Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090212
ISSN 2371-8838 Vol. 9 Num. 2

Multi-scale Self-Attention Convolutional Networks for Skeleton-Based Action Recognition

Yuwen Fang[1,a,*], **Zonghui Wang**[1,b]

[1] *School of Computer and Information Sciences, Chongqing Normal University, Chongqing, China*
[a]2563446200@qq.com, [b]17638253409@163.com
**Corresponding author*

Keywords: Skeletal action recognition; graph convolutional network; temporal self-attention mechanism; multi-scale temporal convolution; spatiotemporal modeling

Abstract: Skeleton-based action recognition is one of the core tasks in the field of video understanding and is widely used in scenarios such as human-computer interaction, intelligent monitoring, and sports analysis. Existing graph convolutional networks (GCNs) effectively model the spatial dependency of joints by constructing a skeletal connection graph, but their temporal modeling usually relies on fixed-window temporal convolution, which makes it difficult to capture the global dynamic associations between distant frames, resulting in the loss of key temporal features in complex actions. To this end, this paper proposes a feature extraction framework based on temporal context enhancement. First, the framework uses GCN to explicitly encode the spatial dependency of skeletal joints and extract spatial features containing physical connection priors; secondly, the local temporal dynamics between adjacent frames are captured through a multi-scale temporal convolution module; on this basis, the self-attention mechanism of the temporal dimension is introduced to model the cross-frame association of the feature sequence output by the temporal convolution, and the key dependencies between distant action frames are adaptively captured through dynamic weight allocation, realizing temporal modeling from local to global. Experimental results on the NTU RGB+D dataset show that the proposed method significantly outperforms the existing advanced models in the task of skeletal action recognition, verifying the effectiveness of the temporal self-attention mechanism in modeling complex action dynamics.

1. Introduction

As a core task in the field of video understanding, skeleton action recognition has important application value in scenarios such as human-computer interaction, intelligent monitoring, and sports analysis. Its core challenge lies in how to efficiently model the spatiotemporal dependencies in human skeleton sequences - the spatial dimension needs to capture the structural correlation of joints, and the temporal dimension needs to capture the dynamic evolution of action sequences. Traditional methods such as graph convolutional networks (GCNs) effectively model the spatial dependencies of joints by constructing a graph structure of skeleton connections, but their temporal modeling part mostly relies on temporal convolutions of fixed windows, which makes it difficult to capture the global dynamic correlation between long-distance action frames, resulting in the loss of key temporal features in complex actions.

In recent years, the self-attention mechanism has shown outstanding advantages in long-range dependency modeling, and can capture the semantic correlation of any position in the sequence through dynamic weight allocation. However, existing studies have mostly directly applied the self-attention mechanism to the spatiotemporal joint modeling of skeleton sequences, ignoring the hierarchical differences between spatial structural features and temporal dynamic features: the spatial dependency of skeletal joints has a clear physical connection prior, while the action evolution in the temporal dimension requires temporal context aggregation based on spatial features. If self-attention is directly applied to the original skeleton sequence, it is easy to blur the spatial structure information and it is difficult to fully utilize the hierarchical features of the skeleton data.

To this end, this paper proposes a feature extraction network architecture based on temporal context enhancement. First, the graph convolutional network (GCN) is used to explicitly encode the spatial dependencies of skeletal joints and extract spatial features containing physical connection priors. Then, the local temporal dynamics between adjacent frames are captured through a multi-scale temporal convolution module, focusing on action details and motion patterns within a short temporal range. On this basis, the temporal dimension self-attention mechanism is introduced to perform cross-frame association modeling on the feature sequence output by the temporal convolution. The key dependencies between long-distance action frames are adaptively captured through dynamic weight allocation, and the local temporal features are extended to global context modeling. The main contributions of this paper are as follows:

●Design a temporal self-attention enhancement module to perform global context modeling on feature sequences in the temporal dimension, effectively capture key dynamic patterns of cross-frame actions, and make up for the shortcomings of traditional temporal models in modeling long-distance dependencies;

●The use of adaptive topology modeling and multi-scale graph convolution methods is beneficial to reducing the number of model parameters and training time;

●Experiments on the NTU RGB+D dataset show that our method significantly outperforms existing advanced models in the skeleton action recognition task, verifying the effectiveness of the temporal self-attention mechanism.

