Pattern Analysis and Applications (2024) 27:118
https://doi.org/10.1007/s10044-024-01319-3

ORIGINAL ARTICLE

Research on decoupled adaptive graph convolution networks based on skeleton data for action recognition

Haigang Deng[1] · Guocheng Lin[2] · Chengwei Li[1] · Chuanxu Wang[2]

Received: 11 January 2024 / Accepted: 1 August 2024

Abstract

Graph convolutional network is apt for feature extraction in terms of non-Euclidian human skeleton data, but its adjacency matrix is fixed and the receptive field is small, which results in bias representation for skeleton intrinsic information. In addition, the operation of mean pooling on spatio-temporal features in classification layer will result in losing information and degrade recognition accuracy. To this end, the Decoupled Adaptive Graph Convolutional Network (DAGCN) is proposed. Specifically, a multi-level adaptive adjacency matrix is designed, which can dynamically obtain the rich correlation information among the skeleton nodes by a non-local adaptive algorithm. Whereafter, a new Residual Multi-scale Temporal Convolution Network (RMTCN) is proposed to fully extract temporal feature of the above decoupled skeleton dada. For the second problem in classification, we decompose the spatio-temporal features into three parts as spatial, temporal, spatio-temporal information, they are averagely pooled respectively, and added together for classification, denoted as STMP (spatio-temporal mean pooling) module. Experimental results show that our algorithm achieves accuracy of 96.5%, 90.6%, 96.4% on NTU-RGB+D60, NTU-RGB+D120 and NW-UCLA data sets respectively.

Keywords Decoupled adaptive graph convolutional network · Residual multi-scale temporal convolution network · Decoupled head of classification lay · Skeleton based action recognition

1 Introduction

Skeleton based behavior recognition is an important research field in computer vision. Compared with the behavior recognition of RGB video, it can filter the non-limb dynamic information interference in the image and has strong robustness in complex background. However, the skeleton data extracted from the video belongs to Non-Euclidean data, and how to effectively extract their features is the core for skeleton behavior recognition.

Deep learning based methods often manually convert skeleton data into a sequence of joint coordinate vectors or pseudo-images for easy feature representation. RNN or CNN based methods [1–5] are effective in processing Euclidean data, but can not explicitly processing skeleton graphic data. For this reason, graph convolutional network (GCNs) is designed to obtain its non-Euclidean data features, typically, Spatio-Temporal Graph Convolutional Networks (ST-GCNs) [6] become the popular ways. Their core idea is to combine GCN graph convolution with spatiotemporal convolution, which sets the configuration of skeleton topological graph manually based on the physical structure of human body, and then define its graph adjacency matrix to extract spatial–temporal feature for skeleton sequences. In short, the quality of feature representation is mainly dependent on how well designing on the adjacency matrix in ST-GCNs.

How to construct ST-GCNs mostly relies on the artificial prior knowledge, which usually leads to the defects of rigid structure and solidified correlation information between nodes. Researchers have a number of improved algorithms

Haigang Deng and Guocheng Lin are contributed equally to this work.

✉ Chengwei Li
chengweili@hit.edu.cn

✉ Chuanxu Wang
wangchuanxu_qd@163.com

[1] School of Instrument Science and Engineering, Harbin Institute of Technology, Harbin, Hei Long Jiang, China

[2] School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shan Dong, China

Published online: 18 September 2024

for this. Among them, the two-flow adaptive graph convolution structure (2 s-AGCNs) [7] uses the Non-Local idea to design the adaptive adjacency matrix, and obtains the autocorrelation matrix by calculating the relationship between the two matrix vectors in time or space. In the graph convolution operation, the adaptive adjacency matrix can dynamically converge node information.

However, AGCNs can only obtain the adjacency matrix of a single channel by non-local operation, which is unreasonable to use it to aggregate information at a specific node with multiple channels feature. To overcome this flaw, inspired by the decoupling idea proposed by [8] and the multi-headed self-attention proposed by [9–11], we think that decoupling can be considered as an multi-attention process, which can decompose one complicated feature into multi-feature norms corresponding to several intra-class variations. Therefore, the decouple adaptive graph convolutional structure layer is proposed in this paper, and its structure can also be regarded as a variant of Transformer Encode. Its relevant topologies can be obtained through multi-head adaptive convolution, and the expression of different behaviors on nodes in dynamic situations can be well described comprehensively. Further, in the temporal convolution layer, we adopt the idea of multi-scale temporal convolution network MSTCN [12], and modify its model structure to the proposed residual multi-scale time convolution, thus, it can achieve a larger field of view and better granularity. In addition, in the final average pooling layer, the spatio-temporal feature map is decomposed into three dimensions of space, time and space–time, their loss functions of the three dimensions are added together and then the inverse derivation is obtained, finally they are merged as a whole for behavior classification.

In short, our innovations can be summarized as follows:

1 A Decoupled Adaptive Graph Convolutional Network (DAGCN) layer is proposed, by which, an adaptive spatial aggregation kernel is designed to obtain multi-channel data, and learns their topology of GCN layers and skeleton data in different channels through self-attention mechanism, so as to extract the multi-leveled feature information of skeleton data more effectively.
2 A new Multi-Scale Time Convolution Network (MSTCN) is proposed, which uses residual multi-scale convolution kernel and dilated convolution to achieve a larger field of view.
3 Spatial Temporal Mean Pooling (STMP) is proposed in the final average pooling layer, that is, the spatial–temporal feature is decoupled into three components as spatial, temporal, and spatial–temporal feature, they are added after individually averaged pooling for classification, which helps to improve the classification accuracy.

2 Related work

Compared with RGB image sequences as inputs for individual behavior recognition, a human skeleton representation only requires the three-dimensional coordinates and confidence score of each joint, which is usually less than 30 joints, so skeleton data can alleviate storage burden and computation cost significantly. Moreover, Human skeleton data can be easily obtained by multi real-time extractors, such as Openpose, HR net, Google PoseNet and Nuitrack. Additionally, compared with RGB image-based action recognition methods, skeleton-based algorithms are more robust to illumination variation, dynamic backgrounds, differences in color, and different subjects etc. So skeleton-based action recognition becomes a main stream on computer vision field. Generally, it undergoes three typical developing stages, which are Convolutional Neural Networks (CNN) based methods, Recurrent Neural Networks (RNN) based method, and Graph Neural Networks (GNNs) based method, especially with graph convolution networks (GCNs) framework.

2.1 CNN-based skeleton data action recognition

Specifically, 2D CNN can not capture temporal feature from a skeleton sequence, so CNN-based methods usually convert it into a “spatially arranged map”. Du et al. [13] parallelly transformed the node coordinates on the time axis into a matrix “image”, and solved the problem of inconsistent length of action sequences through this normalization. Finally, the “image” was input to a CNN network for feature extraction and recognition. Wang et al. [14] proposed the human skeletal joint trajectory motion, by projecting the coordinate trajectory of motion nodes into HSV space, the space–time information was transformed into a multi-view joint trajectory. Finally, features were extracted through pretraining network with ImageNet for classification processing. As the type of an action is mostly determined by its local joint movements, Zhu et al. [15] proposed a cuboid model for action recognition based on skeleton data. Its cuboid arrangement strategy is to obtain a representation of cuboid actions by organizing paired displacements between all body joints, which was suitable to apply deep CNN models to analyse actions.

In short, the above CNN-based methods have explored the ways to process skeleton data with convolution framework, and gained rather good performances, but their network structures are weak to to well describe temporal information within skeleton sequence.

2.2 RNN-based skeleton data action recognition

RNN and its relevant variants, such as LSTM and GRU, are applied to improve skeleton-based action recognition by their better temporal information modelling. Du et al. [16] proposed a hierarchical recursive neural network for skeleton data action recognition. They divided the human skeleton into five parts based on the physical structure of the human body, and modeled them using five subnets. The representations extracted from these subnets were layered and fused as inputs to higher levels. Via a fully connected layer and a softmax layer on the final representation, they could obtain action classifications. Liu et al. [17] introduced a global context aware LSTM network to handle skeleton data. Their first LSTM layer encodes the skeleton sequence and generates an initial global context representation for the action sequence. The second layer achieves sequence attention representation by using global contextual memory units to perform attention on the input. Then, attention representation is used to refine the global context. By paying attention to multiple iterations, gradually the global context memory is improved. Finally, they use refined global contextual information for classification. Wei et al. [18] proposed a new high-order joint relative motion feature (JRMF) and a new human bone tree RNN network (HST-RNN). The JRMF of each skeleton joint was composed of the relative position, velocity, and acceleration of that joint and all its descendants. It better described the instantaneous state of skeleton joints than joint position. The HST-RNN network was constructed using the same tree structure as the human skeletal joints. Each node in the tree was a gated recurrent unit (GRU), representing a skeleton joint. Its child nodes and corresponding JRMF outputs were connected and fed into each GRU. This network combined low-level features and extracted high-level features from leaf nodes to root nodes in a hierarchical manner based on the body structure of the human body.

In all, compared to the above CNN-based algorithms, RNN-based methods are good at describing temporal dynamics of joint sequences, but they are difficult to learn correlations of skeletal joints in the spatial domain. To this end, Graph Neural Networks (GNNs) based methods are developed.

