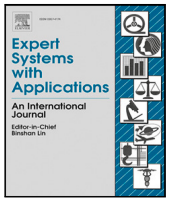
Expert Systems With Applications 257 (2024) 124980

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Multi-granular spatial-temporal synchronous graph convolutional network for robust action recognition

Chang Li [a], Qian Huang [a,*], Yingchi Mao [a], Xing Li [b], Jie Wu [c]

[a] College of Computer Science and Software Engineering, Hohai University, Nanjing, Jiangsu, 211100, China
[b] College of information Science and Technology and College of Artificial Intelligence, Nanjing Forestry University, Nanjing, Jiangsu, 210037, China
[c] Department of Computer and Information Sciences, Temple University, PA, Philadelphia, 19122, USA

ARTICLE INFO

Keywords:
Action recognition
Graph convolutional networks
Spatial–temporal modeling
Multi-granular analysis

ABSTRACT

Graph Convolutional Networks (GCNs) have shown great potential in skeleton-based human action recognition. However, due to the diversity and complexity, modeling human actions as general graphs and capturing discriminative spatial–temporal motion patterns is challenging. Besides, the inevitable interference, especially occlusion, impairs the robustness of existing methods that depend on complete skeletons. To solve these problems, we propose a Multi-Granular Spatial-Temporal Synchronous Graph Convolutional Network (MSS-GCN). Firstly, we investigate three partition strategies: attribute, activity, and mixed partition strategy to optimize the weight-sharing mechanism of GCNs, which facilitates the novel Extended Adaptive Graph Convolution (EAGC) module. Secondly, we elaborate on a Multi-sliced Spatial–temporal Graph (MSTG) for multi-granular action modeling. Thirdly, we present a Synchronized Slice Encoder (Syn-STE) to simultaneously embed spatial and temporal action patterns. Then, we design Multi-granular Spatial–temporal Encoders (Multi-STE) with multi-branch Syn-STE to generate multi-granular context. The extensive experiments verified that MSS-GCN is more robust and outperforms benchmarks on NTU-RGB+D, NTU-RGB+D 120, and NW-UCLA datasets.

1. Introduction

Human action recognition is an active topic in computer vision, which has been widely applied in medical monitoring, elderly health-care, smart home, and intelligent human–computer interaction. Human action recognition aims to automatically interpret the action semantics conveyed by multi-modal data, such as RGB videos, depth videos, and skeleton sequences. Among them, skeleton-based action recognition has drawn increasing attention due to its compactness of representation and robustness to distractions, including appearances, illuminations, viewpoints, and surroundings.

Human skeleton sequences can be collected by sensors, e.g., Kinect (Deng, He, Zhang, & Wang, 2022) or pose estimation algorithms (Li, Zhang, Zhang, & Xiao, 2023), where each skeleton consists of 2D/3D coordinates of several joints. Early skeleton-based methods always rely on hand-crafted descriptors classified as geometric descriptors (Evangelidis, Singh, & Horaud, 2014), kinetic descriptors (Yang & Tian, 2014), and statistical descriptors (Tang, Li, Wang, & Wang, 2018). With deep learning development, end-to-end networks like RNN, CNN, and their variants are introduced for human action recognition (Avola et al., 2020; Hou, Li, Wang, & Li, 2018; Xia, Li, & Luo, 2022; Zhang et al., 2018). To sum up, the commonality of the above methods is that the joints are simply stacked into time series or pseudo images in Euclidean space. As a result, they neglect the irregular topological structure inherent in human skeletons, limiting the capability of action modeling.

Skeleton sequences can be treated as isomorphic spatial–temporal graphs in non-Euclidean space, where bones in skeletons are considered as spatial edges and the same joints in consecutive frames are connected as temporal edges. Therefore, Graph Convolutional Networks (GCNs) have reflected a growing prospect due to their advantages of processing graph-type data like skeletons. Yan et al. (Yan, Xiong, & Lin, 2018) first proposed the ST-GCN for skeleton-based action recognition, which is effective and lays a foundation for GCN-based approaches. Shi et al. (Shi, Zhang, Cheng, & Lu, 2019a) investigated the kinematic dependency between joints and proposed a directed acyclic graph (DAG) for skeleton modeling. Liu et al. (Liu, Zhang, Chen, Wang, & Ouyang, 2020) focused on aggregating the information from multi-range neighbors and alleviated the domination of near joints. However, these GCNs with

* Corresponding author.
E-mail addresses: lichang@hhu.edu.cn (C. Li), huangqian@hhu.edu.cn (Q. Huang), yingchimao@hhu.edu.cn (Y. Mao), lixing@njfu.edu.cn (X. Li), jiewu@temple.edu (J. Wu).

https://doi.org/10.1016/j.eswa.2024.124980
Received 11 April 2024; Received in revised form 4 July 2024; Accepted 1 August 2024
Available online 5 August 2024

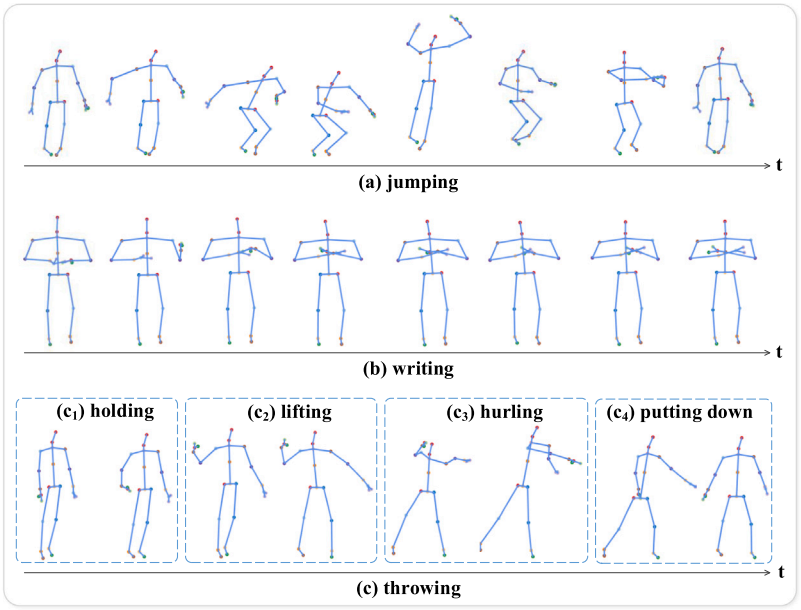


Fig. 1. Example of skeleton sequences for various actions. (a) Jumping is a significant motion that involves a wide range of spatial changes. (b) Writing, on the other hand, is a subtle movement. (c) Throwing is decomposable and can be intuitively divided into sub-actions (c_1-c_4).

predefined graphs fail to picture action-specific dependencies between joints, thus impairing their generality for action recognition.

Motivated by this, some GCNs with dynamic structures are exploited. Shi et al. (Shi, Zhang, Cheng, & Lu, 2019b) embedded the Gaussian function to calculate the relationship of joints. Li et al. (Li, Mao, Huang, Zhu, & Wu, 2023) refined the topology by part-wise correlation modeling and mapping functions. In addition, some researchers further designed multi-stream ensembled networks to improve the action representation ability of GCNs through decision-level fusion of various flows (Cheng, Zhang, He, Cheng, & Lu, 2021). This way can effectively enhance motion information, but it is undeniable that too many streams will create a computational burden. Therefore, designing a general model to analyze disparate actions effectively is still challenging because human movements involve complicated and diverse spatial–temporal concurrency between joints. As shown in Fig. 1, jumping involves a wide range of spatial–temporal changes throughout the body, but writing is mainly influenced by the subtle movements of hands. Besides, some motions are decomposable, such as throwing, which can be divided into holding, lifting, hurling, and putting hands down. However, existing GCN-based methods ignore this diversity in action representation, compromising their performance and robustness. The specific manifestation is that existing approaches are tricky to recognize actions from incomplete skeletons caused by external disturbances, especially occlusion.

Upon analysis, we conclude that the following reasons contribute to the above defects. (1) GCN-based approaches always utilize spatial configuration partition strategy coined in ST-GCN to learn the weights for various neighborhoods. This strategy only involves the simple physical structure but disregards joint kinematic dependency restricted by skeletons, limiting the representation capability of GCNs. (2) Most existing methods first perform Spatial Graph Convolution (SGC) to extract spatial features and then feed them into a Temporal Graph Convolution (TGC) module to capture degraded spatial–temporal patterns. This factorized paradigm hinders the synchronous transfer of spatial–temporal information between joints, thus failing to capture the complex spatial–temporal dynamics for human action recognition. (3) Both SGC and TGC are onefold fixed-size local operations that fail to capture discriminative features for multi-range actions, such as writing and jumping, which is exacerbated in the case of occlusion.

To overcome these limitations, we propose the Multi-Granular Spatial-Temporal Synchronous Graph Convolutional Network (MSS-GCN) for human action recognition. The pipeline is shown in Fig. 2. Firstly, we suggest three partition strategies: attribute, activity, and mixed partition strategy according to the hinged kinematics constraints in skeletons to enrich the weight-sharing mechanism of SGC. Secondly, we endow joints with different diffusion intensities, i.e., different affinity fields, through flexible sliced spatial–temporal graphs. Then, we present a Synchronized Slice Encoder (Syn-STE) to simultaneously embed spatial and temporal action patterns. Thirdly, we design Multi-granular Spatial–temporal Encoders (Multi-STE) to capture the spatial–temporal motion patterns with multiple ranges. By coupling the above effort, MSS-GCN can yield general action representation and multi-granular features, thus showing superior performance and robustness on three public datasets.

In general, the contributions of this work are summarized as follows:

- We investigate three partition strategies: activity, attribute, and mixed partition, according to joint kinematic dependency restricted by skeletons. On this basis, we derive the EAGC module, which has the optimal weight-sharing mechanism and is compatible with other GCN-based approaches.
- We construct the Multi-sliced Spatial–temporal Graph (MSTG) according to the local action semantics to model skeleton sequences more fine-grained and flexible, considering the diversity and complexity of human movements.
- We propose the Multi-STE module to capture and fuse multi-granular motion patterns, which can extract comprehensive and discriminative features even for occluded skeletons.
- We design a Multi-Granular Spatial-Temporal Synchronous Graph Convolutional Network (MSS-GCN), and its effectiveness and robustness have been validated by extensive experiments on three public datasets: NTU-RGB+D, NTU-RGB+D 120, and NW-UCLA.

The remainder of the paper is organized as follows: Section 2 introduces existing studies related to this work. Section 3 details the proposed MSS-GCN and its vital modules. The extensive experimental results are presented in Section 4. In addition, we visualize the features in Section 5 and discuss the partitioning strategy and robustness of MSS-GCN. Finally, the conclusion of our work is described in Section 6.

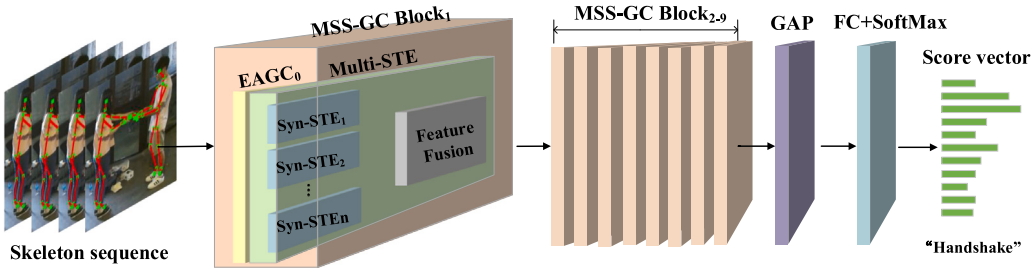


Fig. 2. The pipeline of the proposed MSS-GCN. It stacks nine MSS-GC layers, each containing EAGC and Multi-STE modules. The Multi-STE module has a multi-branch structure composed of Syn-STE modules followed by feature fusion. After the latest MSS-GC layer, the Global average pooling (GAP) and Fully connected (FC) layer with SoftMax is settled. Then, the action with the highest value in the score vector is outputted as the final result.

