Scilight

Transactions on Graph Intelligence and Network Applications
https://www.sciltp.com/journals/tgina

Article

MG^{2}CL: Multi-Granularity Graph Contrastive Learning for Skeleton-Based Action Recognition

Zhize Wu [1,2], Pengsong Wang [1], Xin Chen [1], Thomas Weise [1], Fei Liu [3,4] and Shengwei Ji [1,4,*]

[1] School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China
[2] State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei 230601, China
[3] Innovation School of Artificial Intelligence, Hefei University of Technology, Hefei 230009, China
[4] Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230009, China
* Correspondence: jisw@hfuu.edu.cn

How To Cite: Wu, Z.; Wang, P.; Chen, X.; et al. MG^{2}CL: Multi-Granularity Graph Contrastive Learning for Skeleton-Based Action Recognition. *Transactions on Graph Intelligence and Network Applications* **2026**. https://doi.org/10.53941/tgina.2026.100007

Received: 4 June 2026
Revised: 5 September 2026
Accepted: 16 September 2026
Published: 22 September 2026

Abstract: Skeleton-based action recognition has gained increasing attention as a robust alternative to RGB-based methods due to its resilience to background clutter and lighting variations. Existing approaches have achieved notable progress, particularly with graph-based models that capture topological and temporal patterns of human joints. However, most methods still struggle to model relationships between non-naturally connected joints and fail to fully exploit the fine-grained spatio-temporal characteristics of skeleton sequences, limiting generalization and recognition accuracy. To address these limitations, we propose MG^{2}CL, a Multi-Granularity Graph Contrastive Learning framework designed to enhance skeleton representation learning. Our approach employs a multi-granularity graph structure to model both short- and long-range joint interactions through complementary body-part relations. Furthermore, we introduce a novel spatio-temporal masking strategy for data augmentation, encouraging the model to learn more diverse and informative patterns. A semantic-level memory bank, built upon the multi-granularity graph, is integrated to reinforce the model's ability to distinguish subtle action variations. Extensive experiments show that MG^{2}CL achieves competitive performance on widely used benchmark datasets, including NTU RGB+D, NTU RGB+D 120, and Northwestern-UCLA. These results demonstrate the framework's strong generalization and discriminative capability. Our findings suggest that incorporating multi-granularity structural information and contrastive learning principles can lead to more robust and flexible skeleton-based action models. MG^{2}CL offers a promising direction for future work in representation learning for spatio-temporal graph data, with potential applications in human-computer interaction, surveillance, and healthcare.

Keywords: graph contrastive learning; spatial–temporal masking strategy; skeleton–based action recognition

1. Introduction

Understanding human actions from visual data is a key task in computer vision, with wide applications in areas such as human-computer interaction and abnormal behavior detection [1,2]. Thanks to deep learning, action recognition based on RGB videos has achieved impressive results [3]. However, these methods are easily affected by background clutter, lighting changes, and variations in clothing [4], which limits their reliability in real-world settings. Skeleton-based action recognition offers a more robust alternative by focusing on the 2D or 3D coordinates of human joints [5–7]. This representation reduces noise and simplifies motion patterns, making it well suited for action understanding. Recent works have shown that learning both spatial structure and temporal dynamics from skeleton sequences can lead to stronger performance [8].

Despite notable progress, many skeleton-based methods still rely on static, predefined physical graphs—where only physically connected joints are treated as neighbors [9]. This fixed topology limits the model's capacity to capture non-naturally connected and distant joints that may be semantically crucial for certain actions. As a result, models often struggle to distinguish actions with subtle joint differences, such as "brushing teeth" versus "combing hair", leading to misclassification.

To overcome the limitations of static graphs, researchers have explored data-driven graph reconstruction. In fact, replacing predefined static topologies with adaptive structures learned directly from data has proven highly effective in capturing complex spatial-temporal dependencies across various domains, such as handling dynamic environments in multi-stream time-series modeling [10]. In the context of skeleton data, one line of work focuses on decoupling the skeleton graph into general and individual topologies, combining multi-level graph convolution with spatial incentive modules to enhance adaptability [11]. Another approach integrates hierarchical decomposition with attention-guided aggregation to jointly model adjacent and distant relationships [12]. While these strategies improve flexibility, they often introduce high coupling between modules, increasing complexity and reducing scalability.

In parallel, contrastive learning and self-attention mechanisms have been incorporated to strengthen contextual and global dependencies in skeleton data [13–16]. For instance, contrastive objectives have been used to enhance inter-node correlations, while self-attention enables long-range interaction modeling. Additionally, skeleton-specific augmentations—such as random body part occlusion and frame mirroring—have been proposed to enrich training diversity [17]. Nevertheless, existing methods still face challenges in effectively capturing relationships between distant joints and exploiting the inherent spatio-temporal structure of skeleton sequences, ultimately hindering their generalization and fine-grained recognition capability, as shown in Figure 1.

Given the growing demand for fine-grained and robust action recognition, multi-granularity representation has gained attention as a promising direction. By analyzing action cues at multiple levels of granularity, this paradigm aims to enrich semantic information and improve discriminability. Existing methods have explored coarse-to-fine decomposition or parallel extraction of coarse and fine features to achieve multi-granularity modeling [18, 19]. However, these approaches often rely on dual-stream architectures, incurring high computational overhead. Efforts to reduce complexity have focused on lightweight fusion designs combining adaptive graph and multi-scale temporal convolutions [20]. Despite these advancements, current solutions generally adopt static granularity partitions and lack mechanisms to exploit the varying importance of different body parts across diverse actions.

In this work, we introduce a flexible and efficient framework called Multi-Granularity Graph Contrastive Learning (MG^{2}CL). Our approach builds a multi-granularity graph representation that captures both local and global joint interactions across complementary body-part levels, allowing the model to learn richer spatio-temporal patterns, as shown in Figure 1. The central observation is that action-discriminative cues are not always located at individual joints or fixed physical connections; they often emerge from relations among functional body parts at different granularities. Based on this observation, we combine multi-granularity body structures with graph contrastive learning so that structural semantics can guide representation learning at both the instance level and the semantic level. We also propose a spatio-temporal masking strategy to enhance data diversity and a memory module that emphasizes key motion patterns to support fine-grained recognition.