2.3 GNN-based skeleton data action recognition

The human skeleton is essentially a kind of graphic structure data, Graph Neural Networks (GNNs), especially graph convolution networks (GCNs) and its variants, demonstrate their power to discover the intrinsic relations between skeleton joints. In the following, we mainly summarize three kinds of methods concerning feature extraction from skeleton data, which are primitive GCN related methods; correlation modeling among disconnected far joints; adaptive adjacent matrix based algorithms.

2.3.1 Primitive GCN related methods

At early stage, GCN as a basic network is explored and exploited for skeleton data feature extraction. Typically, Si et al. [19] proposed an Attention enhanced Graph Convolutional LSTM network (AGC-LSTM) for skeleton data action recognition. They designed a linear layer converting the coordinates of each joint into spatial features, and then connected this spatial feature with the difference between two consecutive frames, denoted as augmented features; further applied three AGC-LSTM layers to model its inherent spatiotemporal characteristics, finally, obtained the global features of all joints and the local features for action prediction. Yan et al. [20] proposed Spatial Temporal Graph Convolutional Networks (ST-GCN), namely, they took the spatial topology in traditional GCN as the basic structure, then new continuous temporal step edges were constructed from the each nodes, aiming to form a multi-layered spatiotemporal graph convolutional architecture, which could automatically achieve the integration and description of skeleton data information along the spatial dimension and time dimension simultaneously. This is a milestone achievement in skeleton data action recognition. But, there are some disadvantages for primitive GCN to process skeleton data, such as, its graphical topology is fixed, and is weak to model correlations between disconnected far joints, which is not optimal for spatial feature representation of body joints.

2.3.2 Correlation modeling among disconnected far joints

Traditional adjacent matrix in GCN is of limited perceive field, it can only measure correlations among neighbouring joints, but not good at describing relations among far none connected joints, which is also essential and important to fully represent the dynamics of an action. This problem has aroused much attention in the research of skeleton-based action recognition.

Lee et al. [21] proposed a Hierarchical Decomposition graph (HD Graph) convolutional network, aiming to well model relationships between distant nodes. Specifically, HD Graph connects all joints in a set of adjacent layered links, and describe their relationships between these joints, including meaningful adjacent nodes and remote joint nodes. In addition, joint correlations, which could not be captured with HD Graph, were considered by their Spatial EdgeConv layer based on hierarchical attention oriented aggregation solution. Also, Yang et al. [22] proposed a hybrid diffusion graph convolution model, aiming to combine diffusion graphs and graph convolutions to achieve more comprehensive information diffusion. Specifically, they used diffusion maps to facilitate information diffusion between nodes with similar features in the feature space. Then, using graph convolution models to promote information spreading between adjacent nodes through adjacency matrices; Importantly, their hybrid diffusion graph decomposes all joints through a hierarchical decomposition structure, which can capture the dependency relationships between distant joints.

Similarly, Zhang et al. [23] designed Spatial Transformer Blocks and Directional Temporal Transformer Blocks to model skeleton data in both spatial and temporal dimensions. Namely, their Temporal Transformer generated a global attention map that spanned the entire skeleton sequence to capture relative changes in poses along the time dimension, thereby better perceiving the relative positional relationships between frames. On the other hand, the design of spatial transformer was to capture the spatial relationships between joints in spatial domain, but they used a shared adjacency matrix when dealing with multi-channel, i.e., the aggregation of joint features in different channels was carried out with the same topology structure, which limited the learning of topological relationships between distant far joints. Therefore, researchers attempt to design an adaptive adjacency matrix to adapt to the dynamic changes of various actions.

2.3.3 Adaptive adjacent matrix designing

Due to the flexibility and high accuracy of adaptive adjacent matrix in modeling spatial feature for skeleton-based action recognition, it becomes an attractive research spot recently.

Wei et al. [24] proposed combining static and dynamic hypergraphs to model skeleton data. Specifically, their dynamic hypergraphs consisted of two important components: dynamic joint weights and dynamic topology. Dynamic joint weighting could assign different coefficients based on the movement distance of each joint point, which was more conducive to the aggregation of joint features; The dynamic topology structure improved the flexibility of the topology structure. Compared with traditional skeleton maps [19, 20], this scheme can generate more topological structures for different samples and deeply mine the implicit association information between joint points. But the network has a high computational cost in the process of obtaining dynamic hypergraphs. In addition, assigning joint weights in the hypergraph based on the distance of joint movement may lead to biased weight issues, that is, if some joints move significantly more than others, the weight associated with these joints may increase. Therefore, the model overly focuses on joints with higher weights, which affects the overall performance of the model.

From another perspective, Shi et al. [7] proposed a dual stream adaptive graph convolutional network (2 s-AGCN), in which they designed three sub matrices named A, B, and C. The submatrix A was defined as the physical structure of the human body. The submatrix B is data-driven learning entirely based on training data and worked as a global graphic. The submatrix C is a data dependent graph, which is specific for each sample. Finally, the sum of A, B, and C is used for spatial feature extraction.

But, 2 s-AGCN uses temporal convolution network (TCN) with fixed receptive fields to simulate joint-level motion in a specific time range, ignoring the benefits of modeling multi-level motion patterns in dynamic receptive fields. Differently, Duan et al. [25] proposed a Dynamic Group Time ConvNet with different receptive fields to model spatial–temporal feature from skeleton data. For spatial feature mining, they designed their spatial matrix consisting of one static sub-matrix and two dynamic matrices, among them, the static sub-matrix was learned based on the whole training data-set, and proved to play dominant role in spatial feature extraction; on the other hand, the other two dynamic matrices were learned in a data-dependent manner, which were channel invariant and channel specific respectively, and worked as auxiliary components in spatial feature description. In terms of temporal feature representation, different form TCN in 2 s-AGCN [7], they modeled joint-level and skeleton-level features in parallel using multiple sets of temporal kernels. However, this method is only used for time modeling under the same joint, and can not effectively capture the correlation between different joints in continuous frames.

Comparably, Shi et al. [26] proposed a Multi-Stream Attention-enhanced Adaptive Graph Convolutional Neural network (MS-AAGCN), which is the upgraded version of 2 s-AGCN [7] with the same authors. Firstly, for final spatial adjacent matrix definition, they did not use sub-matrix A, instead, they chose B+αC, the parameterized coefficient α was learned and updated under their Gating mechanism. Secondly, for temporal feature modelling, they designed a Spatial–Temporal channel attention module (STC-attention module), which consisted of three sub-modules: spatial attention module, temporal attention module and channel attention module. Similar to [26], Shi et al. [27] also used the attention mechanism to model the spatio-temporal correlation among joints, without considering the position relationship between joints, and without relying on any graph topologies. Specifically, their DSTA-Net (Decoupled Spatial–Temporal Attention Network) network used spatio-temporal attention decoupling module to decompose the spatio-temporal features into spatial features and temporal features, then they modeled the spatio-temporal dependence of the joints respectively through the self-attention mechanism.

Although DSTA-Net enhances the global modeling ability of joint data, it adopts attention mechanism to extract the spatial and temporal features of joints in space and time dimensions, it does not take into account the differences of skeleton data in space and time dimensions. This difference refers to the temporal dimension of joint sequences emphasizing more characteristics related to movement, for example, the speed, frequency, and rhythm of movement of different body parts may be different. To this end, we propose our decoupled adaptive graph convolutional networks, it is explained as follows.

3 The proposed method

As shown in Fig. 1, our algorithm framework inherits the serial structure of ST-GCN, which consists of two parts: Backbone module and Decouple Head. The skeleton sequence with the N number of batches, C channels, T time frames and V nodes is arranged into a spatiotemporal map and input to the backbone module. Our backbone consists of two sub-modules: DAGCN and RMTCN. The DAGCN can learn the graph convolution kernel adaptively and extract the spatial features of multiple channels effectively, while the RMTCN module is designed to effectively extract multi-scale temporal information.

The feature information extracted by Backbone includes spatial and temporal information, the previous methods, such as AGC-LSTM [19], ST-GCN [20], and 2 s-AGCN [7], MS-AAGCN [26], etc., only simply take Mean pooling to merge the information, which causes the loss of detailed spatio-temporal feature. To this end, Decoupled Head is designed to decompose the spatio-temporal feature into three components as space, time and space–time, they are respectively averaged and then added together for classification. The new average pooling layer can effectively reduce the noise, and weaken the sensitivity to abnormal data, which can improve the robustness and stability of the model. The above three key parts are introduced in detail as follows.

3.1 Decoupled adaptive graph convolutional network

In the GCN-based skeleton behavior recognition, when designing its graphic convolutional kernel, it is crucial to measure the relations among adjacent nodes, as well as important cross-correlations and mutual influences among non-adjacent nodes. Take "running" for an example, there are close interactions and coordination between the hands and feet, even though they are not physically adjacent joints. For more details to explain their coordination, when a person steps with the left foot, usually his right hand will wave forward to ensure the balance of the body. However, present algorithms [4, 6, 21–23] fail to capture this correlation diversity when constructing the adjacency matrix, which results in an inability to accurately represent the movement of an action. Therefore, a new model structure is proposed as shown in Fig. 2, which can obtain more abundant dynamic association information. The specific method is as follows.