2. Related Work

2.1. Graph Neural Networks

Graph Neural Networks (GNNs) aggregate neighborhood information through message passing mechanisms, showing unique advantages in non-Euclidean data modelling, for example, in social network scenarios, GNNs can not only analyze explicit connections between users, but also mine potential behavioral associations; in human action recognition, GNNs break through the dependence of traditional convolutional networks on rigid topological structures by fusing the spatial dependence of skeletal nodes with the temporal correlation of motion trajectories. At present, GNNs based on graph convolution are mainly divided into two categories: spectral-based methods and spatial-based methods. The spectral method is based on the theory of graph signal processing, and its core idea is to map the geometric characteristics of the graph structure to the frequency domain space. The node signal is spectrally decomposed through the graph Fourier transform, and the frequency domain filter is designed with the help of the spectral properties of the Laplace matrix. The spatial method adopts a more intuitive topology-aware design to simulate the local information transmission process directly on the original graph structure. It draws on the local receptive field mechanism of the traditional convolutional neural network (CNN), defines the learnable neighborhood aggregation function of the node, and iteratively updates the node representation layer by layer.

2.2. Skeleton-based Action Recognition

Although traditional human action recognition methods rely on manual feature design and template matching, which can achieve certain results in specific scenarios, their generalization ability is limited and they are difficult to adapt to complex and changing action patterns. In recent years, deep learning methods have been able to automatically mine potential spatiotemporal structural information with data-driven feature learning capabilities. However, methods based on recurrent neural networks (RNNs) often ignore key spatial structural information in the process of serializing joint coordinates into vectors, resulting in loss of position information; while methods based on convolutional neural networks (CNNs) usually map skeletal key points to "pseudo-images" for modeling. Although they can extract local spatial features, they have certain limitations in modeling long-term series dynamics.

The topological connection characteristics of human skeletons are naturally compatible with graph structures. This advantage has promoted the development of skeletal action recognition methods based on graph convolutional networks (GCNs). A milestone work in this field is the spatiotemporal graph convolutional network (ST-GCN) proposed by Yan et al. in 2018. In the spatial dimension, the adjacency relationship is constructed through the physical connection of joint points. In the temporal dimension, cross-frame node connections are used to replace traditional optical flow calculations, thereby achieving efficient joint modeling of the spatiotemporal dynamic characteristics of skeletal motion.

On this basis, a series of improved methods have emerged. For example, [1] et al. used graph distance to encode the topology of the physical connection of the skeleton, and introduced persistent homology analysis to characterize the dynamics of the action-specific system, in order to solve the problem of loss of skeleton connection topology information caused by jointly optimizing the adjacency matrix and model weights in skeleton action recognition. [2] et al. proposed a deformable graph convolutional network (DeGCN) to address the problems of fixed information aggregation mode, lack of intra-class change perception, and insufficient redundant connection processing capabilities in traditional graph convolutional networks. By learning the variable sampling positions on the spatiotemporal graph, the model can adaptively capture highly discriminative joint features and construct a dynamic receptive field. At the same time, the temporal features are defined in a continuous latent space, which fits the continuous nature of human motion.

