ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

DAA-GCN: A Dynamic Adaptive Attention Graph Convolutional Network for Robust Skeleton-Based Action Recognition

Rundong Zhou[1,a,*], Jiawei Wang[1,b]

[1]School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, China
[a]1318844563@qq.com, [b]wzybq123@163.com
**Corresponding author*

Abstract: *Skeleton-based action recognition has attracted increasing attention due to its compact data representation and robustness against appearance variations. Although Graph Convolutional Networks (GCNs) have demonstrated strong performance by modeling spatial dependencies among joints, they still face challenges in capturing long-range interactions and multi-scale temporal dynamics. To address these limitations, we propose a novel framework, DAA-GCN (Dynamic Adaptive Attention Graph Convolutional Network), which comprises two core modules: the Spatio-Temporal Adaptive Feature Extractor (STAFE) and the Multi-Perspective Fusion Graph Attention (MPFGA). STAFE integrates long-range spatio-temporal graph convolutions with multi-branch temporal convolutions to effectively capture both short-term and long-term motion patterns, while MPFGA enhances feature representations by combining global self-attention with local additive attention, thereby balancing global context with local structural information. We evaluate the proposed DAA-GCN on two benchmark datasets, NTU RGB+D 60 and NTU RGB+D 120, under both cross-subject and cross-view/setup protocols. Experimental results show that DAA-GCN consistently outperforms state-of-the-art methods, and ablation studies further confirm the effectiveness of each module in the overall architecture. In summary, DAA-GCN presents a robust and scalable solution for skeleton-based action recognition, with promising applications in human-computer interaction, video surveillance, and healthcare monitoring.*

Keywords: *Skeleton Recognition, Graph Convolution, Attention Mechanism, Feature Fusion*

1. Introduction

As one of the core tasks in the field of computer vision, human skeleton based action recognition has important application value in human-computer interaction, intelligent monitoring and medical assistance. Early research methods mainly relied on manually designed features, such as joint motion trajectories, relative positions of joint points, and temporal dynamic descriptors. Although these methods have achieved some results, their feature expressiveness is limited due to manual design constraints and a lack of robustness to viewpoint changes and noise interference[1].

The rise of deep learning has revolutionized this field, and convolutional neural networks (CNNs) have shown significant advantages in spatio-temporal feature extraction of skeleton sequences through end-to-end learning. Representative works such as ST-GCN introduced spatial-temporal graph modeling to joint association coding for the first time, and subsequent studies captured multi-scale motion patterns through hierarchical CNN architectures[2]. However, CNNs exhibit inherent limitations when processing non-Euclidean data. Their underlying assumptions, which rely on grid-like (rasterized) input structures, are fundamentally misaligned with the graph-based topology of human skeletal data[3]. In order to adapt CNN input requirements, researchers have to pre-process the skeleton data with rasterization, resulting in distorted anatomical topological relationships and weakened geometric invariance[4].

Graph Convolutional Networks (GCNs) provide a theoretical framework to solve the above problems. By modeling the human skeleton as graph-structured data with joints as nodes and skeletal connections as edges, GCN can directly preserve its topological integrity and achieve adaptive learning of spatial dependencies based on local spectral filtering[5]. Compared to traditional methods, skeleton-based action recognition approaches leveraging GCNs offer three significant advantages: (1) the capability to explicitly encode anatomical priors through a learnable adjacency matrix; (2) effective integration of structural information via hierarchical graph propagation mechanisms; and (3) seamless modeling of

ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

spatiotemporal dynamics by incorporating temporal convolutional modules[6]. In recent years, the discriminative power of GCN-based models has been further enhanced through the introduction of attention mechanisms, multi-scale feature aggregation strategies, and other advanced techniques[7]. These improvements have led to state-of-the-art performance on widely used benchmark datasets, such as NTU RGB+D and Kinetics-Skeleton[8].

Nevertheless, several critical challenges remain in the current approach. First, joint occlusions in complex scenes frequently disrupt the continuity of feature propagation paths. Second, the large spatial distances between certain joint pairs hinder effective information exchange, thereby reducing action recognition accuracy[9]. Third, there is an inherent trade-off between computational efficiency and recognition performance, particularly in real-time deployment scenarios[10].