Firstly, DAGCN input data $X_{in} \in R^{N\times C\times T\times V}$ is normalized via the BN module; Secondly, three 1×1 convolutions are used for feature embedding to obtain three third-order tensors respectively as $X_q, X_k, X_v \in R^{H\times T\times V}$. The adaptive adjacency matrix $A_{att} \in R^{H\times V\times V}$ can be obtained according to the non-local operation shown in formula (1), it is merged with the predefined adjacency matrix A_0, which is derived from the human body physical connected joints, addressed as strong correlation matrix or explicit correlation matrix, as used in ST-GCN related methods. Thirdly, the merged adjacency matrix performs adaptive graph convolution operation, as shown in formula (2), to obtain unbiased spatial feature representation. Finally, SENet is used to assign weights to the channel layer and output the final spatial feature results.

formula (1) and (2) are explained as follows respectively:

$$A_{att}^{(m)} = soft\max\left(\frac{\left(X^{(m)}\cdot W_q^{(m)}\right)\left(X^{(m)}\cdot W_k^{(m)}\right)^T}{\sqrt{d_k^{(m)}}}\right),\quad (\mathrm{m} = 1, \ldots, \mathrm{M}) \tag{1}$$

$$f_{out} = \left(A_{att}^{(m)} + A_0\right)\cdot\left(X^{(m)}\cdot W_v^{(m)}\right) \tag{2}$$

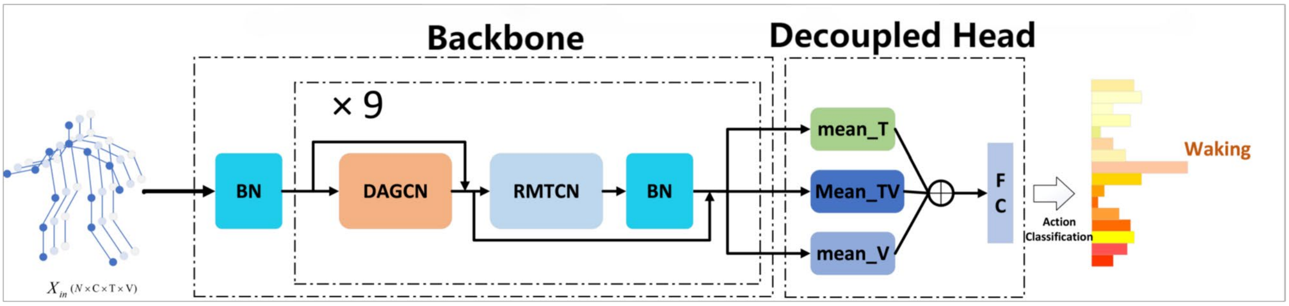


Fig. 1 The overall structure of DAGCN-RMTCN algorithm

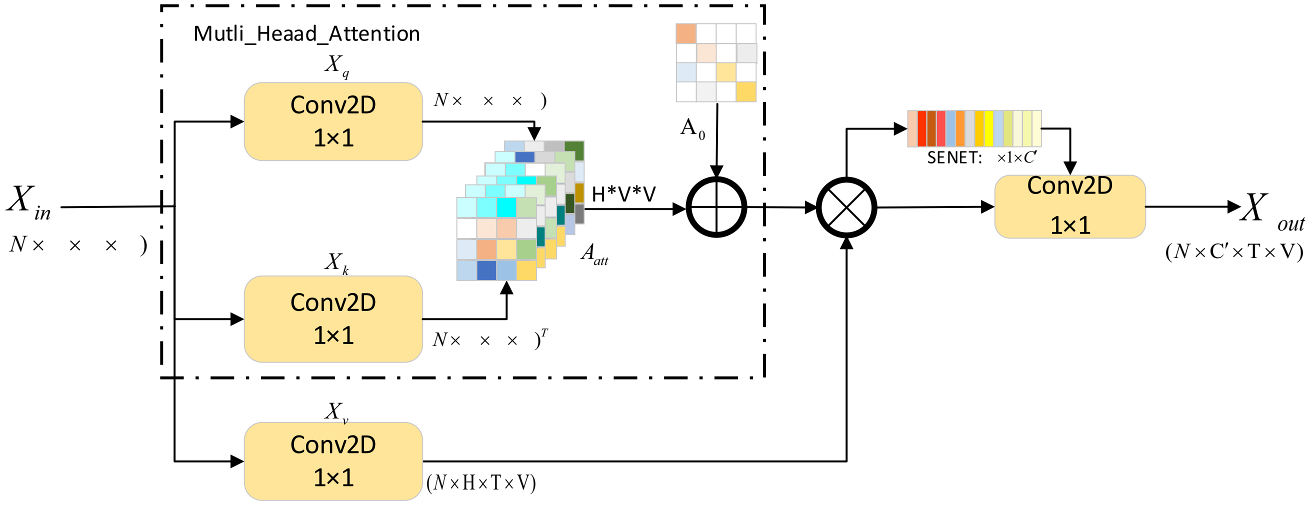

Mutli_Heaad_Attention
X_q
Conv2D 1×1
$N\times \times \times)$
A_0
X_{in}
$N\times \times \times)$
H*V*V
A_{att}
X_k
Conv2D 1×1
$N\times \times \times)^T$
SENET: $\times 1\times C'$
Conv2D 1×1
X_{out}
$(N\times C'\times T\times V)$
X_v
Conv2D 1×1
$(N\times H\times T\times V)$

Fig. 2 Decoupled Adaptive Graph Convolutional Module

where $A_{att} \in R^{H\times V\times V}$ essentially is the multi-head adjacency matrix, $X_{in} \in R^{N\times C\times T\times V}$ is the input data; $W_q \in R^{H\times C_{in}\times 1\times 1}$,$W_k \in R^{H\times C_{in}\times 1\times 1}$ and $W_v \in R^{H\times C_{in}\times 1\times 1}$ are expressed as the weight parameters of the 1×1 convolution kernel respectively, they are multiplied with X to get three tensors respectively as $X_q, X_k, X_v \in R^{H\times T\times V}$; H is the number of multi-heads of the attention mechanism, we use the number of channels transformed by Conv2D 1×1 as its number; and d_k is the number of nodes; m is the index number of layers in the intermediate process, M equalizes 9 as indicated in Fig. 1. The output of DAGCN in Fig. 2 is X_{out} with the dimension of $N \times C\prime \times T \times V_{,}$ $C\prime$ is the output channel number.

In summary, Compared with the adaptive adjacency matrix of AGCN [7] and MS-AAGCN [26], this paper has a decoupling method to design multi-channel adaptive adjacency matrices. It can figure out a more diversified adaptive adjacency matrix, which can obtain more connection relations of nodes under the behavior change, so that the graph has a stronger ability to convolve and aggregate node information. Further, compared with the traditional Transformer method [23], our method proposed in this paper is simpler and more flexible in reducing the original two-dimensional skeleton sequences to one-dimensional sequences for adaptive adjacency matrix calculation.

3.2 Residual multi-scale time convolutional network module

Traditional Temporal Convolution Network (TCN) uses k×1 convolution kernel to aggregate temporal information in the time dimension of data. Due to the fixed scale of its convolution kernel, it is difficult to extract temporal features comprehensively. Classical multi-scale time convolution based on Inception structure can obtain more time sequence information, but there are some problems. For example, Liu et al. [28] proposed MS-TCN to obtain long distance time information by setting different void rates. Although the range of sensitivity field is expanded, the correlation between the information obtained by long distance convolution is weak due to sparse sampling input information of void convolution, thus obtaining incoherent global information. Take the "standing up" action for an instance, its duration time is short, i.e. 0.5 s or so. If employing dilated convolution, only the information at 0.1 s and 0.3 s can be sampled, but missing the most important information at 0.2 s. Therefore, inspired by Res2net and Inception structures, residual multi-scale time convolution (RMTCN) is proposed in this paper as Fig. 3. Its specific methods are explained as follows.

As shown in Fig. 3, RMTCN mainly consists of two scales of convolution $\text{kernel} = \{3 \times 1, 5 \times 1\}$ and one $\text{dilation} = \{1 \times 2\}$ kernel, which can reach the receptive field of 7×1 and 13×1 respectively. Besides, we add residual structure to solve the problem of network degradation. In this paper, the DAGCN output features are equally divided into $\text{S} = 4$ sub-groups as the inputs to the four sub-branches in RMTCN respectively, that is, each subgroup contains a quarter of channel numbers. Among them, the first three sub-groups are convoluted by 1×1, 3×1, 5×1, kernels respectively; while the forth sub-group is processed via a 3×1 Max Pooling. Finally, these four sub-outputs are concatenated as a whole spatial–temporal feature into next module of STMP for classification.

