Eng Int Syst (2026) 5: 647–652
© 2026 CRL Publishing Ltd

Engineering
Intelligent Systems

Recognition of Movement Features of Different Ethnic Dances: Research on Motion Image Information

Fan Yang*

Yan'an University, Yan'an, Shaanxi 716000, China

The inheritance and teaching of ethnic dances can be innovated by introducing digital technology. This paper studied the recognition of movement features of different ethnic dances. It optimized the spatiotemporal graph convolutional network (ST-GCN) based on skeleton information in motion images, developed a multi-stream ST-GCN (MS-STGCN) model, and conducted experiments using NTU-RGB+D 60/120 datasets and a self-constructed ethnic dance dataset. The results showed that for the NTU-RGB+D 60 dataset, the accuracy of the MS-STGCN model was 92.6% (X-Sub) and 96.8% (X-View); for the NTU-RGB+D 120 dataset, its accuracy was 89.7% (X-Sub) and 89.9% (X-Set), outperforming the other algorithms; it achieved an accuracy of 90.6% for the ethnic dance dataset. The findings demonstrate the reliability of the MS-STGCN model in recognizing the features of ethnic dance movements, which can be further applied in practice.

Keywords: spatiotemporal graph convolutional network; ethnic dance; motion image; recognition

1. INTRODUCTION

Ethnic dance is a very important aspect of culture (Deng, 2022). During the long process of development and inheritance, ethnic dance has mostly been passed down and promulgated through oral instruction. The methods used to record ethnic dance are relatively simple, although there are certain difficulties in teaching them to the next generation. The introduction of digital technology can enable the digital archiving of ethnic culture, provide assistance for the standardized teaching of ethnic dance movements, and support the artistic innovation of ethnic dance (Dong et al., 2024). With the advancement and maturation of technologies such as sensors and computer vision (Liu et al., 2024), movement feature recognition based on motion image information has attracted increasingly extensive research (Lian et al., 2025). Dong et al. (2023) collected human motion images from athletes, extracted and fused features, and used an improved deep neural network as a classifier. They discovered that this method had a high recognition rate and an overhead time of only 1.2 seconds for human motion images. Sun et al. (2021) designed a wearable device to obtain electromyographic signals and joint angle signals of the human body, identified four human gait movement patterns through the k-nearest neighbor algorithm, and obtained an overall accuracy of 91.45%. Jiu et al. (2024) developed a multi-sensor network fusion and conditional random field method, which achieved good results on Human 3.6M. Xue et al. (2025) collected surface electromyography signals, extracted features using multivariate autoregressive, and then used gradient boosting decision tree for recognition. They discovered that the method achieved an accuracy of 92.78% in recognizing

*Corresponding address: No. 580, Shrine Road, Baota District, Yan'an, Shaanxi 716000, China. Email: yfanfy@outlook.com

different human in-hand motion. Ethnic dance movements contain long rhythmic cycles, and the differences between the movements can be significant or minor; therefore, higher requirements are placed on recognition algorithms. Based on a spatiotemporal graph convolutional network, this current study designed a movement feature recognition algorithm(ST-GCN) by extracting skeleton information from motion images of different ethnic dances, and by conducting experiments on different datasets, verified that the proposed approach is effective, providing theoretical support for the archiving, teaching, and creation of ethnic dances.

2. MOVEMENT FEATURE RECOGNITION BASED ON A SPATIOTEMPORAL GRAPH CONVOLUTIONAL NETWORK

2.1 Human Skeleton Information

There are challenges such as environmental interference and differences in human appearance in the field of human motion recognition. To address them, researchers have conducted extensive studies on data input modalities, and skeletal information is one of the advantageous methods. Skeleton information refers to the coordinates of multiple skeletal joints in space and time, which can be obtained from two-dimensional motion images using computational vision methods, in addition to being collected by sensors. Compared with information such as red green blue (RGB) and optical flow, skeleton information is insensitive to background, appearance, and noise (Lu et al., 2025). It has a small data dimension, reducing computational and storage overhead (Bian et al., 2024). Moreover, skeleton information is a natural graph structure data, which is very suitable for processing using graph neural networks and has been extensively applied to movement feature recognition.