2.3. Self-Attention Mechanism

In recent years, the self-attention mechanism, as a core component of Transformer, has achieved remarkable results in fields such as computer vision and natural language processing. It has become an important technology for sequence modeling by capturing long-distance dependencies in sequences and adaptively allocating weights to enhance feature extraction capabilities. In the field of action recognition based on skeleton keypoints, [3][4][5][6] achieved excellent results.[7] et al. embed the joint set of the same partition into a unified token before inputting the attention module for feature modeling. They only rely on the symbolic partition strategy and do not introduce a dedicated attention mechanism for a specific partition. Aiming at the possible redundant calculation problem between two joints, [8] et al. designed a sparse self-attention mechanism, which effectively captures the spatial correlation between joint points through sparse matrix multiplication and greatly reduces the computational cost. To this end, this paper first explicitly models the spatial dependencies of skeleton key points through a graph convolutional network (GCN) to extract the spatial features of the skeleton structure; then captures the local temporal dynamic information through a temporal feature extraction module; and finally introduces a self-attention mechanism to re-aggregate the global features in the temporal dimension and adaptively capture the long-range dependencies across time steps.

3. Methods

3.1. Preliminaries

The connection relationship between the key points of the human skeleton is modeled as a graph structure $G = (V, E)$, The vertices V represent the keypoints of the human body, using $V = (v_1, v_2, , , , , v_N)$ to indicate. E is the set of bone edges represented by the adjacency matrix $A \in \mathbb{R}^{N \times N}$,where the element a_{ij} takes a value of 1 or 0, denoting whether v_1 and v_2 are connected. Given a sequence of the skeleton, is represented as $X \in \mathbb{R}^{C \times T \times V}$, where C , T and V mean the number of channels, frames, and joints, respectively. The feature aggregation process of the entire graph is as follows:

$$X^{(l+1)} = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} W^{(l)} X^{(l)}) \tag{1}$$

Where D is the degree matrix of A. W are the graph convolution weights. denotes the input features to the $(l+1)_{th}$ layer and X^{l+1} denotes the output features to $(l+1)_{th}$ the layer. σ is the activation function.

3.2. Multi-scale adaptive convolution Module

In order to more effectively model the spatial relationship between skeletal key points, we proposed a multi-scale adaptive graph convolution module (MS-AGC). Traditional graph convolution methods rely on fixed human topological structures to construct adjacency matrices. However, in practical applications, different categories of actions often involve different joint interaction patterns. Fixed graph structures are difficult to flexibly adapt to the diversity of actions, which easily leads to limited information transmission paths. For example, the action of running relies more on the global coordination between the joints of the limbs, so the connection between the legs and arms is particularly critical; while waving is concentrated in the hand area, and the fine-grained structure between its key points is more worthy of attention. This shows that in action recognition, different actions should focus on graph structure modeling of different parts. To this end, we introduce a learnable adjacency matrix at multiple scales to dynamically adjust the connection strength between nodes, thereby achieving more expressive and adaptive spatial modeling capabilities. As shown in Figure 1

Given an input feature $X \in C \times T \times V$, we first divide it along the channel dimension by 1×1Conv to obtain multiple sub-feature fragments $\{X_1, X_2, \cdots, X_n\}$, each of which has a shape of $\mathbb{R}^{C_i \times T \times V}$. Then, we apply a spatially adaptive graph convolution operation to each fragment and obtain the corresponding output. All sub-outputs are concatenated along the channel dimension to form the final output feature, that is:

$$O_{\text{out}} = [Y(X_1) \,||\, Y(X_2) \,||\, Y(X_3) \,||\, Y(X_4)] \tag{2}$$

Where represents the connection operation, Y represents the adaptive graph convolution operation, and O_{out} represents the total output. In order to improve the stability of model training and accelerate the convergence speed, this study introduces the residual connection mechanism, followed by batch normalization processing and ReLU activation function mapping, and finally obtains the output representation. That is:

$$Z = \sigma(b_n(O_{out} + res)) \tag{3}$$

Where res represents residual link, $b_n(\cdot)$ are the batch normalization, $\sigma(\cdot)$ is the activation function ReLU.