2. Related work

2.1. GCN-based action recognition

Unlike the RNN-based methods that regard the skeleton data as a sequence of time series (Avola et al., 2020; Zhang et al., 2018) or the CNN-based methods that consider the skeleton data as pseudo-images (Hou et al., 2018; Xia et al., 2022), GCN-based approaches construct the spatial–temporal graph for action modeling. In the skeleton graph, the spatial edges connect the joints like human bones and temporal edges link the identical joints in the consecutive frames. Compared with images or time series, operating convolution on graphs is challenging due to its non-Euclidean nature and lack of rigid arrangement. Yan et al. (2018) solved this problem and proposed Spatial Temporal Graph Convolutional Networks (ST-GCN). At this point, GCN is extended to skeleton-based action recognition.

Single-stream methods. According to the characteristics of skeletons, Yan et al. (2018) investigated the sampling function and partition strategies and proved to be effective for performing convolution on skeleton graphs. To exploit higher-order dependencies, Li et al. (2019) introduced actional-structural graph convolution to capture actional links and structural links between joints. Considering the complementarity between the graph node and edge, Zhang et al. (Zhang, Xu, Tian, & Tao, 2020) devised a graph edge convolutional neural network as a complement to existing GCNs. Besides, Song et al. (Song, Zhang, Shan, & Wang, 2021) cascaded temporal difference and relative coordinate of joints to improve the robustness, and measured the activation degrees of skeleton joints by the class activation maps (CAM). However, their action representation ability is limited by the monotonous input data stream.

Ensembled-stream methods. Unlike the above single-stream approaches, ensembled-stream methods have multiple streams, each adopting diverse action information as input and ensembled for decision fusion. Shi et al. (2019b) generated the lengths and directions of bones as the second-order information of the skeleton data. They proposed a two-stream framework to model both the first-order and the second-order information simultaneously, showing notable improvement in the recognition accuracy. Liu et al. (Liu, Gao, Khan, Qi, & Guan, 2021) devised a multi-stream network including two static feature streams, i.e., the relative coordinate of the joints and bone direction, and the dynamic feature stream, i.e., temporal displacements between two consecutive frames. In addition, Cheng et al. (2021) utilized four streams as input data and proposed the ShiftGCN++, where the joint stream is the original coordinates, the bone stream represents the difference between adjacent joints, the joint motion stream and bone motion stream indicates the joint and bone difference between adjacent frames, respectively.

Topology optimization. On the basis of ensembled-stream framework, many researchers further improve the representation ability of each stream by optimizing the graph topology in GCNs. Shi et al. (2019a) designed the directed acyclic GCN based on the kinematic dependency between the joints and bones, in which joints are directed to each other by incoming and outgoing edges. Due to the fixed topology, the predefined models like this lack the generality to new samples. To capture action-specific topology, Huang et al. (Huang, Huang, Ouyang, Wang, & Assoc Advancement Artificial, Intelligence, 2020) devised the part relation block with graph pooling operators to get the body parts relationship. Li et al. (Li, Huang, & Mao, 2023) constructed the directed diffusion graph to emphasize spatial–temporal information fusion between joints. To refine the topology, Chen et al. (Chen, Zhang, et al., 2021) exploited the channel-specific topologies through a channel-wise modeling function as a generic prior and refined it in an end-to-end way. Considering that the pairwise topology above ignores the high-order correlation between joints, Zhou et al. (2022) constructed a hypergraph and designed a hypergraph self-attention module for action representation. Nonetheless, the above methods ignore the potential temporal correlation between joints, thus limiting the action representation capability of GCNs. To solve this issue, Plizzari et al. (Plizzari, Cannici, & Matteucci, 2021) designed a Temporal Self-Attention module (TSA) to model inter-frame joint correlations, supplementing spatial body parts dependency. Wu et al. (2024) built multiple hypergraphs and updated the weights of joints for salient regions of the actions. However, the above disassembled spatial–temporal methods hinder the transmission of spatial–temporal synchronized information, making it intractable to capture comprehensive action concurrences, and their universality to depict diverse actions is limited.

2.2. Multi-granular analysis in action recognition

Human behavior involves complex contexts and various manifestations. To improve the representation capability of GCNs, many researchers tend to employ multi-granular tricks in human action recognition. Li et al. (2016) modeled frame and video stream as different granularities and boosted action recognition by generating the hierarchical multi-granular motion representations. Pan et al. (Pan, Chen, & He, 2023) treated the supernodes as critical nodes for graph pooling and then used the 3-clique algorithm to coarsen the aggregated features repeatedly. Zhang et al. (2019) introduced pose-guided interactions as the fine-grained semantics of senses and regional cues to capture social relations in human activity. Chen et al. (Chen, Zhou, et al., 2021) proposed a DualHead-Net that jointly pictures the coarse- and fine-grained skeleton motion patterns by high- and low-temporal resolutions. Shu et al. (Shu, Xu, Zhang, & Tang, 2022) defined local and context granularity, which represents one-joint and partial-joint skeleton sequences. Besides, Huang et al. (Huang, Guo, Peng, & Xia, 2023) proposed the Hypergraph-convolution Transformer to capture fine-grained motion patterns that appear on the human face and body. In addtion, Liu et al. (Liu, Zhang, et al., 2020) achieved multi-granular with respect to space, and Chen et al. (Chen, Zhang, et al., 2021) obtained multiple granularities from the view of time. However, the previous methods only capture a single time or space granularity and fail to extract multi-granular spatial–temporal features, which deteriorates their robustness. In this paper, we extend the granularity at the spatial–temporal level and perform feature fusion through a multi-branch network structure, allowing for a more comprehensive and robust representation of human actions.

3. Methods

3.1. Preliminaries

Notations. A skeleton sequence can be represented as a spatial–temporal graph defined as $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{v_1,v_2,\ldots,v_V\}$ is the joint set, and $\mathcal{E}$ is the spatial–temporal edge set. The spatial edges are human bones reflected by adjacency matrix $A\in\mathbb{R}^{V\times V}$, where $A_{i,j}$ reflects the edge that exists between v_i and v_j. Therefore, the spatial neighborhood of v_i can be defined as $\mathcal{N}_S(v_i)=\{v_j \mid A_{i,j}\neq 0\}$. The corresponding joints in two consecutive frames are connected as temporal edges. Given the time interval Γ, the temporal neighborhood of $v_{t,i}$ is represented as $\mathcal{N}_\mathcal{T}(v_{t,i})=\{v_{q,i}\|q-t|\leq\lfloor\Gamma/2\rfloor\}$. The input of GCN is $X\in\mathbb{R}^{C\times T\times V}$, where C, T, and V denote the number of channels, frames, and joints, respectively. T and V are determined by the particular dataset. Since the original joints are represented by 2D or 3D coordinates, C is initialized to 2 or 3. The above notations apply throughout.

Spatial graph convolution. As shown in Fig. 1, the skeleton graph lacks the inherent rigid arrangement that exists in images, which poses a significant hurdle to establishing the correspondence between neighboring joints and weights. To solve this problem, ST-GCN et al. Yan et al. (2018) designed the partition strategy and mapping function to realize the weight-sharing mechanism and then extend classical convolution to skeleton data. Specifically, they partitioned joints into K subsets and associated each joint with a unique weight vector according to its subset index. On this basis, the spatial graph convolution can be formulated as:

$$f_{\text{out}}(v_i)=\sum_{v_j\in\mathcal{N}_{v_i}}\frac{1}{Z_{i,j}}f_{\text{in}}(\mathcal{P}(v_i,v_j))\cdot W(\mathcal{M}(v_i,v_j)) \tag{1}$$

where $Z_{i,j}$ denotes the cardinality of subset $S_{i,k}$ that contains v_j. It equilibrates the contribution of each subset. f_{in} is the input feature map. The partition function $\mathcal{P}$ and the mapping function $\mathcal{M}$ serves to assign an appropriate subset index of v_j. W is the K weight vector associated with K subsets. Therefore, the partition strategy is crucial, as it determines the convolution kernel size and information aggregation of joints.

Temporal graph convolution. The temporal edges can be interpreted as the trajectories of the joints during time intervals intuitively. Specifically, each trajectory records the essential dynamics of the identical joint. The temporal graph convolution, a special convolution operation in CNNs, is introduced to capture this information, which can be written as

$$X_\mathcal{T}^{(l+1)}=\text{Conv}\,2D[\Gamma\times 1](X^{(l)}) \tag{2}$$

where $\Gamma\times 1$ is the kernel size of the 2D convolution. $X^{(l)}$ is the input to the l_{th} hidden layer. Typically, the cross-spacetime motion patterns are captured by stacking the SGC module and TGC module, i.e., alternately executing Eq. (1) followed by Eq. (2).

3.2. Partition strategies for GCNs

How to effectively help GCNs extract discriminative information from spatial–temporal skeletal graphs is a crucial issue for skeleton-based human action recognition. Xu et al. (Xu, Ye, Zhong, & Xie, 2022) has provided theoretical evidence that GCNs are essentially an extension of CNNs. From this perspective, we revisit graph convolution operation and find that it is effective to optimize the partition strategy to enhance the representative capability of GCNs. In CNNs, the local connectivity is directly established through the natural grid structure of images, and each pixel has the same number of neighbors. The learned weight matrix is element-wise multiplied with the corresponding regularly arranged pixel values. Then, the convolutional kernel slides from the top-left to the bottom-right to update pixel values and eventually obtain the output feature map. In other words, the relation between pixels and weights is determined by the grid structure of the image itself.

On the contrary, the skeletal graph is flexible, with each joint having varying neighbors arranged irregularly, and thus the pertinence between joints and weights is absent. To solve this problem, partition strategies are introduced, which not only define the kernel size of SGC, i.e., the length of the weight vector but also establishes the correspondence between joints and weights. Therefore, just as the importance of the convolutional kernel in CNNs, the partition strategy is crucial for GCNs. However, most existing methods rely on the limited partition strategy (Figure 3(a-c)) coined in ST-GCN (Yan et al., 2018), which only counts on the physical distance of joints and underutilizes the articulated nature of human skeletons. In this paper, we present three novel partition strategies and introduce the EAGC module to enhance the representation capacity of GCNs.

Attribute partition strategy. The human skeleton is a hinge structure in that one joint moves around another joint with bone. On this basis, the human skeleton can be constructed as a directed graph, where the directed edge means the dependency between joints, i.e., $v_i\rightarrow v_j$ indicates that v_j moves around node v_i. As shown in Fig. 3(d), the joints are divided into source and target sets, i.e., $K=2$. Mathematically, the mapping function $\mathcal{M}$ is defined as:

$$\mathcal{M}(v_i,v_j)=\begin{cases}0, & \text{if } v_i\rightarrow v_j\\ 1, & \text{if } v_j\rightarrow v_i\end{cases} \tag{3}$$

Activity partition strategy. Each joint has a unique function in action execution, and its activity affects its contribution to actions. As depicted in Fig. 3(e), hands and feet are the most active, and their commonality is that they are both leaf nodes. Inspired by this, we employ out-degree as a measure and divide the joints into three subsets: active subset, medium subset, and silent subset. Formally,

$$\mathcal{M}(v_i,v_j)=\begin{cases}0, & \text{if } D(v_j)=0\\ 1, & \text{if } D(v_j)=1\\ 2, & \text{if } D(v_j)\geq 2\end{cases} \tag{4}$$

where $D(v_j)$ is the out-degree of v_j. According to the activity partition strategy, the optimized graph convolution kernel can assign weights to joints based on their contribution to the action, bringing better modeling capacity and recognition performance.