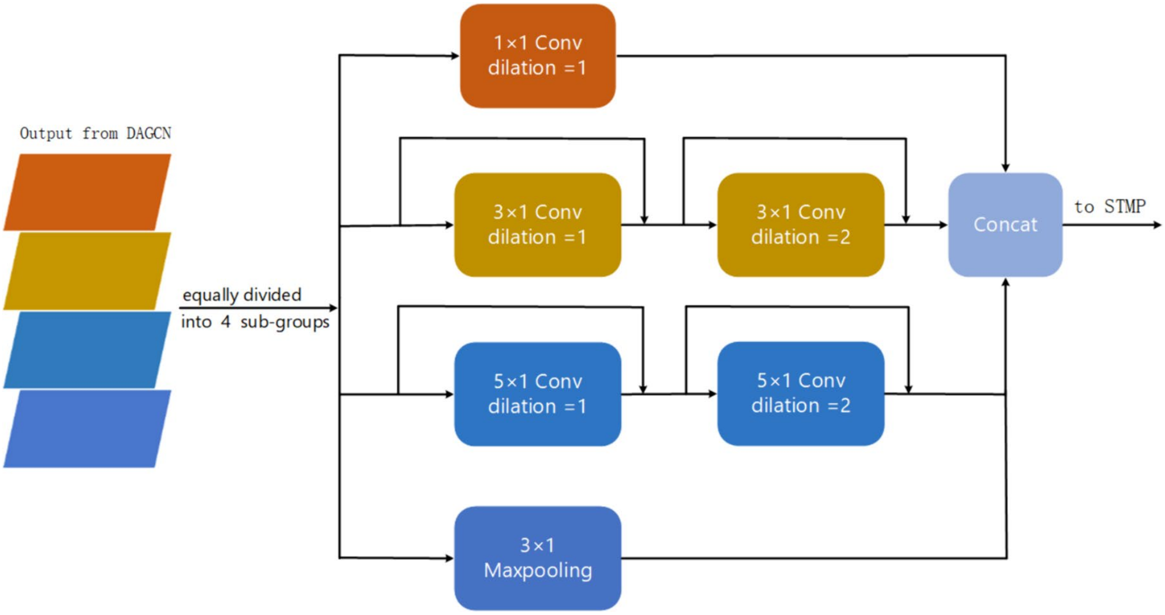


Fig. 3 Residual multi-scale time convolution network (RMTCN)

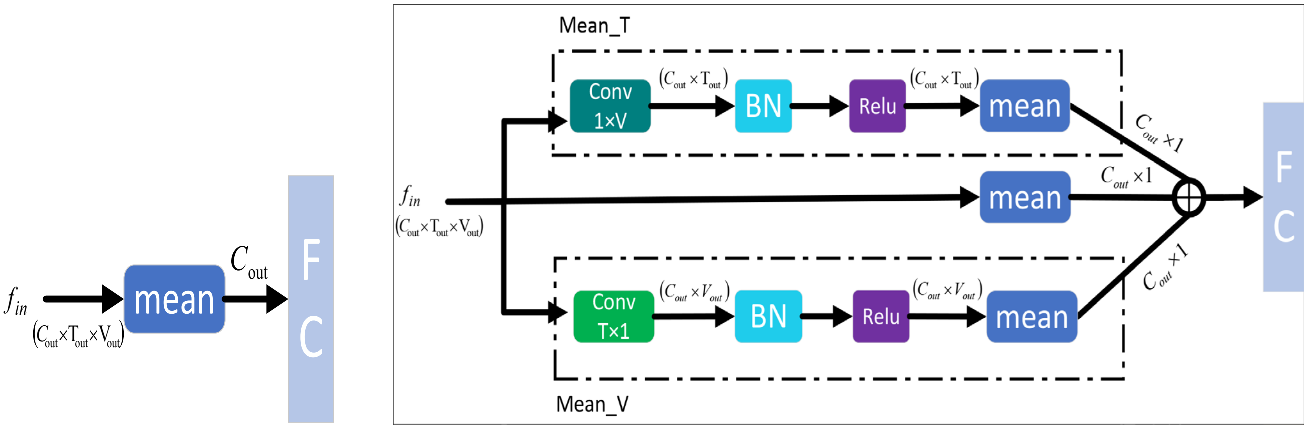
Different from concurrent structures that use different convolution kernels and dilated convolutions to improve multi-scale capability, we use hierarchical residual-like connections on each branch, which has the advantage that the multi-scale structure does not have the problem of gradient disappearance and less information loss as the depth of the network increases. Our RMTCN uses the standard residual mechanism to fuse low-level and high-level temporal features. The specific formulas are as follows:

$$F_{i,j} = Conv\left(K_i \times 1,\ d_i\right)\left(X_{ij}\right) + F_{(i,j-1)i=1,2,\ldots,S-1j=1,2} \tag{3}$$

$$F_{S-1} = pool\left(Conv(3 \times 1,\ d = 1)\left(X_{S-1}\right)\right) \tag{4}$$

$$F = Concat\left(\left[F_0,\ F_1, \ldots,\ F_{S-1}\right]\right) \tag{5}$$

Among them, $F_{i,j}$ stands for the feature map obtained by the K_i convolution and the d_i dilation rate. F_{S-1} is the Max Pooling and F represents the multi-scale feature map.


f_{in}
$(C_{out}\times T_{out}\times V_{out})$
mean
C_{out}
FC
Mean_T
Conv 1×V
$(C_{out}\times T_{out})$
BN
Relu
$(C_{out}\times T_{out})$
mean
$C_{out}\times 1$
f_{in}
$(C_{out}\times T_{out}\times V_{out})$
mean
$C_{out}\times 1$
Conv T×1
$(C_{out}\times V_{out})$
BN
Relu
$(C_{out}\times V_{out})$
mean
$C_{out}\times 1$
FC
Mean_V

（a） （b）

Fig. 4 Classifier based on STMP (spatio-temporal mean pooling) module

In short, Using dilation convolution can further increase the receptive field of the convolution kernel, while reducing the number of parameters and computational complexity. By adjusting the sampling rate of the dilation convolution, the receptive field size of the convolution operation can be flexibly controlled to adapt the requirements of feature extraction at different scales.

3.3 Spatio-temporal decoupling classifier module

The operation of mean pooling on spatio-temporal features in classification layer, as shown in Fig. 4a, will result in losing information and degrade recognition accuracy. In order to retain more spatio-temporal information and highlight the difference of spatio-temporal characteristics in reverse derivation after summation of these three loss functions, we propose a spatio-temporal decoupling classifier based on STMP module, as shown in Fig. 4b.

In backbone, DAGCN and RMTCN modules only aggregate information in space and time domain respectively, we suppose their output feature map as $Z \in R^{C_{out} \times T_{out} \times V_{out}}$. In the classifier, we modify the traditional global average pooling into the average pooling of space, time and space–time in parallel. That is, $Z \in R^{C_{out} \times T_{out} \times V_{out}}$ is decomposed as the spatial sub-feature $Z_V \in R^{C_{out} \times V_{out}}$ by a $T_{out} \times 1$ convolution operation, and temporal sub-feature $Z_T \in R^{C_{out} \times T_{out}}$ by a $V_{out} \times 1$ convolution operation, and spatial–temporal feature $Z_{TV} \in R^{C_{out} \times T_{out} \times V_{out}}$ as unchanged. because Γ number of spatial–temporal features are disassembled into the same number of spatial feature and temporal feature respectively, so their combining loss function can be written out as follows:

$$\mathrm{H(p,q)} = -\frac{1}{3}\Big[\sum_{\mathrm{i=1}}^{\Gamma} \mathrm{p(Z_{Vi})log(q(Z_{Vi}))} + \sum_{\mathrm{i=1}}^{\Gamma} \mathrm{p(Z_{Ti})log(q(Z_{Ti}))} + \sum_{\mathrm{i=1}}^{\Gamma} \mathrm{p(Z_{TVi})log(q(Z_{TVi}))}\Big] \tag{6}$$

In the process of parameter updating and gradient calculation by backpropagation algorithm, parameter updating will be optimized in three dimensions respectively, which can help spatio-temporal features obtain richer feature information.

Among them, the temporal sub-feature undergoes a $1 \times \mathrm{V}$ convolution kernel, BN layer and Relu output, and averagely pooled in the time dimension, then we can obtain its temporal averaged output$Z_{TA} \in R^{C_{out}}$. Similarly, we can get spatial averaged pooling output $Z_{VA} \in R^{C_{out}}$. But spatial–temporal feature is directly averaged as$Z_{TVA} \in R^{C_{out}}$. These three outputs are added as a whole for classification. In all, by above processing, more abundant meaningful information are retained in each feature domain, while their feature dimensions are also reduced accordingly.

4 Experimental results and their analysis

The proposed algorithm will be tested in this section, specifically, in Sect. 4.1, three related skeleton datasets, the NTU RGB+D 60 dataset [29], the NTU RGB + D 120 dataset [30], and the Northwestern-UCLA dataset [31], are introduced. In Sect. 4.2, the details of experimental settings and parameter settings are described. In Sect. 4.3, ablation experiments are conducted. Afterwards, in Sect. 4.4, our algorithm is compared with the latest methods, and the results are analyzed in depth. Finally, we give our conclusions and talk about limitations of our method in Sect. 5.

4.1 Skeleton behavior recognition datasets

4.1.1 NTU RGB + D 60

It contains 60 behavior classes, 40 person IDs, and a total of 56,880 RGB+D video samples, including RGB video, depth sequences, 3D skeleton data, and infrared frames. The skeleton information for each sample contains the 3D coordinates of 25 body joints.

The NTU-60 dataset follows two different criteria for dividing the training set and the test set. The first is called Cross-View (X-View) camera, in which, the second and third cameras are divided as training sets with a total of 37,920 samples, and the first camera is treated as test sets with a total of 18,960 test samples. The second criteria is divided by personnel ID, which is called Cross-Subject (X-Sub). According to ID, it is divided into training set and test set. The training set consists of 40,320 training samples and the test set consists of 26,560 test samples.

4.1.2 NTU RGB + D 120

It is an extension of NTU RGB+D 60, which newly adds 57,367 skeleton sequences representing 60 new behaviors for a total of 113,945 videos, and 120 categories from 106 subjects and 32 camera settings.

