Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2024, 11(3):139-151**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Skeleton Action Recognition Based on Multi-Scale Spatial Temporal Convolution Dynamic Gated Transformer

Ziyang Lin*[1], Bo Su[1,2]

[1]School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China,454003
[2]Henan International Joint Laboratory of Direct Drive and Control of Intelligent Equipment, Jiaozuo, China,454003
*Email: linziyang@home.hpu.edu.cn

Abstract In current research on human skeletal action recognition, convolutional networks lack the ability to model global skeletal information, while Transformer networks exhibit weaker learning capabilities for short-term local skeletal information. Moreover, these networks are susceptible to interference from other high-frequency noise when modeling fine-grained skeletal spatial information. To address these issues, a multi-scale spatial temporal convolution dynamic gated Transformer network (MSST-CDGT) is proposed for human skeletal action recognition. Initially, the network performs positional encoding on human skeletal information and introduces a multi-scale spatial dynamic gated Transformer to suppress high-frequency noise in non-spatial dimensions, simultaneously modeling both local multi-scale and global spatial information of the skeletal structure. Subsequently, a multi-scale temporal convolution network is designed to model the temporal sequence information of the skeletal structure. Finally, a multi-stream data fusion framework is constructed to weightedly integrate the multi-stream operation data of the skeletal structure and obtain prediction results. The proposed network achieves accuracies of 92.6% and 96.8% on the NTU-RGB+D 60 dataset across subject and view benchmarks and accuracies of 89.5% and 90.7% on the NTU-RGB+D 120 dataset across subject and setting benchmarks, respectively.

Keywords action recognition; Transformer; dynamic gated; multi-scale spatial; multi-scale temporal

1. Introduction

Human skeletal action recognition has emerged as a research hotspot in the field of computer vision [1], with widespread applications in areas such as intelligent surveillance [2], industrial control [3], autonomous driving [4], rehabilitation training [5], and motion instruction [6]. Algorithms for human skeletal action recognition can be broadly categorized into two main types: traditional methods and deep learning methods. Traditional human skeletal action recognition algorithms rely on manual extraction of action features. For example, Yang et al. [7]proposed a feature representation based on differences in joint positions, while Wang et al. [8]used Speeded Up Robust Features (SURF) descriptors and optical flow to match inter-frame feature points. Although traditional methods have achieved initial success in action recognition, they can only extract simple shallow features, resulting in limited feature representation capabilities.

Deep learning-based human skeletal action recognition algorithms can be broadly categorized into two major types: convolutional networks [9] and Transformer-based networks [10]. For instance, Yan et al. [11] proposed a seminal Spatial Temporal Graph Convolutional Network (ST-GCN), representing human motion skeletal sequences as a spatial temporal graph and designing action recognition models by extending graph networks. Shi et al. [12] argued that the fixed topological structure used by ST-GCN lacks flexibility in modeling and

introduced a Two Stream Adaptive Graph Convolutional Network (2s-AGCN), which significantly enhances recognition performance by fusing skeletal features and keypoint features using a dual-stream structure. Cheng et al. [13] proposed a Shift Graph Convolutional Network (Shift-GCN) to endow the model with flexible receptive fields. In recent years, Transformers have emerged with their unique global modeling capabilities and self-attention mechanisms, showing broader prospects for development. For example, Plizzari et al. [14] introduced a Spatial Temporal Transformer Network (ST-TR) for human skeletal action recognition, which models dependencies between joints using Transformer self-attention algorithms, thereby capturing the entire skeleton's long-range dependencies. Shi et al. [15] proposed a novel Decoupled Spatial-Temporal Attention Network (DSTA-Net), which can model spatial temporal dependencies between joints without requiring the positions of keypoints and connectivity information. Despite the multifaceted explorations in previous human skeletal action recognition tasks, there are still unresolved issues:

(1) Graph convolution-based human skeletal action recognition networks are subject to limitations due to the finite receptive field of convolutional operations, resulting in inherent shortcomings in their network potential. Although Transformers can capture global human joint information, their ability to extract discriminative features from local short-term information is weaker compared to convolutional networks. Therefore, addressing how to model a human skeletal action recognition network that can overcome the shortcomings of both convolutional and Transformer networks while inheriting their advantages is crucial in this research field.

(2) During network modeling, fine-grained spatial information of human skeletal structures is more susceptible to interference from high-frequency noise. Current research has not addressed this issue adequately. Additionally, the predominant use of single-scale modeling methods in current research overlooks the complexity and diversity of human movements. Single-scale networks often lose critical information, and the use of multiple large-scale convolutional kernels increases the network parameter count, leading to increased computational burden.

In response to the aforementioned challenges, a Multi-Scale Spatial Temporal Convolution Dynamic Gated Transformer (MSST-CDGT) network is proposed for human skeletal action recognition. The research contributions are outlined as follows:

(1) A multi-scale spatial dynamic gated Transformer is introduced, effectively suppressing high-frequency noise interference in non-spatial dimensions while modeling both local multi-scale and global spatial information of human skeletal structures.

(2) A multi-scale temporal convolution network is designed, efficiently capturing rich features of human skeletal action sequences in the temporal dimension.

(3) A multi-stream fusion network architecture is constructed to enhance the overall performance of the model. Experimental results demonstrate that the MSST-CDGT network outperforms mainstream human skeletal action recognition networks both domestically and internationally.

2. Relevant Theory

2.1 Spatial Temporal Graph Convolutional Action Recognition

The human skeletal structure can be represented as an undirected graph $G=(V,E)$, where $V=\{v_1,v_2,\cdots,v_N\}$ represents the set of N vertices, and E represents the set of edges, formally represented by the adjacency matrix $\boldsymbol{A}\in\mathbb{R}^{N\times N}$. The elements of the adjacency matrix denote the degree of association between nodes in the topological graph. Continuous human skeletal sequence data can be represented by a feature matrix $\boldsymbol{X}\in\mathbb{R}^{C\times T\times N}$, where C represents the coordinates of skeletal joints, T denotes the number of frames in the skeletal sequence, and N represents the number of skeletal data nodes.

