Multi-Feature Enhancement and Adaptation for Lightweight Human Behavior Recognition

Shuai Ding[a], Liqun Kuang[a,b*], Qingde Li[b], Yongqiang Cheng[c], Huiyan Han[a], Yaming Cao[a]
[a] *School of Computer Science and Technology, North University of China, Taiyuan, China*
b Department of Computer Science, University of Hull, Hull, UK
c School of Computer Science, University of Sunderland, UK
kuang@nuc.edu.cn

***Abstract*—Human behavior recognition is a key research area in the field of computer vision. With the flourishing development of computer vision technology, a large number of new embedded vision devices such as VR/AR helmets and mobile visual robots have emerged. Deploying human behavior recognition models on these devices with limited storage and computational resources has become a challenge. Existing behavior recognition methods struggle to simultaneously balance algorithm accuracy and complexity. In order to significantly reduce the amount of model parameter and computational complexity while maintaining high accuracy recognition, this paper proposes a multi-feature enhancement and adaption model for lightweight human behavior recognition. Firstly, efficient multi-scale attention modules are added to the self-attention graph convolution and multi-scale temporal convolution modules to enhance the features of human skeleton data. Secondly, a multi-feature fusion adaptive module is employed to enhance feature fusion and generalization capabilities. Finally, comparative experiments are conducted on a large-scale skeleton dataset. The results demonstrate that the proposed algorithm outperforms recent SOTA methods in terms of parameters, floating-point operations, and recognition accuracy, providing a lightweight method for accurate human behavior recognition.**

Keywords—human skeleton; behavior recognition; lightweight; multi-scale; feature fusion; graph convolutional network

I. INTRODUCTION

Human behavior recognition [1] is a key research area in the field of computer vision, which is dedicated to realizing automatic recognition of human behavior by studying human body movements and postures. It has been widely used in virtual reality, intelligent security, medical assistance, gesture recognition [2] and human-computer interaction, gradually changing people's lifestyles. According to the different modalities of input data, human behavior recognition can be roughly divided into three forms, i.e., RGB video-based, depth video-based and skeleton data-based [3]. Among these modalities, skeleton data is less susceptible to interference from factors such as light brightness, observation angle, and body occlusion. Therefore, skeleton data-based behavior recognition methods are more robust and popular [4].

In the initial stage of the field of skeleton action recognition, convolutional neural networks and recurrent neural networks were used as normal models. However, these methods have limitations and do not fully utilize the topological information of the skeleton structure. With the introduction of graph convolutional network, various methods began to exploit external topological graph structures[5][6]. Spatial temporal graph convolution network (ST-GCN) [7][8][9][10][11] models the human skeleton as a spatiotemporal graph, using nodes to represent joints and edges to represent the relationship between node connections and time frames. It learns spatiotemporal information and improves the feature expression and generalization ability of the model. Addressing the lack of flexibility in the ST-GCN model and failure to utilize skeletal features, Shi et al. proposed a two-stream adaptive graph convolutional network (2s-AGCN) [12], which uses a two-stream framework to process joint features and skeletal features respectively, while designing an adaptive topological graph to improve the model's flexibility. However, this method only focuses on the natural connections of adjacent joint points but does not consider the impact of certain behaviors on non-directly adjacent nodes. In this regard, Cheng et al. proposed a shift graph convolutional network (Shift-GCN) [13], which extends the expression of spatial adjacency by introducing translation operations and enhances the model's feature representation capability and computational efficiency. This method can effectively handle complex physical constraints and intent relationships. Although Shift-GCN broadens the data flow, it still fails to fully exploit the impact of correlations between physically disconnected nodes on behavior recognition. Therefore, a new action recognition learning framework that combines learning goal of information bottleneck and attention-based graph convolution was proposed [14], which achieves a concise but informative latent representation and captures contextually relevant intrinsic topologies of human behavior. Although GCN-based behavior recognition methods have made significant progress in recognition accuracy, most models have suffered from high complexity, large parameter size, and long running time. In order to solve this problem, graph convolutional neural networks based on lightweight methods have been proposed, such as the semantics-guided neural networks (SGN) [15], which uses a single layer of temporal convolutional network (TCN) [16] to extract time domain features from skeleton data, and ends the training after extracting features once. Although this method reduces the computational cost to some extent, its simple network structure leads to inferior accuracy compared to other models.