The costumes and props for ethnic dances are often complex. Skeleton information can effectively filter out these appearance noises, so that the spatiotemporal movements of the body can be focused on. The acquisition and processing of skeleton information are convenient. Therefore, this study determined the movement features of ethnic dances by extracting skeleton information from motion images.

2.2 Spatiotemporal Graph Convolutional Network

ST-GAN (Tang et al., 2024) defines joint points in the skeleton information as graph nodes, written as $V = \{v_{ti} | t = 1, 2, \ldots, T, i = 1, 2, \ldots N\}$, where v_{ti} is the i-th joint at the t-th frame. The skeletons connecting two joint points are defined as the edges of the graph. The edge set includes two parts, E_S and E_F. $E_S = \{v_{ti} v_{tj} | (i, j) \in H\}$, which is the joint point connection within each frame (Hrepresents the joint set of natural connections). $E_F = \{v_{ti} v_{(t+1)i}\}$, which is the connection of the same joint point in consecutive frames, meaning the time edge. After batch normalization of the skeleton data, the human movement features are extracted from the spatial and temporal dimensions through nine ST-GCN units, the temporal and spatial dimension information is compressed through the global average pooling layer and, finally, the recognition results of the movement features are obtained through the fully connected layer and the Softmax layer. Figure 1 shows the structure of the ST-GCN.

2.3 Multi-Stream, Spatiotemporal Graph Convolutional Network

As the complexity of human movement feature recognition increases, the limitations of ST-GCN gradually become apparent. The fixed adjacency matrix used by ST-GCN can capture only the local features of adjacent joints. The input of a single-joint information flow enables it to intelligently extract joint coordinate features, but it cannot mine high-order features and limits the model's recognition of complex movements. In order to achieve better performance in movement feature recognition for different ethnic dances, this paper designs a multi-stream ST-GCN (MS-STGCN) model and improves the GCN module by using a global adaptive spatial graph convolutional network (GAS-GCN) instead of the original GCN to extract skeleton features, breaking through the limitations of the fixed topology.

The adjacency matrix of the GAS-GCN model consists of the following three parts.

(1) Skeleton physical connection matrix A_k

A_k is consistent with the adjacency matrix defined in ST-GCN. It reflects the stability of the physical connection relationship of the skeletal joints during motion, as follows:

$$A_k(i, j) = \begin{cases} 1, & \text{joint } v_i \text{ and } v_j \text{ has physical connection} \\ 0, & \text{joint } v_i \text{ and } v_j \text{ has no physical connection} \end{cases}.$$

(2) Global enhanced mask matrix B_k

B_k is a learnable common association matrix, essentially a mask matrix for augmented data. B_k is a trainable parameter. By training all sample data, the common weight distribution of the global skeleton features can be obtained, which is continuously updated through backpropagation.

(3) Spatially adaptive adjacency matrix C_k

C_k can adaptively learn the degree of similarity between the joints in the current sample skeleton diagram, thereby measuring the connection strength of each joint:

$$C_{ij} = \frac{e^{\theta(v_i)^T \varphi(v_j)}}{\sum_{j=1}^{N} e^{\theta(v_i)^T \varphi(v_j)}},$$

where $\theta(\cdot)$ and $\varphi(\cdot)$ are learnable linear transformation functions, v_i and v_j are feature vectors of joints i and j.

The output of the n-th layer of GAS-GCN can be written as:

$$f_{out} = \sum_{k=1}^{K} (\lambda A_k + B_k + C_k) f_{in} W_k + Res(f_{in}),$$

where λ is a dynamic proportional coefficient that decreases gradually with training rounds, f_{in} is an input feature, W_k

