a spatial-temporal graph, which uses the physical connections between human joints, and applied the graph convolutional networks(GCN) to skeleton-based action recognition task. ST-GCN uses CNN to extract temporal information, and GCN to extract spatial information, the process of GCN is formulated as (1):

$$f_{out} = \sum_{k}^{K_v} W_k (f_{in} A_k) \odot M_k \tag{1}$$

where M_k is a learnable weight matrix, $\odot$ denotes element-wise product between two matrices, and K_v is set to 3 which divides A into three different partitions. A is the normalized adjacency matrix, $A = \Lambda^{-\frac{1}{2}}(\overline{A} + I)\Lambda^{-\frac{1}{2}}$, where $\overline{A}$ is the adjacency matrix of the human joints' physical connection, I represents self-connection, $\Lambda^{ii} = \sum_j (\overline{A}^{ij} + I^{ij})$ is the normalized diagonal matrix. A is divided into three partitions A_k following the partitions strategies which ST-GCN proposed, $A_1 = I$, and since the human body's physical connection graph is an undirected graph, $\overline{A}$ is divided into two directed graph A_2 and A_3, one points to the center joint of the human body and the other points to the terminal joint. $A_k = \Lambda^{-\frac{1}{2}}\overline{A_k}\Lambda^{-\frac{1}{2}}$, where $\Lambda_k^{ii} = \sum_j (\overline{A_k}^{ij})$.

2s-AGCN introduces a data-driven adaptive graph, the graph convolutional layer of 2s-AGCN is formulated as (2):

$$f_{out} = \sum_{k}^{K_v} W_k f_{in} (A_k + B_k + C_k) \tag{2}$$

where A_k is the same as (1), B_k is a learnable adjacency matrix which replaces the element-wise product in (1), and C_k is an adaptive adjacency matrix which depends on the input data, the calculation of C_k is formulated as (3).

$$C_k = softmax(f_{in}^T W\theta k^T W\phi k f_{in}) : \tag{3}$$

where $W\theta k$ and $W\phi k$ are the weights of 1×1 convolutional layers. In this way, the network can learn an adjacency matrix according to different input data.

B. Multi-Scale Graph Convolutional Module

In the task of skeleton-based action recognition, GCN-based methods represent skeleton data as a spatial-temporal graph. Previous works often extract spatial information on only one scale, without considering feature extraction on multiple scales. To solve this problem, we propose a Multi-Scale Graph Convolutional Module(MSGCM) which downsamples the spatial-temporal graph and aggregates information at three different scales which we called small-scale, medium-scale and large scale.

The MSGCM module conducts 2 downsampling operations to extract additional graphs of different scales from the original graph. The first downsampling operation reduces the spatial dimension of the graph to 10, and the second reduces it to 5. In particular, the module uses average pooling to downsample nodes belonging to the same body part of each scale. Which nodes belong to the same part of each scale are decided by predefined rules which are shown in Fig. 1.

After getting graphs of three different scales, graph convolution layers are performed at three graphs individually. The architecture of each graph convolutional layer is the same as 2s-AGCN. Through the inverse transform of downsampling, the output feature maps are upsampled to the same size as the original feature maps and we add the three outputs together to get the output of the module. The structure of the Multi-Scale Graph Convolutional Module is shown in Fig. 2.

In this way, nodes belonging to the same part on one scale will share a part of the outputs. This multi-scale structure enables sub-modules to focus on their physical scale and aggregate multi-scale information. The MSGCM module's process is formulated as (4):

$$f_{out} = \sum_{g}^{G} F_{up}(\sum_{k}^{K_v} W_k F_{dn}(f_{in})(A_{kg} + B_{kg} + C_{kg})) \tag{4}$$

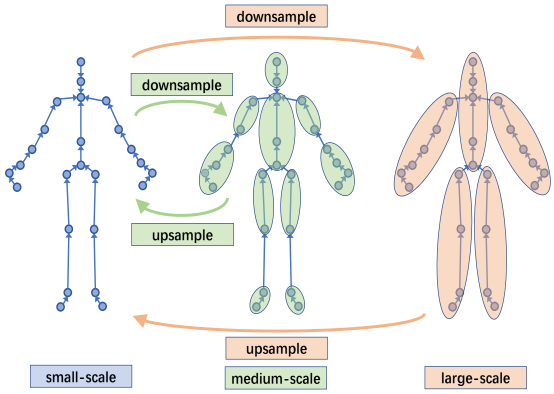

Fig. 1. Downsampling and upsampling method we proposed. The small-scale graph is the original input graph. The medium-scale graph is obtained by the first downsampling operation and the large-scale graph obtained by the second downsampling operation. The colored areas in medium-scale and large-scale are the different parts in these scales, and the nodes in one colored area belong to the same part. The connection between two parts of one scale inherits the connection between the nodes on the edge of them.