In summary, current skeleton behavior recognition network models based on deep learning algorithms often suffer from problems such as high structural complexity, deep layers, and large parameter sizes, heavily reliant on computational resources. Due to limitations in memory resources, processor performance and power consumption, it is difficult to deploy complex models on mobile/embedded devices. In this context, from the perspective of lightweight, we proposes a lightweight behavior recognition model with multi-dimensional feature enhancement and adaption, by introducing two lightweight coding blocks, i.e., multi-dimensional feature enhancement (MDFE) and multi-feature fusion adaptive (MFFA). Specifically, the MDFE coding

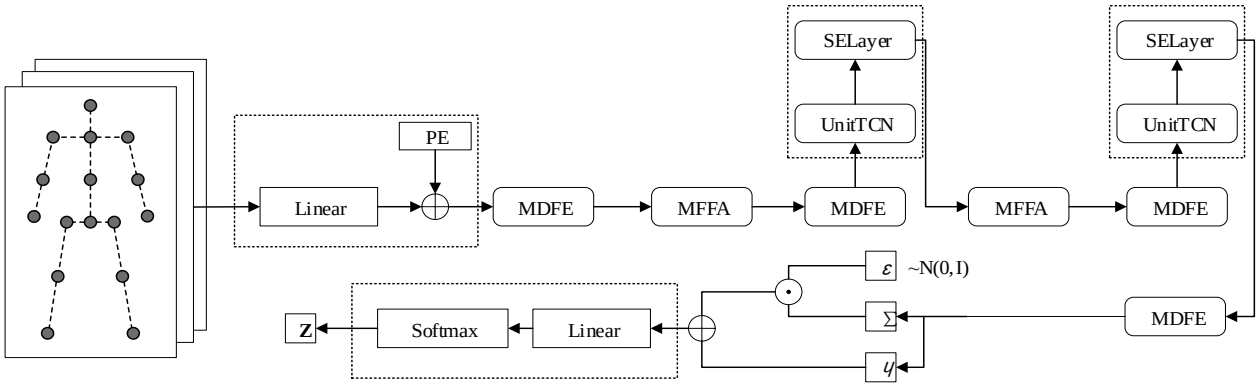

Figure 1 Lightweight Network Structure

block provides rich data features, while the MFFA coding block enhances feature fusion and expression capabilities. This design not only solves the problem of excessive parameter size, but also provides excellent performance in recognition accuracy.

II. METHODOLOGY

A. Lightweight Network Model

We propose a lightweight network model, which is committed to taking into account both lightweight computation and high-precision recognition, as shown in Figure 1.

First, the human skeleton is represented as a graph structure, serving as the input data for the human behavior recognition network model. The joint feature tensors and positional embeddings are used to effectively capture joint features and position information. This process includes specific operations such as data reshaping, adjacency matrix multiplication, feature embedding, position encoding, and batch normalization processing.

Then, we interweave the use of multi-dimensional feature enhancement (MDFE) encoding block and the multi-feature fusion adaptive (MFFA) encoding block to process the extracted features. The MDFE is proposed to provide richer data features, while the MFFA enhance the ability of feature fusion and expression. After the second and third layers of the MDFE block, we apply the unit temporal convolutional network (UnitTCN) and channel attention mechanism (squeeze-and-excitation layer, SELayer) to effectively adjust and enhance the expression capabilities of new feature channels, addressing the mismatch between input and output features caused by the expansion of channel numbers, thereby ensuring effective information transmission and learning between network layers. The combination of these lightweight encoding blocks facilitates the superposition of diverse functional layers, and further improving the model's performance.

Further, by using an auxiliary independent random noise, $\varepsilon \sim N(0,I)$, we sample z as $z = \mu + \sum \varepsilon$, where μ is the mean and ε is a diagonal covariance matrix that inferred from the output of the encoder. In this way, the model can be trained end-to-end through gradient optimization, thus estimating unbiased gradients.

Finally, we use a classifier composed of a single linear layer and SoftMax function to transform the latent variable z into the model parameters of a classification distribution.