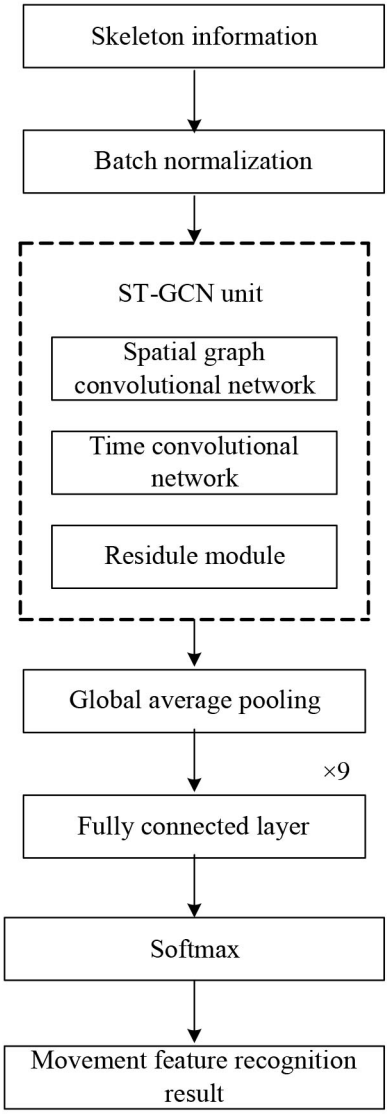


Figure 1 ST-GCN structure.

is a spatial convolution kernel weight, K is the number of convolution kernels, and $Res(\cdot)$ is residual connection, $Res(f_{in}) = ReLU(BN(Conv_{1\times1}(f_{in})))$.

MS-STGCN uses multi-stream information as input because single-stream joint point coordinate data is insufficient. The multi-stream information is:

(1) Joint and bone stream information

It is the three-dimensional coordinates of the joint point. The information of the i-th joint point at the t-th frame can be written as:

$$J_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t}), \quad \forall i \in N, \quad t \in T.$$

The skeleton is represented by vectors of the joint points at both ends. Based on the distance from the joint points to the skeleton'center-of-gravity, the farther points are used as the target points, and the closer points are used as the source points. Then, the skeleton is represented by vectors of the source points pointing to the joint points. The information of skeleton B_i at the t-th frame can be written as:

$$B_{i,t} = (x_{j,t} - x_{i,t}, y_{j,t} - y_{i,t}, z_{j,t} - z_{i,t}), \quad \forall i \in N, \forall t \in T.$$

(2) Joint and skeleton motion stream information

Motion change is indicated by the differences between different frames, i.e., the motion information. The motion stream information of joint point i and skeleton B_i at the t-th frame is:

$$M_{J_{i,t}} = J_{i,t+1} - J_{i,t}, \quad \forall i \in N, \forall t \in T,$$
$$M_{B_i,t} = B_{i,t+1} - B_{i,t}, \quad \forall i \in N, \forall t \in T.$$

(3) Joint and skeleton relative position stream information

The relative position information of the joint to the skeleton is represented by the coordinate difference of adjacent joint point/skeleton based on the sequence of joint point numbers, as follows:

$$RP_{J_{i,t}} = J_{i+1,t} - J_{i,t}, \quad \forall i \in N, \forall t \in T,$$
$$RP_{B_{i,t}} = B_{i+1,t} - B_{i,t}, \quad \forall i \in N, \forall t \in T.$$

Table 1 Experimental environment configuration.

	Parameter
Processor	Intel(R)Xeon(R)Gold 6226R CPU@2.90GHz
Memory	32 GB
Operating system	Ubuntu 16.04
Programming language	Python 3.7
Development platform	PyCharm2024
Deep learning framework	PyTorch 1.7.1

Table 2 Parameter settings.

	Value
Number of training rounds	50
Batch size	32
Learning rate	0.1
Weight decay coefficient	0.0001
Dropout	0.25
Loss function	Cross-entropy loss function

3. RESULTS AND ANALYSIS

3.1 Experimental Settings

Table 1 shows the setting of the experimental environment. MS-STGCN was optimized using stochastic gradient descent (SGD). Parameter settings are displayed in Table 2.