In this study, a novel skeleton behavior recognition model called Dynamic Adaptive Attention Graph Convolution Network (DAA-GCN) is proposed, which aims to fully exploit the spatio-temporal features of human skeleton sequences and enhance the dependencies between long-distance skeleton nodes. Experiments show that the method has some improvement over the existing models on the NTU-RGB+D dataset[11].

2. Methods

2.1 DAA-GCN Framework

The network mainly consists of Spatio-Temporal Adaptive Feature Extractor (STAFE) and Multi-Perspective Fusion Graph Attention (MPFGA) modules with global Pooling and Fully Connected Layer for classification. As shown in Fig. 1, the input skeleton sequence is first processed by three stacked STAFE modules, followed by the MPFGA attention module. The alternate arrangement of STAFE and MPFGA layers allows progressive integration of local and global information.

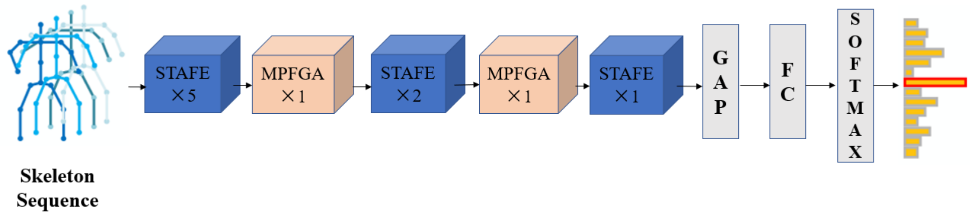

Figure 1: Overview of the DAA-GCN architecture. STAFE: Spatio-Temporal Adaptive Feature Extractor; MPFGA: Multi-Perspective Fusion Graph Attention; FC: Fully Connected Layer; GAP: Global Average Pooling.

The dimensions of the input data are $C_{in} \times T \times N$,where C_{in} represents the number of input channels, T denotes the number of frames per input sequence, and N denotes the number of human joints. The output of the model is a probability distribution over action categories, obtained through a softmax activation function applied to the fully connected layer.

After the skeleton data are preprocessed, the features are firstly learned by the STAFE module, which adopts a GCN-based approach to model the skeleton topology and extract spatio-temporal features by combining the information of the temporal dimension. For the three-part STAFE module in the network architecture, the number of channels is set to 64, 128, and 256, respectively. The first five layers of the network, STAFE, are mainly used for local feature extraction to enable the model to recognize short time-series action patterns; subsequently, the MPFGA module is inserted into the network to enhance the correlation between different skeleton nodes and optimize the feature representation. This module employs a multi-view attention mechanism to fuse information from different scales of the skeleton graph,thus enhancing the model's ability to perceive complex behaviors.

After MPFGA, the network continues to stack STAFE layers to further learn the fused global features. The design of this alternating structure enables the model to integrate information at different layers and fully utilize the long and short-term dependencies. In addition, MPFGA employs an adaptive attention mechanism to dynamically adjust the weights among different nodes, thus optimizing the spatial representation capability of skeleton features.

Finally, after a series of STAFE and MPFGA processing, the features are downscaled by Global

ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

Average Pooling (GAP) to reduce redundant information, and fed into Fully Connected (FC) layer for classification. The classification results are used to output the category probability distribution through the softmax layer to finally predict the behavioral categories.

2.2 Spatio-Temporal Adaptive Feature Extractor (STAFE)

2.2.1 Motivation

One of the core challenges in the task of Skeleton-based Action Recognition (SAR) for the human skeleton lies in how to effectively model the spatial relationships between joints and the dependency of long time-series information[12]. Currently, GCN-based methods have made significant progress in spatial modeling, but the following problems still exist[13].

The first is the insufficient modeling of remote joint point relationships14]. [For traditional GCNs, information propagation is usually based on the predefined topology of the human skeleton, which can only capture the dependencies of local neighboring joints, but cannot effectively model the interactions between remote joints, such as the synergistic movements of the hand and the leg[15].