Skeletal action recognition based on spatial temporal graph convolution consists of a series of stacked modules, each comprising spatial graph convolution for modeling joint spatial features and temporal convolution for modeling joint temporal features. For modeling spatial joint features, we follow the structure proposed by Kipf et al. [9], and the computational process is illustrated as shown in Equation (1):

$$\boldsymbol{Z}_t=\tilde{\boldsymbol{A}}\boldsymbol{X}_t\boldsymbol{W} \tag{1}$$

Where $\boldsymbol{X}_t \in \mathbb{R}^{N\times C}$ represents the input information of the network at time *t*, and $\boldsymbol{Z}_t \in \mathbb{R}^{N\times C'}$ denotes the output information of the network at time *t*.

For the temporal dimension, convolution operations with a kernel size of $K_t \times 1$ are employed to model the temporal features of the skeletal sequence.

2.2 Transformer and Attention Mechanisms

The Transformer consists of an encoder-decoder architecture, aimed at addressing the shortcomings of recurrent neural networks in handling long sequences and achieving data parallelization. The attention mechanism serves as a core component of the Transformer. Its computational process involves several steps: firstly, the input sequence undergoes three linear transformations to convert it into query vectors (***Q***), key vectors (***K***), and value vectors (***V***); secondly, the dot product of ***Q*** and ***K*** is transformed into an attention matrix with varying degrees of attention for each joint using the $softmax(\cdot)$ function; finally, the dot product of the attention matrix and ***V*** yields the final output sequence. The computational process is illustrated by Equation (2):

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{2}$$

Here, d_k represents the dimensionality of the key vectors, and $\sqrt{d_k}$ represents the scaling factor, which serves to mitigate the issue of gradient explosion.

3. Network Structure

3.1 Overall Structure of the MSST-CDGT Network

The overall structure of the proposed MSST-CDGT network, as illustrated in Figure 1, comprises ten basic blocks. Each basic block consists of multiple units: the Multi-Scale Spatial Dynamic Gated Transformer (MSS-DGT) unit and the Multi-Scale Temporal Convolution (MST-C) unit. The output channels of each basic block are 64, 64, 64, 64, 128, 128, 128, 256, 256, and 256, respectively. Doubling the number of channels corresponds to doubling the stride. The human skeleton information $\boldsymbol{X} \in \mathbb{R}^{C\times T\times N}$ is input into the network and then undergoes positional encoding to obtain $\boldsymbol{X}_{\mathrm{PE}} \in \mathbb{R}^{C\times T\times N}$. Subsequently, it is fed into the Multi-Scale Spatial Dynamic Gated Transformer to model local multiscale and global spatial information of the human skeleton while suppressing high-frequency noise interference. Then, the data passes through the Multi-Scale Temporal Convolution unit to model the temporal sequence information of the human skeleton. The network incorporates residual connections, divided into two branches. These branches fuse the positional encoding embedded skeleton information $\boldsymbol{X}_{\mathrm{PE}}$ with the processed information from the Multi-Scale Spatial Dynamic Gated Transformer unit and the Multi-Scale Temporal Convolution unit, respectively. This fusion effectively delivers $\boldsymbol{X}_{\mathrm{PE}}$ to the deeper layers of the network, thereby enhancing the richness of network information and facilitating more stable gradient propagation. Finally, the output is obtained after passing through the Fully Connected layer (FC) and *Softmax* operation, yielding action recognition results.

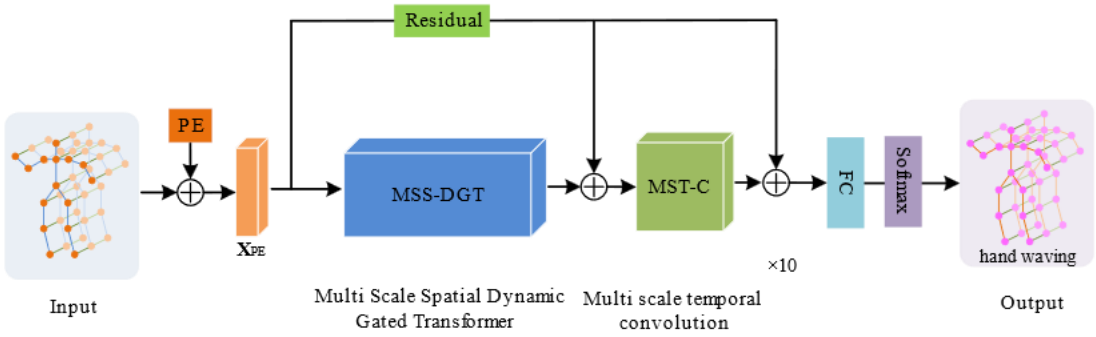

Figure 1: Overall Structure of The MSST-CDGT Network

3.2 Multi-Scale Spatial Dynamic Gated Transformer

In Transformer-based methods for human skeleton action recognition, modeling fine-grained spatial information is more susceptible to the influence of high-frequency noise. Although the attention mechanism can better model global spatial joint dependencies, it lacks the capability to model local spatial joint dependencies. To address this issue in spatial modeling of human skeleton data, we propose a Multi-Scale Spatial Dynamic Gated Transformer (MSS-DGT). MSS-DGT, as depicted in Figure 2, mainly consists of three components: Position Encoding (PE), Dynamic Gated Module (DGM), and Multi-Scale Spatial Transformer (MSS-TR). Each of these components will be elaborated upon below.