3.2 Experimental Dataset

(1) NTU-RGB+D 60 (Shahroudy et al., 2016)

This dataset was constructed using Kinect v2. Forty dancers performed 60 types of daily movements, and 56,880 video samples were collected. Two criteria were used for evaluation: cross-subject (X-Sub), with 20 performers for training and another 20 performers for testing; cross-view (X-View), taking samples from cameras 2 and 3 for training and samples from camera 1 for testing.

(2) NTU-RGB+D 120 (Liu et al., 2020)

This is an extended version of the NTU-RGB+D 60. 106 performers performed 120 types of everyday actions, collected 114,480 video samples. Two criteria were used for evaluation: cross-subject (X-Sub), dividing performers into two groups, one for training and the other for testing; cross-setting (X-Set), dividing according to the location of the camera setup, and ensuring that the samples for training and testing come from different camera locations.

(3) Ethnic dance dataset

Videos were shot in the laboratory to build the ethnic dance dataset. There were 1,800 videos, including six ethnic dances.

① Tibetan dance: It is characterized mainly by quivering knee joints. Dancers move their hands and feet in the same direction, circle clockwise, and manipulate their sleeves through movements such as coiling, lifting, and flinging, with coiling as the core technique.

② Dai dance: It is characterized by subtlety and gracefulness. The dance has graceful curves and slow rhythm. Dancers move the whole body through knee undulation.

③ Yi dance: The basic movements consist of turning over while hooking the feet and leaping side to side. The footwork is complex, involving swings, extensions, and kicks.

④ Uyghur dance: There are many small movements of the head, wrists, and feet, and the second half of the dance often features rapid spins in place.

⑤ Mongolian dance: It is characterized by shoulder shaking, and hand movements during dance are agile and flexible.

⑥ Lisu dance: It has many life-like movements, such as sowing and weeding. The main part of the dance is kicks, stomps, steps, and jumps of the lower limbs, and the arms swing naturally.

For each type of dance, 300 videos were collected and divided into training and test sets in a ratio of 8:2. The videos were adjusted to a resolution of 240 $\times$ 256 and a frequency of 30 frames per second. The Openpose algorithm (Huang et al., 2023) was used to obtain 18 joint points from each frame of video images, including the two-dimensional coordinates and the confidence degree of each joint point. Then, an ethnic dance dataset was established.

3.3 Analysis of Results

The MS-STGCN model was compared with other movement feature recognition methods.

According to Table 3, the X-Sub evaluated algorithms' generalization ability for various people, the X-View evaluated algorithms' robustness for various perspectives, and the X-Set evaluated algorithms' generalization ability for various scenarios. The performance of all the algorithms was the best for X-View, which might be because the difficulty of X-View was relatively low. For the NTU-RGB+D 60 dataset,

Table 3 Results of different movement feature recognition methods for NTU-RGB+D 60/120.

	NTU-RGB+D 60		NTU-RGB+D 120	
	X-Sub/%	X-View/%	X-Sub/%	X-Set/%
Two-stream adaptive GCN (Shi et al., 2019)	88.5	95.1	82.9	84.9
Semantics-guided neural network (Zhang et al., 2020)	89.0	94.5	79.2	81.5
Topology-aware convolutional neural network (Xu et al., 2022)	90.4	94.8	85.4	86.8
Shift GCN (Cheng et al., 2020)	90.7	96.5	85.9	87.6
Decoupling GCN + DropGraph module (Cheng et al., 2020)	90.8	96.6	86.5	88.1
Dynamic GCN (Ye et al., 2020)	91.5	96.0	87.3	88.6
Multi-scale graph 3D convolutional network (Liu et al., 2020)	91.5	96.2	86.9	88.4
Decoupled spatial-temporal attention network (Shi et al., 2020)	91.5	96.4	86.6	89.0
EfficientGCN-B4 (Song et al., 2021)	91.7	95.7	88.3	89.1
Relation-mining self-attention network (Gedamu et al., 2023)	91.8	**96.8**	88.4	89.7
MS-STGCN	**92.6**	**96.8**	**89.7**	**89.9**