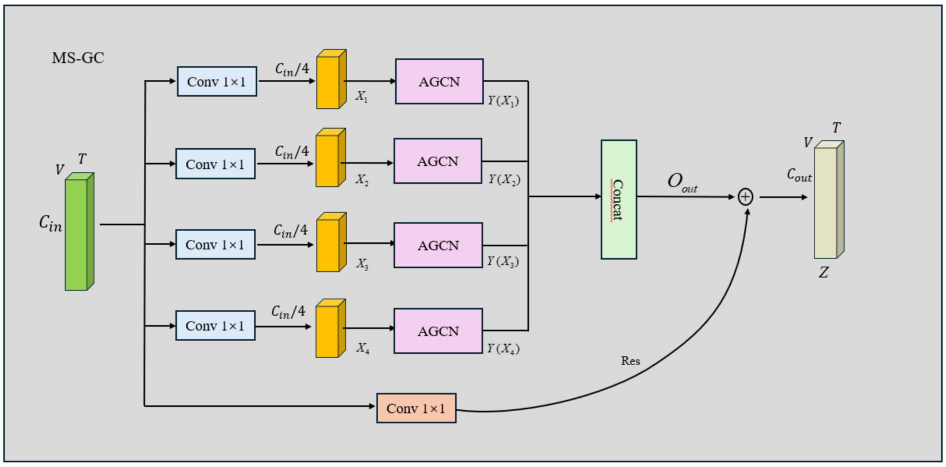

Figure 1: Multi-scale Adaptive Graph Convolution

The above adaptive convolution module is mainly composed of three sub-modules: feature conversion, adaptive adjacency matrix calculation and information aggregation. For the input feature $X \in \mathbb{R}^{C/4 \times T \times V}$, it is first mapped to a high-dimensional space through a linear transformation to enhance the expression ability, that is:

$$X^{'} = M(X) = XW \tag{4}$$

Where $X^{'} \in \mathbb{R}^{C^{'}/_{4} \times T \times V}$ is the transformed high-dimensional feature, and $W \in \mathbb{R}^{C^{'} \times C}$ is a learnable weight matrix. In the adaptive calculation of the adjacency matrix, we first perform average pooling on the input feature $X \in \mathbb{R}^{C/4 \times T \times V}$ in the time dimension to extract static structural features. Then, we compress the input features through two different linear transformation functions $\psi(\cdot)$ and $\varphi(\cdot)$ to reduce the computational complexity. For each pair of joint points (V_i, V_j), the transformed feature (x_i, x_j) can be expressed as:

$$M(\psi(x_i), \varphi(x_j)) = \sigma(\psi(x_i) - \varphi(x_j)) \tag{5}$$

to reflect the correlation of the dynamic structure between nodes. On this basis, the final adaptive adjacency matrix is expressed as:

$$\tilde{A} = A_s + \alpha \bullet Q \tag{6}$$

Among them, A_s is the predefined shared adjacency matrix, Q is the dynamically generated channel-specific adjacency matrix, and α is a learnable parameter used to adjust the influence of the adaptive part. After obtaining high-dimensional features of different scales $\{X_1^{'}, X_2^{'}, \cdots, X_n^{'}\}$ and the corresponding adaptive graph structure $\{Q_1, Q_2, \cdots Q_n\}$, the final graph convolution output is:

$$Z = [Q_1 X_1^{'} \,||\, Q_2 X_2^{'} \,||\, Q_3 X_3^{'} \,||\, Q_4 X_4^{'}] \quad (7)$$

Where || is concatenate operation, since the adjacency matrix Q changes dynamically with the input samples, the whole process constitutes a dynamic graph convolution framework, which can adaptively build graphs according to the semantic features of the input actions, thereby enhancing the model's ability to model structural differences. Its architecture is shown in Figure 2.