Mixed partition strategy. Adjacent joints are strongly coupled, and the distance between joints is proportional to their dependence. Therefore, neighbors should assign weights based on the number of hops, which can be regarded as the connection strength. To this end, we propose a mixed partition strategy, as drawn in Fig. 3(f). Specifically,

$$\mathcal{M}(v_i,v_j)=\begin{cases}2d, & \text{if } H(v_i,v_j)=d, 0\geq d<\theta,\ v_i\rightarrow v_j\\ 2d+1, & \text{if } H(v_i,v_j)=d, 0\geq d<\theta,\ v_j\rightarrow v_i\end{cases} \tag{5}$$

where $H(v_i,v_j)$ is the least hop from v_i to v_j. θ is the maximum distance of the sampling function, which determines the receptive field. When $\theta=1$, the mixed partition strategy is equivalent to the attribute partition strategy. After executing the mapping function, each joint is assigned a value representing the group index to which it belongs. Taking the activity partition strategy as an example, the group number of joint D is $\mathcal{M}(E,D)=0$, meaning that D belongs to the 0_{th} group. The grouping results of all the joints have been annotated in Fig. 3.

3.3. Extended adaptive graph convolution

As depicted in Fig. 3, joints are divided into fixed groups according to the partitioning strategy, and joints in the same group share identical weights in the graph convolution, resolving the different number of neighbors and weight assignment problems in the graph structure. To this end, Eq. (1) is transformed into

$$f_{\text{out}}=\sigma\left(\sum_k^K f_{\text{in}}\left(D_k^{-\frac{1}{2}}\bar{A}_kD_k^{-\frac{1}{2}}\right)W_k\right) \tag{6}$$

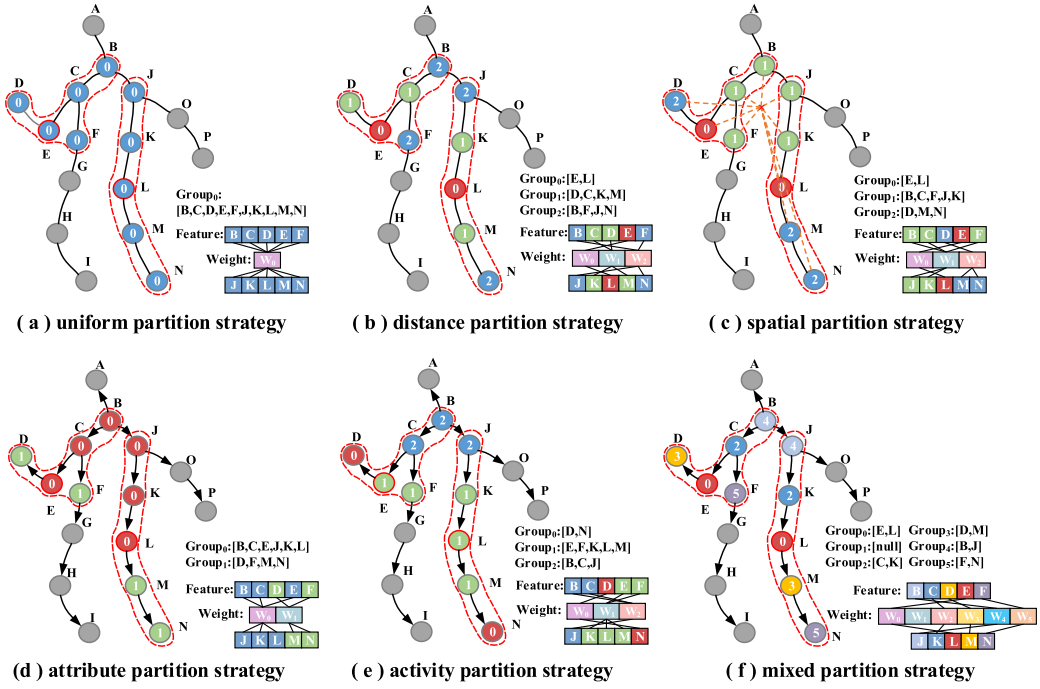


Fig. 3. Illustration of partition strategies. The first row shows the strategies presented in ST-GCN (Yan et al., 2018), and the second row illustrates our proposed strategies. The red dashed line indicates the corresponding neighborhood of the specific root joint E and L. The number marked on a joint indicates its group index obtained by the mapping function $\mathcal{M}$, and joints belonging to the same group share identical learned weights.

where $\bar{A} = A + I$, $\bar{A}_k \in \mathbb{R}^{V\times V}$ and I is the identity matrix which means self-connection. The element of $A_{k,i,j}$ denotes whether v_j is in the k_{th} subset of v_i. D_k is the diagonal degree matrix of A_k. W_k is the weight vector similar to W in Eq. (1) and $k = \mathcal{M}(v_i, v_j)$.

The manually defined topology fails to model intrinsic dependency between unnaturally connected joints (Liu, Zhang, et al., 2020). Therefore, we propose the EAGC module to enhance GCNs' representation capability. As shown in Fig. 4(b), we adopt channel-wise correlation modeling to generate an adaptive adjacency matrix $\widetilde{A}$. Given two joints v_i and v_j with their corresponding C' channel features x_i and x_j, the channel-specific relationships $\widetilde{A} \in \mathbb{R}^{V\times V\times C'}$ is defined as

$$\tilde{A}_{i,j} = \sigma\left(MLP\left(\varphi\left(x_i\right) \| \phi\left(x_j\right)\right)\right) \tag{7}$$

where φ and ϕ is the linear transformation function to reduce feature dimension. $\|$ is denoted as a cascade operation. MLP is the multilayer perceptron. σ is the activation function by which the more relevant joints are emphasized. Then adaptive channel-refined topology can be learned through backpropagation. Specifically,

$$\hat{A} = \mathcal{T}(\bar{A}) + \alpha\tilde{A} \tag{8}$$

where $\mathcal{T}$ is a transformation function to conform $\bar{A}$ and $\tilde{A}$ has the same dimension. α is a trainable parameter. $\hat{A}$ can capture distant joint correlations and automatically construct the semantically-based topology. By combining the above partition strategies and dynamic topology, we devise EAGC modules including EAGC-attr, EAGC-act, and EAGC-mix, respectively. They are compatible with other GCNs and can effectively improve their performance. The analysis is detailed in Sections 4.4 and 5.3.

3.4. Multi-sliced spatial-temporal graph

Human movements are varied and intricate in nature. Some actions are obvious, such as "fall down", while some actions are subtle, such as "writing", and some are decomposable, such as "pick up and throw". Therefore, it is challenging to design a general graph to model and analyze disparate actions effectively. In this paper, we propose the MSTG to enhance the generalization of GCN-based action modeling. Intuitively, an action can be decomposed into different phases, which convey multi-level action semantics. For example, the action such as "throwing" can be decomposed into holding, lifting, hurling, and putting hands down, as illustrated in Fig. 1(c). Therefore, we crop the skeleton sequence along the time dimension to model sub-actions. In addition, the spatial configuration of joints is significant for action recognition, especially subtle movements. To capture more fine-grained motion patterns, we also segment the skeleton to characterize the actions more pertinently. Considering that actions have high spatial–temporal parallelism, we perform these two slicing operations simultaneously to obtain spatial–temporal action slices and model them as spatial–temporal sliced graphs. For clarity, the MSTG for "running" is as sketched in Fig. 4(c).

Mathematically, given a skeleton sequence $X \in \mathbb{R}^{C\times T\times V}$ that is modeled as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the multi-sliced spatial–temporal graph $\mathcal{G}'$ containing n graph slices is defined as $\mathcal{G}' = \{g_1 \cup g_2 \cup \cdots \cup g_n\}$, with $g_i = \{v,\varepsilon\}$, $v \subseteq \mathcal{V}$ and $\varepsilon \subseteq \mathcal{E}$. In other words, $X = \{x_1 \| x_2 \|, \ldots, x_n\}$, $x_i \in \mathbb{R}^{T/P\times V/Q\times C}$, where P and Q is the temporal step and spatial size of slices, respectively. For each slice g_i, we treat bones as spatial edges depicted as $A^i \in \mathbb{R}^{V/Q\times V/Q}$. Unlike the original $\mathcal{G}$, we establish fully connected local temporal edges within each slice. These extended edges serve two purposes: on the one hand, they can capture the temporal changes of individual joints (trajectories); on the other hand, they can capture distant spatial–temporal dependencies between different joints. Drawing upon this foundation, the spatial neighborhood of the m_{th} slice graph is $\mathcal{N}_S^m\left(v_i\right) = \left\{v_j \mid A^m{}_{i,j} \neq 0, v_j \in g_m\right\}$, and the temporal neighborhood of $v_{t,i}$ is represented as $\mathcal{N}_\mathcal{T}^m\left(v_{t,i}\right) = \left\{v_{q,j} \| q - t \| \leq \lfloor T/2P \rfloor, v_j \in g_m\right\}$. To reduce computing complexity and redundant relations, we split the skeleton sequence by non-overlapping windows. The proposed MSTG is more flexible for multi-range action modeling, thus boosting recognition performance.

3.5. Multi-granular spatial–temporal encoders

Synchronized spatial–temporal modeling. Human movements contain highly concurrent spatial–temporal dynamics of joints. This

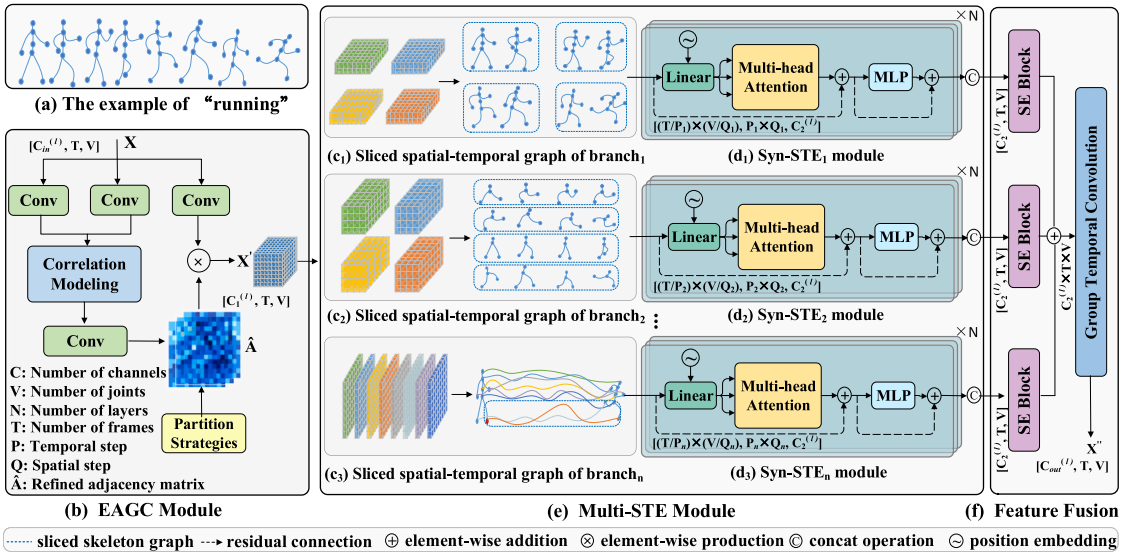


Fig. 4. The detailed structure of the proposed MSS-GC layer. MSS-GC contains EAGC(b) and Multi-STE(e) modules. EAGC is an adaptive GC combined with the proposed strategies. The MSTG for running (a) is illustrated in (e), including multi-granular sliced graphs from multiple branches and then encoded by relevant Syn-STE modules. Feature fusion is designed for extracting multi-granular motion features.