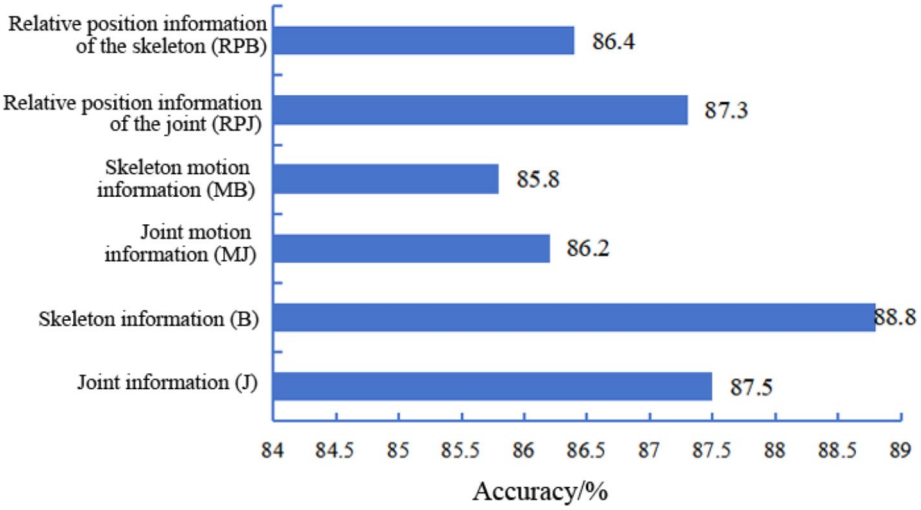


Figure 2 Comparison of single-stream input results.

the MS-STGCN model achieved accuracies of 92.6% for X-Sub and 96.8% for X-View, respectively, outperforming the other algorithms. The RSA-Net method had an accuracy rate of 96.8% for X-View, but 91.8% for X-Sub, lower than the MS-STGCN model. Due to the larger size and more complex data of the NTU-RGB+D 120 dataset, the accuracy rates were generally lower than those for the NTU-RGB+D 60 dataset. The accuracies of the MS-STGCN model for X-Sub and X-Set were 89.7% and 89.9%, respectively, which were superior to the other algorithms.

Then, the results of different input streams were compared using the ethnic dance dataset. From Figure 2, it can be observed that in the single-stream case, the accuracy of movement feature recognition for different ethnic dances was generally low, with the highest reaching 88.8% when using skeleton information. As shown in Table 4, the accuracy improves to some extent in the two-stream situation, where the accuracy of using joint and skeleton information was the highest, reaching 89.7%. In the case of four streams, the accuracy of using joint/skeleton information and joint/skeleton motion information was the highest, reaching 90.4%. Finally, when the data of all streams was input, the recognition was the best, with an accuracy of 90.6%. The findings demonstrate the effectiveness of the MS-STGCN model in the recognition of movement features of ethnic dances.

4. CONCLUSION

This paper developed an MS-STGCN model for movement feature recognition of different ethnic dances based on the skeleton information of motion images. Experiments were conducted using NTU-RGB+D 60/120 datasets and an ethnic dance dataset. According to experimental results, the proposed model achieved superior recognition outcomes than the other algorithms for the NTU-RGB+D 60/120 datasets and an accuracy rate of 90.6% for the ethnic dance dataset. The method can accurately identify the movement features of different ethnic dances and provide support for the digital management of such dances.

Table 4 Comparison of multi-stream input results.

	Accuracy/%
$J + B$	89.7
$M_J + M_B$	87.2
$RP_J + RP_B$	89.6
$J + B + M_J + M_B$	90.4
$J + B + RP_J + RP_B$	89.7
$M_J + M_B + RP_J + RP_B$	89.6
$J + B + M_J + M_B + RP_J + RP_B$	90.6