Secondly, the ability to model temporal information is limited. Existing methods tend to process temporal information through single-scale temporal convolution or temporal pooling, which lacks the ability to capture behavioral patterns at different time scales, leading to insufficient long-time dependency modeling.

Aiming at the above problems STAFE is used to enhance the modeling capability of long-range joints dependencies under the GCN framework, and at the same time improve the capturing effect of long time-series information. The module consists of Long-Range Enhanced Spatio-Temporal Graph Convolution (LREST-GC) and Temporal Convolution Module (TCM).LREST-GC enhances the remote joints dependency modeling by introducing adaptive LREST-GC enhances the information interaction between remote joints by introducing an adaptive adjacency matrix, while TCM adopts a multi-scale temporal convolution structure to realize the fusion modeling of short-time dependence and long-time dependence. The overall structure is shown in Fig. 2. In Fig. 2, the input features are processed in two branches: the upper branch corresponds to LREST-GC, responsible for spatial modeling, and the lower branch is the TCM block designed for multi-scale temporal modeling.

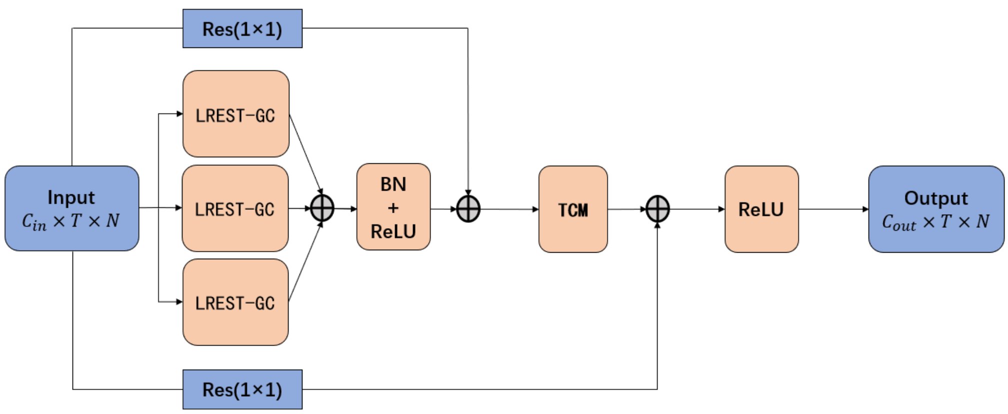

Figure 2: Architecture of the STAFE module. The module consists of LREST-GC (Long-Range Enhanced Spatio-Temporal Graph Convolution) and TCM (Temporal Convolution Module) branches.

The STAFE module achieves efficient spatio-temporal modeling of human skeleton data by combining the capabilities of LREST-GC and TCM. LREST-GC enhances the relationship between remote joints by using adaptive adjacency matrix and adopts a dual-path feature extraction strategy, which achieves effective combination of local and global information. Conversely, TCM leverages a multi-scale temporal convolutional framework to effectively model human skeletal motion over both short and long temporal ranges, while a multi-branch fusion mechanism is incorporated to further enrich temporal information representation. Skeleton motion patterns are modeled from short-time and long-time perspectives, and the expression ability of temporal information is enhanced with the help of a multi-branch fusion mechanism. The whole STAFE module can be optimally trained in an end-to-end framework, and ensures stable propagation of gradients through batch normalization and residual concatenation.

ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

2.2.2 Long-Range Enhanced Spatio-Temporal Graph Convolution (LREST-GC)

LREST-GC is mainly used to enhance the spatial feature modeling capability of skeleton data, breaking through the limitation that traditional GCN can only deal with local neighboring nodes, making the model able to capture the spatial association information of key joints in the global scope. Its structure is shown in Figure 3. In Fig. 3, the top path uses 1×1 Conv and temporal pooling to capture local joint features. The bottom path applies Deformable Graph Convolution (DGC), guided by a Dynamic Adjacency Matrix (DAM), to enhance long-range joint interactions. The refined features are then passed through a tanh activation followed by a refinement block.