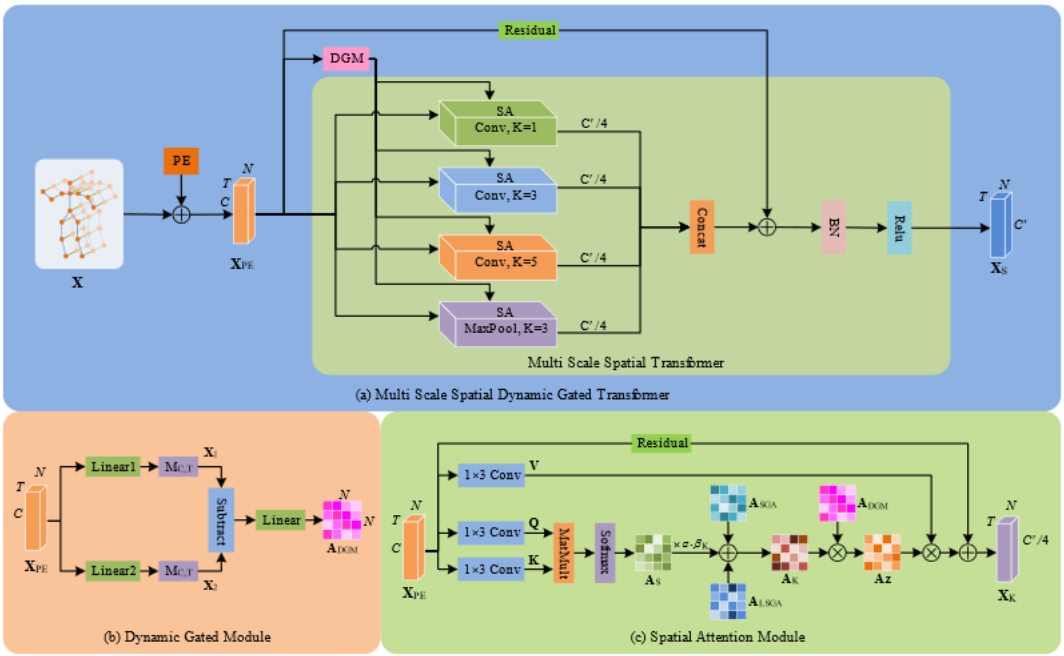

Figure 2: Multi-Scale Spatial Dynamic Gated Transformer Network

3.2.1 Position Encoding

When Transformer processes sequential data in parallel, it overlooks the order of the sequences. For instance, when modeling spatial joint dependencies for the input three-dimensional skeleton sequence $\boldsymbol{X}$, only the spatial positions of each joint are considered. To leverage the order information, a uniform position vector is generated for all frames, and it is overlaid with the original skeleton sequence to allocate joint position information. The computation of the position vector is illustrated in Equation (3), where sinusoidal and cosine functions with different frequencies are utilized for initializing the position vector.

$$\begin{cases} PE(p,2i) = \sin\left(p/10000^{2i/C_{\text{in}}}\right) \\ PE(p,2i+1) = \cos\left(p/10000^{2i/C_{\text{in}}}\right) \end{cases} \tag{3}$$

Where p and i represent the dimensions of the joint position and the position encoding vector, respectively.

3.2.2 Dynamic Gated Module

In the process of network modeling, the quality of human skeleton spatial information modeling directly influences the network performance. Due to the susceptibility to interference from other high-frequency noise when processing fine-grained spatial information, in order to suppress such non-spatial dimension high-frequency noise contained in human skeleton information, inspired by the research of Lim et al. [16] on Gated Residual Networks (GRN), a Dynamic Gated Module (DGM) is proposed and introduced into the spatial modeling sub-network of the proposed action recognition network. This allows the network to learn spatially related information of human skeletons with higher quality, corresponding to the DGM module in the upper left of Figure 2(a). The specific structure of the DGM module is illustrated in Figure 2(b). For the input $\boldsymbol{X}_{\text{PE}}$, after

two different linear transformations followed by average pooling in the temporal and channel dimensions, $\boldsymbol{X}_1 \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{1\times N}$ are obtained, as shown in Equation (4):

$$\begin{cases} \boldsymbol{X}_1 = \text{Mean}_{(\text{T,C})}\left(\text{linear}_1\left(\boldsymbol{X}_{\text{PE}}\right)\right) \\ \boldsymbol{X}_2 = \text{Mean}_{(\text{T,C})}\left(\text{linear}_2\left(\boldsymbol{X}_{\text{PE}}\right)\right) \end{cases} \tag{4}$$

Then, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are element-wise subtracted according to Equation (5), followed by a Linear transformation to obtain their attention matrix $\boldsymbol{A}_{\text{DGM}} \in \mathbb{R}^{N\times N}$.

$$\boldsymbol{A}_{\text{DGM}} = \text{Linear}\left(\text{Subtract}\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right)\right) \tag{5}$$

The proposed dynamic gated module does not completely block the transmission of specific information within the module. Instead, it dynamically processes information in the time and channel dimensions, suppressing the interference of high-frequency noise on the spatial information modeling process. While preserving the contributions of other dimensions to the network, it enhances the modeling quality of spatial dimension information.

3.2.3 Multi-Scale Spatial Transformer

To address the shortcomings of both convolutional and Transformer-based human skeleton action recognition networks, inspired by the characteristics of convolutional operations and the attention mechanism of Transformers, a multi-scale spatial Transformer network is proposed. As illustrated in the green box in Figure 2(a), the input vector $\boldsymbol{X}_{\text{PE}}$, obtained after position encoding, is fed into four spatial attention modules (SA) of different scales. Each SA module involves operations for computing $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, with the parameter K representing the scale size. When K=1, it models global joint dependencies, while K=3 and 5 are used to model local joint dependencies at corresponding scales. Here, we provide a detailed description of the second SA module, as shown in Figure 2(c). Initially, $\boldsymbol{X}_{\text{PE}}$ undergoes three separate 1×3 convolutions to derive $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, with the computation process depicted in Equation (6):

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \text{Conv}_{1\times 3}\left(\boldsymbol{X}_{\text{PE}}\right) \tag{6}$$

After obtaining the query vector $\boldsymbol{Q}$ and the key vector $\boldsymbol{K}$, the spatial attention matrix $\boldsymbol{A}_{\text{S}} \in \mathbb{R}^{N\times N}$ is computed through Equation (7).

$$\boldsymbol{A}_{\text{S}} = Softmax\left(\boldsymbol{Q}\boldsymbol{K}^{\text{T}}\right) \tag{7}$$