REFERENCES

1. Bian, C., Lyu, W., &Feng, W. (2024). Skeleton-Based Human Action Recognition: History, Status and Prospects. *Journal of Computer Engineering & Applications*, *60*(20), 1.
2. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., &Lu, H. (2020). Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition. *16th European Conference*, 536–553.
3. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., &Lu, H. (2020). Skeleton-Based Action Recognition with Shift Graph Convolutional Network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 180–189.
4. Deng, Y. L. (2022). Formation and Direction of Chinese School of Dance Studies. *Dance Studies in China: Selected Writings from the Journal of Beijing Dance Academy*, 25–53
5. Dong, Y., Nie, X., Yang, W., &Xie, Y. (2024). An Interactive Teaching System for Ethnic Minority Dance Based on Deep Learning. *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 434–440.
6. Dong, Z., &Wang, X. (2023). An improved deep neural network method for an athlete's human motion posture recognition. *International Journal of Information and Communication Technology*, *22*(1), 45–59.
7. Gedamu, K., Ji, Y., Gao, L., Yang, Y., &Shen, H.T. (2023). Relation-mining self-attention network for skeleton-based human action recognition. *Pattern Recognition*, *13*, 109455.
8. Huang, W. C., Shih, C. L., &Fan, F. C. P. (2023). OpenPose Technology Based Yoga Exercise Guidance Functions by Hint Messages and Scores Evaluation for Dynamic and Static Yoga Postures. *Journal of Advances in Information Technology*, *14*(5), 1029–1036.
9. Jiu, Y., Niu, Q., &Feng, J. (2024). Human Motion Trajectory Perception and Recognition Based on Multi Sensor Network Fusion with Conditional Random Fields. *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 1–5.
10. Lian, Y., He, F., &Wang, S. (2025). Classroom Teaching Behaviour Analysis and Teaching Quality Evaluation Design Basedon Deep Learning. *Engineering Intelligent Systems*, *33*(6), 689–699.
11. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., &Kot, A. (2020). 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.
12. Liu, Z.,Bai, Y., Hou, B., Ning, K., Liu, X., &Zhang, J. (2024). Investigation of AI-based Image RecognitionTechnology Combined with Sensor Technology for Power Grid Quality and Safety.*EngineeringIntelligent Systems*, *32*(5), 533–542.
13. Liu, Z., Zhang, H., Chen, Z., Wang, Z., &Ouyang, W. (2020). Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 140–149.
14. Lu, K., Liu, L., Wang, X., &Ren, S. (2025). PHMG: Prompt-based Human Motion Generation for action recognition. *Image and Vision Computing*, *163*, 13.
15. Shahroudy, A., Liu, J., Ng, T. T.,&Wang, G. (2016). NTU RGB+D: A Large-Scale Dataset for 3D Human Activity Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019.
16. Shi, L., Zhang, Y., Cheng, J., &Lu, H. (2019). Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12018–12027.
17. Shi, L., Zhang, Y., Cheng, J., &Lu, H. (2020). Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition. Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, 38–53.
18. Song, Y. F., Zhang, Z., Shan, C., &Wang, L. (2021). Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 1474–1488.
19. Sun, B., Cheng, G., Dai, Q., Chen, T., Liu, W., &Xu, X. (2021). Human motion intention recognition based on EMG signal and angle signal. *Cognitive Computation and Systems*, *3*(1), 37–47.
20. Tang, C., Yang, L., Cao, G., Du, J., &Zhang, Q. (2024). ST-GCN: EEG Emotion Recognition via Spectral Graph and Temporal Analysis with Graph Convolutional Networks. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3748–3751.
21. Tun, K. C., Tun, H. M., &Win, K. K. K. (2024). Human Motion Recognition Using Temporal Foot-Lift Features Extracted from a Small Number of Skeleton Data Frames and Multi Classifiers. *Engineering Journal*, *28*(7), 41–51.
22. Xu, K., Ye, F., Zhong, Q., &Xie, D. (2022). Topology-Aware Convolutional Neural Network for Efficient Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*,*36*(3), 2866–2874
23. Xue, Y., Ru, F., Du, H., Yin, K., Li, P., &Ju, Z. (2025). SEMG-Based Complex Human In-Hand Motion Recognition for Dexterous Robotic Manipulation. *IEEE Access*, *13*, 51042–51053.
24. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., &Tang, H. (2020). Dynamic GCN: Context-enriched Topology Learning for Skeleton-based Action Recognition. *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, 55–63.
25. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., &Zheng, N. (2020). Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1109–1118.