property is particularly obvious at the local scale. For example, an action consists of multiple sub-actions in which joints are highly spatial–temporal correlated. However, such critical synchronous information between joints is disregarded by factorized paradigms of existing GCNs. The proposed MSTG facilitates localizing these highly concurrent joints. Based on this, we design the Syn-STE module to encode their synchronized spatial–temporal motion patterns, as shown in Fig. 4(d). For each sliced graph, we rearranged all the joints as a sequence by flattening the slice in the order of spatial and then temporal dimension. After that, each joint can be regarded as the token in Transformer (Vaswani et al., 2017). Then, we apply the Multi-Head Attention (MHA) mechanism to estimate the global spatial–temporal dependency between the sliced joints, which re-encodes the joints and facilitates spatial–temporal information diffusion across the MSTG simultaneously. Particularly, the linear embedded joint sequences are denoted as Q, K, and V, and then the scaled dot-production is conducted to compute dependency. For each attention head, the attention is conducted as follows

$$\mathcal{W}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{C'}} + B\right)\mathrm{V} \tag{9}$$

where B is the relative position bias (Liu, Lin, et al., 2021) to preserve the original structure. $\sqrt{C'}$ is the scaled factor, and C' is the dimension of Q, K, V. The h-head attention weight W of i_{th} sliced graph is obtained by

$$W_i = \psi\left(\mathrm{concat}\left(w_1, w_2, \ldots, w_h\right)\right) \tag{10}$$

where w_i is the attention vector of i_{th} head. ψ is the linear mapping function. MHA allows the model to extract discriminative features from different representation subspaces. To this end, each MSTG is encoded as x_i', and $x_i' = x_i \cdot W_i$. Then, every head attention is concatenated and encoded for final embedding of x_i'. That is

$$X' = \zeta\left(\mathrm{concat}\left(x_1', x_2', \ldots, x_n'\right)\right) \tag{11}$$

where ζ is the linear transformation employed for channel consistency. The Syn-STE module exploits the spatiotemporal synchronous characteristics of actions, thus effectively extracting discriminative features for human action recognition.

Multi-Granular analysis. To endow the model with general represent capability for various actions, many researchers tend to introduce multi-granular analysis into GCNs (Chen, Zhang, et al., 2021; Liu, Zhang, et al., 2020). However, these methods are limited because they only focus on one aspect, overlooking the fact that human motion is the collection of skeletons with various spatial transformations over time. In biomechanics, the human body can be represented as an articulated system of rigid segments connected by joints, and human motion can be considered a continuous evolution of the spatial configuration of these rigid segments (Vemulapalli, Arrate, & Chellappa, 2014). Therefore, the key point of multi-granular action analysis is to describe spatial and temporal dynamics at different semantic levels simultaneously.

In this study, we regard the size of the spatial–temporal sliced graph as the granularity factor, which controls the transmission scope and strength of spatial–temporal synchronization information. As displayed in Fig. 4(e), we propose the Multi-STE with a multi-branch structure, and each Syn-STE branch has specific granularity for feature embedding and fusion. On this basis, each branch provides granular-specific action semantics combined for a more comprehensive representation of actions.

Multi-Granular Feature Fusion. After considering that certain occlusion situations can result in the degradation of spatial–temporal slices and that the action semantics communicated by different slices are complementary, we explored four distinct fusion strategies outlined in Fig. 5. Specifically, we introduce the SE block (Hu, Shen, & Sun, 2018) for feature selection, where the added features are augmented, as shown in Fig. 5(a). In addition, we try to add SE for every branch to fuse granularity-specific semantics. We proposed Multi-SE-Add fusion (Fig. 5(b)) which firstly performs attention encoding separately and then conducts add fusion. Besides, as drawn in Fig. 5(c), α-Add fusion dynamically adjusts the importance of different branches by constructing a learnable vector $\mathbf{b}$, but it does not work well. Vary to the above methods, SE-W-Add fusion (Fig. 5(d)) only retains the channel-wise attention of each branch obtained by SE and adds them to derive the global weight matrix W. Then the feature fusion is implemented by conducting element-wise multiplication between W and the primary input X. Considering the continuity of spatial–temporal information, we employ group convolution followed by a ReLU layer to realize cross-slice feature aggregation and sparse multi-granular motion representation. In addition, the residual connection is employed to mitigate the vanishing gradient problem. See Section 4.4 for the evaluation of the above fusion strategies.

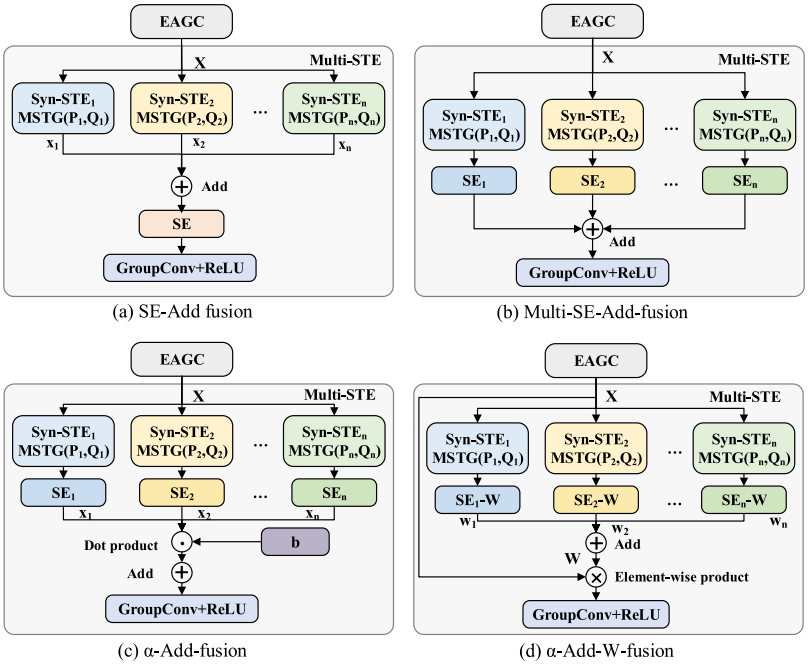


Fig. 5. The illustration of multi-granular feature fusion strategies. We propose four fusion strategies and classify them into two categories: intermediate fusion (a), and semantic fusion (b, c, d).

4. Experiments and discussions

4.1. Datasets

NTU-RGB+D. NTU-RGB+D (Shahroudy, Liu, Ng, & Wang, 2016) contains 56,880 action samples categorized into 60 classes. The actions are conducted by 40 volunteers. The skeleton sequences are captured by three Microsoft Kinect v2 cameras from three views. There are two popular benchmarks: (1) cross-subject (C-Sub): training data comes from half of the subjects, and testing data comes from the other. (2) cross-view (C-View): the training set comes from camera IDs 2 and 3, and the testing set comes from camera ID 1.

NTU-RGB+D 120. NTU-RGB+D 120 (Liu, Shahroudy, et al., 2020) is an extension of NTU-RGB+D, which has 120 action classes and 114,480 samples. The samples are collected in various locations and backgrounds denoted as 32 setups. In addition to the original cross-subject (C-Sub), the cross-setup (C-Set) evaluation is introduced, where the training set comes from samples with odd setup IDs, and the testing set comes from the rest.

NW-UCLA. NW-UCLA (Wang, Nie, Xia, Wu, & Zhu, 2014) is a multi-view dataset captured by three Kinect cameras. It contains 1494 video clips and covers 10 action labels. Each action is performed by 10 subjects. We follow the same evaluation protocol in Chen, Zhang, et al. (2021): the samples captured by the first two cameras are grouped as a training set, and the residual makes a testing set.

4.2. Training details

Unless otherwise stated, the model MSS-GCN and its variants are trained by the Stochastic Gradient Descent (SGD) with 0.9 momentums for a total of 80 epochs under the PyTorch deep learning framework. The standard cross-entropy loss is employed. We apply a warmup strategy in the first 5 epochs for training stability. The global parameters, such as weight decay, initial learning rate, and learning rate decay is set to 0.0004, 0.1, and 0.1, respectively. The learning rate is linearly scaled down at specific steps. For NTU-RGB+D and & NTU-RGB+D 120 dataset, the step is 30, 40, 50, and the batch size is set to 64. For the NW-UCLA dataset, the step is 50, 70, and the batch size equals 16. In addition, inputs are preprocessed with normalization and translation following Chen, Zhang, et al. (2021). We adopt a two-stream framework to train models using joint and bone data and ensemble their results literately.

Table 1
The accuracy (%) of MSS-GCN embedded single Syn-STE with various slices and head numbers.

Parameter ($[P, Q]$)	Head (H)	Accuracy (%)
[4, 5]	4	89.49
[4, 25]	4	**90.52**
[8, 25]	4	90.08
[16, 25]	4	88.88
[32, 25]	4	88.60
[4, 25]	2	88.52
[4, 25]	8	88.34

4.3. Parameter selection

The granularity of MSTG. For MSS-GCN, the size of the spatial–temporal slice is critical, which not only determines the intensity of the spatial–temporal information diffusion but also the granularity of motion features. As the decisive factors affecting the slices, P and Q in Section 3.4 are essential. In this work, we conduct incremental experiments to obtain the optimal configuration. Firstly, we explore the influence of $[P, Q]$ on MSS-GCN with a single Syn-STE branch. As shown in Table 1, the slice size can significantly affect the model accuracy, reflecting the importance of fine-grained analysis of human actions. The highest accuracy of 90.52% is achieved when $[P, Q]$ matches $[4, 25]$. In addition, the number of heads (H) in MHA also impacts the performance. Considering the trade-off between complexity and accuracy, we set H to 4.

Then, we fix the above parameters and explore the Multi-STE with various combinations. Specifically, we design a two-branch structure where the window of one branch is settled to [4, 25], which is the optimal setting in the single-branch regime. And then, we explore several window combinations of various parameters [P, Q]. As shown in Table 2, when $[P, Q]$ of the two branches is the same, i.e., they all equal

Table 2
The accuracy (%) of two-branch MSS-GCN with various combinations of MSTGs.

$[P, Q]$	[4, 5]	[4, 25]	[8, 25]	[1, 25]	[64, 1]	[64, 5]
[4, 25]	89.57	89.46	89.71	90.28	**90.62**	90.04

Table 3
The accuracy (%) of models with different partition strategies.

Methods	Uniform	Distance	Spatial	Activity	Attribute	Mixed
ST-GCN (Yan et al., 2018)*	80.84	83.33	83.65	**84.89**	82.19	85.01
AGCN (Shi et al., 2019b)*	88.29	88.51	88.84	89.00	88.94	**89.83**
DD-GCN (Li et al., 2023)	89.71	90.04	90.13	**90.52**	90.10	90.18
MSS-GCN	89.96	90.10	90.25	**90.62**	90.06	90.30

Those marked with * are methods we reproduced.