There are also two different standards for dividing the training and testing sets in NTU-RGB+d120. The first type is Cross subject (X-sub120), which is allocated based on different subject groups. The training set contains 63,026 samples, and the test set contains 50,922 samples. The second method is Cross setting (X-set120): based on the camera ID, the samples are divided into a training set and a testing set, where even numbered cameras capture samples as the training set and odd numbered cameras capture samples as the testing set. Specifically, the training set contains 5447 samples, and the testing set contains 59,477 samples.

4.1.3 Northwestern-UCLA

The Northwestern-UCLA dataset consists of three Kinect cameras capturing video content simultaneously. It has a total collection of 1494 video clips covering 10 action categories. Each action was performed by 10 different subjects. According to the method provided in the literature [26] as an evaluation protocol: training data containing the first two cameras, test data including other camera.

4.2 Experimental parameter Settings

All experiments are programmed under Pytorch framework and conducted with the above three data sets. That is, in terms of NTU RGB+D 60 and the NTU RGB+D 120 dataset, the proposed model is trained with a total of 100 epochs with batch size of 32 and SGD as the optimizer, its learning rate is set to 0.1 at the beginning and then reduced by a factor of 10 at epochs {60, 90}, the weight decay is set to 5E-4. For Northwestern UCLA dataset, its batch size, epochs, learning rate, weight decay and reduced step are set to 64, 110, 0.2, 4E-4 and {90, 100} respectively. In addition, we use an improved version of focal loss to suppress the weights of easy to classify samples, and increase the weights of difficult to classify samples, that is, when PT (positive truth) is close to 1, (1-PT) becomes smaller, the weight in calculating the loss is also smaller, thus we can reduce the influence of easily classified samples on the total loss.

4.3 Ablation experiments

In this section, the Decouple Adaptive Graph Convolution module (DAGCN), Residual Multi-scale Temporal Convolution module (RMTCN), and Spatio-Temporal Mean Pooling module (STMP) are test to evaluate their performance respectively with behavior prediction accuracy. The NTU RGB+D 60 data set is chosen in the experiments, with X-View and X-subject criterion respectively to divide training and testing samples. The performances of ST-GCN [20], AGCN [7] and its counterpart of 2S-AAGCN [26] are compared as baselines. The details are as follows.

4.3.1 DAGCN Ablation test

We choose AGCN [7] as the baseline and replace its adaptive graph convolution kernel with our DAGCN. The experiments are conducted on dataset of X-View and X-Subject respectively with streams of joint and bone (JB-streams), so the symbol of JB-AGCN-DAGCN (ours) is used to clearly express this testing conditions. For fair evaluations, we compare the similar ablation tests of 2S-AAGCN [26], which is also take AGCN [7] with JB-streams as the baseline. Besides, we also compare our results with those of ST-GCN [20]. The results as shown in Table 1.

Table 1 DAGCN ablation tests based on 2S-AGCN baseline in the NTU RGB+D 60

Method	X-View Acc (%)	X-Subject Acc (%)
JB-ST-GCN [20]	88.3	81.5
JB-AGCN-B [26]	93.6	86.4
JB-AGCN-C [26]	93.5	86.1
JB-AGCN-BC [26]	94.1	87.0
JB-AGCN-BC-G [26]	94.4	87.4
JB-AGCN-DAGCN wo/A_0	93.8	86.7
JB-AGCN-DAGCN (ours)	94.9	88.0
Our (DAGCN+RMTCN+STMP)-JB	96.4	90.7

In terms of X-View criteria, in order to verify the role of the pre-defined human structure based graph convolution kernel A_0, we remove A_0 from DAGCN, that is, DAGCN wo/A_0, then it is transplanted into AGCN [20] baseline, symbolized as JB-AGCN-DAGCN wo/A_0 in Table 1, its performance is better than JB-ST-GCN [20], but decreases by 1.1% in accuracy compared to JB-AGCN-DAGCN (ours) with kernel A_0, which confirms the importance of shared topology A_0 in enhancing explicit association information among joints.

Comparing to ablations of MS-AAGCN [26], whose adjacency matrices of the graph are divided into global graph B and individual graph C, its ablations are also based on AGCN [20] baseline with JB-streams. Specifically, JB-AGCN-B [26] is the ablation for sub-graph of B, similarly, JB-AGCN-C [26] is the ablation of sub-graph of C, JB-AGCN-BC [26] is for B+C; JB-AGCN-BC-G [26] is for B+αC, where the parameterized coefficient α is learned and updated under their gating mechanism, denoted as G. As shown in Table 1, JB-AGCN-DAGCN (ours) can surpass all the ablation results of MS-AAGCN [26]. Especially, better than JB-AGCN-BC-G [26] with 0.5% higher in accuracy, which is the optimal combinations of adjacency matrices of B and C under its gating mechanism G.

In terms of X-subject criteria, our JB-AGCN-DAGCN wo/A_0 is better than ST-GCN [20], but inferior to JB-AGCN-DAGCN (ours), which is with physically defined joint adjacent matrix A_0. Similarly, comparing to ablations of MS-AAGCN [26], JB-AGCN-DAGCN (ours) can perform better than the best ablation test of JB-AGCN-BC-G [26] with 0.6% higher in accuracy.

By comparing above experimental results, it can be concluded that our DAGCN module has a significant effect on improving the recognition of skeleton behavior, because it

Table 2 RMTCN ablation tests in baseline AGCN in the NTU RGB + D 60

NTU RGB + D60		
Model factorized pathway	X-View Acc (%)	X-subject Acc (%)
JB-AGCN [7]	95.1	88.5
JB-AAGCN-STC [26]	95.1	88.0
AGCN + RMTCN (factorized pathway)		
With RMTCN (k = 3, d = (1, 2))	93.8	87.2
With RMTCN (k = 5, d = (1, 2))	94.0	87.7
With 2 RMTCN pathways k = (3, 5), d = (1, 2)	94.3	88.2
DAGCN + RMTCN (factorized pathway)		
With RMTCN (k = 3, d = (1, 2))	94.0	88.7
With RMTCN (k = 5, d = (1, 2))	94.3	89.4
With 2 RMTCN pathways k = (3, 5), d = (1, 2)	95.2	90.2

can effectively capture the rich internal spatial correlation among skeleton joints, even they are not physically connected. At the same time, the use of basic topology A_0 can further enhance the representation of explicit association information.

4.3.2 RMTCN Ablation test

To verify the performance of our RMTCN in obtaining temporal features, AGCN [7] is again used as the baseline, its original temporal convolution (TCN) layer is replaced by our RMTCN, and the parameters and components of RMTCN are adjusted according to the experimental comparisons to gain a better recognition accuracy. Additionally, MS-AAGCN [26], as a better counterpart of AGCN [7], it also designed similar attention mechanism, named Spatial Temporal Channel attention module (STC), its adaptive adjacent matrices B+αC combining with STC ablation is also tested in joint and bone stream, which is denoted as JB-AAGCN-STC [26] in Table 2, which is compared with our DAGCN+RMTCN combination. The detailed experimental results are analyzed as follows.

In terms of X-View criteria, for RMTCN with k = 3 and k = 5 kernels, it can achieve the temporal view of the 7 × 1 and 13 × 1 respectively, when configuring their 2-RMTCN-Pathway into AGCN [7] and cascading it after our DAGCN respectively, the accuracy can be improved obviously compared to that of single pathway. Further, our DAGCN+RMTCN does slightly better than AGCN [7] and its counterpart ablation of MS-AAGCN-STC [26].

In terms of X-subject criteria, similarly, 2-RMTCN-Pathway configured with AGCN and DAGCN respectively, can do better than singe RMTCN pathway works. Noticeably, our DAGCN+RMTCN with 2-RMTCN-Pathway shows obvious superiority compared to AGCN [7] and its counterpart ablation of JB-AAGCN-STC [26], which is 0.7% and 1.2% respectively higher in accuracy.

Table 3 STMP Experiment Comparisons with MP

Method	X-view Acc (%)	X-subject Acc (%)
ST-GCN [20]	88.3	81.5
JB-AGCN [7]	95.1	88.5
JB-AGCN [7]+STMP	95.3	89.2
JB-DAGCN	93.9	88.1
JB-DAGCN+STMP	94.1	88.7
JB-AAGCN [26]	96	89.4
JB-(DAGCN+RMTCN+MP)	95.2	90.2
JB-(DAGCN+RMTCN+STMP)	96.4	90.7

4.3.3 STMP Ablation test

Our STMP test is based on AGCN [7] baseline with joint and bone streams in NTU RGB+D 60, also 2-stream JB-AAGCN [26] and ST-GCN [20] are compared as shown in Table 3.