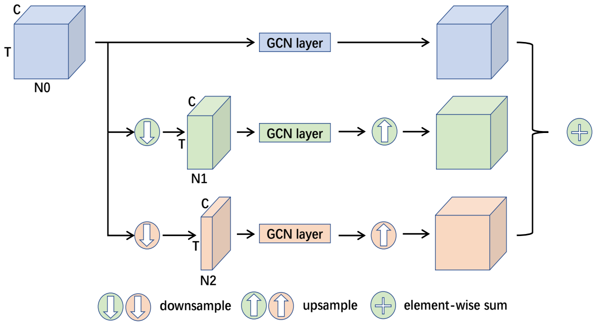

Fig. 2. The structure of the Multi-Scale Graph Convolutional Module(MSGCM), the MSGCM consists of three steps: (1)Downsample: using predefined rules to downsample the spatial-temporal graph. (2)Graph convolution: including three individual graph convolutional layers. (3)Upsample: conducting the inverse transformation of downsampling, and aggregating the outputs together.

where G is set to 3, F_{up} and F_{dn} are the upsampling and downsampling operations which is illustrated in Fig. 1.

C. Multi-Scale Selective Fusion Module

In the Multi-Scale Graph Convolutional Module(MSGCM) we proposed, the module fuses information from different scales by upsampling and adding. Considering the connection between graph convolution and attention mechanism, the MSGCM module can be regarded as applying the attention mechanism in the spatial dimension. In order to apply the channel-wise attention mechanism and selectively fuse multi-scale information we designed a Multi-Scale Selective Fusion Module(MSSFM).

We denote the multi-scale outputs of MSGCM as X_1 X_3, and firstly fuse them by element-wise adding. The MSSFM module only focuses on information aggregation in the channel dimension, the module conducts global average pooling to aggregate the spatial-temporal information. After that, we used a "One Squeeze and Multi-Scale Excitation" architecture to learn the channel-wise attention matrices of A_1 A_3. One Squeeze: we use one fully connective layer followed by a ReLU operation to squeeze the channels to half of the original. Multi-Scale Excitation: three individual fully connective layers are conducted to excite multi-scale channel-wise attention matrices with the original channel size. After that, we stack the multi-scale attention matrices and conduct a softmax operation across the scale dimension to generate scale-wise attention weights. The process is formulated as follows:

$$A_n^c = \frac{exp(M_n^c)}{\sum_{k=1}^{3} exp(M_k^c)} \tag{5}$$

where n is the scale dimension, c is the channel dimension, A_n^c is the learned scale attention matrix. The addition of A_n^c along the scale dimension equals 1. Finally, we aggregate the multi-scale by element-wise multiplication which is defined in (6). And the structure of the MSSFM module is shown in Fig. 3.

$$F = \sum_{n=1}^{3} (A_n \cdot X_n) \tag{6}$$

where F is the final output feature map of the MSSFM module, and $\cdot$ is the element-wise multiplication.