B. MDFE Coding Block

The lightweight encoding block of MDFE is designed to provide rich data features for subsequent operations, as shown

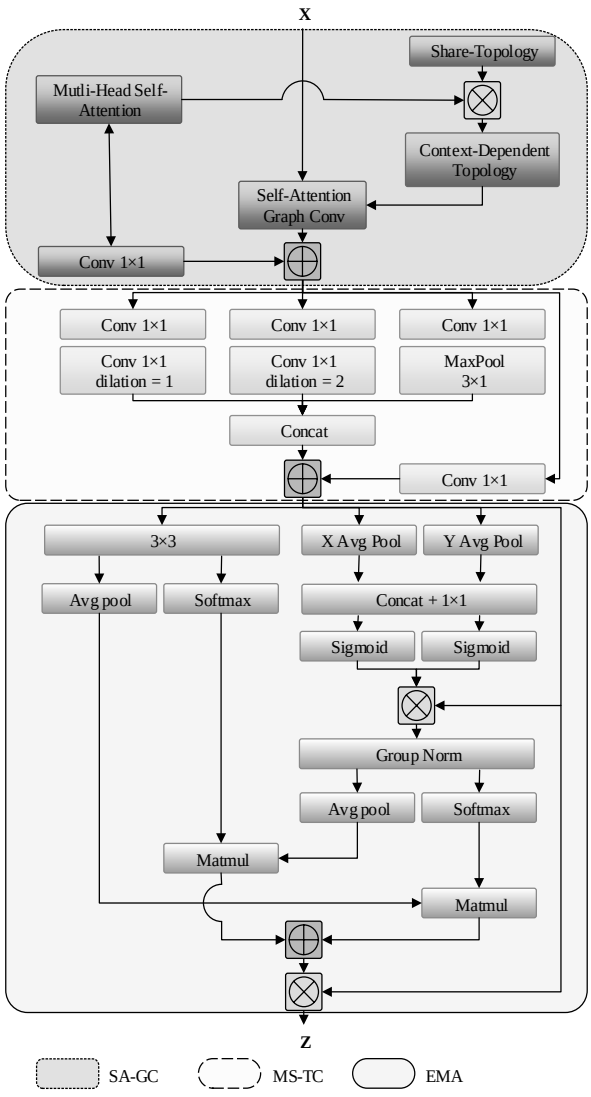

Figure 2 MDFE coding block

in Figure 2. It consists of three core modules, including the self-attention based graph convolution (SA-GC) module for spatial modeling, the multi-scale temporal convolution (MS-TC) module for temporal modeling, and the efficient multi-scale attention (EMA) module [17].

Specifically, SA-GC infers intrinsic topology by self-attention of joint features, and this topology is used as vertex information of graph convolution. MS-TC employs convolution branches with three different kernel sizes and dilation rates to perform multi-branch convolution operations on temporal features. While SA-GC mainly focuses on local relationships of skeletal connections, inferring correlations between bones through self-attention mechanism, it may ignore global information. MS-TC is utilized for extracting multi-scale temporal features, but in some cases, key features at various scales may not be fully captured. Therefore, the EMA module is introduced to adaptively process and weight input features, enabling a more comprehensive extraction of multi-scale feature information.

Further, the EMA module divides input features into multiple groups according to the number of channels, and performs various operations on the features in each channel group, including global adaptive pooling, 1x1 and 3x3 convolution operations, and a SoftMax weighting mechanism. First, the feature information within each channel is integrated through 1x1 and 3x3 convolution operations. Subsequently, global adaptive pooling and the SoftMax weighting mechanism are employed to weight the features within each channel, generating a feature representation with an attention mechanism, thereby improving feature consistency and expression capability. The EMA module aims to extract features from a more global and detailed perspective, focusing on operations within channels. It emphasizes weighted processing of features to acquire more comprehensive multi-scale information. This approach enhances the global-local expression and consistency of features, thereby improving the capture of local relationships and multi-scale features by the SA-GC and MS-TC modules.

C. MFFA Coding Block

Due to the consecutive use of MDFE, there might be an issue of over-extraction of multi-dimensional features, making it challenging to effectively fully integrate these features. Therefore, an encoding block called MFFA, as shown in Figure 3, is introduced. It performs adaptive feature fusion on the already extracted multi-dimensional features, thereby utilizing the multi-dimensional feature information more effectively. By using MDFE and MFFA alternately, a balance can be established between feature extraction and feature integration. This alternating approach helps enhance the model's comprehensive utilization of feature information from different scales and sources, thereby strengthening the model's feature representation and fusion capabilities.