Table 4
Comparison of feature fusion strategies.

Fusion strategy	Accuracy	Parameter
Concat fusion	89.86%	3.04M
Add fusion	89.46%	2.65M
SE-Add fusion	89.89%	2.75M
α-SE-Add-fusion	90.06%	2.85M
SE-W-Add fusion	90.10%	2.85M
Multi-SE-Add fusion	**90.62**%	2.85M
Multi-SE-Add w/o GroupConv	87.60%	2.62M

to [4, 25], the accuracy is only 89.46%, which illustrates that extracting multi-granular features stimulates performance. We find that too small slices may neglect global information. That is why the supplement branch with $[P, Q]$ =[4, 5] fails to refine recognition results. Therefore, the integrity of action semantics should be emphasized while considering the complementarity of multi-branch structures. When the $[P, Q]$ in the extra branch is [64, 1], the accuracy is the highest, reaching 90.62%. The possible reason is that these two branches capture the global information in both time and space dimensions, thus extracting multi-granular spatial–temporal details simultaneously. We also tried to design a three-branch structure and set $[P, Q]$ to [4, 25], [64, 1], and [1, 25], respectively. However, the accuracy is only 89.82%, and the additional branch leads to a computational burden. Therefore, it is worth studying the appropriate grade for multi-granular action analysis.

4.4. Ablation study

Comparison of partition strategies. We compare the proposed three partition strategies against other strategies in ST-GCN (Yan et al., 2018). Table 3 demonstrated that our strategies are competitive and compatible with other GCNs. The SGC (Yan et al., 2018) incorporated activity, attribute, and mixed strategy can enhance the accuracy up to 4.05%, 1.35%, and 4.17%, respectively. For AGCN (Shi et al., 2019b), our three partition strategies are all optimal, among which the mixed partition strategy raises the result by nearly 1% compared with the spatial partition strategy. Overall, the activity and mixed strategies work better, and we apply the activity partition policy by default considering its advantage. The above results confirm our hypothesis: optimizing partition strategies can effectively improve the performance of GCN-based methods. It is worth noting that the three partition strategies we proposed are general and can be embedded in existing GCN-based methods to improve the accuracy effectively. We discuss the properties of these partitioning strategies in detail in Section 5.3.

Evaluation of feature fusion strategies. We compare several multi-granular feature fusion strategies, and the experimental results are shown in Table 4. It can be seen that the fusion strategy with the SE module is generally better than simple fusion strategies, i.e., Concat fusion and Add fusion. Further, we divide other strategies into intermediate fusion strategies, i.e., SE-Add and semantic fusion, as depicted in Fig. 5. The fusion strategy combined with the SE block is preferable to the intermediate fusion strategy because each branch captures the action features from different granularities, and multiple SE blocks corresponding to each branch can emphasize the action semantics related to granularities, extracting more discriminative features. Among them, the accuracy of the Multi-SE-Add fusion strategy is the best, reaching 90.62%. Compared with Multi-SE-Add, the α-SE-Add strategy with weight vector and the SE-W-Add strategy with the global attention weight are 0.56% and 0.52% lower, respectively, and bring computational burden. Furthermore, by analyzing the parameters of these strategies in Table 4, we prove that Multi-SE-Add can balance the accuracy and complexity and outperform other strategies. In addition, we find that it is necessary to fuse the multi-granular temporal information through the GroupConv module. Otherwise, the performance will be reduced by 3.02% for MSS-GCN.

Table 5
Comparison of various models under different component configurations.

Methods	Layers	Accuracy(%)
ST-GCN* w/ EAGC	10	85.55 ↑ 1.90
AGCN* w/ EAGC	10	89.39 ↑ 0.55
ST-GCN* w/ Syn-STE	10	87.85 ↑ 4.20
AGCN* w/ Syn-STE	10	89.53 ↑ 0.69
ST-GCN* w/ Multi-STE	10	88.75 ↑ 5.10
AGCN* w/ Multi-STE	10	90.01 ↑ 1.17
MSS-GCN w/o Multi-STE	9	89.73 ↓ 0.89
MSS-GCN w/o Multi-STE	10	89.81 ↓ 0.81
MSS-GCN w/o 1 Syn-STE	9	90.11 ↓ 0.51
MSS-GCN w/o 1 Syn-STE	10	90.52 ↓ 0.10
MSS-GCN w/o EAGC	9	90.25 ↓ 0.37
MSS-GCN	9	**90.62**

Effectiveness of components. Table 5. displays the results of the ablation experiments on the proposed components. Based on these results, we can conclude that (1) the EAGC, Syn-STE, and Multi-STE modules are practical and can be transferred to existing GCN-based approaches to stimulate their performance. Replacing the SGC and AGC modules in ST-GCN and AGCN with EAGC can increase the accuracy by 1.99% and 0.55%, respectively. The performance of ST-GCN and AGCN is boosted regardless of whether Syn-STE or Multi-STE is embedded. (2) Experiments on ST-GCN, AGCN, and MSS-GCN show that the accuracy of Multi-STE is better than that of single Syn-STE. As expected, capturing multi-granular action features benefits for better performance. (3) MSS-GCN only needs to stack nine MSS-GC layers to achieve the optimal recognition effect. Therefore, embedding EAGC and Multi-STE is computationally cost-effective and efficient for GCN-based action recognition.

4.5. Comparison with the state-of-the-art

To prove the advantages of the proposed MSS-GCN, we compare it with existing methods on the NTU-RGB+D, NTU-RGB+D 120, and NW-UCLA. The experimental results are reported in Tables 6–7. We classify the listed methods into four types according to their backbones, i.e., RNN-based, CNN-based, Transformer-based, and GCN-based methods. It can be seen that our method not only outperforms GCN-based methods but also outperforms other types, especially RNN-based methods like AGC-LSTM (Si, Chen, Wang, Wang, & Tan, 2019) and AMCGC-LSTM (Xu et al., 2021). As shown in Table 6, MSS-GCN achieves the highest performance on the NW-UCLA dataset, which indicates that it can effectively model daily actions and abnormal behaviors. In Table 7, MSS-GCN exceeds many competitive GCN-based methods on NTU-RGB+D and NTU-RGB+D 120 datasets. To be fair, the self-supervised methods based on contrastive learning (e.g., AimCLR (Guo et al., 2022), ConGT (Pang, Lu, & Lyu, 2023), ActCLR (Lin, Zhang, & Liu, 2023) and HiCLR (Zhang, Lin, & Liu, 2023)) we report are the results after finetune. As we can see, MSS-GCN not only outperforms STHG-DAN

Table 6
Comparisons of the number of ensembled streams (E-S), FLOPs (G) and number of parameters (M) and the top-1 accuracy (%) with the methods on the NW-UCLA dataset.

Type	Methods	E-S	FLOPs	Para.	Accuracy (%)	Venue
RNN	AGC-LSTM (Si et al., 2019)	2	–	–	93.3	CVPR'19
	AMCGC-LSTM (Xu et al., 2021)	1	–	–	87.9	JIOT'21
CNN	SLnL+rFA+ML (Hu, Cui, & Yu, 2020)	2	–	–	93.5	TMM'20
Transformer	ConGT (Pang et al., 2023)	2	–	–	85.3	TMM'23
GCN	Shift-GCN (Cheng, Zhang, He, et al., 2020)	4	0.7	1.23	94.6	CVPR'20
	DC-GCN+ADG (Cheng, Zhang, Cao, et al., 2020)	4	3.6	9.8	95.3	ECCV'20
	RGCA (Yao, Zhao, Xie, Ye, & Liang, 2021)	1	–	–	85.3	ICME'21
	CTR-GCN (Chen, Zhang, et al., 2021)	4	2.3	5.7	96.5	ICCV'21
	ShiftGCN++(Li et al., 2023)	4	0.1	0.4	95.0	TIP'21
	GCN-HCRF (Liu, Gao, et al., 2021)	3	–	–	91.5	TMM'21
	FGCN (Yang, Yan, et al., 2022)	2	–	–	95.3	TIP'22
	Graph2Net (Wu, Wu, & Kittler, 2022)	2	0.6	1.6	95.3	TCSVT'22
	CrossMoCo (Zeng, Liu, Liu, & Chen, 2023)	2	–	–	87.6	TMM'23
	DD-GCN (Li et al., 2023)	2	5.7	2.8	96.7	ICME'23
	SaPR-GCN (Li, Mao, et al., 2023)	4	1.3	2.1	96.6	TCSVT'23
	MSS-GCN (ours)	2	3.9	2.2	**96.8**	–

(Wu et al., 2024) introduces multiple spatial–temporal hypergraphs constructed of multi-view human body joints but also exceeds ACE-ens (Qin et al., 2024) that fuses higher-order features in the form of angular encoding. Besides, our method can make a good trade-off between accuracy and computational burden. On the NTU-RGB+D 60 dataset, the FLOPs of our method is 6% and 20% of that of ST-TR (Plizzari et al., 2021) and ACE-ens (Qin et al., 2024), respectively, even that we utilize multi-stream ensemble that inevitably increases the computational burden. Besides, the overall parameters of our method is 26% and 35% of that of the single-stream methods PL-GCN (Huang et al., 2020) and MTT-AGCN (Kong, Bian, & Jiang, 2022), respectively. To summarize, the extensive experiments fully demonstrate the superiority of MSS-GCN thanks to its components: EAGC and Multi-STE modules. In addition, these components are compatible with existing GCN-based methods and can be easily ported to obtain accuracy gains.

5. Discussion

In this section, we further discuss the number of layers, the robustness of MSS-GCN, and the properties of partition strategies. Experiments are all conducted on the NTU-RGB+D dataset with the cross-subject setup.

5.1. Number of layers

Most existing GCN-based methods adopt a ten-layer network structure, and each layer includes spatial and temporal graph convolution. We find that when using multi-granular tricks to enrich the motion representation, only nine layers are needed to extract enough discriminative features for action recognition, which shows the powerful representation ability of MSS-GCN. We further illustrate feature differences of various layers by t-SNE dimensionality reduction. As shown in Fig. 6, the original input data is projected into three clusters, each containing a mix of samples from different classes. The middle fifth layer extracts the shallow unified feature space.

Although the sample distribution is still chaotic, the distribution is relatively uniform. It is worth noting that, compared with the features obtained in the tenth layer, the features in the ninth layer have a more considerable inter-class distance and a more precise decision boundary. In other words, increasing the depth of neural networks is not always beneficial because the network may learn the noise in the training data rather than the underlying data structure. Therefore, the ideal depth of the network should be determined experimentally to ensure that it can learn valuable features while avoiding overfitting.

5.2. Results on the occluded data

To illustrate the robustness of MSS-GCN, we report its performance on the occluded data in Table 8. To be fair, all the listed methods follow the same experimental setup of Song et al. (2021) as default. Fig. 7 depicts the corresponding occlusion strategy, taking cheer-up action of NTU-RGB+D dataset as an example. Specifically, we evaluate the robustness from two perspectives: body occlusion and time occlusion. (1) Body occlusion is the joint-level data degradation, which occludes joints with the same index for all action samples. We adopt the joint index officially given by NTU-RGB+D dataset, as shown in Fig. 7(b), and divide the human body into five parts, namely the left arm(5, 6, 7, 8, 22, 23), right arm(9, 10, 11, 12, 24, 25), two hands(22, 23, 24, 25), two legs(13, 14, 15, 16, 17, 18, 19, 20), and trunk(1, 2, 3, 4, 21). After that, we select one of the parts at a time and set the coordinates of the covered joints to 0 to simulate the scene where the part is occluded. On this basis, we can ensure that the missing joints are the same when all action samples occlude the same part, which guarantees fairness in all cases. Figure 7 (c) illustrates the occlusion of the left arm.