Specifically, in terms of X-View criteria, the classification accuracy of AGCN+STMP and DAGCN+STMP have been improved significantly by replacing their previous global averaging pooling layer with our spatiotemporal averaging pooling module (STMP). In particular, the recognition accuracy of our proposed model DAGCN+RMTCN+STMP is nearly 1.2% raised compared with that of DAGCN+RMTCN+MP (mean pooling), which is a great improvement. Also, compared to 2-stream JB-AAGCN [26], our DAGCN+RMTCN+MP, without STMP module, is inferior to it with 0.8% lower in accuracy, but our

Table 4 Comparisons on NTU-RGBD 60 with the state-of-the-arts

NTU-RGB+D60				
Methods	Year	Mode	X-Sub Acc (%)	X-View Acc (%)
ST-GCN [20]	2018	2 ensemble	81.5	88.3
Ind-RNN [32]	2018	2 ensemble	81.8	88.0
HCN [33]	2018	2 ensemble	86.5	91.1
2 s-AGCN [7]	2019	2 ensemble	88.5	95.1
AGC-LSTM [19]	2019	2 ensemble	89.2	95.0
2S-AAGCN [26]	2020	2 ensemble	89.4	96
SGN [34]	2020	2 ensemble	89.0	94.5
ST-TR [10]	2021	2 ensemble	89.9	96.1
SNAS-GCN [35]	2023	2 ensemble	89.0	95.0
MADT-GCN [36]	2024	2 ensemble	89.9	96.1
Shift-GCN [37]	2020	4 ensemble	90.7	96.5
MS-AAGCN [26]	2020	4 ensemble	90.0	96.2
DC-GCN+ADG [8]	2020	4 ensemble	90.8	96.6
PA-ResGCN-B19 [38]	2020	4 ensemble	90.9	96.0
Dynamic GCN [39]	2020	4 ensemble	91.5	96.0
MDKA-GCN [40]	2023	4 ensemble	92.1	96.8
EfficientGCN [41]	2023	4 ensemble	91.7	95.7
LKA-GCN [42]	2023	4 ensemble	90.7	96.1
MGCF-Net [43]	2024	4 ensemble	92.7	96.8
MSS-GCN [44]	2024	4 ensemble	92.7	96.9
MADT-GCN [36]	2024	4 ensemble	90.4	96.5
Ours (Joint-AGCN)		Joint	89.2	95.3
Ours (Bone-AGCN)		Bone	89.0	95.0
Ours(2 s)		2 ensemble	90.7	96.4
Ours(4 s)		4 ensemble	91.4	96.5

DAGCN+RMTCN+STMP is better than it with a rising of 0.4% in accuracy.

In terms of X-subject criteria, similarly, AGCN+STMP and DAGCN+STMP show quite better performances than they do with the previous global averaging pooling layer. Also, our 2-stream of DAGCN+RMTCN with STMP is 0.5% better than it without STMP. Compared to JB-AAGCN [26], our 2-stream of DAGCN+RMTCN+STMP is 1.3% raised in accuracy, which is a significant improvement.

4.4 Comparisons with other advanced algorithms

At present, most advanced skeleton-based methods generally adopt the four-streamed input framework, namely joint, joint motion, bone, bone motion. So we compare the performances with them on the following three data sets, NTU RGB+D60, NTU RGB+D120, and NW-UCLA with input of four streams, but also partially including some methods with two-stream input. The details are shown as follows.

4.4.1 Performance comparisons on NTU-RGB + D60

Accuracy statistics of various mainstream algorithms are displayed in Table 4, which are performed on X-Sub and X-View sub-data sets of NTU RGB+D60 respectively. In order to make a thorough comparison, we test our (DAGCN+RMTCN+STMP) method correspondingly in single-stream, double-stream and four-stream, as indicated in Table 4 with the different number of ensembles.

Compared with similar two-streamed algorithms [10, 19, 20, 33–34], and even the well-known methods of 2 s-AGCN [7], 2S-AAGCN [26], and newly published methods of [35, 36], Ours(2 s) performs better than all of them.

In terms of four-streamed methods, ours is better than [8, 26, 37, 38, 42, 36] in X-Sub tests, and also superior to [26, 38, 39, 41, 42] in X-View test, while equals to [37, 36] in X-View test. It is noteworthy that methods of [40, 43, 44] are all better in both X-Sub and X-View tests in NTU-RGB+D60, but they are all inferior to ours in both X-Sub and X-View tests in NTU-RGB+D120 as shown in Table 5, the reasons are analyzed as follows.

Table 5 Comparisons on NTU-RGBD 120 with the state-of-the-arts

NTU-RGB+D120				
Methods	Year	Mode	X-sub Acc (%)	X-set Acc (%)
2S-AGCN [7]	2019	2 ensemble	82.9	84.9
SGN [34]	2020	2 ensemble	79.2	81.5
MS-G3D [28]	2020	2 ensemble	86.9	88.4
ST-TR [10]	2021	2 ensemble	82.7	84.7
MADT-GCN [36]	2024	2 ensemble	85.4	87.4
4S-Shift-GCN [37]	2020	4 ensemble	85.9	87.6
DC-GCN+ADG [8]	2020	4 ensemble	86.5	88.1
DSTA [27]	2020	4 ensemble	86.6	89.0
PA-ResGCN-B19 [38]	2020	4 ensemble	87.3	88.3
Dynamic GCN [39]	2020	4 ensemble	87.3	88.6
MST-GCN [45]	2021	4 ensemble	87.5	88.8
CTR-GCN [46]	2021	4 ensemble	88.9	90.6
TA-CNN [47]	2022	4 ensemble	85.4	86.8
FG-STFormer [48]	2022	4 ensemble	89.0	90.6
Info-GCN [49]	2022	4 ensemble	89.4	90.7
EfficientGCN [41]	2023	4 ensemble	88.3	89.1
LKA-GCN [42]	2023	4 ensemble	86.3	87.8
MDKA-GCN [40]	2023	4 ensemble	87.9	89.4
GSTLN [50]	2023	4 ensemble	88.1	89.3
MGCF-Net [43]	2024	4 ensemble	88.7	90.4
MSS-GCN [44]	2024	4 ensemble	88.9	90.6
MADT-GCN [36]	2024	4 ensemble	86.5	88.2
Ours(2 s)		2 ensemble	87.2	89.1
Ours(4 s)		4 ensemble	89.3	90.6

The method of [40] aimed to fully explore the connections between non-adjacent joints. Specifically, they proposed a dilated convolution attention module to introduce high-level semantics (joint type and frame index) of joints to enhance the spatial feature representation ability. While for temporal feature boosting, it designed a pyramid partition attention module applied to two cascaded TCNs (Temporal Convolution Networks). In terms of the method [43], it proposed a multi-grained clip focus network (MGCF-Net). Namely, a skeleton sequence was divided into multiple clips, each clip was encoded with spatial temporal convolution layers. Further, the action information of intra-clip and inter-clip were explored based on multi-headed self-attention mechanism. As for the scheme of [44], a multi-scale structural graph convolutional network (MSS-GCN) was designed, which could learn a structural graph topology to enhance feature representation and capture semantic correlation between vertices. Besides, a graph weight annealing (GWA) method was proposed to adjust weights between root and neighboring vertices, aiming to mitigate the problem of over-smoothed feature extraction in GCN-based methods. Afterwards these spatial features were undergone a spatio-temporal blocks for action predicting.

Table 6 Comparisons on NW-UCLA with the state-of-the-arts

NW-UCLA			
Methods	Year	Mode	Acc (%)
2S-AGC-LSTM [19]	2019	2 ensemble	93.3
TS-LSTM [51]	2021	2 ensemble	89.2
4S-Shift-GCN [37]	2020	4 ensemble	94.6
DC-GCN+ADG [8]	2020	4 ensemble	95.3
CTR-GCN [46]	2021	4 ensemble	96.5
Info-GCN [49]	2022	4 ensemble	96.6
TA-CNN [47]	2022	4 ensemble	96.1
FR Head [54]	2023	4 ensemble	96.8
GSTLN [50]	2023	4 ensemble	94.8
MSS-GCN [44]	2024	4 ensemble	96.6
MATR-GCN [52]	2024	4 ensemble	96.3
CSR-Net [53]	2024	4 ensemble	96.5
Ours(2 s)		2 ensemble	94.1
Ours(4 s)		4 ensemble	96.4

In short, comparatively, the above three methods work similarly as our modules of DAGCN and RMTCN do in enhancing the spatial temporal feature extraction, but differently, we have additional STMP module, which can further boost their discriminability. So, in the more challenging dataset of NTU-RGB+D120, our method performs better than [40, 43, 44], which is more persuasive than tests in NTU-RGB+D60, and proves our method is more robust and powerful in recognizing ambiguous actions.

4.4.2 Performance comparisons on NTU-RGB + D120

Experimental comparisons are made on X-Sub and X-Set subsets of NTU RGB+D120 respectively, as shown in Table 5. Because NTU RGB+D120 dataset is new, the most recent algorithms are four-streamed experiments, compared with the results of the existing publicly available two-flow experiments [7, 10, 28, 34, 36], the accuracy of our model with two streams is significantly better than all of them.

In terms of four-ensembled mode testing, only the method [49] is 0.1% higher in accuracy than ours in both X-Sub and X-Set testing, while the rest sixteen methods in Table 5 are all inferior to our proposed algorithm. The merits of method

[49] was derived based on its complex network. Firstly, it introduced a learning objective based on information-bottleneck theory, which could encode implicit and general latent representation from a sequence of the skeleton, this bridged the input-level physical information and action semantics. Secondly, a self-attention based graph convolution module was designed in encoding stage, aiming to extract the intrinsic graph structure among skeleton joints. Lastly, this method proposed to utilize the relative positions information between joints, and forming a multi-modal skeleton representation, which could further provide complementary spatial information of a joint and drastically improved its recognition performance.