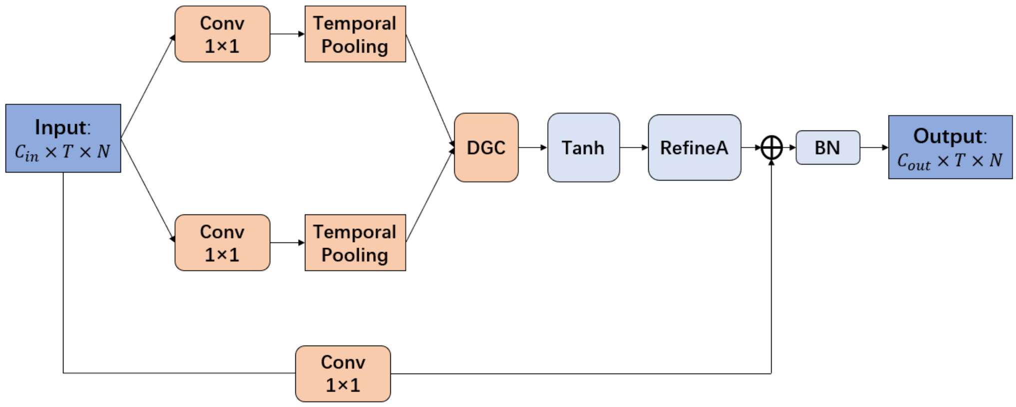

Figure 3: Illustration of the LREST-GC module, which includes dual-path processing, adaptive graph learning, and feature enhancement steps.

Adaptive long-range modeling approach: the DGC module is employed to model the association weights between remote joints in a learnable manner. Specifically, we break the limitation of predefined topology by constructing DAM to compute remote dependencies using feature similarity.

Dual-Path Feature Extraction Method:LREST-GC uses a dual-path structure for feature extraction. Path 1 uses 1×1 convolution (Conv 1×1) + Temporal Pooling to enhance the representation of local joint point information. Path 2 adopts Deformable DGC for dynamically adjusting the adjacency matrix so that information between remote joints can be efficiently interacted.

Nonlinear Transformation and Feature Enhancement Methods: After extracting the features by DGC, we introduce the hyperbolic tangent (Tanh) activation function to enhance the nonlinear expression ability of the features, and refine the features by RefineA module to make the information propagation more robust.

For the input skeleton data $X \in \mathbb{R}^{C_{\text{in}} \times T \times N}$ where C_{in} is the number of input channels, T is the time step, N is the number of nodes, LREST-GC propagates the information through the adaptive adjacency matrix A_{dyn} with the following equation:

$$\mathrm{H} = \sigma\left(\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{A}_{\mathrm{dyn}}^{(\mathrm{k})} \mathrm{X} \mathrm{W}^{(\mathrm{k})}\right) \tag{1}$$

Where, K denotes the extended neighborhood size of the graph convolution, $W^{(k)}$ is the learned weight matrix, $\sigma(\cdot)$ is the activation function, W_1, W_2 are learnable parameters used to dynamically compute the similarity between nodes. The adaptive neighbor matrix A_{dyn} is computed by.

$$\mathbf{A}_{\mathrm{dyn}} = \mathrm{Softmax}(\mathrm{X}\mathrm{W}_1\mathrm{W}_2^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}) \tag{2}$$

2.2.3 Temporal Convolution Module (TCM)

TCM aims to enhance the temporal modeling capability of skeleton data to avoid information loss in the temporal dimension. To capture behavioral patterns at different time scales, we designed the Multi-scale TCM architecture, as shown in Fig. 4. In Fig. 4, the upper left branch captures short-term patterns using a dilation=1 convolution. The second branch, to its right, targets long-term patterns via dilation=2. The third branch (bottom left) performs temporal max pooling, while the final branch (bottom right) is a shortcut connection for preserving original features.

ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

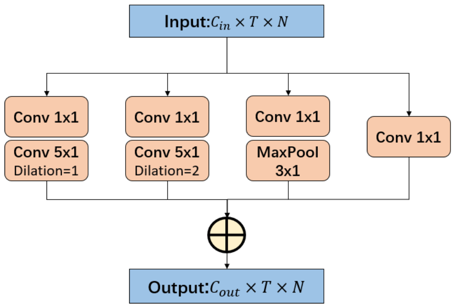

Figure 4: Temporal Convolution Module (TCM) with four branches: short-range, long-range, temporal pooling, and residual shortcut.

Short-time dependency modeling branch: 1×1 convolution + 5×1 dilation convolution (Dilation=1) is used to capture behavioral changes over a short time span.

Long time dependent modeling branch: 1×1 convolution + 5×1 dilation convolution (Dilation=2) is used to capture motion patterns over longer time spans.

Time Pooling Branch: 1×1 Convolution + 3×1 Maximum Pooling is used to reduce the redundancy of information in the time dimension and improve the temporal modeling capability of the model.

Shortcut connection branch: 1×1 convolution is used to directly transfer the original feature information to avoid the information loss caused by long time-series modeling.

Given the input feature $H \in \mathbb{R}^{C_{\text{mid}} \times T \times N}$, the TCM is computed as follows.

$$\mathrm{H'} = \sum_{\mathrm{i}=1}^{4} \mathrm{W_i} * \mathrm{H_i} \tag{3}$$

Where $*$ denotes the convolution operation, W_i represents the weight of each branch, and H_i is the feature output of the corresponding branch. Eventually, we use fusion to integrate the outputs of all branches in order to form the final temporal feature representation.

2.3 Multi-Perspective Fusion Graph Attention (MPFGA)

In skeleton-based human behavior recognition tasks, GCNs are widely used for feature extraction from skeleton data due to their advantages in spatial domain modeling. However, traditional GCNs mainly rely on local neighbor information for feature aggregation, which leads to limitations in capturing long-distance node relationships and makes it difficult to effectively model global dependency information over long distances.

To address this problem, we propose a Multi-Perspective Fusion Graph Attention (MPFGA) mechanism, which combines Global Self-Attention (GSA) and Graph Additive Attention (GAA), which combine GSA and GAA for modeling global dependencies and local fine-grained features, respectively. In this way, MPFGA can take into account feature interactions between long-distance nodes while preserving local neighborhood information, thus enhancing the representation of skeleton features.

Given a skeleton sequence denoted as Figure $G = (V, E)$ where V is the set of skeleton joints, E shows the connection relationship between joints, and the input feature matrix is $X \in \mathbb{R}^{N \times C}$ where N is the number of joints and C is the feature dimension. In order to effectively capture remote dependencies and local structure information, we define the MPFGA mechanism as follows:

$$\mathrm{H} = \alpha \cdot \mathrm{GSA(X)} + \beta \cdot \mathrm{GAA(X)} \tag{4}$$

Among them, α and β are learnable parameters that control the contribution of GSA and GAA, respectively, enabling the network to dynamically adjust the weighting of global and local information according to the needs of different scenarios during the training process.

GSA implements global dependency modeling by calculating the attention weights between all nodes as follows:

$$\mathrm{GSA}(X) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

Where $Q = W_q X, K = W_k X, V = W_v X$ are the query matrix, key matrix and value matrix respectively, W_q, W_k, W_v are the learnable parameters, and d_k is the scaling factor for stabilizing the gradient.

GAA is mainly used for the aggregation of local neighborhood information and learns local structural information by explicitly computing the importance scores of neighboring nodes as follows:

$$e_{ij} = \sigma\left(W_a [X_i \parallel X_j]\right) \tag{6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{7}$$

$$GAA(X) = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} X_j \tag{8}$$

Where W_a is the learnable weight, $\parallel$ denotes feature splicing, $\sigma(\cdot)$ is the nonlinear activation function, and $\mathcal{N}(i)$ denotes the set of neighbors of node i. The GAA is able to explicitly model the local neighborhood information, enabling the model to focus on the feature interactions between closer joints in the skeleton structure.