Extensive experiments on three widely used benchmarks—NTU RGB+D [21], NTU RGB+D 120 [22], and Northwestern-UCLA [23]—show that MG^{2}CL achieves competitive performance in terms of accuracy and generalization.

The main contributions of this work are:

- A multi-granularity graph structure that captures complementary body-part relations across both local and long-range dependencies.
- A spatio-temporal masking strategy that creates diverse skeleton samples to improve learning robustness.
- A semantic-level memory module that incorporates body-granularity semantics into contrastive learning and enhances the model's ability to distinguish subtle differences in actions.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the proposed method, Section 4 presents experiments, and Section 5 concludes the paper and outlines future directions.

2. Related Work

Skeleton-based action recognition has emerged as a promising alternative to RGB-based methods. In this section, we review related work from three perspectives: graph-based modeling, hierarchical representations, and contrastive learning.

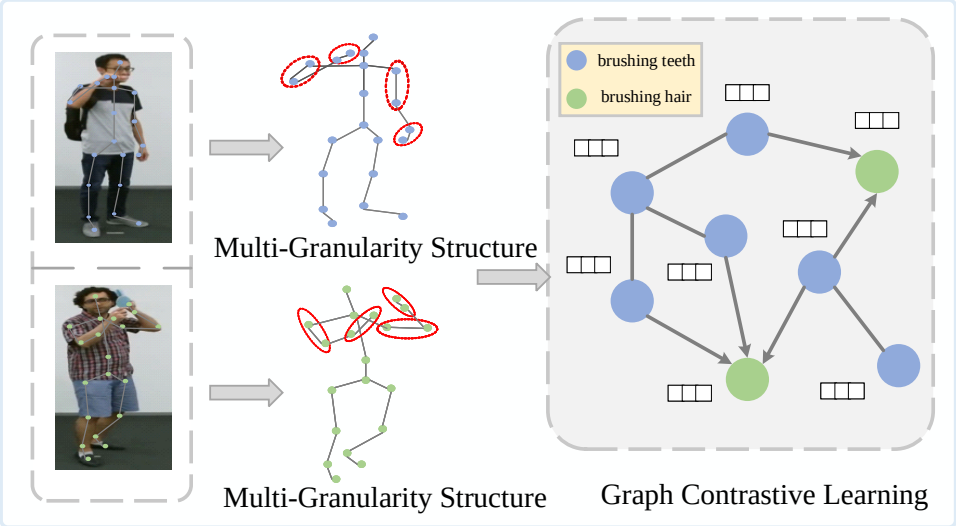

Figure 1. Our approach integrates the proposed Multi–Granularity Structure with Graph Contrastive Learning to capture locally similar features. This enables the model to bring feature points of the same category closer together while pushing those of different categories further apart in the feature space. The multi–granularity graph structure remains highly applicable within the contrastive learning framework, enhancing the transferability of the model.

2.1. Skeleton-Based Graph Modeling

Graph-based models have emerged as a dominant paradigm in skeleton-based action recognition, representing skeletons as graphs where joints are modeled as nodes and physical connections as edges. ST-GCN [9] pioneered the application of spatial-temporal graph convolutional networks (ST-GCN) for end-to-end learning; however, its fixed topology limited the capacity to model non-local dependencies. To overcome this limitation, MS-G3D [24] introduced multi-scale temporal modeling and disentangled spatial-temporal processing, enabling richer motion dynamics to be captured. CTR-GCN [25] further enhanced model adaptability by dynamically learning channel-wise topologies during training. Building upon these foundations, subsequent studies explored data-driven graph construction and attention mechanisms to facilitate dynamic topology learning. For example, LG-SGNet [26] proposed a Local-Global Graph Convolutional Network (LG-GCN) for parallel spatial modeling, and integrated it with a Local-Global Temporal Convolutional Network employing multi-head self-attention to capture global temporal dependencies. DEGCN [27] introduced an explicit dependency modeling mechanism to capture semantic correlations beyond physical connectivity. Skateformer [7] incorporated transformer-based modules into skeleton graphs, enabling long-range dependency learning across non-adjacent joints. SkeletonMAE [6] leveraged masked autoencoders during pretraining to enhance structural feature learning. Additionally, SelfGCN [15] combined graph convolution and self-attention in parallel to jointly model local and global features.

Despite these advances, many graph-based models still rely on static or predefined adjacency matrices, limiting their capacity to capture semantic relationships among physically distant yet functionally correlated joints. Such constraints reduce the expressive power of these models in recognizing complex and subtle actions. Moreover, most existing graph models primarily focus on node-level or pairwise connectivity, often overlooking the multi-granularity structural semantics inherent in the skeleton hierarchy. These limitations highlight the need for more flexible and semantically enriched graph representations that can effectively capture structural semantics across multiple granular levels.

2.2. Hierarchical Representations

Hierarchical representations aim to capture structured semantics across different body levels, including joints, body parts, and the entire body. PL-GCN [28] introduced part-level pooling and unpooling operations to learn latent representations of body parts for action recognition. Building on this idea, ML-STGNet [11] integrated joint-level, part-level, and body-level graphs to enrich hierarchical semantic modeling. HD-GCN [12] proposed a hierarchical decomposition strategy combined with attention-guided aggregation to jointly model both adjacent and distant relationships. PSUMNet [29] employed a part-streaming mechanism to facilitate partial information flow and enhance intra-part interactions across different body parts. These hierarchical approaches have improved intra-part interactions and cross-part connectivity, thereby contributing to more effective semantic representation learning.

Nevertheless, most hierarchical models treat different body levels independently and lack explicit mechanisms for cross-level feature interaction. Such separation limits the fusion of multi-level semantic cues, which is critical for fine-grained action recognition. Moreover, existing hierarchical models are primarily designed under supervised learning settings and are not readily adaptable to self-supervised or contrastive learning frameworks. Bridging hierarchical semantics with contrastive objectives remains an open challenge in skeleton-based action recognition.