Specifically, the MFFA block is designed based on SA-GC, and incorporates a residual attention feature fusion (RAFF) module and a channel attention module, squeeze-and-excitation layer (SELayer) [18]. The SELayer is introduced to enable better matching between input and output features while expanding the number of channels, and can adaptively learn the weights and importance between feature channels, effectively adjusting and enhancing the expressive capability of new feature channels. The RAFF module consists of residual connection operation and a multi-scale channel attention module (MS-CAM) [19], serving as a crucial module

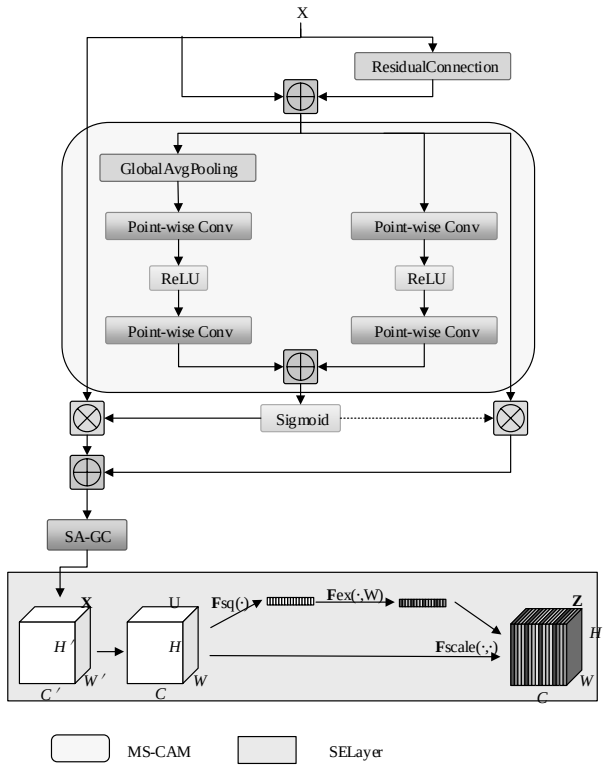

Figure 3 MFFA module

for feature fusion. Its objective is to merge local and global features through residual connections and attention mechanisms. The calculation process of the RAFF module is as shown in formula (1), (2), and (3),

$$Z = MS(Y) \otimes X + (1 - MS(Y)) \otimes Y \tag{1}$$

$$Y = X + \mathrm{Re}s(X) \tag{2}$$

$$MS(Y) = \sigma(l(Y) \oplus g(Y)) \tag{3}$$

First, in order to prevent the issues of vanishing or exploding gradients, the residual connection (ResidualConnection) is used to add the original input features and the features processed by convolution, batch normalization and ReLU activation function. Then, feature fusion is performed through two branches. One is the local feature attention branch , which extracts local information of the input features, and the other is the global feature attention branch, which maps the features to the global perspective through global average pooling and extracts the global information. Finally, using the Sigmoid function, weighting coefficients are computed, and these coefficients are utilized to perform weighted fusion of the original features and the processed features, generating a new feature representation. This weighted fusion strategy helps the model better learn and utilize the correlation between different features, thereby improving the model's feature representation capability and consistency.

III. Experimental Results And Analysis

A. Dataset

In this study, we validated our experimental model using the NTU-RGB+D [20] dataset, which contains over 56,000 video samples from 40 different subjects, totaling 4 million frames. It encompasses 60 action classes, divided into three categories, i.e., daily life (40 actions, such as drinking and eating), health-related (9 actions, such as sneezing and falling),

Figure 4 An example of the NTU-RGB+D dataset

and mutual interaction (11 actions, such as punching and kicking). The dataset was collected from 40 different subjects, aged between 10 to 35 years old. An example of the video sample is shown in Figure 4. The dataset is divided into training and test sets according to their IDs to conduct evaluation experiment, with the training set being approximately 2.5 times the test set data.