The spatial attention matrix $\boldsymbol{A}_{\text{S}}$ represents specific dependencies among the joints within this branch. To encourage the network to learn universal dependencies between joints, local spatial global attention (LSGA) matrices $\boldsymbol{A}_{\text{LSGA}}$ and spatial global attention (SGA) matrices $\boldsymbol{A}_{\text{SGA}}$ are added for each branch. Each $\boldsymbol{A}_{\text{LSGA}}$ within a branch is unique and is used to learn universal dependencies between joints in the corresponding branch. As for $\boldsymbol{A}_{\text{SGA}}$, it is shared among all branches to learn universal dependencies between joints across branches. It is initialized using a normal distribution and optimized along with the network. Finally, the obtained $\boldsymbol{A}_{\text{S}}$, $\boldsymbol{A}_{\text{LSGA}}$, and $\boldsymbol{A}_{\text{SGA}}$ are combined through Equation (8) to yield the attention matrix $\boldsymbol{A}_{\text{K}} \in \mathbb{R}^{N\times N}$ for this branch.

$$\boldsymbol{A}_{\text{K}} = \alpha \cdot \beta_{\text{K}} \cdot \boldsymbol{A}_{\text{S}} + \boldsymbol{A}_{\text{LSGA}} + \boldsymbol{A}_{\text{SGA}} \tag{8}$$

In this equation, α and β_{K} respectively denote scalars shared across branches and specific to each branch, utilized to regulate the strength of $\boldsymbol{A}_{\text{S}}$, optimized alongside the network.

For the attention matrix $\boldsymbol{A}_{\text{K}}$, it is multiplied with the matrix $\boldsymbol{A}_{\text{DGM}}$ obtained from the dynamic gated module to achieve dynamic correction of the attention matrix $\boldsymbol{A}_{\text{K}}$. This yields the corrected attention matrix $\boldsymbol{A}_{\text{Z}} \in \mathbb{R}^{N\times N}$ belonging to this branch. Afterwards, it is multiplied with the value vector $\boldsymbol{V}$, and then added to $\boldsymbol{X}_{\text{PE}}$ for fusion, resulting in the final output $\boldsymbol{X}_{\text{K}} \in \mathbb{R}^{C'/4\times T\times N}$ of this spatial attention module. The computation process is illustrated as equation (9):

$$\boldsymbol{X}_{\text{K}} = \boldsymbol{A}_{\text{K}} \cdot \boldsymbol{A}_{\text{DGM}} \cdot \boldsymbol{V} + \boldsymbol{X}_{\text{PE}} \tag{9}$$

In summary, the multi-scale spatial dynamic gated Transformer encodes the original human skeleton sequence $\boldsymbol{X}$ into $\boldsymbol{X}_{\text{PE}}$ through position encoding. Subsequently, it is inputted into both the dynamic gated module and the multi-scale spatial Transformer for learning skeleton spatial features. The attention matrices from the dynamic

gated module and the multi-scale spatial attention module are fused, followed by concatenation of the output features from the four different scales of spatial attention modules. Then, they are fused with $\boldsymbol{X}_{\mathrm{PE}}$ through addition. After batch normalization (BN) and *ReLU* activation functions, the final output $\boldsymbol{X}_{\mathrm{S}} \in \mathbb{R}^{C' \times T \times N}$ of the multi-scale spatial dynamic gated Transformer is obtained.

3.3 Multi-Scale Temporal Convolution

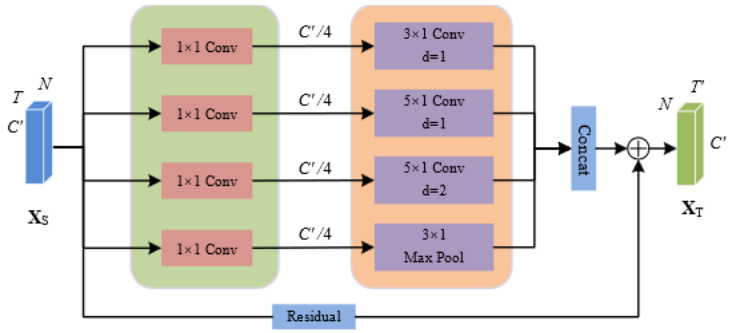

Figure 3: Multi-Scale Temporal Convolution Network

In the task of human skeleton action recognition, different actions performed by individuals correspond to sequences of different lengths. To enable the network to finely learn the temporal sequence features of different actions, a multi-scale temporal convolution (MST-C) network is designed. As illustrated in Figure 3, MST-C comprises one branch of 3×1 convolution, two branches of 5×1 convolution with different dilation factors, and one branch of 3×1 max pooling. Each branch undergoes channel division using 1×1 convolution. The computation process is depicted in Equation (10):

$$\begin{cases} \boldsymbol{X}_{\mathrm{T}}^{1} = f_{3\times1}^{d=1}\left(f_{1\times1}\left(\boldsymbol{X}_{\mathrm{S}}\right)\right) \\ \boldsymbol{X}_{\mathrm{T}}^{2} = f_{5\times1}^{d=1}\left(f_{1\times1}\left(\boldsymbol{X}_{\mathrm{S}}\right)\right) \\ \boldsymbol{X}_{\mathrm{T}}^{3} = f_{5\times1}^{d=2}\left(f_{1\times1}\left(\boldsymbol{X}_{\mathrm{S}}\right)\right) \\ \boldsymbol{X}_{\mathrm{T}}^{4} = \mathrm{MaxPool}_{3\times1}\left(f_{1\times1}\left(\boldsymbol{X}_{\mathrm{S}}\right)\right) \end{cases} \tag{10}$$