MPFGA combines GSA and GAA to balance remote dependencies and local structural features in skeleton behavior recognition. GSA enables the model to capture features of remote joints, compensating the limitation that traditional GCNs can only deal with local neighborhoods, whereas GAA enhances the modeling of local information through additive attention to ensure the integrity of local skeleton relationships. By introducing learnable parameters, the model can dynamically adjust the weight of global and local information fusion to adapt to different types of behavioral patterns. This design not only improves the model's ability to capture long-distance features, but also avoids the interference of irrelevant information that may be introduced by global attention, which improves the recognition accuracy and generalization performance, and provides a more robust feature modeling capability for the recognition of complex human behaviors.

3. Experimental Results And Analysis

3.1 Datasets

NTU-RGB+D 60 and its extended version NTU-RGB+D 120 are large-scale datasets for human movement recognition. The original dataset, NTU-RGB+D 60, contains 56,800 motion sequences covering 60 different motion categories, and each skeleton data consists of 25 3D joints. In contrast, the extended dataset NTU-RGB+D 120 is further enlarged in size, containing 113,945 motion sequences and extending to 120 motion categories. Additionally, the data sources of NTU-RGB + D 120 are more diverse, covering 106 subjects and offering three camera views, while the original NTU-RGB + D 60 dataset only has two views.

In data partitioning, both datasets adopt X-Sub (Cross-Subject Split), which trains and tests on different subjects. NTU-RGB+D 120 further incorporates X-Set (Cross-Setup Split), which splits data based on camera setups, to boost diversity and challenge. These features make NTU-RGB+D 60 and NTU-RGB+D 120 two of the most widely used benchmark datasets in human action recognition research.

3.2 Training Details

The proposed model in this study is implemented using the PyTorch 1.9.1 deep learning framework and runs on a high-performance computing platform. The platform is equipped with a CUDA 11.1 computing environment with NVIDIA RTX 4090 GPUs to provide powerful computational capabilities and efficient deep learning training support.

In the training process, we use the Stochastic Gradient Descent (SGD) optimization algorithm combined with Nesterov momentum with the momentum parameter set to 0.9. Backpropagation typically uses the Cross-Entropy loss function. The Cross-Entropy loss function serves as the objective function to measure the model's classification performance. Meanwhile, in order to prevent overfitting, we introduce the Weight Decay mechanism and set its hyperparameter to 0.0004.

In terms of data processing and training strategy, we divide the dataset into small batches with a Batch Size of 16 for training, with a total of 65 batches. The initial value of the learning rate is set to 0.1, and Learning Rate Decay is performed during the training process. The specific strategy is: at the 35th and

55th batches, the learning rate is multiplied by a decay factor of 0.1 to improve the convergence and stability of the model.

Each sample contained a maximum of 300 frames and each sample in the dataset contained a maximum of two human individuals. For samples containing only a single individual, we use Zero Padding to complete the data to ensure uniformity of data input and effective training of the model.

3.3 Comparative Experiments

DAA-GCN was tested against other models and the results were compared by the metrics of NTU-RGB+D 60 and NTU-RGB+D 120 as shown in Tables 1 and 2.

Table 1: Comparison of accuracy (%) on NTU-RGB+D 60 dataset under X-Sub and X-View settings.

Model	NTU-RGB+D 60 X-Sub(%)	NTU-RGB+D 60 X-View(%)
ST-GCN	84.2	90.5
Pe-GCN	85.5	93.6
Js-AGCN	85.7	93.4
Bs-AGCN	86.5	94.1
STGR-GCN	86.9	93.8
AS-GCN	87.8	94.2
2s-AGCN	89.2	95.3
DGNN	90.1	96.2
CTR-GCN	91.4	96.5
DAA-GCN (ours)	91.8	97.2

Table 2: Comparison of accuracy (%) on NTU-RGB+D 120 dataset under X-Sub and X-Setup settings.