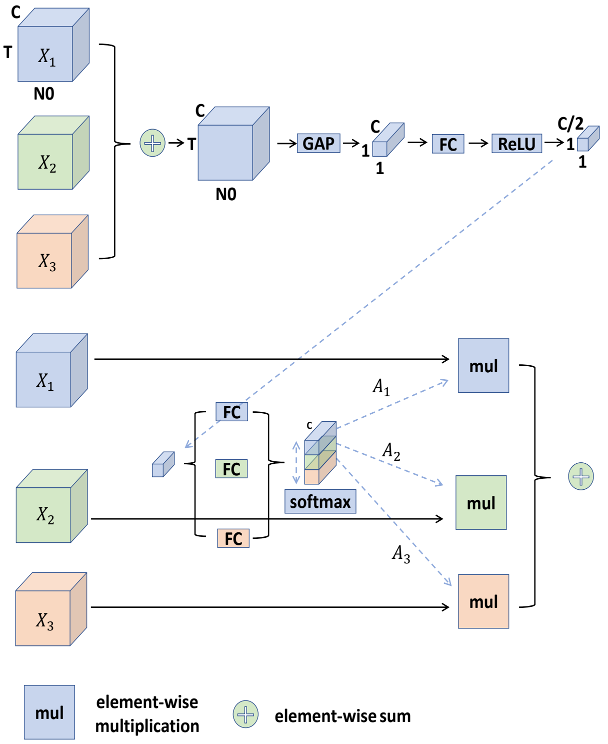

Fig. 3. The structure of Multi-Scale Selective Fusion Module(MSSFM). X_1 X_3 are multi-scale output features from MSGCM module. A_1 A_3 are the channel-wise attention matrices

D. *Network Architecture*

With the Multi-Scale Graph Convolutional Module(MSGCM) and Multi-Scale Selective Fusion Module(MSSFM), we propose a novel Multi-Scale Adaptive Graph Convolutional Network for skeleton-based action recognition. We process spatial dimension with the MSGCM and MSSFM followed by the batch normalization and ReLU in sequence. Then we process temporal dimension with $K_t \times 1$ CNN. These two blocks with a residual connection form a basic module of our network. The basic module is shown in Fig. 4.

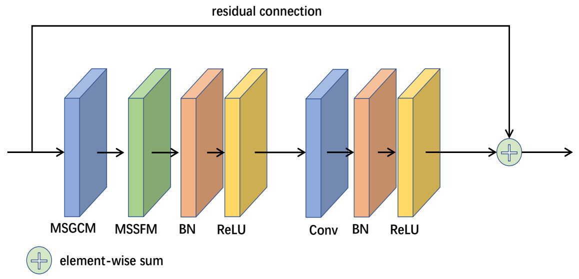

Fig. 4. The basic block of the proposed network. The MSGCM module is followed by an MSSFM module, and the element-wise sum fusion in the MSGBM module is replaced by adaptive fusion in the MSSFM module. The Conv layer is a $K_t \times 1$ convolutional layer. The MSSFM module and the convolutional layer is followed by a BN layer and a ReLU layer. There is a residual connection in each basic module.

We stack 10 basic modules to capture spatial-temporal information, then a global average pooling layer and a fully connected layer are performed to get the output. The channels of each modules are 64, 64, 64, 64, 128, 128, 128, 256, 256, 256. The output is conducted a softmax operation and output the classification. The overall of Multi-Scale Adaptive Graph Convolutional Network is shown in Fig. 5.

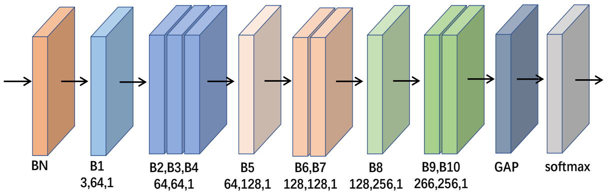

Fig. 5. The overall structure of the Multi-Scale Adaptive Graph Convolutional Network. The whole network is consists of 10 basic blocks. The three numbers of each block represent the input channels, the output channels and the stride. GAP is the global average pooling operation.

IV. Experiments

A. *Dataset and Evaluation Metrics*

NTU-RGBD: NTU-RGBD [2] is a benchmark dataset in skeleton-based action recognition. It contains 60 action classes and 56880 video samples with 3D skeletal data captured by three Kinect V2 cameras. These clips are performed by 40 volunteer actors. The dataset is divided into two benchmarks in different ways: 1).Cross-view: divided by the id of the camera, the training set consists of 37920 videos captured by camera 2 and 3, the validation set is consists of 18960 videos captured by camera 1. 2)Cross-subject: divided by the id of actors, the training set consists of 40320 videos performed by 20 actors, the validation set consists of 16560 performed by the other 20 actors.

Evaluation Metrics: The evaluation is conducted top-1 and top-5 accuracy.

Data Augmentation: To reduce the influence of the input data distribution, we use the central node's coordinates of the first frame in the temporal dimension as the coordinate origin to normalize the data. Then, in order to reduce the influence of the angle of view, we rotate the joints' coordinates. In particular, we first rotate the coordinates to align the left and right shoulder line with the horizontal direction, then we align the spine with the vertical direction. We use joint and bone two-stream framework following 2s-AGCN [16].