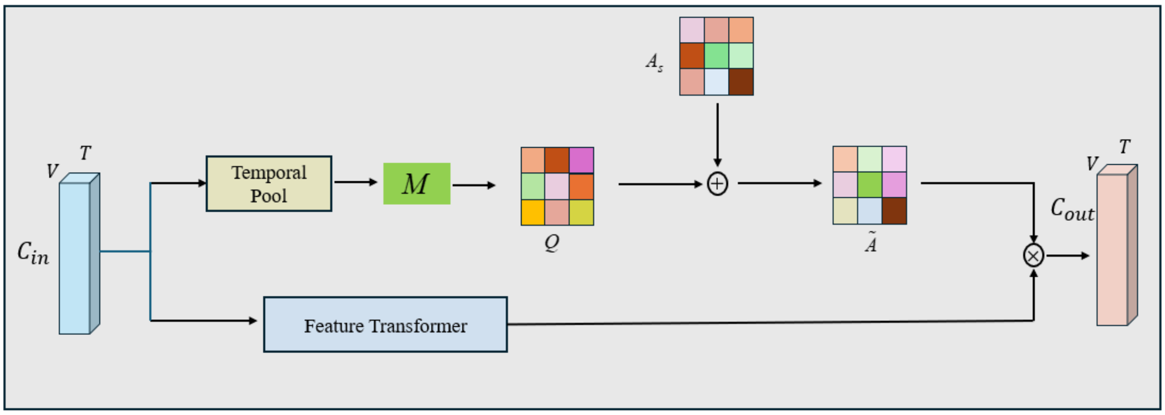

Figure 2: Feature transformation and adaptive topology generation

3.3. Hybrid Attention Module

In order to more fully model the dynamic patterns of joints changing over time in action sequences, we introduced a multi-scale temporal convolution module (MSTC); in addition, in order to capture the temporal dependencies between joints from a global perspective, we further combined a temporal modeling branch based on an attention mechanism, thereby improving the model's perception of complex temporal dynamics, as shown in the Figure 3. First, for the input feature $X \in \mathbb{R}^{C \times T \times V}$, we use 4 different receptive fields $(dilation = 1, dilation = 2, dilation = 3.dilation = 4)$ and reduce the channel size through 4 parallel $1 \times 1 Conv$ branches to change the feature to $X \in \mathbb{R}^{C/4 \times T \times V}$. For the first two branches, we further introduce three trainable parameter matrices W^Q、 W^K、 W^V to generate the corresponding query vectors, key vectors, value vectors, Q , K , $V \in \mathbb{R}^{C^{'} \times T \times V}$. Subsequently, the generated Q , K , V features are rearranged and reshaped to obtain Q_1, K_1, V_1 to meet the computational requirements of multi-head attention. Specifically, it is divided into groups, (where represents the number of heads), and in each group, the query vector Q and the key vectors K are dot-producted, and the attention map on the global time dimension is obtained by combining normalization and activation function processing. Then, the obtained attention map is weighted multiplied with the corresponding Value vector. The output features generated by all heads (multi-head) are then concatenated in the channel dimension to integrate the information from different attention subspaces to form the final multi-head attention output. The above process is expressed as:

$$Out_n = softamx(\frac{Q^n K^{nT}}{\sqrt{C}})V^n, n \in \{1, 2, 3, 4\} \quad (8)$$

Finally, the global temporal features generated by the self-attention mechanism are fused with the local dynamic features extracted by multi-scale temporal convolution, so as to simultaneously model the local details and global dependencies of the action sequence in the same representation space, further improving the model's ability to express and discriminate temporal behavior patterns. It can be expressed as:

$$Z = [Out_1 \,||\, Conv_1 \,||\, Out_2 \,||\, Conv_2 \,||\, Conv_3 \,||\, Conv_4] \tag{9}$$

Among them, *Conv* indicates that the convolution operation is performed at the time level to extract local temporal dynamic features.