Where $f_{3\times1}^{d=1}(\cdot)$, $f_{5\times1}^{d=1}(\cdot)$, and $f_{5\times1}^{d=2}(\cdot)$ represent convolution operations with kernel sizes of 3, 5, and 5, respectively, and dilation factors of 1, 1, and 2. $f_{1\times1}(\cdot)$ denotes standard 1×1 convolution operation, and the final branch employs 3×1 MaxPool operation to extract salient information along the temporal dimension. The final feature $\tilde{\boldsymbol{X}}_{\mathrm{T}} \in \mathbb{R}^{C' \times T' \times N}$, with different receptive fields, is obtained by concatenating the results from the four branches, as depicted in Equation (11):

$$\tilde{\boldsymbol{X}}_{\mathrm{T}} = \mathrm{Concat}\left(\boldsymbol{X}_{\mathrm{T}}^{1}, \boldsymbol{X}_{\mathrm{T}}^{2}, \boldsymbol{X}_{\mathrm{T}}^{3}, \boldsymbol{X}_{\mathrm{T}}^{4}\right) \tag{11}$$

Finally, the residual connection is introduced to merge $\boldsymbol{X}_{\mathrm{S}}$ with $\tilde{\boldsymbol{X}}_{\mathrm{T}}$, resulting in the final output $\boldsymbol{X}_{\mathrm{T}} \in \mathbb{R}^{C' \times T' \times N}$ of the MST-C network. The residual connection preserves the original features of the temporal frames while better propagating gradient information, thereby enhancing network stability. Multi-scale temporal convolution reduces network parameter count while increasing the network's receptive field, allowing for richer human skeletal information to be captured in the temporal dimension.

3.4 Multi Stream Fusion Architecture

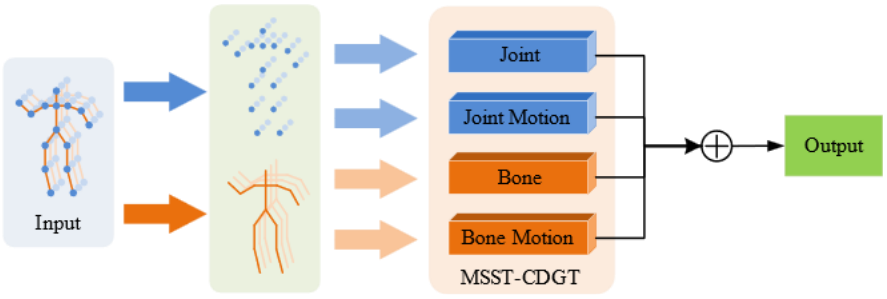

Figure 4: Multi Stream Data Fusion Network

In the task of human skeletal action recognition, relying solely on individual human joint data may not fully exploit the network's performance potential. The skeletal and motion information associated with human joint data are complementary in nature, and effectively combining them is advantageous for action recognition. Therefore, a framework comprising four data streams, as illustrated in Figure 4, is constructed to form a multi-stream fusion network architecture. The first data stream utilizes the original skeletal joint coordinates as input, referred to as the joint stream. The second data stream represents the joint motion stream, depicting the coordinate differences between adjacent frames in time. The third data stream employs the differences in spatial coordinates of skeletal joints as input, known as the bone stream. The fourth data stream utilizes the bone motion stream, representing the disparities in skeletal information between adjacent frames in time. Subsequently, the softmax scores of multiple streams are weighted and fused to obtain the final result.

4. Experimental Results and Analysis

4.1 Datasets

To validate and analyze the performance of the MSST-CDGT network, experiments were conducted on the NTU-RGB+D 60 [17] and NTU-RGB+D 120 [18] datasets. NTU-RGB+D 60 is one of the commonly used large-scale datasets in human action recognition research, comprising 56,880 samples and over 4 million frames of data. It includes two different evaluation benchmark protocols: Cross-Subject (X-Sub) and Cross-View (X-View). In the Cross-Subject benchmark protocol, 40 individuals participating in the test were evenly divided into training and testing groups, with 60 action classes executed. In the Cross-View benchmark protocol, data captured by three cameras were used, where data from camera 1 were used for testing, and data from cameras 2 and 3 were used for training. NTU-RGB+D 120 is an extended and upgraded version of NTU-RGB+D 60, containing 120 categories and 114,480 samples. Slightly differently, the evaluation benchmark protocols are Cross-Subject (X-Sub) and Cross-Setup (X-Set). For the Cross-Subject benchmark protocol, 106 individuals participating in the test were evenly divided into training and testing groups. For the Cross-Setup benchmark protocol, 32 different setups were completed by changing the positions and backgrounds, with corresponding numbers assigned. Samples with even numbers were used for training, and samples with odd numbers were used for testing.

4.2 Experimental Setup

The experimental setup utilized an Intel(R) Xeon(R) Gold 6348 CPU, 2 NVIDIA GeForce RTX 3080 Ti GPUs, running on a Windows 10 Professional Workstation Edition. PyTorch deep learning framework was employed for the experiments, programmed in Python, and utilizing CUDA and cuDNN libraries. Each behavioral sample for the experiment was uniformly sampled with a sequence length of 100 for the skeleton input. A warm-up strategy was implemented for the first 5 epochs of the experiment. The optimization strategy utilized a stochastic gradient descent (SGD) algorithm with a weight decay of 0.0004 and Nesterov momentum of 0.9. The batch size was set to 64, with a total of 85 training epochs. The initial learning rate was set to 0.1, and it was reduced to one-tenth of its value at epochs 35, 55, and 75. The evaluation metric used was the action recognition accuracy (Acc), calculated as shown in Equation (12), where N_{corr} and N_{total} represent the number of correctly predicted instances and the total number of instances, respectively.

$$f = N_{\text{corr}} / N_{\text{total}} \tag{12}$$

4.3 Ablation Experiments

To validate and analyze the performance of various aspects of the MSST-CDGT network, comprehensive ablation experiments were conducted on the NTU-RGB+D 60 dataset using joint data under the cross-subject (X-Sub) benchmark protocol.

4.3.1 Impact of Different Spatial and Temporal Scales on the Network

To verify the impact of different spatial and temporal scales on the network, experiments were conducted on the MSST-CDGT network under various spatial temporal scales to analyze its accuracy, network parameter count, and computational load. Observing Table 1, it was found that the network, with multi-scale spatial temporal modeling, achieved the highest accuracy of 90.7% while maintaining relatively low network parameter count and computational load. This indicates that the proposed multi-scale spatial temporal convolution dynamic gated Transformer network can effectively capture features at different scales, learn rich contextual action information, enhance network robustness, and better adapt to noise variations at different scales, thereby ultimately improving overall network performance.