4.4.3 Performance comparisons on NW-UCLA

The experimental comparisons are performed on the NW-UCLA dataset, as shown in Table 6. Ours(2 s) is obviously better than two-flow methods [19, 51]. In terms of four-streamed mode, our method is obviously better than [8, 37, 47, 50], and slightly superior to [52]. On the other hand, methods [46, 49, 44, 53] are slightly better than ours with 0.1%-0.2% higher in accuracy. Noteworthily, the method [54] is with 0.4% better in accuracy than ours, the reasons are analyzed as follows.

In the method [54], a Feature Refinement module (FR Head) was designed to improve the recognition performance of ambiguous actions. Specifically, contrastive learning was adopted to constrain the distance between confident action samples and ambiguous action samples. Besides, the raw feature map was decoupled into spatial and temporal components for efficient feature enhancement. Finally, this module was imposed on different stages of GCNs to build a multi-level refinement for an efficient and robust skeleton representations.

In short, the performance gap between the superior methods and ours is not large, which is implicit that our model still has potential and room for improvement. we could explore optimizing model architecture, parameter tuning, and feature extraction to improve our performance on the NW-UCLA dataset.

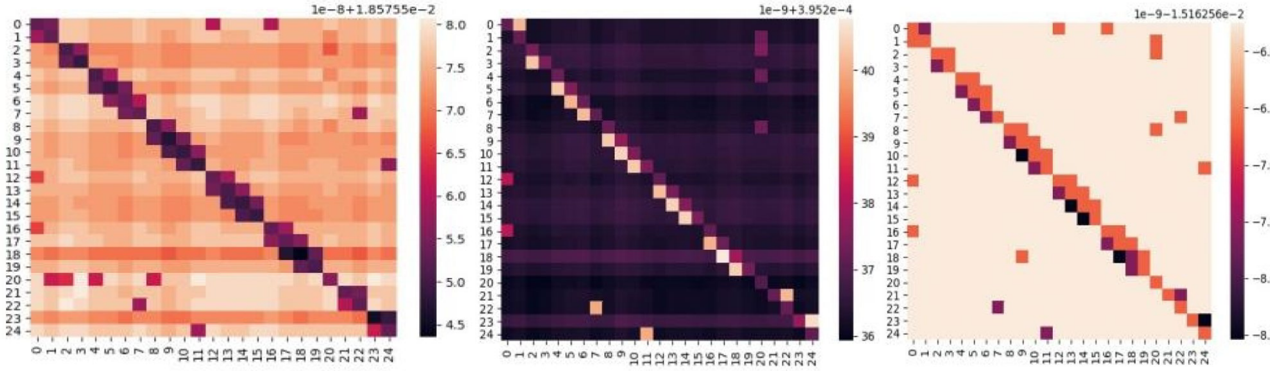
4.5 Visual analysis of adjacency matrix of DAGCN

The core of skeleton data behavior recognition based on GCN is to reasonably construct adjacency matrices, aiming to fully extract the spatial information of multi-level skeleton joints. Taking walking behavior in the NTU RGB+D60 data set as an example, in Fig. 5, we show the visualization results of three groups of typical adjacency matrices generated by our DAGCN algorithm under the multi-head self-attention mechanism.

Since each skeleton sample in the NTU RGB+D60 dataset consists of 25 body joints, so the adjacency matrix is of 25×25. Among them, Fig. 5b is the adjacency matrix based on the physical structure of the human body joints, which highlights the strong correlation between two physically connected joints; while on the diagonal, the auto-correlation of each joint is weak.

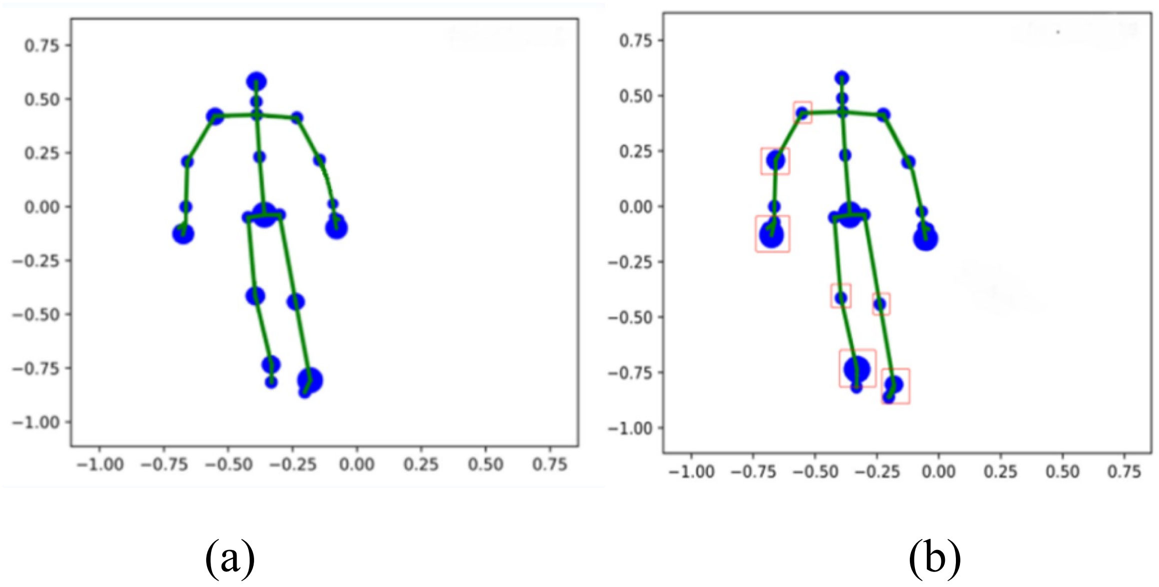
In contrast, from another point of view, we show a quite different adjacency matrix in Fig. 5a, in which, the self-correlation of a joints is high, while the cross-correlation among its disconnected joints is weak. Only sparsely appeared in Fig. 5a, there are several strong cross-association, such as, between the 11th joint and the 24th joint, the 22th and the 7th, the 0th and the 16th.

The more sparsely strong cross-association between disconnected joints is shown in Fig. 5c, indicating that at the current attention level, the average cross-correlation level



(a) (b) (c)

Fig. 5 Visualization of Adjacency Matrices constructed with DAGCN



(a)

(b)

Fig. 6 Visualization of key joints shifting in the walking action

is very weak, only a few disconnected joints are of certain strong associations.

In short, DAGCN can obtain different adjacency matrices by learning multiple self-attention levels, which can give a comprehensive representation of behavior dynamics and posture shifting from one to another.

4.6 Visualization of key joints exchanging in different postures

As shown in Fig. 6, the weights of the adjacency matrix are normalized in order to visualize the changes of different key joints under different postures while in walking action, that is, (a) and (b) are two gestures of "walking", their important joints are marked with different sized solid circles, the bigger, the more important. The joins in the red box in Fig. 6b can clearly display the differences from those correspondingly in Fig. 6a. For an example, when a person steps forward with the left foot (left ankle joint), usually his right hand (right twist joint) also will wave forward to ensure the balance of the body, as shown in Fig. 6b, these two joints cooperate each other, they have strong association, also they are key joints. Anyway, the visualization of key joints shifting indicates the power of diverse adjacency matrices to explore different level of information hidden in dynamic actions.

5 Conclusion and future work

In this section, there are two parts, firstly, we make an thorough comparisons with baseline method AGCN [7] and its counterpart method [26] to further sum up our conclusion. Secondly we talk about the our limitations and future work.

5.1 Similarities and differences among our method with similar baseline algorithms

For the baseline methods of AGCN [7] and its counterpart MS-AAGCN [26], We have talked about their theory, advantages and disadvantages in Sect. 2 of related work, and made recognition accuracy comparisons with our method in Sect. 4 of experimental results and their analysis. Here, we sublimate their main similarities and differences.

They all propose a kind of method to design adaptive adjacent matrices, wish to solve the problems in existing GCN-based methods, that is, the topology of the graph is set manually, and it is fixed over all layers and input samples. This surely is not optimal for the hierarchical GCN and diverse samples in action recognition tasks.

AGCN [7] and MS-AAGCN [26] are counterparts proposed by the same authors, their adaptive adjacent matrices consist of three sub-matrices, named A, B, C. Among them, sub-matrix A is defined with the physical structure of the human body; sub-matrix B is completely data-driven learned according to the training data, which can learn graphs fully targeted to the recognition task; sub-matrix C is a data-dependent graph which learn a unique graph for each sample. For AGCN [7], their final adaptive adjacent matrix is the sum of the three sub-matrices; but for MS-AAGCN [26], they proposed a gating mechanism to fuse $B + \alpha C$, where α is unique for each layer and is learned in the training process.

Similarly, our adaptive adjacent matrices contain a matrix A_0, which is similar to the above sub-matrix A determined with the physical structure of the human body. But different from them, we use multi-head self-attention mechanism to generate more than three adaptive adjacent matrices, actually 64 matrices are applied, which can roundly discover more potential correlations among skeleton joints.

On the contrary, we do not have sub-matrix B, which is a global graph completely learned according to the training data. In the view of our points, sub-matrix B is general and may benefit for all types of actions. But in our method, we have matrix A_0, which can work as a general and basic global graph to guide a better spatial feature extraction.