Model	NTU-RGB+D 120 X-Sub(%)	NTU-RGB+D 120 X-Set(%)
ST-GCN	79.3	81.1
2s-AGCN	82.9	84.9
Js-AGCN	84.4	86.8
Bs-AGCN	85.3	87.2
STGR-GCN	87.2	88.1
MS-G3D	86.9	88.4
MST-GCN	87.5	88.8
Info-GCN	88.5	89.7
CTR-GCN	88.9	90.6
DAA-GCN (ours)	89.6	91.3

3.4 Ablation Experiment

Table 3: Ablation results showing the contribution of LREST-GC (A), TCM (B), and MPFGA (C) modules.

Model	NTU-RGB+D 60 X-Sub(%)
A	73.5
B	72.1
C	69.4
A+B	83.2
A+C	84.5
B+C	81.6
A+B+C	91.8

We conducted ablation experiments to assess the contribution of each module to model performance. Key modules including the LREST-GC graphical convolution module, the TCM temporal convolution module, and the MPFGA attention module were systematically removed or replaced. Comparative analysis of the results identified the specific role and performance impact of each module in skeletal behavior recognition. For a fair comparison, all experiments were conducted on the same dataset, NTU-RGB+D 60, with identical training settings and hyperparameter conditions. Some of the modules removed from the experiments were processed with 1×1 convolutional modules instead. The experimental results are shown in Table 3, where the LREST-GC module is A, the TCM module is B,

ISSN 2616-5775 Vol. 8, Issue 6: 19-26, DOI: 10.25236/AJCIS.2025.080603

and the MPFGA module is C.

4. Conclusions

This paper proposed DAA-GCN, a skeleton-based action recognition model that integrates STAFE and MPFGA modules to capture both local and long-range spatio-temporal dependencies. Experiments on NTU-RGB+D 60 and 120 show that DAA-GCN outperforms existing methods. Ablation studies confirm the effectiveness of each component. The model offers a robust and efficient solution for skeleton-based action recognition. Future work will aim to improve computational efficiency and adapt the model to more complex real-world scenarios.

References

[1] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Network," 2017.

[2] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018.

[3] Ren, Bin, et al. "A survey on 3d skeleton-based action recognition using learning method." Cyborg and Bionic Systems 5 (2024): 0100.

[4] Peng, Kunyu, et al. "Navigating open set scenarios for skeleton-based action recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 38. No. 5. 2024.

[5] Song, Yi-Fan, et al. "Constructing stronger and faster baselines for skeleton-based action recognition." IEEE transactions on pattern analysis and machine intelligence 45.2: 1474-1488 (2022).

[6] Zhou, Huanyu, Qingjie Liu, and Yunhong Wang. "Learning discriminative representations for skeleton based action recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10608-10617. 2023.

[7] Su, Kun, Xiulong Liu, and Eli Shlizerman. "Predict & cluster: Unsupervised skeleton based action recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9631-9640. 2020.

[8] Wu, L., Zhang, C. and Zou, Y., 2023. SpatioTemporal focus for skeleton-based action recognition. Pattern Recognition, 136, p.109231,2023.

[9] Tang, Yansong, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. "Deep progressive reinforcement learning for skeleton-based action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5323-5332. 2018.

[10] Feng, Liqi, Yaqin Zhao, Wenxuan Zhao, and Jiaxi Tang. "A comparative review of graph convolutional networks for human skeleton-based action recognition." Artificial Intelligence Review (2022): 1-31.

[11] Shi, Lei, et al. "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks." IEEE Transactions on Image Processing 29: 9532-9545 (2020).

[12] Du, Yong, Yun Fu, and Liang Wang. "Representation learning of temporal dynamics for skeleton-based action recognition." IEEE Transactions on Image Processing 25.7 : 3010-3022 (2016).

[13] Cao, Congqi, et al. "Skeleton-based action recognition with gated convolutional neural networks." IEEE Transactions on Circuits and Systems for Video Technology 29.11: 3247-3257 (2018).

[14] Liu, Jun, et al. "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates." IEEE transactions on pattern analysis and machine intelligence 40.12 (2017): 3007-3021.

[15] Li, Chuankun, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. "Skeleton-based action recognition using LSTM and CNN." In 2017 IEEE International conference on multimedia & expo workshops (ICMEW), pp. 585-590. IEEE, 2017.