2.3. Contrastive Learning

Contrastive learning has emerged as a powerful unsupervised approach for skeleton-based action recognition, aiming to learn discriminative representations by contrasting positive and negative sample pairs. CrosSCLR [30] proposed a single-view and cross-view contrastive learning framework to improve representation quality under different viewpoints. SkeletonGCL [13] introduced dual memory banks to simultaneously capture instance-level and semantic-level contextual information during training. SCD-Net [31] employed spatio-temporal structure masking to disentangle spatial and temporal cues while refining data augmentation strategies. Masked feature reconstruction [32] focused on recovering key motion features from masked skeleton inputs to enhance model robustness. JMDA [33] utilized joint spatio-temporal mixing modules with feature alignment to diversify training samples and improve generalization. Skelbumentations [17] designed real-world-inspired skeleton augmentations, such as random part occlusion and frame mirroring, to increase sample diversity. Recent studies have also explored integrating contrastive learning with graph structures. For example, Pang et al. [14] combined graph contrastive learning with topology refinement to enhance global dependency modeling. Similarly, Tian et al. [16] employed contrastive objectives to strengthen inter-node correlations and contextual dependencies.

Despite these advances, most contrastive learning methods focus primarily on instance-level or single-level representations and lack mechanisms for coordinating features across hierarchical semantic levels. Moreover, existing frameworks underexplore structural constraints to guide contrastive objectives, limiting their ability to leverage multi-level semantics and structural relationships inherent in skeleton data.

In summary, prior research has significantly advanced skeleton-based action recognition through graph modeling, hierarchical representations, and contrastive learning. However, two key challenges remain unresolved. First, existing methods struggle to jointly model local and global interactions among semantically relevant yet physically distant joints. Second, they lack strategies to integrate hierarchical semantic representations within contrastive learning frameworks. Addressing these challenges requires a unified approach that seamlessly combines multi-granularity graph structures with contrastive objectives.

To this end, we propose a novel framework that integrates a multi-granularity graph structure, a spatio-temporal masking strategy, and a semantic-level memory bank. This design facilitates effective cross-level interactions and enhances feature discrimination through multi-level contrastive learning. Our approach improves model robustness and fine-grained recognition performance for skeleton-based action recognition. By bridging structural modeling and contrastive learning under a unified perspective, our framework offers a new solution for learning semantically enriched and structurally aware representations from skeleton data. Compared with hierarchical GCN methods such as HD-GCN, MG^{2}CL does not only decompose the skeleton into hierarchical physical structures for supervised feature aggregation. Instead, it uses multi-granularity body partitions as structural priors and further introduces them into the contrastive learning objective. Compared with graph contrastive methods such as SkeletonGCL, MG^{2}CL does not only contrast instance-level or contextual representations. It explicitly constructs semantic-level contrastive relations among coarse-grained, upper fine-grained, and lower fine-grained body representations, thereby linking body-level structural semantics with contrastive representation learning.

3. Method

We now first introduce the overall framework in Section 3.1. In Section 3.2, we provide a detailed description of our Multi-Granularity graph structure. In Section 3.3, we develop our graph contrastive learning module. Finally, in Section 3.4, we present a novel Spatio-temporal data augmentation strategy.

3.1. Overall Framework

The overall architecture of our Multi-Granularity Graph Contrastive Learning (MG^{2}CL) framework is comprised of two main branches: the GCN Encoder and the Momentum Encoder, as illustrated in Figure 2. Our approach integrates a novel multi-granularity graph structure with enhanced data augmentation strategies within the dual-path encoder, which enables efficient feature extraction and representation learning. Each branch is equipped with memory banks at different granularity levels, allowing for the extraction of both instance-level and semantic-level features. In this work, the body-part partitions used to define coarse-grained, upper fine-grained, and

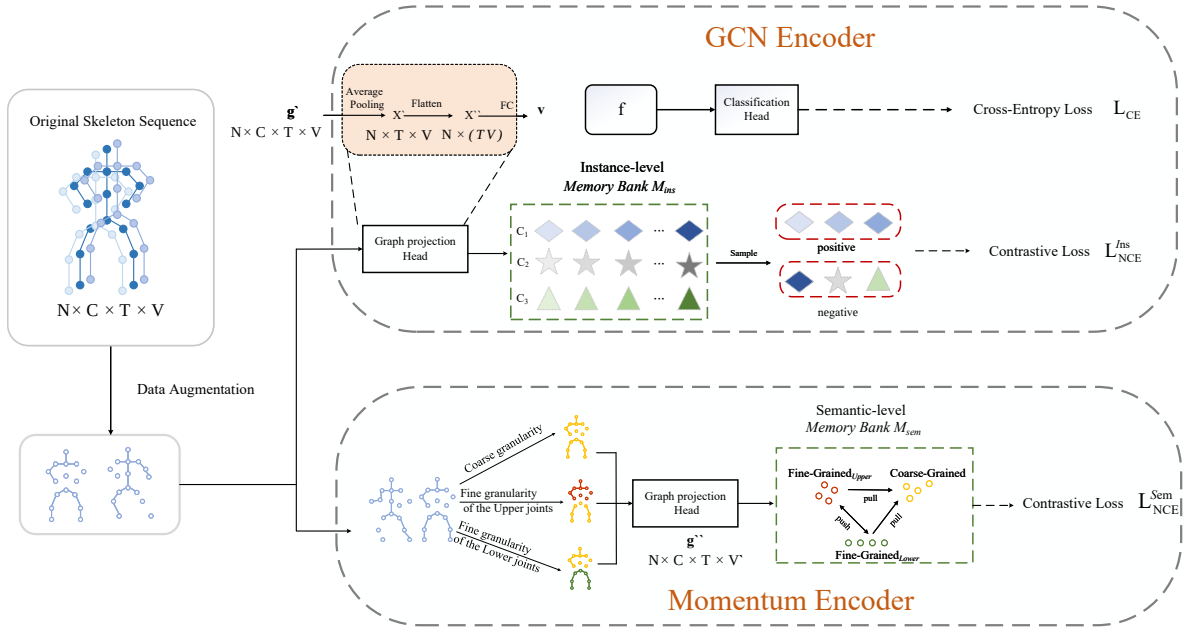

Figure 2. Framework of MG^{2}CL. The original skeleton sequence, enhanced by data augmentation, is fed into a dual-branch encoder. The upper branch represents the GCN Encoder constructing the Instance-level memory bank, while the lower branch represents the Momentum Encoder constructing the semantic-level memory bank.

lower fine-grained structures are predefined. The dynamic aspect of MG^{2}CL lies in the learned feature interactions, the momentum update of the semantic-level memory bank, and the action-dependent contribution of different granularity relations during representation learning.