From Table 8(top half), our method achieves optimal accuracy except for the case of trunk occlusion, where MSS-GCN is 0.7% lower than our previous method SaPR-GCN (Li, Mao, et al., 2023). It indicates that part-based methods can effectively cope with trunk occlusion interference. In addition, it is worth noting that the robustness of MSS-GCN is significantly improved compared to ST-GCN (Yan et al., 2018). In the case of left-arm occlusion, the accuracy is improved dramatically by 35.4%, which benefits from multi-granular action modeling and feature fusion. (2) Time occlusion is skeleton-level data degradation where all samples occlude with the same proportion of skeletons. As there may be instances where the human body is entirely occluded, we use frame occlusion to simulate this scenario. We normalize action samples to time series with the same length to ensure that a fair number of frames will be degraded for all actions at the same occlusion ratio. Given a sequence with T frames and p proportion of occluded frames, we zero out skeleton data for $T\times p/2$ frames before and after the middle frame. Fig. 7 (d) and (e) provide examples of occluded 10% and 20% frames of cheering up. We report the experimental results under various proportions of frame occlusion in Table 8(bottom half), where it can be seen that MSS-GCN achieves the best performance. The reason is that multi-granular spatial–temporal slices cover behavior information of varying time intervals, and the multi-granular features extracted based on these can dispair the impact caused by frame loss. In general, improving the fine-grained modeling of skeletons and extracting multi-granular features can effectively enhance the robustness of action recognition.

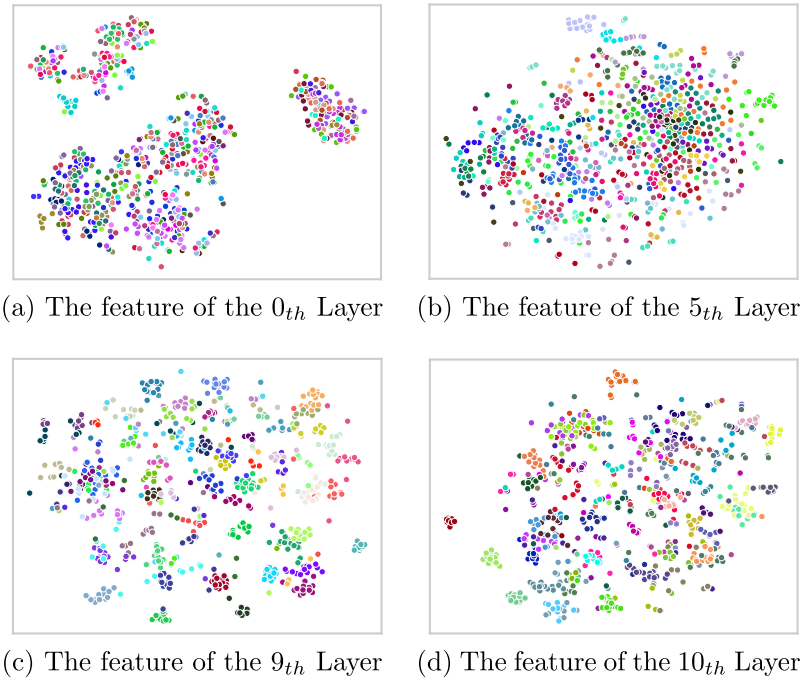


Fig. 6. The feature visualization by t-SNE. The points of different colors represent various categories.

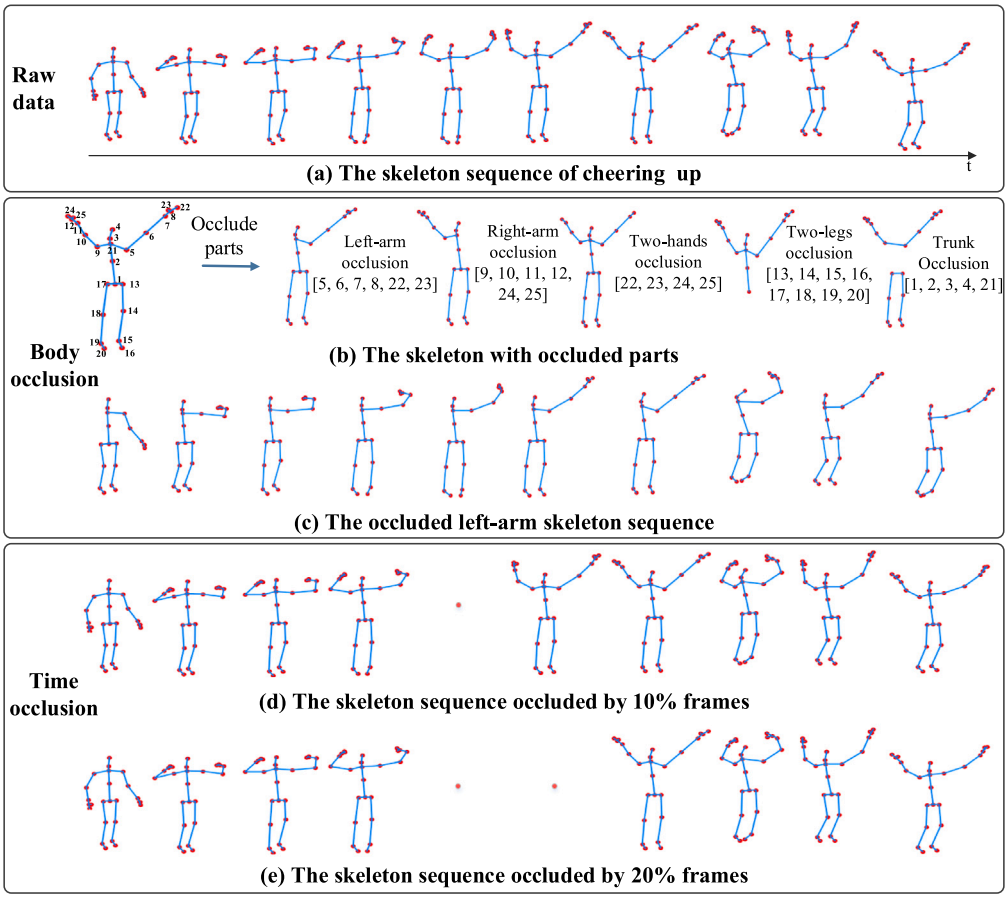


Fig. 7. Example of occlusion cases. Taking the action cheering up with ten selected frames as an example, we show the raw joint data, the body occlusion, and the time occlusion settings of the NTU-RGB+D dataset. Body occlusion is the joint-level data degradation, which occludes joints with the same index for all action samples. Time occlusion is skeleton-level data degradation where all samples occlude with the same proportion of skeletons.

Table 7
Comparisons of number of ensembled streams (E-S), FLOPs (G) and number of parameters (M) and the top-1 accuracy (%) with the methods on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets.

Type	Methods	E-S	FLOPs	Para.	NTU-RGB+D		NTU-RGB+D 120		Venue
					X-sub	X-view	X-sub	X-set	
RNN	AGC-LSTM (Si et al., 2019)	2	–	–	89.2	95.0	–	–	CVPR'19
	AMCGC-LSTM (Xu et al., 2021)	1	–	–	80.1	87.6	71.7	72.4	JIOT'21
CNN	LDT-NET (Yin, He, Soomro, & Yuan, 2023)	1	–	0.4	82.3	89.1	–	–	ESWA'23
	RGB+D+DFN (Li, Hou, Li, Ding, & Wang, 2024)	2	–	–	91.8	96.5	–	–	ESWA'24
Trans- former	DSTA-Net (Shi, Zhang, Cheng, & Lu, 2020)	4	64.7	4.1	91.5	96.4	86.6	89.0	ACCV'20
	ST-TR (Plizzari et al., 2021)	2	259.4	12.1	90.3	96.3	84.3	86.7	CVIU'21
	ConGT (Pang et al., 2023)	2	–	–	84.6	91.6	79.4	80.5	TMM'23
GCN	ST-GCN (Yan et al., 2018)	1	16.3	3.1	81.5	88.3	70.7	73.2	AAAI'18
	2s-AGCN (Shi et al., 2019b)	2	37.3	6.9	88.5	95.1	82.9	84.9	CVPR'19
	AS-GCN (Li et al., 2019)	1	26.8	9.5	86.8	94.2	77.9	78.5	CVPR'19
	NAS-GCN (Peng, Hong, Chen, & Zhao, 2020)	2	72.3	13.0	89.4	95.7	–	–	AAAI'20
	MS-G3D (Liu, Zhang, et al., 2020)	2	48.9	6.4	91.5	96.2	86.9	88.4	CVPR'20
	Shift-GCN (Cheng, Zhang, He, et al., 2020)	4	10.0	2.8	89.7	96.0	85.9	87.6	CVPR'20
	PL-GCN (Huang et al., 2020)	1	–	20.7	89.2	95.2	–	–	AAAI'20
	RA-GCN (Song et al., 2021)	3	32.8	6.2	87.3	93.6	81.1	82.7	TCSVT'21
	ShiftGCN++(Cheng et al., 2021)	4	1.7	1.8	90.5	96.3	85.6	87.2	TIP'21
	CTR-GCN (Chen, Zhang, et al., 2021)	4	7.9	5.8	92.4	96.8	88.7	90.1	ICCV'21
	Graph2Net (Wu et al., 2022)	2	9.9	1.6	90.1	96.0	86.0	87.6	TCSVT'22
	MKE-GCN (Yang, Wang, Gao, & Song, 2022)	3	–	–	91.8	96.2	89.0	90.3	ICME'22
	MTT-AGCN (Kong et al., 2022)	1	32.4	15.6	89.3	95.8	82.0	83.8	LSP'22
	FGCN (Yang, Yan, et al., 2022)	2	–	–	90.2	96.3	85.4	87.4	TIP'22
	SMotif-GCN (Wen et al., 2022)	1	–	–	90.5	96.1	87.1	87.7	TPAMI'22
	AimCLR (Guo et al., 2022)	3	1.7	2.5	88.2	93.9	82.1	84.6	AAAI'22
	ML-STGNET (Zhu, Shuai, Liu, & Liu, 2023)	2	–	–	91.9	96.2	88.6	90.0	TIP'23
	TA-HGCN-FC (Huang, Qin, et al., 2023)	2	–	–	90.8	96.4	87.0	88.4	TCSVT'23
	ActCLR (Lin et al., 2023)	3	1.7	2.5	88.2	93.9	82.1	84.6	CVPR'23
	DD-GCN (Li et al., 2023)	2	17.4	5.7	92.6	96.9	88.9	90.2	ICME'23
	HiCLR (Zhang et al., 2023)	3	3.5	4.7	90.4	95.7	85.6	87.5	AAAI'23
	SaPR-GCN (Li, Mao, et al., 2023)	4	6.6	8.3	92.4	96.4	88.7	90.3	TCSVT'23
	EfficientGCN(B4)(Song, Zhang, Shan, & Wang, 2023)	1	15.24	2.0	91.7	95.7	88.3	89.1	TPAMI'23
	SkeAttnCLR (Hua et al., 2023)	3	10.4	9.2	89.4	94.5	83.4	92.7	IJCAI'23
	STHG-DAN (Wu et al., 2024)	3	5.2	2.7	91.2	96.5	88.7	89.8	PR'24
	ACE-ens (Qin et al., 2024)	2	78.0	5.8	91.6	96.3	88.2	89.2	TNNLS'24
	Bs-MSS-GCN(ours)	1	7.9	2.7	90.6	95.7	86.9	87.6	–
	Js-MSS-GCN(ours)	1	7.9	2.7	90.2	95.3	86.2	87.1	–
	MSS-GCN(ours)	2	15.8	5.4	**92.9**	**97.0**	**89.1**	**90.5**	–