Beside, we have RMTCN and STMP modules, while MS-AAGCN [26] contains a spatial–temporal channel attention module (STC-attention module), which helps their model pay more attention to important joints, frames and features. We think their STC-attention module will be very helpful to raise our accuracy if replacing our RMTCN module. On the other hand, our STMP module may bring about improvements if configured into MS-AAGCN.

5.2 Limitations and future work

This algorithm has the following shortcomings. Firstly, the overall network structure determines that the extraction of behavioral features of the skeleton data is carried out asynchronously, that is, the spatial features are extracted by DAGCN module firstly, and then the inter-frame time sequence features are extracted by RMTCN. This results in computational inefficiency.

Further, compared with the spatial characteristics, inter-frame temporal feature is more important to differ action categories, however, this asynchronous cascaded structure will inhibit the overall temporal feature capturing. Besides, our accuracy needs to be improved. it is partially due to the spatial feature description, determined by the number of self-attention heads setting in DAGCN module empirically; also, inspired by spatial–temporal channel attention module (STC-attention module) proposed in MS-AAGCN [26], our RMTCN can be optimized to more focus on discriminative characteristics for a better action representation. So, our future work is to improve DAGCN, especially RMTCN, for enhancing the none local feature extraction and also gradually optimize its calculation cost.

Acknowledgements All authors agree to submit the manuscript with the name list appeared in the title page.

Author contributions Haigang Deng and Guocheng Lin designed the Decoupled Adaptive Graph Convolutional Network and Residual Multi-scale Temporal Convolution Network, and parepared corresponding tests; Haigang Deng also wrote Sect. 4.1, 4.2, 4.3; and Guocheng Lin wrote Sect. 4.4, 4.5, 4.6. Chengwei Li wrote introduction Section, related work and conclusion section. Chuanxu Wang wrote Sect. 3.1 and 3.2. Wenting Xu wrote Sect. 3.3.

Funding National Natural Science Foundation of China, 61672305.

Data availability Our code and data were uploaded onto Githuab at the address: https://github.com/linguo123/DAGCN_behavior_recognition.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. Huang J, Xiang X, Gong X, Zhang B (2020) Long-short graph memory network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 645–652
2. Sheng L, Tingting J, Tiejun H, Yonghong T (2020) Global co-occurrence feature learning and active coordinate system conversion for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp 586–59416
3. Du Y, Fu Y and Wang L (2015) Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) IEEE, pp 579–583
4. Li C, Zhong Q, Xie D and Pu S (2017) Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) IEEE, pp 597–600
5. Zhu A, Wu Q, Cui R, Wang T, Hang W, Hua GAND, Snoussi H (2020) Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN. Neurocomputing 414:90–100
6. Papadopoulos K, Ghorbel E, Aouada D et al. (2021) Vertex feature encoding and hierarchical temporal modeling in a spatio-temporal graph convolutional network for action recognition. In: 25th International Conference on Pattern Recognition (ICPR). IEEE, pp 452–458
7. Shi L, Zhang Z, Cheng J and Lu H (2019) Two stream adaptive graph convolutional networks for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 12026–12035
8. Cheng K, Zhang Y, Cao C, Shi L, Cheng J and Lu H (2020) Decoupling gcn with dropgraph module for skeleton-based action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV)

9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst. https://doi.org/10.1007/978-3-030-58586-0_32
10. Plizzari C, Cannici M, Matteucci M (2021) Skeleton-based action recognition via spatial and temporal transformer networks. Comput Vis Image Underst 208:103219
11. Wang Q, Peng J, Shi S et al. (2021) Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition. arXiv preprint arXiv:2110.13385
12. Sekaran RS, Pang YH, Ling GF et al. (2022) MSTCN: a multi-scale temporal convolutional network for user independent human activity recognition. F1000Research. https://doi.org/10.12688/f1000research.73175.2
13. Du Y, Fu Y, Wang L (2015) Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, pp 579–583
14. Wang P, Li Z, Hou Y et al. (2016) Action recognition based on joint trajectory maps using convolutional neural networks. In: Proceedings of the 24th ACM international conference on Multimedia. pp 102–106
15. Zhu K, Wang R, Zhao Q, Cheng J, Tao D (2020) A cuboid CNN model with an attention mechanism for skeleton-based action recognition. IEEE Trans Multimedia 22(11):2977–2989. https://doi.org/10.1109/TMM.2019.2962304
16. Du Y, Wang W and Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1110–1118
17. Liu J, Wang G, Duan L-Y, Abdiyeva KAND, Kot AC (2017) Skeleton-based human action recognition with global contexta-ware attention LSTM networks. IEEE Trans Image Process 27(4):1586–1599
18. Wei S, Song Y and Zhang Y (2017, September) Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 91–95
19. Si C, Chen W, Wang W, Wang L and Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
20. Sijie S, Xiong Y and Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol 32. no 1
21. Lee J, Lee M, Lee D et al. (2023) Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp 10444–10453
22. Yang Z, Li K, Gan H et al. (2023) HD-GCN: A Hybrid Diffusion Graph Convolutional Network. arXiv preprint arXiv:2303.17966
23. Zhang Y, Wu B, Li W et al. (2021) STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp 3229–3237
24. Wei J, Wang Y, Guo M, et al. (2021) Dynamic hypergraph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv: 2112.10570
25. Haodong D et al. (2022) DG-STGCN: dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv preprint arXiv:2210.05895
26. Shi L, Zhang Y, Cheng J et al (2020) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans Image Process 29:9532–9545
27. Shi L, Zhang Y, Cheng J et al. (2020) Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of the Asian Conference on Computer Vision
28. Liu Z, Zhang H, Chen Z, Wang Z and Ouyang W (2020) MS-G3D: disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 143–152
29. Shahroudy A, Liu J,Ng T-T and Wang G (June 2016) Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition
30. Liu J, Shahroudy A, Perez ML, Wang G, Duan L-Y, Chichung AK (2019) Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell 42:2684
31. Wang J, Liu Z, Ying Wu, Yuan J (2013) Learning actionlet ensemble for 3D human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927
32. Li S, Li W, Cook C, Zhu C and Gao Y (2018) Independently recurrent neural network (indrnn): building a longer and deeper rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 5457–5466
33. Li C, Zhong Q, Xie D et al. (2018) Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv preprint arXiv:1804.06055, pp 786–792
34. Zhang P, Lan C, Zeng W, Xing J, Xue J and Zheng N (2020) Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 1112–1121
35. Jiang Y, Yu S, Wang T, Sun Z, Wang S (2023) Skeleton-based human action recognition based on single path one-shot neural architecture search. Electronics 12(14):3156
36. Yu X et al (2024) Skeleton-based action recognition based on multidimensional adaptive dynamic temporal graph convolutional network. Eng Appl Artif Intell 127:107210
37. Cheng K, Zhang Y, He X, Chen W, Cheng J and 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
38. Song Y-F, Zhang Z, Shan C, and Wang L (2020) Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp 1625–1633
39. Ye F, Pu S, Zhong Q, Li C, Xie D and Tang H (2020) Dynamic gcn: context-enriched topology learning for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp 55–63
40. Shu Y, Li W, Li D, Gao K, and Jie B (2023, October) Multi-scale dilated attention graph convolutional network for skeleton-based action recognition. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer Nature Singapore, Singapore. pp 16–28
41. Ong YF, Zhang Z, Shan C et al (2023) Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans Pattern Anal Mach Intell 45:1474–1488
42. Liu Y, Zhang H, Li Y, He K, Xu D (2023) Skeleton-based human action recognition via large-kernel attention graph

convolutional network. IEEE Trans Visual Comput Graph 29(5):2575–2585

43. Qiu H, Hou B (2024) Multi-grained clip focus for skeleton-based action recognition. Pattern Recogn 148:110188
44. Jang S, Lee H, Kim WJ, Lee J, Woo S and Lee S (2024) Multi-scale structural graph convolutional network for skeleton-based action recognition. In: IEEE transactions on circuits and systems for video technology. https://doi.org/10.1109/TCSVT.2024.3375512
45. Chen Z, Li S, Yang B et al (2021) Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. Proc AAAI Conf Artif Intell 35(2):1113–1122
46. Chen Y, Zhang Z, Yuan C, et al. (2021) Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp 13359–13368
47. Xu K, Ye F, Zhong Q et al (2022) Topology-aware convolutional neural network for efficient skeleton-based action recognition. Proc AAAI Conf Artif Intell 36(3):2866–2874
48. Gao Z, Wang P, Lv P, Jiang X, Liu Q, Wang P and Li W (2022) Focal and global spatial-temporal transformer for skeleton-based action recognition. In: Proceedings of the Asian Conference on Computer Vision. pp 382–398
49. Chi H, Ha M- H, Chi S et al. (2022) Infogcn: representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 20186–20196
50. Dai M et al (2023) Global spatio-temporal synergistic topology learning for skeleton-based action recognition. Pattern Recognition 140:109540
51. Lee I, Kim D, Lee S (2021) 3-D human behavior understanding using generalized TS-LSTM networks. IEEE Trans Multimed 23:415–428. https://doi.org/10.1109/TMM.2020.2978637
52. Hu H et al. (2024) Multi-scale Adaptive Graph Convolution Network for Skeleton-based Action Recognition. IEEE Access
53. Yu Z et al. (2024) Cross-scale spatiotemporal refinement learning for skeleton-based action recognition. IEEE signal processing letters
54. Zhou H, Liu Q and Wang Y (2023) Learning discriminative representations for skeleton based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 10608–10617

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.