Before being processed by the dual-branch encoder, the original skeleton sequence $g \in \mathbb{R}^{N\times C\times T\times V}$ undergoes spatio-temporal masking within the data augmentation module. This augmentation step transforms the sequence into an augmented version $g' \in \mathbb{R}^{N\times C\times T\times V}$, which enhances the model's generalization ability. By simulating the variability of real-world scenarios, this augmentation process serves as a critical component of our methodology, improving the robustness of the learned representations.

The Graph Convolutional Encoder is responsible for constructing the instance-level memory bank by mapping the skeleton sequence into a feature space via the Graph Projection Head. In this feature space, samples from the same class are pulled closer together, while those from different classes are pushed apart, facilitating discriminative learning.

In parallel, the Momentum Encoder utilizes a skeleton reconstruction unit to generate features at various levels of granularity. These features serve as the foundation of the semantic-level memory bank. Within each class, features from different body-part granularities are coordinated to preserve complementary semantic information, enabling a more refined representation of action semantics.

Through this dual-branch encoder design, our framework leverages both instance-level and semantic-level feature learning to capture subtle and significant variations in skeleton-based action recognition. This architecture combines graph modeling with contrastive learning in a way that links structural cues across body-part granularities to semantic representation learning.

3.2. Multi-Granularity Graph Structure

We draw inspiration from the hierarchical graph structure of ML-STGNet [11] and propose a multi-granularity graph structure with three levels of skeleton segmentation. Our approach divides the human body into three levels of granularity: coarse-grained, upper fine-grained, and lower fine-grained, as shown in Figure 3. The coarse-grained level segments the body into six major parts: head, torso, upper limbs, hands, lower limbs, and feet. This level captures broad action patterns, such as "jumping" or "standing", which involve large-scale body movements. At the upper fine-grained level, the upper limbs are subdivided into upper arms and forearms to capture detailed arm movements. This level helps the model differentiate between actions like "brushing teeth" and "combing hair", which involve distinct wrist movements. Similarly, the lower limbs are divided into thighs and calves at the lower fine-grained level, capturing subtle variations in leg movements, such as those seen in "kicking" versus "running." These complementary divisions allow the model to capture both coarse motion patterns and fine local movements. In this way, our multi-granularity graph structure provides body-part representations that are aligned with action

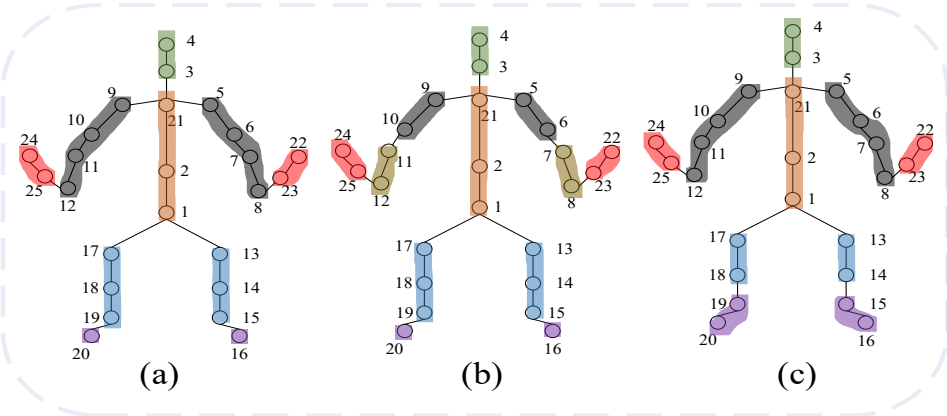

Figure 3. Definition of a multi-granularity graph structure. (**a**) Coarse granularity representation of the human body; (**b**) Fine granularity representation of the upper joints of the human body; (**c**) Fine granularity representation of the lower joints of the human body. The numbers denote the joint indices of the skeleton.

semantics at different levels of detail. By segmenting the body into different granularity levels, our approach helps the model learn more informative structural relationships for skeleton-based actions.

3.3. Graph Contrastive Learning

The graph contrastive learning module is composed of the GCN encoder and Momentum encoder, as illustrated in Figure 2. Within the GCN encoder, we construct an instance-level memory bank to store the feature representations of augmented skeleton sequences, enabling differentiation between samples from different categories. In parallel, the Momentum encoder employs a semantic-level memory bank that stores feature representations at different body-part granularities, enriching the semantic structure learned by the model. Through our ablation studies, we found that the InfoNCE loss function outperforms commonly used alternatives. The InfoNCE loss works by calculating the similarity between the anchor and both positive and negative samples, effectively pulling samples from the same category closer together while pushing those from different categories apart. This loss function facilitates learning by reinforcing the model's ability to distinguish between different action categories. In the following sections, we will provide a detailed breakdown of the module components and the processes within each encoder.

3.3.1. Graph Convolution Network Encoder

The GCN encoder maps the input skeleton sequence into a feature vector $f(x)$, which is then fed into the classification head to produce the predicted class-probability vector $p = (p_1, \ldots, p_K)$ for K action classes. Given the one-hot ground-truth label $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_K)$, the multi-class cross-entropy loss is defined as follows:

$$L_{CE}(p, \hat{y}) = -\sum_{c=1}^{K} \hat{y}_c \log p_c, \tag{1}$$

where p_c denotes the predicted probability of class c, $\hat{y}_c$ is the corresponding one-hot ground-truth label, and K is the total number of action classes.

The GCN encoder constructs an instance-level memory bank $\mathcal{M}_{Ins}$. Initially, the augmented skeletal sequences are fed into the Graph Projection Head. This module processes the input vector $g' \in \mathbb{R}^{N\times C\times T\times V}$. After average pooling along the channel dimension, we obtain $X' \in \mathbb{R}^{N\times T\times V}$. The temporal and joint dimensions are then flattened to obtain $X'' \in \mathbb{R}^{N\times(TV)}$. Finally, a fully connected layer is employed to project the flattened feature vectors into a unified feature space, thereby establishing the instance-level memory bank $\mathcal{M}_{Ins}$. Throughout the construction process, the InfoNCE loss function is utilized, and the instance-level contrastive loss L_{NCE}^{Ins} is defined as Equation (2).

$$L_{NCE}^{Ins} = -\sum_{v^+\in\mathcal{N}_{Ins}^+} \log \frac{\exp\left(\mathrm{sim}(v, v^+)/\tau\right)}{\exp\left(\mathrm{sim}(v, v^+)/\tau\right) + \sum_{v^-\in\mathcal{N}_{Ins}^-} \exp\left(\mathrm{sim}(v, v^-)/\tau\right)}, \tag{2}$$

Here, $\mathcal{N}_{Ins}^+$ and $\mathcal{N}_{Ins}^-$ denote the positive and negative sample sets retrieved from the instance-level memory bank, respectively.