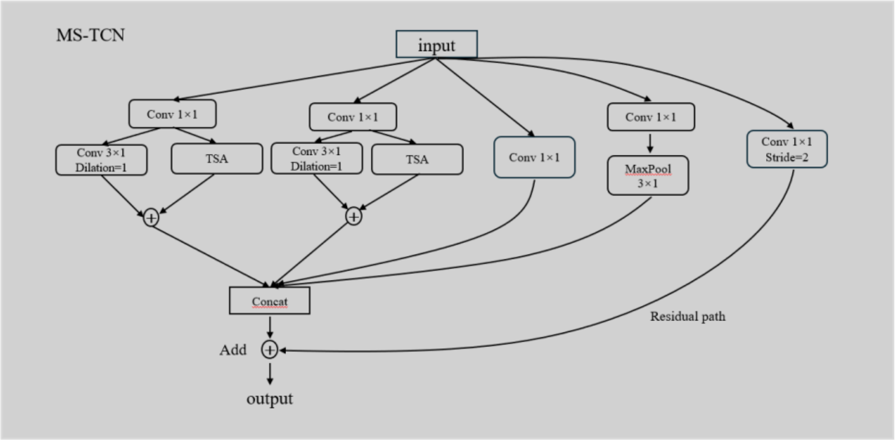

Figure 3: Multi-scale Self-Attention Temporal Convolution

3.4. Overall architecture

The model consists of 10 blocks, as shown in the Figure 4. The number of input channels of each block is 3, 64, 64, 64, 64, 128, 128, 128, 256, 256. The temporal channel is halved in the 5th and 8th blocks with stride=2.

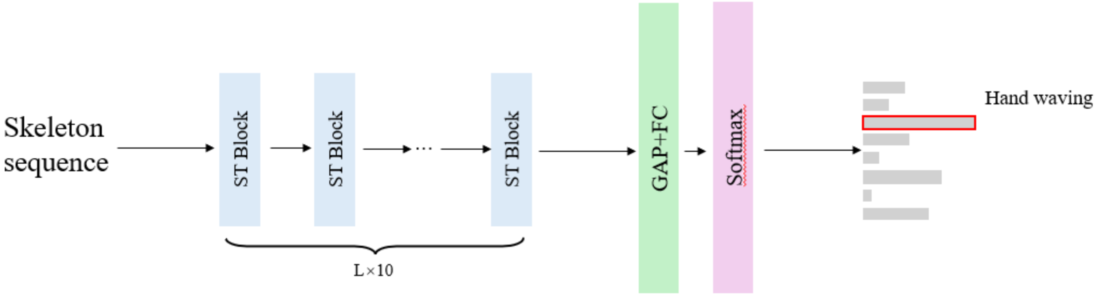

Figure 4: The overall architecture of our model

4. Experiments

To verify the effectiveness of the model, this paper evaluates the Top-1 and Top-5 indicators on the NTU RGB +D 60 dataset.

4.1. Dataset

NTU RGB+D 60 is a large indoor dataset widely used in skeletal action recognition research. It contains 60 human action categories, covering 40 daily behaviors, 9 health-related actions, and 11 interactive actions. The dataset was completed by 40 subjects, with a total of 56,880 action samples, collected synchronously from three horizontal viewing angles (-45 °, 0 °, 45 °) using a Microsoft Kinect v2 device. Each sample contains up to two human skeletons, each consisting of 25 joint nodes. The dataset provides two standard test settings: Cross-Subject (X-Subject) and Cross-View (X-View). In the X-Sub setting, 40,230 samples collected by 20 subjects are used as the training set, and 16,560 samples collected by the remaining 20 subjects are used as the test set; in the X-View setting, samples collected by cameras 2 and 3 are used for training, and the data of camera 1 is used for model evaluation. These two evaluation methods effectively measure the generalization of the model under different subjects and viewpoint changes.

4.2. Implementation Details

This model is trained using the stochastic gradient descent (SGD) optimization algorithm, with momentum set to 0.9, L2 regularization introduced to prevent overfitting, and weight decay coefficient set to $\lambda = 0.0004$. The training is performed for 65 epochs, with an initial learning rate of 0.1, and the learning rate is decayed by multiplying it by 0.1 in a piecewise manner at the 35th and 55th epochs. The entire training process is completed on an NVIDIA A6000 GPU.