B. *Ablation Study*

We conduct ablation experiments with the cross-view benchmark of the NTU-RGBD dataset. We use 2s-AGCN architecture as our baseline network. Based on it, we first explore the effectiveness of the Multi-Scale Graph Convolutional Module(MSGCM) and Multi-Scale Selective Fusion Module(MSSFM), finally analyze the effect of each proposed module.

Effect of Multi-Scale Graph Convolutional Module. We evaluate the network with MSGCM compared with baseline in the joint stream, bone stream and the final results. The results are shown in TABLE I.

TABLE I
Ablation Experiments about Multi-Scale Graph Convolutional Module on NTU-RGBD dataset(Top-1 accuracy).

Methods	Accuracy
Baseline-2s-AGCN-J	93.7
Baseline-2s-AGCN-B	93.2
Baseline-2s-AGCN	95.1
MSGCM-J	94.1
MSGCM-B	94.1
MSGCM	95.5

Effect of Multi-Scale Selective Fusion Module. We evaluate the network with MSGCM and MSSFM compared with the network only with MSGCM in the joint stream, bone stream and the final results. The results are shown in TABLE II.

Comparison Analysis: In this experiment, we explore the effectiveness of each module to the whole network. Compared

TABLE II
ABLATION EXPERIMENTS ABOUT MULTI-SCALE SELECTIVE FUSION MODULE ON NTU-RGBD DATASET(TOP-1 ACCURACY).

Methods	Accuracy
MSGCM-J	94.1
MSGCM-B	94.1
MSGCM	95.5
MSGCM-MSSFM-J	94.4
MSGCM-MSSFM -B	94.3
MSGCM-MSSFM	95.7

with the baseline network, we achieve 0.4% improvement only with MSGCM and 0.6% with two proposed modules.

C. Comparison with previous method

To evaluate the performance of our method, we compare our final model with previous methods on the NTU-RGBD dataset. The results are shown in TABLE III. It shows that our model achieves competitive performance compared with the previous methods. We outperform the 2s-AGCN baseline with 0.3% improvement in cross-subject benchmark and 0.6% improvement in cross-view benchmark.

TABLE III
COMPARISONS OF THE VALIDATION ACCURACY WITH PREVIOUS METHODS ON THE NTU-RGBD DATASET.

Methods	X-sub	X-view
Lie Group [23]	50.1	82.8
H-RNN [1]	59.1	64.0
ST-LSTM [3]	69.2	77.7
VA-LSTM [6]	79.,4	87.6
ST-GCN [15]	81.5	88.3
DPRL [28]	83.5	89.8
PB-GCN [29]	87.5	93.2
AS-GCN [30]	86.8	94.2
AGC-LSTM [8]	89.2	95.0
2s-AGCN [16]	88.5	95.1
MSGCN	88.8	95.7

D. Conclusion

In this paper, we propose a Multi-Scale Adaptive Graph Convolutional Network for skeleton-base action recognition. We design a Multi-Scale Graph Convolutional Module to extract spatial features of different scales by downsampling and upsampling the spatial-temporal graph. In the meantime, we design a Multi-Scale Selective Fusion Module under the attention mechanism to adaptively fuse features of different scales. The effectiveness of MSGCM and MSSFM are evaluated on the validation set of cross-subject and cross-view benchmarks of the NTU-RGBD dataset. The results show that the proposed network achieves competitive performance on the NTU-RGBD dataset.

ACKNOWLEDGMENT

This work was supported by the Nation Science Foundation Project [grant numbers 61471055].

REFERENCES

[1] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[3] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[4] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 499–508.

[5] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[6] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2117–2126.

[7] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5457–5466.

[8] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[9] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *arXiv preprint arXiv:1705.08106*, 2017.

[10] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[11] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 597–600.

[12] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[13] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[14] A. Hernandez Ruiz, L. Porzi, S. Rota Bulò, and F. Moreno-Noguer, "3d cnns on distance matrices for human action recognition," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1087–1095.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[16] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[18] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[19] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[21] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[22] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1290–1297.

[23] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[28] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5323–5332.

[29] K. Thakkar and P. Narayanan, "Part-based graph convolutional network for action recognition," *arXiv preprint arXiv:1809.04983*, 2018.

[30] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.