3.3.2. Momentum Encoder

In the Momentum encoder, the augmented skeletal sequence $g' \in \mathbb{R}^{N\times C\times T\times V}$ is passed through the proposed skeletal reconstruction unit module to capture features at different granularities in the spatial dimension, resulting in an output vector $g'' \in \mathbb{R}^{N\times C\times T\times V'}$. Here, V' denotes the number of reconstructed body-part nodes, whereas V denotes the number of joints in the original skeleton sequence. Similar to the GCN encoder, the output vector is then mapped through the Graph Projection Head to align the two branches in the same space. With the introduction of the skeletal reconstruction unit, each joint is associated with body-part representations at different granularities, thereby preserving richer semantic information.

When constructing the semantic-level memory bank $\mathcal{M}_{Sem}$, the feature vectors at different body-part granularities are organized to preserve complementary semantic information. Coarse-grained representations provide body-level context, while upper- and lower-body fine-grained representations preserve more localized motion cues within the same semantic space. This design allows the model to capture both shared and complementary information among different granularities. The construction of the semantic-level loss function L_{NCE}^{Sem} is as follows:

$$L_{NCE}^{Sem} = -\sum_{v^+\in\mathcal{N}_{Sem}^+} \log \frac{\exp\left(\mathrm{sim}(v,v^+)/\tau\right)}{\exp\left(\mathrm{sim}(v,v^+)/\tau\right) + \sum_{v^-\in\mathcal{N}_{Sem}^-}\exp\left(\mathrm{sim}(v,v^-)/\tau\right)}, \tag{3}$$

Similarly, $\mathcal{N}_{Sem}^+$ and $\mathcal{N}_{Sem}^-$ denote the positive and negative sample sets retrieved from the semantic-level memory bank, respectively. For both contrastive objectives, samples with the same action label as the anchor are treated as positive samples, whereas samples with different action labels are treated as negative samples. The similarity function is defined as $\mathrm{sim}(u,v) = u^{\mathsf{T}}v$ after ℓ_2 normalization of both feature vectors, which is equivalent to cosine similarity. The temperature coefficient is set to $\tau = 0.8$.

The update of $\mathcal{M}_{Sem}$ follows a momentum updating strategy, as illustrated in the following equation:

$$m_{sem}^{c^*} \leftarrow \alpha m_{sem}^{c^*} + (1-\alpha)v, \tag{4}$$

Here, $m_{sem}^{c^*}$ represents the semantic prototype of class c^*, v denotes the anchor feature after projection, and α is the momentum coefficient used to update the semantic-level memory bank. In our experiments, we set $\alpha = 0.9$. This means that 90% of the previous semantic prototype and 10% of the current projected anchor feature are used during each update.

The total loss function L is composed of both the contrastive loss and the classification loss. Specifically, the contrastive loss is further divided into two parts: the instance-level contrastive loss $L_{\mathrm{NCE}}^{\mathrm{Ins}}$ and the semantic-level contrastive loss $L_{\mathrm{NCE}}^{\mathrm{Sem}}$. The classification loss is denoted by L_{CE}. Therefore, the total loss function L can be explicitly represented as follows:

$$L = L_{CE} + L_{NCE} = L_{CE} + L_{NCE}^{Ins} + L_{NCE}^{Sem}. \tag{5}$$

3.4. Data Augmentation

Human activities captured by cameras often face challenges such as obstruction, low light conditions, or extremely brief durations. To bolster the capacity of models in discerning specific behaviors amidst complex environments, we drew inspiration from Skelbumentations [17] and crafted the spatiotemporal masking data augmentation module depicted in Figure 4. This module encompasses fundamental skeleton preprocessing tasks like cropping and rotation.

Building on our multi-granularity graph structure, a spatial masking strategy employing random part masking and a temporal masking strategy employing a randomly selected consecutive frame window were employed. By interlinking basic skeleton preprocessing, random part masking, and random frame masking, the spatio–temporal masking strategy was formulated. The pseudo-code for spatial and temporal masking is outlined in Algorithm 1.

The original skeleton sequence $P \in \mathbb{R}^{C\times T\times V}$ is the input, where C signifies the number of channels, T represents the total video frames, V denotes the number of joints, and B denotes the set of body parts including left leg, right leg, left arm, right arm, and back. The algorithm uses two independently sampled temporal windows for spatial and temporal masking. The parameters spatial_size and temporal_size denote the lengths of the corresponding masking windows. The variables spatial_start and spatial_end define the window used for spatial masking, whereas temporal_start and temporal_end define the window used for temporal masking. The variable part denotes a randomly selected body part from the set B.

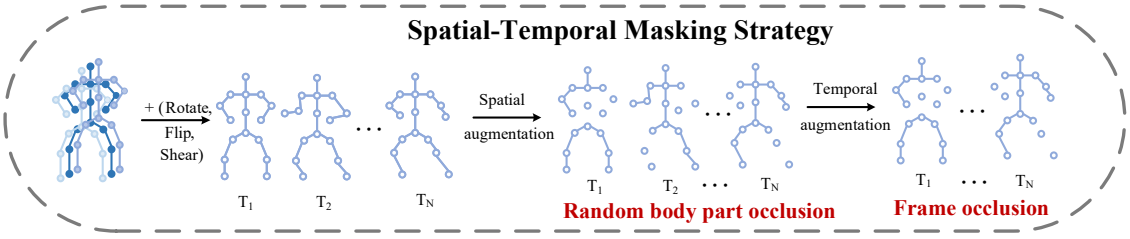

Figure 4. Spatio-temporal masking strategy. The original skeleton sequence undergoes traditional augmentation operations, followed by serial application of random body part occlusion and frame occlusion to obtain the enhanced sequence.