Table 1: Performance of Networks at Different Scales

Networks of Different Scales	Params (M)	FLOPs (G)	Acc (%)
Single-scale spatial temporal network	3.04	3.30	90.2
Single-scale spatial network	1.24	1.39	90.2
Single-scale temporal network	3.69	4.13	90.3
Multi-scale spatial temporal network	1.90	2.22	90.7

4.3.2 Impact of Dynamic Gated Module on the Network

To validate the impact of the proposed dynamic gated module on network performance, experiments were conducted by integrating the dynamic gated module into the constructed multi-scale spatial transformer network. Observing Table 2, it was found that the multi-scale spatial dynamic gated transformer achieved a 1% increase in recognition accuracy compared to the multi-scale spatial transformer network, with only an additional parameter count of 0.08M. This indicates a significant improvement in network performance due to the dynamic gated module. The reason lies in its ability to suppress high-frequency noise in other dimensions during the process of human skeleton spatial modeling, allowing the network to focus more on learning spatial information and thus achieve higher-quality modeling of human skeleton spatial information.

Table 2: Performance of Different Networks

Method	Params (M)	Acc (%)
Multi-Scale Spatial Transformer	1.82	89.7
Multi-Scale Spatial Dynamic Gated Transformer	1.90	90.7

4.3.3 Experiments on the Validity of Individual Modules

To validate the effectiveness of each module in the proposed MSST-CDGT network, experiments were conducted by individually removing the position encoding module, dynamic gated module, global attention module, local attention module, and spatial attention module to assess their respective impacts on the overall network performance. The experimental results are shown in Table 3, where "√" indicates the inclusion of the corresponding module in the network, and "×" indicates the removal of the module from the network.

Observing Table 3, it was found that the network incorporating all modules (Network 5) achieved an accuracy of 90.7%. When the position encoding module was removed from Network 1, the accuracy decreased by 0.4% compared to Network 5, indicating that encoding joint sequence information can effectively enhance network performance. Removing the global attention module from Network 2 resulted in a decrease in accuracy by 0.8% compared to Network 5, demonstrating that the global attention module effectively models global joint dependency relationships, thereby improving overall network performance. Removing the local attention module from Network 3 led to a decrease in accuracy by 0.2% compared to Network 5, indicating that the local

attention module enhances the network's ability to learn multi-scale local details and effectively improves overall network performance. When the spatial attention module was removed from Network 4, the accuracy decreased by 0.7% compared to Network 5, highlighting the effective contribution of the spatial attention module to overall network performance.

Table 3: Effectiveness Experiments of Each Module in the MSST-CDGT Network

Networks	Position Encoding	Global Attention	Local Attention	Spatial Attention	Acc (%)
1	×	√	√	√	90.3
2	√	×	√	√	89.9
3	√	√	×	√	90.5
4	√	√	√	×	90.0
5	√	√	√	√	90.7

4.4 Multi Stream Data Fusion Experiment

To validate the effectiveness of the constructed multi-stream data fusion framework, experiments were conducted on the NTU-RGB+D 60 dataset and NTU-RGB+D 120 dataset, comparing the performance of network ensembles trained with different data fusion approaches. Multiple data fusion comparative experiments were performed on the proposed MSST-CDGT network: the first using joint stream, the second using bone stream, the third employing a dual-stream fusion (2S) combining joint and bone streams, and the fourth incorporating a four-stream (4S) fusion including joint, bone, joint motion, and bone motion streams. As shown in Table 4, with an increase in the number of fused data streams, the overall performance of the network improved accordingly. This indicates that the constructed multi-stream data fusion framework enhances the diversity of feature learning in the network, effectively boosting its overall performance.

Table 4: Multi Stream Data Fusion Experiments on the NTU-RGB+D 60 and NTU-RGB+D 120 Datasets

Data Types	NTU-RGB+D 60		NTU-RGB+D 120	
	X-Sub (%)	X-View (%)	X-Sub (%)	X-Set (%)
Joint	90.7	95.7	85.0	86.8
Bone	90.5	95.0	86.8	87.9
2S	92.2	96.6	89.1	90.4
4S	92.6	96.8	89.5	90.7

4.5 Visualization of Experimental Analysis

4.5.1 Visualization of the Attention Weighting Matrix

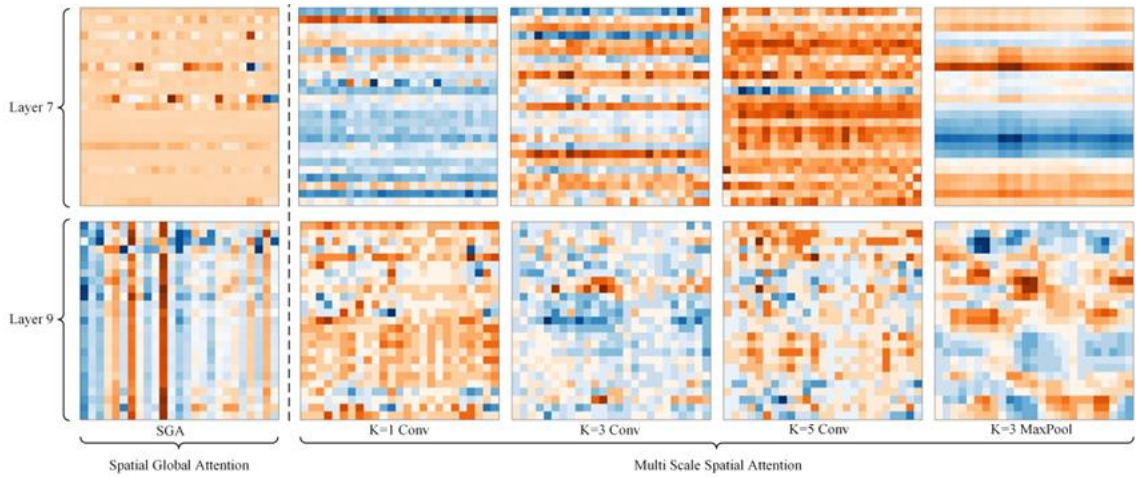

Figure 5: Attention Weight Matrix Visualization

To validate the effectiveness of the proposed MSST-CDGT network in modeling local multiscale and global information, visual experiments were conducted to analyze the attention weight matrices in the network. This aimed to intuitively demonstrate the degree of attention that the network's attention mechanism pays to each