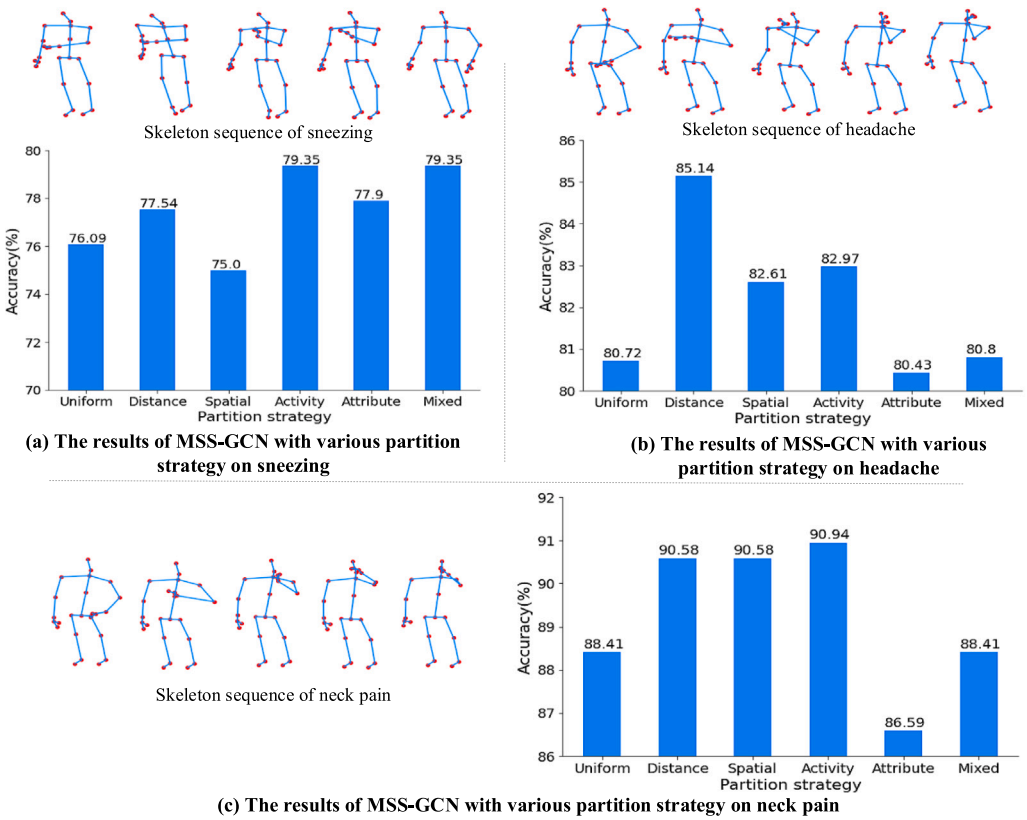
Js is the raw joint stream, and Bs is the bone stream.

5.3. Characteristics of the partition strategy

The partition strategy directly influences the weight-sharing mechanism in graph convolution, determining how the neighborhoods aggregate their information. Hence, the partition strategy plays a crucial role in GCNs. We derive several variants of MSS-GCN by keeping other network structures unchanged and solely modifying the partition strategy. Building upon this, we further explore the characteristics of different partition strategies mentioned in Section 3.2, as depicted in Fig. 8. We discover that the activity, attribute, and mixed partition strategies exhibit superior performance in recognizing actions with limited motion information, like sneezing, with the activity partition strategy being optimal (see Fig. 8(a)). Besides, Fig. 8(b-c) illustrates that distance, activity, and spatial partition strategies outperform others when distinguishing highly identical movements such as headache and neck pain. After conducting extensive analysis, we have concluded that optimizing partitioning strategies based on the inherent skeleton structure of the human body can significantly enhance the performance of GCN-based action recognition methods. Integrating joint attributes into the partition strategy can provide additional prior knowledge for instantaneous action recognition. Additionally, factoring in the relative distances between joints in the partition strategy can highlight spatial configuration, enabling a more in-depth analysis of subtle actions. Researchers can select GCNs with suitable partitioning strategies for specific applications to achieve better performance. Here are the guidelines. (1) Graph Convolutional Networks (GCNs) that incorporate a partition strategy optimized by joint properties (e.g. mixed and activity strategy) have shown improved ability to recognize transient and abnormal actions. These methods can be utilized in public security, intelligent elderly care, and similar fields. (2) GCNs with partitioning strategies that emphasize relative distance between joints (e.g., distance and spatial strategy) are often effective for fine-grained and similar action analysis. We suggest that these methods can be applied to human resource management, behavioral psychology, and related areas.

6. Conclusion

In this work, we modeled human skeleton sequences as multi-sliced spatial–temporal graphs to represent diverse actions and mitigate occlusion interference. We presented two practical components compatible with GCNs, i.e., EAGC and Multi-STE, for capturing discriminative motion patterns. EAGC optimized the weight-sharing mechanism of graph convolution by extended partition strategies and enhanced the representation capability of GCNs. Multi-STE emphasized spatial–temporal synchronization and multi-granular analysis with various motion slices. On this basis, we proposed MSS-GCN, a robust framework that generates multi-granular and synchronized spatial–temporal features impaired by previous factorized paradigms. The extensive experiments on the NTU-RGB+D, NTU-RGB+D 120, and NW-UCLA datasets demonstrated that MSS-GCN outperforms existing methods. In future work, we will focus on fine-grained action analysis via guiding deep learning models to emphasize subtle patterns in specific body parts, such as hands and heads.



(a) The results of MSS-GCN with various partition strategy on sneezing

(b) The results of MSS-GCN with various partition strategy on headache

(c) The results of MSS-GCN with various partition strategy on neck pain

Fig. 8. Results of variants of MSS-GCN with various partition strategies. We analyze the characteristics of six partition strategies on specific action classes, including instantaneous and similar actions, and report the three cases above.

Table 8
Performance in the case of occlusion.

Methods	Accuracy (%)				
	Body occlusion				
	Left-arm	Right-arm	Hands	Legs	Trunk
ST-GCN† (Yan et al., 2018)	71.4	60.5	62.6	77.4	50.2
2s-AGCN† (Shi et al., 2019b)	72.4	55.8	82.1	74.1	71.9
RA-GCNv1† (Song, Zhang, & Wang, 2019)	73.4	60.4	73.5	81.8	70.6
RA-GCNv2† (Song et al., 2021)	74.5	59.4	74.2	83.2	72.3
CTR-GCN* (Chen, Zhang, et al., 2021)	83.2	81.9	85.5	78.5	80.7
SaPR-GCN* (Li, Mao, et al., 2023)	81.2	79.5	82.7	82.8	**86.3**
DD-GCN* (Li et al., 2023)	76.7	62.5	77.8	76.4	79.5
MSS-GCN	**83.5**	**82.5**	**86.6**	**85.1**	85.6
Methods	Time occlusion				
	10%	20%	30%	40%	50%
ST-GCN† (Yan et al., 2018)	69.3	57.0	44.5	34.5	24.0
2s-AGCN† (Shi et al., 2019b)	74.8	60.8	49.7	38.2	28.0
RA-GCNv1† (Song et al., 2019)	85.9	81.9	75.0	66.3	40.6
RA-GCNv2† (Song et al., 2021)	83.9	76.4	66.3	53.2	38.5
CTR-GCN*(Chen, Zhang, et al., 2021)	86.1	83.2	79.1	72.8	65.6
SaPR-GCN* (Li, Mao, et al., 2023)	83.1	81.9	78.3	72.2	64.5
DD-GCN* (Li et al., 2023)	86.8	84.7	81.0	76.1	69.7
MSS-GCN	**87.2**	**85.0**	**81.4**	**76.2**	**69.9**

† represents the results referred to in other papers.
* denotes the methods we reproduced and trained by the bone stream.

CRediT authorship contribution statement

Chang Li: Methodology, Investigation, Data curation, Writing – original draft, Validation, Visualization. **Qian Huang:** Conceptualization, Software, Writing – review & editing, Supervision. **Yingchi Mao:** Resources, Supervision, Writing – review & editing, Project administration. **Xing Li:** Methodology, Investigation, Visualization. **Jie Wu:** Conceptualization, Software, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under grant KYCX23_0753, the Fundamental Research Funds for the Central Universities under grant B230205027, the Key Research and Development Program of China under grant 2022YFC3005401, the Key Research and Development Program of China, Yunnan Province under grant 202203AA080009, the 14th Five-Year Plan for Educational Science of Jiangsu Province under grant D/2021/01/39, and the Jiangsu Higher Education Reform Research Project under grant 2021JSJG143.

References

Avola, D., Cascio, M., Cinque, L., Foresti, G. L., Massaroni, C., & Rodolà, E. (2020). 2-d skeleton-based action recognition via two-branch stacked LSTM-RNNs. *IEEE Transactions on Multimedia*, *22*(10), 2481–2496. http://dx.doi.org/10.1109/TMM.2019.2960588.

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *International conference on computer vision* (pp. 13339–13348). http://dx.doi.org/10.1109/ICCV48922.2021.01311.

Chen, T., Zhou, D., Wang, J., Wang, S., Guan, Y., He, X., et al. (2021). Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *29th ACM international conference on multimedia* (pp. 4334–4342). http://dx.doi.org/10.1145/3474085.3475574.

Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., & Lu, H. (2020). Decoupling GCN with DropGraph module for skeleton-based action recognition. In *European conference on computer vision* (pp. 536–553). http://dx.doi.org/10.1007/978-3-030-58586-0_32.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 180–189). http://dx.doi.org/10.1109/CVPR42600.2020.00026.

Cheng, K., Zhang, Y., He, X., Cheng, J., & Lu, H. (2021). Extremely lightweight skeleton-based action recognition with shiftgcn++. *IEEE Transactions on Image Processing*, *30*, 7333–7348. http://dx.doi.org/10.1109/TIP.2021.3104182.

Deng, Z., He, S., Zhang, M., & Wang, Y. (2022). A skeleton posture transfer method from kinect capture. In *2022 international conference on virtual reality, human-computer interaction and artificial intelligence* (pp. 150–155). http://dx.doi.org/10.1109/VRHCIAI57205.2022.00033.

Evangelidis, G., Singh, G., & Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In *2014 22nd international conference on pattern recognition* (pp. 4513–4518). http://dx.doi.org/10.1109/ICPR.2014.772.

Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., & Ding, R. (2022). Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI conference on artificial intelligence* (pp. 762–770). http://dx.doi.org/10.1609/aaai.v36i1.19957.

Hou, Y., Li, Z., Wang, P., & Li, W. (2018). Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(3), 807–811. http://dx.doi.org/10.1109/TCSVT.2016.2628339.

Hu, G., Cui, B., & Yu, S. (2020). Joint learning in the spatio-temporal and frequency domains for skeleton-based action recognition. *IEEE Transactions on Multimedia*, *22*(9), 2207–2220. http://dx.doi.org/10.1109/TMM.2019.2953325.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). http://dx.doi.org/10.1109/CVPR.2018.00745.