Algorithm 1 Spatio-Temporal Masking Strategy

Require: Skeleton sequence $P \in \mathbb{R}^{C\times T\times V}$, body parts set B
Ensure: Augmented skeleton sequence P
spatial_size $\leftarrow$ Random integer in $[1, T-1]$
spatial_start $\leftarrow$ Random integer in $[1, T - \text{spatial\_size} + 1]$
spatial_end $\leftarrow$ spatial_start + spatial_size $- 1$
part $\leftarrow$ Choose a random body part $b \in B$
for $i =$ spatial_start **to** spatial_end **do**
    $P[:, i, \text{part}] \leftarrow 0$ ▷ Spatial masking
end for
temporal_size $\leftarrow$ Random integer in $[1, T-1]$
temporal_start $\leftarrow$ Random integer in $[1, T - \text{temporal\_size} + 1]$
temporal_end $\leftarrow$ temporal_start + temporal_size $- 1$
for $i =$ temporal_start **to** temporal_end **do**
    $P[:, i, :] \leftarrow 0$ ▷ Temporal masking
end for
return P

The spatial masking strategy masks the joints belonging to a randomly selected body part within the interval from spatial_start to spatial_end. Subsequently, temporal masking is applied to a separately sampled consecutive-frame window from temporal_start to temporal_end, where all joints are masked. Thus, the spatial and temporal masking operations are performed on independently sampled temporal intervals, and frames outside the temporal masking window are retained. Ultimately, the processed skeleton sequence P is returned.

4. Experiments

To verify the effectiveness of our method, we conduct extensive experiments on three important benchmark datasets and report the Top-1 accuracies.

4.1. Datasets

The NTU RGB+D [21] dataset contains 60 action classes and 56,880 samples, including 40 daily action classes, 9 health-related action classes, and 11 interaction action classes, performed by 40 subjects aged between 10 and 35. It was captured using Microsoft Kinect v2 sensors from three different camera angles and the data was represented in the form of 3D skeleton information. Two standards were used to split the dataset into training and testing sets: (1) cross-subject (X–Sub): 20 designated subjects were used for training and the rest were used for testing. (2) cross-view (X–View): the data captured by one camera was used for training and the data captured by the other two cameras were used for testing.

NTU RGB+D 120 [22] is currently the largest dataset with 3D joint annotations. It extends the NTU RGB+D dataset by adding 57,600 skeleton sequences and 60 additional action classes, resulting in 114,480 samples with 120 action classes captured by three cameras from 106 volunteers. The dataset is again split into training and testing sets using two different methods: (1) cross-subject (X–Sub): the training data comes from actions performed by 53 volunteers, and the testing data comes from the rest of the volunteers. (2) cross-setup (X–Set): the training data comes from samples with even IDs, and the testing data comes from samples with odd IDs.

The Northwestern-UCLA [23] dataset consists of 1494 video clips with 10 action classes performed by 10 different subjects. The data is captured from multiple camera views using three Kinect cameras. The training samples are sourced from two of them, while the testing data is from the other.

4.2. Experimental Setup

The experiments were conducted using the PyTorch deep learning framework, running on NVIDIA GeForce RTX 3090 GPUs. The baseline model employed was CTR–GCN [25]. Training was carried out for 65 epochs across all three datasets, with an initial learning rate set to 0.1. A learning rate decay factor of 0.1 was applied at the 35th and 55th epochs. To stabilize the training process, a warm-up strategy was implemented during the first 5 epochs.

The InfoNCE loss function was used for both the semantic and instance-level memory banks, and model optimization was performed using stochastic gradient descent (SGD) with a momentum parameter of 0.9.

For reproducibility, the temperature coefficient in both contrastive objectives was set to $\tau = 0.8$. The momentum coefficient used to update the semantic-level memory bank was set to $\alpha = 0.9$. The capacity of each memory bank was set to the number of training samples in the corresponding dataset. For each anchor feature, up to 128 samples with the same action label were selected as positive samples, while up to 512 samples with different action labels were selected as negative samples. The instance-level memory bank stores projected instance features for contrastive retrieval, whereas the semantic-level memory bank maintains class-level semantic prototypes updated using the momentum rule described in Section 3.3.2.

For the NTU RGB+D [21] and NTU RGB+D 120 [22] datasets, we adhered to the data preprocessing steps detailed in CTR–GCN [25], utilizing a batch size of 64. For the Northwestern–UCLA dataset [23], we maintained consistent preprocessing while adjusting the batch size to 16.

Following the standard multi-stream setting for skeleton-based action recognition, we use four input streams, namely the joint stream, the bone stream, the joint-motion stream, and the bone-motion stream. The predictions from these four streams are fused to obtain the final recognition result.

4.3. Comparison with State-of-the-Art Methods

In this section, we evaluate the performance of our method by integrating it with GCN architectures on three benchmark datasets. The results, summarized in Table 1, demonstrate consistent improvements in accuracy and indicate that the proposed design remains competitive while keeping the model lightweight.

One of the most notable results is the integration of our MG^{2}CL architecture with CTR-GCN [25]. This combination achieved accuracy improvements of 0.8%, 0.9%, and 0.6% on the NTU RGB+D [21], NTU RGB+D 120 [22], and Northwestern-UCLA [23] datasets, respectively, compared to the CTR-GCN baseline.

Additionally, our method demonstrates competitive performance against other recent advancements in skeleton-based action recognition, such as the Multi-Level Spatial-Temporal Graph Network (ML-STGNet) [11], Graph Contrastive Learning (SkeletonGCL) [13], Efficient-B4 [34], and Self-Attention Graph Convolution Network (SelfGCN) [15].

Compared with the CTR-GCN baseline, MG^{2}CL increases the parameter count from 5.80 M to 7.20 M and the FLOPs from 7.9 G to 8.7 G. This additional cost mainly comes from the multi-granularity representation branch and the semantic-level memory module. The increase is moderate relative to the performance gains across NTU RGB+D, NTU RGB+D 120, and Northwestern-UCLA, and the resulting model remains comparable to or lighter than several recent methods in Table 1. These results suggest that MG^{2}CL provides a reasonable trade-off between representational capacity and computational complexity.