4.3. Comparison of the state art

We combine joint and motion. Our model is compared with the latest models in Table 1 on the NTU RGB+D dataset. The experimental results show that the proposed model reaches the current state-of-the-art level in Top-1 indicators and shows a strong performance advantage.

Table 1: Comparison of the Top-1 accuracy with the state-of-the-art methods on the NTU RGB+D dataset

Methods	**NTU-RGB+D**	
	X-Sub(%)	**X-View(%)**
ST-GCN[9]	81.5	88.3
2s-AGCN[10]	88.5	95.1
Shift-GCN[11]	90.7	96.5
MS-G3D[12]	91.5	96.2
ST-TR[13]	89.9	96.1
Dynamic GCN[14]	91.5	96.0
DSTANet[15]	91.5	96.4
MST-GCN[16]	91.5	96.6
Ours	**91.8**	**96.8**

5. Conclusions

In view of the shortcomings of traditional temporal convolution in modeling long-distance temporal dependencies, this paper proposes a novel spatiotemporal modeling framework for skeletal action recognition. This method models spatial dependencies through graph convolutional networks, extracts local dynamic features through multi-scale temporal convolution modules, and integrates temporal self-attention mechanisms to achieve global temporal context modeling, achieving a good balance between local detail perception and long-range dynamic capture. In addition, the introduction of adaptive topological modeling and multi-scale design effectively improves the efficiency and expressiveness of the model. Experiments on the NTU RGB+D dataset show that the proposed method significantly outperforms existing mainstream methods in terms of recognition accuracy, demonstrating the wide application potential and research value of this method in the field of complex action modeling.

References

[1] Zhou Y, Yan X, Cheng Z Q, et al. Blockgcn: Redefine topology awareness for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 2049-2058.

[2] Myung W, Su N, Xue J H, et al. Degcn: Deformable graph convolutional networks for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2024, 33: 2477-2490.

[3] Qin X, Cai R, Yu J, et al. An efficient self-attention network for skeleton-based action recognition[J]. Scientific Reports, 2022, 12(1): 4111.

[4] Wang Q, Shi S, He J, et al. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition[C]//2023 IEEE International Conference on Big Data (BigData). IEEE, 2023: 936-945.

[5] Shi F, Lee C, Qiu L, et al. Star: Sparse transformer-based action recognition[J]. arXiv preprint arXiv:2107.07089, 2021.

[6] Choi J, Wi H, Kim J, et al. Graph convolutions enrich the self-attention in transformers![J]. Advances in Neural Information Processing Systems, 2024, 37: 52891-52936.

[7] Pang Y, Ke Q, Rahmani H, et al. Igformer: Interaction graph transformer for skeleton-based human interaction recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 605-622.

[8] Shi F, Lee C, Qiu L, et al. Star: Sparse transformer-based action recognition[J]. arXiv preprint arXiv:2107.07089, 2021.

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv:1801.07455, 2018.

[10] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12026–12035, 2019.

[11] Cheng K, Zhang Y F, He X Y, et al. Skeleton-based action recognition with shift graph convolutional network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 180-189

[12] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 143–152, 2020.

[13] C. Plizzari, M. Cannici, M. Matteucci, Skeleton-based action recognition via spatial and temporal transformer networks, Comput. Vis. Image Underst. 208-209 (2021) 103219

[14] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In Proceedings ofthe 28th ACM International Conference on Multimedia, pages 55–63, 2020.

[15] L. Shi, Y. Zhang, J. Cheng, H. Lu, Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, in: Revised Selected Papers of the Asian Conf. on Computer Vision (ACCV'20), Part V, Springer, Cham, Switzerland, 2020, pp. 3853.

[16] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in: AAAI Conf. on Articial Intelligence (AAAI'21), IAAI'21, EAAI'21, AAAI, RedHook, NY, USA, 2021, pp. 11131122.