Hua, Y., Wu, W., Zheng, C., Lu, A., Liu, M., Chen, C., et al. (2023). Part aware contrastive learning for self-supervised action recognition. In *32th international joint conference on artificial intelligence* (pp. 855–863). http://dx.doi.org/10.24963/ijcai.2023/95.

Huang, H., Guo, X., Peng, W., & Xia, Z. (2023). Micro-gesture classification based on ensemble hypergraph-convolution transformer. In *2023 international IJCAI workshop on micro-gesture analysis for hidden emotion understanding, vol. 3522* (pp. 1–9).

Huang, L., Huang, Y., Ouyang, W., Wang, L., & Assoc Advancement Artificial, Intelligence (2020). Part-level graph convolutional network for skeleton-based action recognition. In *AAAI conference on artificial intelligence, vol. 34* (pp. 11045–11052). http://dx.doi.org/10.1609/aaai.v34i07.6759.

Huang, Z., Qin, Y., Lin, X., Liu, T., Feng, Z., & Liu, Y. (2023). Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(4), 1868–1883. http://dx.doi.org/10.1109/TCSVT.2022.3217763.

Kong, J., Bian, Y., & Jiang, M. (2022). MTT: Multi-scale temporal transformer for skeleton-based action recognition. *IEEE Signal Processing Letters*, *29*, 528–532. http://dx.doi.org/10.1109/LSP.2022.3142675.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2019). Actional-structural graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 3590–3598). http://dx.doi.org/10.1109/CVPR.2019.00371.

Li, C., Hou, Y., Li, W., Ding, Z., & Wang, P. (2024). DFN: A deep fusion network for flexible single and multi-modal action recognition. *Expert Systems with Applications*, *245*, Article 123145. http://dx.doi.org/10.1016/j.eswa.2024.123145.

Li, C., Huang, Q., & Mao, Y. (2023). DD-GCN: Directed diffusion graph convolutional network for skeleton-based human action recognition. In *IEEE international conference on multimedia and expo* (pp. 786–791). http://dx.doi.org/10.1109/ICME55011.2023.00140.

Li, C., Mao, Y., Huang, Q., Zhu, X., & Wu, J. (2023). Scale-aware graph convolutional network with part-level refinement for skeleton-based human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 4311–4324. http://dx.doi.org/10.1109/TCSVT.2023.3334872.

Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., & Luo, J. (2016). Action recognition by learning deep multi-granular spatio-temporal video representation. In *ACM international conference on multimedia retrieval* (pp. 159–166). http://dx.doi.org/10.1145/2911996.2912001.

Li, Q., Zhang, Z., Zhang, F., & Xiao, F. (2023). HRNeXt: High-resolution context network for crowd pose estimation. *IEEE Transactions on Multimedia*, *25*, 1521–1528. http://dx.doi.org/10.1109/TMM.2023.3248144.

Lin, L., Zhang, J., & Liu, J. (2023). Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proc. IEEE conf. comput. vis. pattern recognit.* (pp. 2363–2372). Vancouver, BC, Canada: IEEE, http://dx.doi.org/10.1109/CVPR52729.2023.00234.

Liu, K., Gao, L., Khan, N. M., Qi, L., & Guan, L. (2021). A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition. *IEEE Transactions on Multimedia*, *23*, 64–76. http://dx.doi.org/10.1109/TMM.2020.2974323.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (pp. 9992–10002). http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Liu, J., Shahroudy, A., Perez, M., Wang, G., et al. (2020). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(10), 2684–2701. http://dx.doi.org/10.1109/TPAMI.2019.2916873.

Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Computer vision and pattern recognition* (pp. 140–149). http://dx.doi.org/10.1109/cvpr42600.2020.00022.

Pan, H., Chen, Y., & He, Z. (2023). Multi-granularity graph pooling for video-based person re-identification. *Neural Networks*, *160*, 22–33. http://dx.doi.org/10.1016/j.neunet.2022.12.015.

Pang, C., Lu, X., & Lyu, L. (2023). Skeleton-based action recognition through contrasting two-stream spatial-temporal networks. *IEEE Transactions on Multimedia*, 1–14. http://dx.doi.org/10.1109/TMM.2023.3239751.

Peng, W., Hong, X., Chen, H., & Zhao, G. (2020). Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *The thirty-fourth AAAI conference on artificial intelligence*. http://dx.doi.org/10.1609/aaai.v34i03.5652.

Plizzari, C., Cannici, M., & Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, *208*, http://dx.doi.org/10.1016/j.cviu.2021.103219.

Qin, Z., Liu, Y., Ji, P., Kim, D., Wang, L., McKay, R. I., et al. (2024). Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(4), 4783–4797. http://dx.doi.org/10.1109/TNNLS.2022.3201518.

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Computer vision and pattern recognition* (pp. 1010–1019). http://dx.doi.org/10.1109/CVPR.2016.115.

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019a). Skeleton-based action recognition with directed graph neural networks. In *Computer vision and pattern recognition* (pp. 7904–7913). http://dx.doi.org/10.1109/CVPR.2019.00810.

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019b). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Computer vision and pattern recognition* (pp. 12018–12027). http://dx.doi.org/10.1109/CVPR.2019.01230.

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *ACCV* (pp. 38–53). http://dx.doi.org/10.1007/978-3-030-69541-5_3.

Shu, X., Xu, B., Zhang, L., & Tang, J. (2022). Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*, http://dx.doi.org/10.1109/TPAMI.2022.3222871.

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 1227–1236). http://dx.doi.org/10.1109/CVPR.2019.00132.

Song, Y.-F., Zhang, Z., Shan, C., & Wang, L. (2021). Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(5), 1915–1925. http://dx.doi.org/10.1109/TCSVT.2020.3015051.

Song, Y.-F., Zhang, Z., Shan, C., & Wang, L. (2023). Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 1474–1488. http://dx.doi.org/10.1109/TPAMI.2022.3157033.

Song, Y.-F., Zhang, Z., & Wang, L. (2019). Richly activated graph convolutional network for action recognition with incomplete skeletons. In *2019 IEEE international conference on image processing* (pp. 1–5). http://dx.doi.org/10.1109/ICIP.2019.8802917.

Tang, C., Li, W., Wang, P., & Wang, L. (2018). Online human action recognition based on incremental learning of weighted covariance descriptors. *Information Sciences*, *467*, 219–237. http://dx.doi.org/10.1016/j.ins.2018.08.003.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. In *Neural information processing systems* (pp. 6000–6010). http://dx.doi.org/10.5555/3295222.3295349.

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3D skeletons as points in a Lie group. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 588–595). http://dx.doi.org/10.1109/CVPR.2014.82.

Wang, J., Nie, X., Xia, Y., Wu, Y., & Zhu, S. (2014). Cross-view action modeling, learning, and recognition. In *Computer vision and pattern recognition* (pp. 2649–2656). http://dx.doi.org/10.1109/CVPR.2014.339.

Wen, Y., Gao, L., Fu, H., Zhang, F., Xia, S., & Liu, Y.-J. (2022). Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009–2023. http://dx.doi.org/10.1109/TPAMI.2022.3170511.

Wu, Z., Ma, N., Wang, C., Xu, C., Xu, G., & Li, M. (2024). Spatial–temporal hypergraph based on dual-stage attention network for multi-view data lightweight action recognition. *Pattern Recognition*, *151*, Article 110427. http://dx.doi.org/10.1016/j.patcog.2024.110427.

Wu, C., Wu, X., & Kittler, J. (2022). Graph2Net: Perceptually-enriched graph learning for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(4), 2120–2132. http://dx.doi.org/10.1109/TCSVT.2021.3085959.

Xia, R., Li, Y., & Luo, W. (2022). LAGA-Net: Local-and-global attention network for skeleton based action recognition. *IEEE Transactions on Multimedia*, *24*, 2648–2661. http://dx.doi.org/10.1109/TMM.2021.3086758.

Xu, S., Rao, H., Peng, H., Jiang, X., Guo, Y., Hu, X., et al. (2021). Attention-based multilevel co-occurrence graph convolutional LSTM for 3-D action recognition. *IEEE Internet of Things Journal*, *8*(21), 15990–16001. http://dx.doi.org/10.1109/JIOT.2020.3042986.

Xu, K., Ye, F., Zhong, Q., & Xie, D. (2022). Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *AAAI conference on artificial intelligence* (pp. 2866–2874). http://dx.doi.org/10.1609/aaai.v36i3.20191.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI conference on artificial intelligence* (pp. 7444–7452). http://dx.doi.org/10.1609/aaai.v32i1.12328.

Yang, X., & Tian, Y. (2014). Effective 3D action recognition using EigenJoints. *Journal of Visual Communication and Image Representation*, *25*(1), 2–11. http://dx.doi.org/10.1016/j.jvcir.2013.03.001.

Yang, S., Wang, X., Gao, L., & Song, J. (2022). MKE-GCN: Multi-modal knowledge embedded graph convolutional network for skeleton-based action recognition in the wild. In *International conference on multimedia and expo* (pp. 01–06). http://dx.doi.org/10.1109/ICME52920.2022.9859787.

Yang, H., Yan, D., Zhang, L., Sun, Y., Li, D., & Maybank, S. J. (2022). Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, *31*, 164–175. http://dx.doi.org/10.1109/TIP.2021.3129117.

Yao, H., Zhao, S.-J., Xie, C., Ye, K., & Liang, S. (2021). Recurrent graph convolutional autoencoder for unsupervised skeleton-based action recognition. In *2021 IEEE international conference on multimedia and expo* (pp. 1–6). http://dx.doi.org/10.1109/ICME51207.2021.9428403.

Yin, M., He, S., Soomro, T. A., & Yuan, H. (2023). Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. *Expert Systems with Applications*, *226*, Article 120080. http://dx.doi.org/10.1016/j.eswa.2023.120080.

Zeng, Q., Liu, C., Liu, M., & Chen, Q. (2023). Contrastive 3D human skeleton action representation learning via CrossMoCo with spatiotemporal occlusion mask data augmentation. *IEEE Transactions on Multimedia*, *25*, 1564–1574. http://dx.doi.org/10.1109/TMM.2023.3253048.

Zhang, J., Lin, L., & Liu, J. (2023). Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proc. AAAI conf. artif. intell.* (pp. 3427–3435). http://dx.doi.org/10.1609/aaai.v37i3.25451.

Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., & Mei, T. (2019). Multi-granularity reasoning for social relation recognition from images. In *2019 IEEE international conference on multimedia and expo* (pp. 1618–1623). http://dx.doi.org/10.1109/ICME.2019.00279.

Zhang, X., Xu, C., Tian, X., & Tao, D. (2020). Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(8), 3047–3060. http://dx.doi.org/10.1109/TNNLS.2019.2935173.

Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D., et al. (2018). Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. *IEEE Transactions on Multimedia*, *20*(9), 2330–2343. http://dx.doi.org/10.1109/TMM.2018.2802648.

Zhou, Y., Cheng, Z.-Q., Li, C., Geng, Y., Xie, X., & Keuper, M. (2022). Hypergraph transformer for skeleton-based action recognition. http://dx.doi.org/10.48550/arXiv.2211.09590, arXiv preprint arXiv:2211.09590.

Zhu, Y., Shuai, H., Liu, G., & Liu, Q. (2023). Multilevel spatial–temporal excited graph network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, *32*, 496–508. http://dx.doi.org/10.1109/TIP.2022.3230249.