4.4. Ablation Study

In this section, we will validate the effectiveness of the modules of our method, with all experiments conducted and tested on the NTU RGB+D 60 dataset.

4.4.1. Effectiveness of Multi-Granularity Graph Structure

To evaluate the effectiveness of the multi-granularity graph structure, we conducted experiments using the NTU RGB+D 60 dataset, with the results summarized in Table 2. As detailed in Section 3.2, our multi-granularity graph structure reconstructs the skeleton graph by pruning non-essential joints during action execution. This process reduces the model's parameter scale without compromising its performance. In fact, experimental results indicate that the pruning procedure leads to performance improvements rather than a decline. Particularly, when focusing on the lower body joints with finer granularity, we observe significant performance enhancements. Notably, this approach resulted in a 0.27% increase in accuracy, all while reducing the overall parameter count. This outcome highlights the effectiveness of the multi-granularity graph structure in improving representation quality while reducing model size.

Table 1. Comparison of accuracy, parameter count, and floating-point operations between the proposed method and recent state-of-the-art methods on the NTU RGB+D 60, NTU RGB+D 120, and Northwestern-UCLA datasets. The best accuracy results are shown in bold.

Method	Year	NTU RGB+D 60		NTU RGB+D 120		NW-UCLA	Params (M)	FLOPs (G)
		X-Sub (%)	X-View (%)	X-Sub (%)	X-Set (%)			
ST-GCN [9]	AAAI 2018	81.5	88.3	70.7	73.2	-	3.10	16.3
2s-AGCN [35]	CVPR 2019	91.5	95.9	88.2	89.6	95.3	7.10	6.8
PL-GCN [28]	AAAI 2020	89.2	95.0	-	-	-	3.50	15.2
MS-G3D [24]	CVPR 2020	91.5	96.2	86.9	88.4	-	2.80	48.8
MST-GCN [36]	AAAI 2021	91.5	96.6	87.5	88.8	-	12.00	-
PSUMNET [29]	ECCV 2022	92.9	96.7	89.4	90.6	96.7	2.80	2.7
InfoGCN [37]	CVPR 2022	92.7	96.9	89.4	90.7	96.6	6.40	7.4
ML-STGNet [11]	TIP 2022	91.9	96.2	88.6	90.0	96.8	11.52	-
Efficient-B4 [34]	TPAMI 2022	92.1	96.1	88.7	88.9	96.8	2.00	15.2
SkeletonGCL [13]	ICLR 2023	93.1	97.0	89.5	91.0	96.9	-	-
HD-GCN [12]	ICCV 2023	93.0	97.0	**89.8**	**91.2**	96.6	6.72	6.4
SAN-GCN [16]	TMM 2023	92.1	96.2	88.7	90.1	96.8	9.64	-
TD-GCN [8]	TMM 2024	92.8	96.8	-	-	**97.4**	5.48	6.8
SelfGCN [15]	TIP 2024	93.1	96.6	89.4	91.0	96.9	6.80	10.5
TR-GCN [38]	TAI 2024	90.9	96.8	87.9	88.9	-	8.34	41.65
LG-SGNet [26]	PR 2025	93.1	96.7	89.4	91.0	96.8	6.8	-
JMDA [33]	TOMM 2025	92.9	96.6	89.2	90.9	-	7.2	8.8
CTR-GCN [25]	ICCV 2021	92.4	96.8	88.9	90.6	96.5	5.80	7.9
MG^{2}CL (CTR-GCN w/ Ours)	-	**93.2**	**97.1**	**89.8**	**91.2**	97.1	7.20	8.7

Table 2. Ablation experiments with different granularity graph on joint modality across NTU RGB+D 60 X–Sub criterion.

Model	Params (M)	Accuracy (%)
Baseline	1.47	90.14
w/ Coarse-grained	1.38	90.25
w/ Upper fine-grained	1.42	90.27
w/ Lower fine-grained	1.43	90.41

4.4.2. Transferability of Multi-Granularity Graph Structure

In this section, we examine the transferability of the proposed multi-granularity graph structure across various architectures, emphasizing its versatility and effectiveness beyond the CTR–GCN framework [25]. We extend our analysis to three other state-of-the-art methods: 2s–AGCN [35], MST–GCN [36], and InfoGCN [37]. This comprehensive evaluation aims to assess the adaptability and robustness of the multi-granularity graph structure in enhancing the performance of different models. For this analysis, we conducted extensive experiments on the NTU RGB+D 60 dataset using the cross-subject (X–Sub) evaluation protocol. The results, based on the fusion of four distinct data streams—joint, bone, joint motion, and bone motion—are summarized in Table 3. These findings provide a holistic view of the impact of integrating the multi-granularity graph structure into various models. Our approach not only improves recognition accuracy but also reduces model complexity. Specifically, 2s–AGCN [35] saw a reduction in parameters from 7.0 M to 6.4 M. MST–GCN [36] experienced a decrease from 12.0 M to 11.2 M. InfoGCN [37] achieved a reduction from 6.4 M to 5.8 M. These results emphasize the transferability of the multi-granularity graph structure. In addition to improving recognition performance, it offers a favorable trade-off between representation capacity and model complexity.

4.4.3. Effectiveness of Spatial-Temporal Masking Strategy

This section examines the impact of various data augmentation strategies on model performance, as summarized in Table 4. In particular, we assess the combination of spatial and temporal masking applied to different sequences, following traditional data augmentation techniques. Our results demonstrate that the sequence "Traditional Data Augmentation → Spatial Masking → Temporal Masking" yields the best performance, improving accuracy by 0.7% compared to the baseline. This outcome highlights the synergistic effect of combining traditional augmentation with advanced masking strategies. In contrast, when traditional augmentation operations are excluded and only spatial-temporal masking is applied, accuracy decreases by 0.3% relative to the baseline. This emphasizes the critical role of traditional cropping techniques as a foundation for the effective application of spatial and temporal masking, ensuring maximum performance gains.

Table 3. Fusion results of the joint, bone, joint-motion, and bone-motion streams on the NTU RGB+D 60 X–Sub dataset.