B. *Experimental setup*

In this study, the initial feature channel number is set to 16, and each sample is adjusted to 64 frames [21], and the batch_size is set to 128. The optimizer with a momentum coefficient of 0.9 is used to update the model to minimize the total loss. To normalize the learned latent features, the mean of the variational distribution is set to zero. Additionally, the mean of the conditional variational distribution for each action category is set to a randomly generated orthogonal vector, and appropriately scaled to triple its original size. During the training stage, the averages of latent features and the conditional latent features are estimated by averaging the latent vectors in small batch data, respectively. The label smoothing technique [22] is adopted and the value is set to 0.1 to prevent the model from being overconfident in predictions. Then the loss function is as defined in formula (4),

$$L_{TOTAL} = L_{CLS} + \lambda_1 L_{mMMD} + \lambda_2 L_{cmMMD} \quad (4)$$

Here, L_{CLS} quantifies the empirical loss between the output of the prediction network and the true labels, L_{mMMD} represents the marginal maximum mean discrepancy loss, and L_{cmMMD} represents the conditional marginal maximum mean discrepancy loss. λ_1 and λ_2 are the weight coefficients of L_{mMMD} and L_{cmMMD}, and were set to 0.0001 and 0.05, respectively. The experimental environment is listed in Table 1.

TABLE I. EXPERIMENTAL SETUP

Environment name	**Specific Configuration**
Memory	6 4GB
Operating system	Ubuntu18.0.4
GPU	NVIDIA RTX 3090
CPU	Intel(R) Xeon(R) W-2245
Python/Pytorch	3.7/1.8.0
CUDA	11.1

C. *Experimental results and analysis*

In order to verify the effectiveness of our proposed method, experiments and comparisons were conducted on the NTU RGB+D dataset with other seven SOTA methods for skeleton-based behavior recognition. The experimental results are presented in Table 2.

TABLE II. RECOGNITION ACCURACY AND COMPLEXITY OF NTU-RGB+D DATASET

Method	Year	Parameters ($\times 10^6$)	Flops ($\times 10^9$)	Accuracy (%)
Shift-GCN [13]	2020	4.54	10	90.7
MS-G3D [5]	2020	6.4	48.88	**91.5**
MST-GCN [9]	2021	2.82	16.03	89.0
HST-GCNS [10]	2022	2.00	-	89.5
STF-Net [7]	2023	6.8	-	91.1
LST-GCN [8]	2023	1.62	17.54	90.8
TFC-GCN [11]	2023	**0.18**	1.9	87.9
Ours	2024	0.64	**0.57**	90.9

The results demonstrate that our model is highly competitive in terms of accuracy, parameter number, and computational complexity compared to recent SOTA methods. The proposed method achieve the fastest computation speed, and the flop is only 0.57×10^9. Specifically, compared to the TFC-GCN method, although our method has 3.5 times more parameters, its computational complexity is only 30%, while achieving a higher accuracy by 3.0%. Then, compared to LST-GCN, our method's parameter size is only 2/5 of LST-GCN, with computational complexity at only its 1/30, and a recognition accuracy improvement of 0.1%. Moreover, in comparison to the remaining five comparison methods, i.e., Shift-GCN, MS-G3D, MST-GCN, HST-GCNS, and STF-Net, our approach achieves a parameter reduction of 3 to 10 times and computational complexity decrease of 17 to 85 times, while maintaining similar accuracy.

In general, our proposed method achieves an excellent performance among these SOTA methods. It obtains competitive performance in recognition accuracy while significantly reducing both network parameter size and computational complexity. The scatter plots comparing our algorithm with recent methods in terms of parameter size and computational complexity are presented in Figures 5 and 6, respectively. These figures provide a more intuitive and clear demonstration that our algorithm exhibits significant advantages in terms of parameter size, computational complexity, and recognition accuracy.