human joint during its operation. Taking the "phone call" action as an example, visualizations were performed on the spatial global attention and multiscale spatial attention matrices at the 7th and 9th layers of the network during action recognition. As shown in Figure 5, the origin of each attention matrix is located at the top left corner, with the vertical axis representing different human joints following the scheme proposed by Shahroudy et al. [17], and the horizontal axis representing consecutive time frames. The attention matrices are mainly colored in orange and blue, where deeper shades of orange indicate higher attention to the corresponding joint, while deeper shades of blue indicate lower attention. Observation of Figure 5 reveals that the spatial global attention matrices exhibit consistent coloration among different joints at the same time frame. This is because spatial global attention allows the network to consider dependencies with all other joints when computing information for a specific joint, enhancing the global modeling of the human skeleton and facilitating better recognition of various actions. Additionally, observation of the multiscale spatial attention matrices shows varying degrees of color change for different human joints across different scales. This is because the attention mechanism allocates different degrees of attention to each human joint at different scales, focusing more on joints with larger variations to accurately identify differences between various human actions. The experiments demonstrate the effectiveness of multiscale spatial attention in the MSST-CDGT network for different-scale human joints and the effectiveness of spatial global attention in globally modeling dependencies among all human joints.

4.5.2 Human Action Recognition Visualization

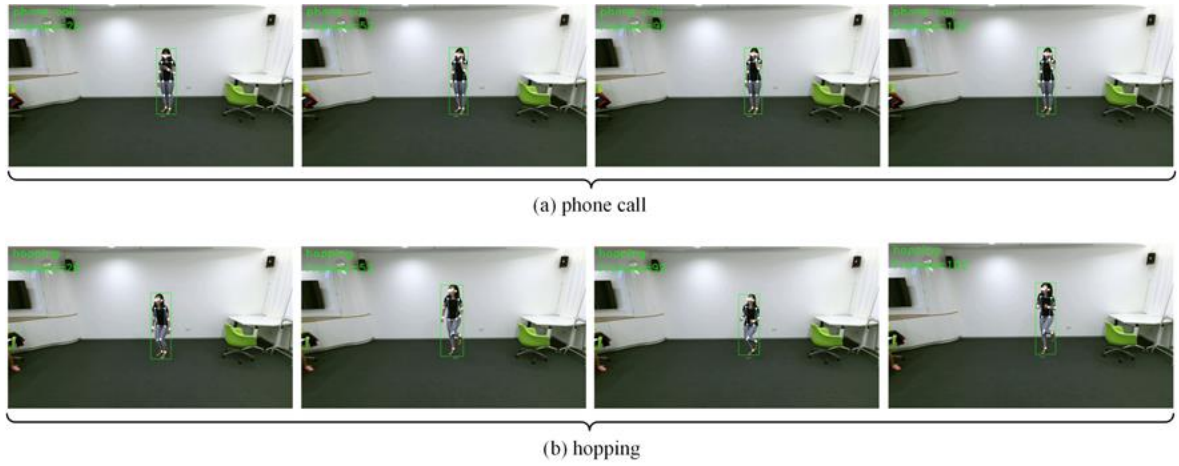

Figure 6: Experiment Video Frames for Human Action Recognition

To validate the practical effectiveness of the proposed MSST-CDGT network in action recognition, visual experiments were conducted for human action recognition. Firstly, a human pose estimation network [19] was utilized to extract human joint data from videos in COCO format [20], thus obtaining complete human skeleton information from the videos. Subsequently, the obtained human skeleton information was fed into the MSST-CDGT network for action recognition. Randomly selected videos of two actions, "phone call" and "hopping", were chosen for the human action recognition visual experiment, and the visual experiment results were provided. Observing the video frames extracted from the action recognition visual experiment in Figure 6, both actions, "phone call" in Figure 6(a) and "hopping" in Figure 6(b), were correctly recognized, with the recognition results displayed in the top left corner of the videos. The visual experiment results for human action recognition further validate the effectiveness of the MSST-CDGT network.

4.6 Comparisons to the State-of-the-Art

To validate the performance of the proposed MSST-CDGT network, comparisons were made with state-of-the-art networks on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets. The selected comparison networks were advanced networks corresponding to the respective benchmarks. Observing Table 5, it can be noted that on the cross-subject X-Sub benchmark protocol of the NTU-RGB+D 60 dataset, the MSST-CDGT network achieved an accuracy of 92.6%, while on the cross-view X-View benchmark protocol, it achieved an accuracy of 96.8%. These experimental results were respectively 2.3% and 0.5% higher than the ST-TR [14] network and 1.1% and

0.4% higher than the DSTA-Net [15] network on the two benchmark protocols. Additionally, compared to the graph convolution-based Efficient GCN [30] network, the proposed network achieved higher accuracies of 0.5% and 0.7%. Furthermore, to further validate the generalization ability of the MSST-CDGT network, experiments were conducted on the larger-scale NTU-RGB+D 120 dataset. As shown in Table 5, the MSST-CDGT network achieved accuracies of 89.5% and 90.7% on the cross-subject X-Sub and cross-setup X-Set benchmark protocols, respectively. These experimental comparison results were all higher than those of mainstream human skeletal action recognition networks in recent years.