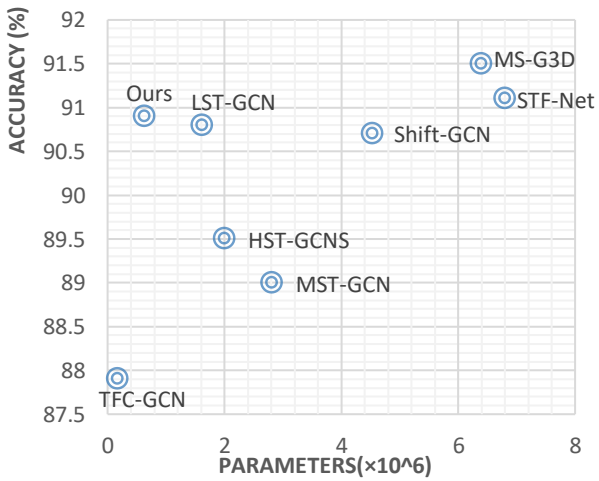

Figure 5 Recognition accuracy and parameter size

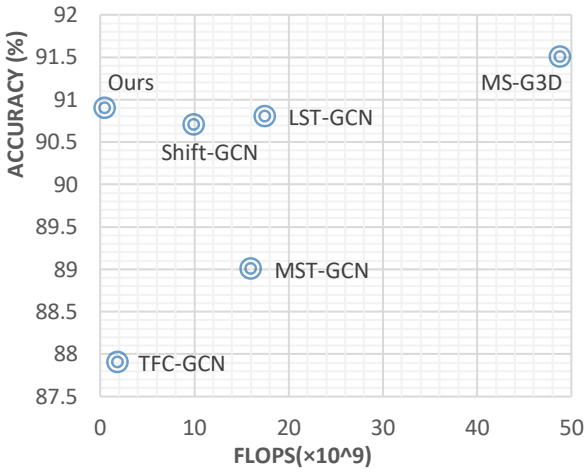

Figure 6 Recognition accuracy and computational complexity

IV. Conclusion

This paper proposed a lightweight human behavior recognition model that achieves comprehensive feature extraction, detailed spatial modeling, channel-level information integration, and residual iterative attention feature fusion through the cross combination of multidimensional feature enhancement and adaptive multi-feature fusion. While improving recognition accuracy, the model effectively controls the temporal and spatial complexity. However, to make the model as lightweight as possible, we reduce the number of feature channels and perform lightweight design on network structure. In future work, further research could focus on enhancing the model's ability to extract deep features while maintaining model complexity. Additionally, continuous lightweight improvements to the network model could be studied without affecting the recognition accuracy as much as possible.

Acknowledgment

Our work is supported in part by the Shanxi Province Science and Technology Major Special Project (No. 202201150401021), the Shanxi Province Science and Technology Achievements Transformation Guidance Special Project (No. 202104021301055), and the Natural Science Foundation of Shanxi Province (202303021211153 and 202203021222027).

References

[1] Hu K, Jin J, Zheng F, et al. Overview of behavior recognition based on deep learning[J]. Artificial Intelligence Review, 2023, 56(3): 1833-1865.

[2] Yang L, Huang J, Feng T, et al. Gesture interaction in virtual reality[J]. Virtual Reality & Intelligent Hardware, 2019, 1(1): 84-112.

[3] Sun Z, Ke Q, Rahmani H, et al. Human action recognition from various data modalities: A review[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(3):3200-3225.

[4] Ren Z, Zhang Q, Cheng J, et al. Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition [J]. Neurocomputing, 2021, 433: 142-153.

[5] Liu Z, Zhang H, Chen Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 143-152.

[6] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2018.

[7] Wu L, Zhang C, Zou Y. SpatioTemporal focus for skeleton-based action recognition[J]. Pattern Recognition, 2023, 136: 109231.

[8] Xing Y, Zhu J, Li Y, et al. An improved spatial temporal graph convolutional network for robust skeleton-based action recognition[J]. Applied Intelligence, 2023, 53(4): 4592-4608.

[9] Chen Z, Li S, Yang B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2021: 1113-1122.

[10] Mostafa A, Peng W, Zhao G. Hyperbolic spatial temporal graph convolutional networks[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 3301-3305.

[11] Wang K, Deng H. TFC-GCN: Lightweight Temporal Feature Cross-Extraction Graph Convolutional Network for Skeleton-Based Action Recognition[J]. Sensors, 2023, 23(12): 5593.

[12] Shi L, Zhang Y, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12026-12035.

[13] Cheng K, Zhang Y, He X, et al. Skeleton-based action recognition with shift graph convolutional network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 183-192.

[14] Chi H, Ha M H, Chi S, et al. Infogcn: Representation learning for human skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20186-20196.

[15] Zhang P, Lan C, Zeng W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1112-1121.

[16] Kim T S, Reiter A. Interpretable 3d human action analysis with temporal convolutional networks[C]//2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, 2017: 1623-1631.

[17] Ouyang D, He S, Zhang G, et al. Efficient Multi-Scale Attention Module with Cross-Spatial Learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[19] Dai Y, Gieseke F, Oehmcke S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3560-3569.

[20] Shahroudy A, Liu J, Ng T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1010-1019.

[21] Chen Y, Zhang Z, Yuan C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13359-13368.

[22] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.