Table 5: Comparative Experiments with State-of-the-Art Networks on the NTU-RGB+D 60 and NTU-RGB+D 120 Datasets

Method	NTU-RGB+D 60		NTU-RGB+D 120		Year
	X-Sub (%)	X- View (%)	X-Sub (%)	X-Set (%)	
GAT [21]	89.0	95.2	84.0	86.1	2023
ST-TR [14]	90.3	96.3	85.1	87.1	2021
MADT-GCN [22]	90.4	96.5	86.5	88.2	2024
TA-CNN [23]	90.7	95.1	85.7	87.3	2022
LKA-GCN [24]	90.7	96.1	86.3	87.8	2023
Shift-GCN [13]	90.7	96.5	85.9	87.6	2020
STF-Net [25]	91.1	96.5	86.5	88.2	2023
Hybrid-Net [26]	91.4	96.9	87.5	89.0	2023
DSTA-Net [15]	91.5	96.4	86.6	89.0	2020
MST-GCN [27]	91.5	96.6	87.5	88.8	2021
JPA-DESTGCN [28]	91.6	96.9	87.5	88.5	2024
Dual Head-Net [29]	92.0	96.6	88.2	89.3	2021
Efficient-GCN [30]	92.1	96.1	88.7	88.9	2022
MSST-CDGT	**92.6**	**96.8**	**89.5**	**90.7**	**2024**

5. Conclusion

The study introduces a multi-scale spatial temporal convolution dynamic gated Transformer network for human skeleton action recognition. It comprises multi-scale spatial dynamic gated Transformer units and multi-scale temporal convolution units. The multi-scale spatial dynamic gated Transformer unit enables global and local multi-scale spatial modeling of human skeleton information and utilizes dynamic gated modules to suppress high-frequency noise interference in the network. The multi-scale temporal convolution unit efficiently extracts rich skeleton temporal sequence features in the time dimension. Finally, through a multi-stream fusion architecture, the network weights and fuses the data of multiple skeleton streams, enhancing the overall performance of the network and producing the final output. Through comprehensive experiments, the effectiveness of the MSST-CDGT network has been verified, surpassing current mainstream human skeleton action recognition networks in overall performance. Future work will focus on integrating more modalities of information to further improve the network's recognition accuracy and robustness.

References

[1]. Gupta, S. C., Kumar, D., & Athavale, V. (2021, June). A review on human action recognition approaches. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) (pp. 338-344). IEEE.

[2]. Sathya, R., Mythili, M., Ananthi, S., Asitha, R., Vardhini, V. N., & Shivaani, M. (2023, December). Intelligent Video Surveillance System for Real Time Effective Human Action Recognition using Deep Learning Techniques. In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1826-1831). IEEE.

[3]. Yang, C. L., Hsu, S. C., Hsu, Y. W., & Kang, Y. C. (2023). HAR-time: Human action recognition with time factor analysis on worker operating time. International Journal of Computer Integrated Manufacturing, 36(8), 1219-1237.

[4]. Liang, K., Wang, J., & Bhalerao, A. (2022, October). Lane Change Classification and Prediction with Action Recognition Networks. In European Conference on Computer Vision (pp. 617-632). Cham: Springer Nature Switzerland.
[5]. Zheng, L., Sun, Y., Zeng, J., Liu, Y., & Zhou, L. (2023, October). Stroke Rehabilitation Action Recognition Based on MY_SlowFast. In 2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP) (pp. 124-128). IEEE.
[6]. Hongwei, W. (2023, June). Real-time Swimming Posture Image Correction Framework based on Novel Visual Action Recognition Algorithm. In 2023 8th International Conference on Communication and Electronics Systems (ICCES) (pp. 1714-1718). IEEE.
[7]. Yang, X., & Tian, Y. L. (2012, June). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops (pp. 14-19). IEEE.
[8]. Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of the IEEE international conference on computer vision (pp. 3551-3558).
[9]. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
[10]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[11]. Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
[12]. Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12026-12035).
[13]. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 183-192).
[14]. Plizzari, C., Cannici, M., & Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding, 208, 103219.
[15]. Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In Proceedings of the Asian conference on computer vision.
[16]. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764.
[17]. Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019).
[18]. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence, 42(10), 2684-2701.
[19]. Dubey, S., & Dixit, M. (2023). A comprehensive survey on human pose estimation approaches. Multimedia Systems, 29(1), 167-195.
[20]. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.
[21]. Zhang, J., Xie, W., Wang, C., Tu, R., & Tu, Z. (2023). Graph-aware transformer for skeleton-based action recognition. The Visual Computer, 39(10), 4501-4512.
[22]. Xia, Y., Gao, Q., Wu, W., & Cao, Y. (2024). Skeleton-based action recognition based on multidimensional adaptive dynamic temporal graph convolutional network. Engineering Applications of Artificial Intelligence, 127, 107210.

[23]. Xu, K., Ye, F., Zhong, Q., & Xie, D. (2022, June). Topology-aware convolutional neural network for efficient skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 3, pp. 2866-2874).
[24]. Liu, Y., Zhang, H., Li, Y., He, K., & Xu, D. (2023). Skeleton-based human action recognition via large-kernel attention graph convolutional network. IEEE Transactions on Visualization and Computer Graphics, 29(5), 2575-2585.
[25]. Wu, L., Zhang, C., & Zou, Y. (2023). SpatioTemporal focus for skeleton-based action recognition. Pattern Recognition, 136, 109231.
[26]. Yang, W., Zhang, J., Cai, J., & Xu, Z. (2023). HybridNet: Integrating GCN and CNN for skeleton-based action recognition. Applied Intelligence, 53(1), 574-585.
[27]. Chen, Z., Li, S., Yang, B., Li, Q., & Liu, H. (2021, May). Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 2, pp. 1113-1122).
[28]. Lu, J., Huang, T., Zhao, B., Chen, X., Zhou, J., & Zhang, K. (2024). Dual Excitation Spatial-temporal Graph Convolution Network for Skeleton-Based Action Recognition. IEEE Sensors Journal.
[29]. Chen, T., Zhou, D., Wang, J., Wang, S., Guan, Y., He, X., & Ding, E. (2021, October). Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In Proceedings of the 29th ACM international conference on multimedia (pp. 4334-4342).
[30]. Song, Y. F., Zhang, Z., Shan, C., & Wang, L. (2022). Constructing stronger and faster baselines for skeleton-based action recognition. IEEE transactions on pattern analysis and machine intelligence, 45(2), 1474-1488.