Model	Params (M)	Accuracy (%)
2s–AGCN [35]	7.0	91.5
2s–AGCN w/ Multi-granularity	6.4	92.3
MST–GCN [36]	12.0	91.5
MST–GCN w/ Multi-granularity	11.2	92.0
InfoGCN [37]	6.4	92.3
InfoGCN w/ Multi-granularity	5.8	92.8

Table 4. Ablation experiments with the data augmentation on joint modality across NTU RGB+D 60 X–Sub criterion. ✓ and ✗ indicate that the corresponding operation is used and not used, respectively. The best result is shown in bold.

Conventional Augmentation	Masking Operation	Accuracy (%)
✗	✗	89.89
✓	✗	90.14
✓	Spatial	90.35
✓	Temporal	90.27
✓	Temporal + Spatial	90.45
✓	**Spatial + Temporal**	**90.62**
✗	Spatial + Temporal	89.63

4.4.4. Analysis of Loss Function

We analyze the performance of two contrastive loss functions, Triplet Loss and Info–NCE Loss, using the NTU RGB+D 60 dataset with the X–Sub evaluation protocol for joint modality outcomes. The instance-level loss L_{NCE}^{Ins} and semantic-level loss L_{NCE}^{Sem} were constructed using these two loss functions. The results are presented in Table 5. Initially, when both L_{NCE}^{Ins} and L_{NCE}^{Sem} were constructed using either the Triplet Loss or the Info–NCE Loss, the combination of Info–NCE Loss at both levels resulted in a 0.36% improvement in accuracy. In comparison, using Triplet Loss at both levels led to a 0.23% increase in accuracy. Furthermore, applying the Triplet Loss to the instance level ($\mathcal{M}_{Ins}$) and the Info–NCE Loss to the semantic level ($\mathcal{M}_{Sem}$) resulted in a 0.3% improvement. Interestingly, reversing this configuration did not significantly affect performance.

In conclusion, the results indicate that Info–NCE Loss outperforms Triplet Loss for both $\mathcal{M}_{Ins}$ and $\mathcal{M}_{Sem}$, making it the more effective choice overall.

4.4.5. Analysis of Memory Bank

We evaluate the performance of different memory bank constructions, based on experiments conducted on the joint modality data of the X–Sub evaluation protocol on the NTU RGB+D 60 dataset. The results are summarized in Table 6. When compared to the baseline model, incorporating an instance-level memory bank led to a performance improvement of approximately 0.1%. Building a semantic-level memory bank resulted in a slightly higher improvement of 0.2%. Moreover, the combination of both instance-level and semantic-level memory banks brought a more substantial performance enhancement of 0.34%. These findings highlight the effectiveness of integrating both memory banks into our framework, as they provide a compounded boost to overall model performance. Notably, this integration improves the model's ability to accurately recognize actions with local similarities.

4.5. Visualization

To better understand the improvements brought by our MG^{2}CL framework, we visualize the adjacency matrices for the "sitting" action in Figure 5. In these matrices, darker colors indicate stronger associations between joints. The baseline model predominantly captures interactions at the hip joint and torso level, which are typical for the "sitting" action. However, our MG^{2}CL approach goes a step further by identifying meaningful relationships between joints below the knee and in the arms, which provide more semantic relevance for the action, enhancing the model's grasp of spatial-temporal dependencies.

In addition, we selected 15 actions from the NTU RGB+D 60 dataset to compare the performance of the baseline model with that of our MG^{2}CL framework. The results of these experiments, performed using the NTU RGB+D 60 X–Sub evaluation protocol, are shown in Figure 6. These actions, such as "brushing hair"

Table 5. Ablation experiments with different loss functions on joint modality across NTU RGB+D 60 X–Sub criterion. ✗ indicates that the corresponding loss is not used. The best accuracy result is shown in bold.

Instance Level	Semantic Level	Accuracy (%)
✗	✗	89.89
Triplet Loss	Triplet Loss	90.12
Info-NCE Loss	Info-NCE Loss	**90.25**
Triplet Loss	Info-NCE Loss	90.18
Info-NCE Loss	Triplet Loss	90.17

Table 6. Ablation experiments with instance-level and semantic-level memory banks on joint modality across NTU RGB+D 60 X–Sub criterion. The best accuracy result is shown in bold.

Model	Accuracy (%)	Params (M)
Baseline	90.14	1.47
w/ Instance	90.25	1.62
w/ Semantic	90.34	1.73
w/ Instance + Semantic	**90.48**	1.81

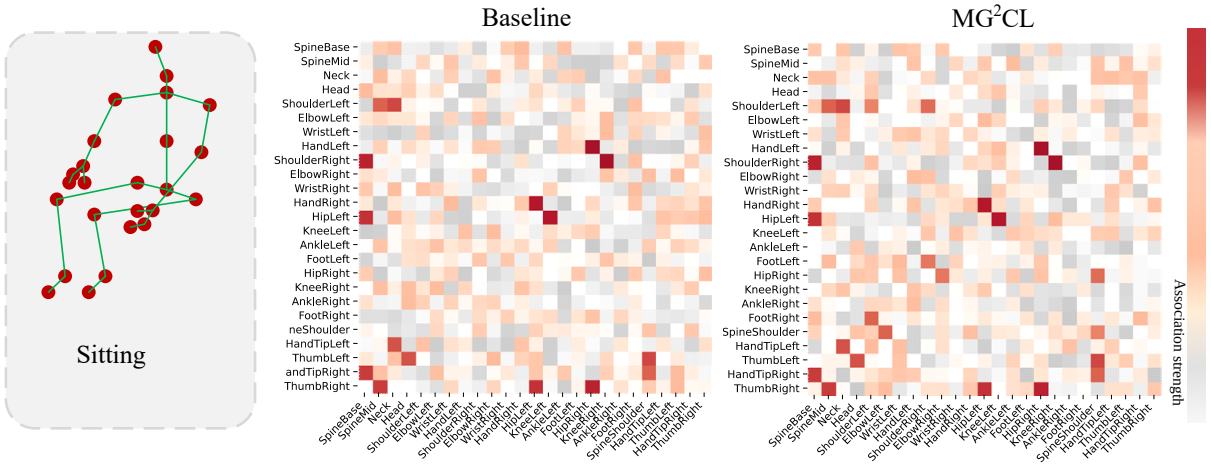

Figure 5. Visualization comparison of the adjacency matrix for the "sitting" action between the baseline model and MG^{2}CL.

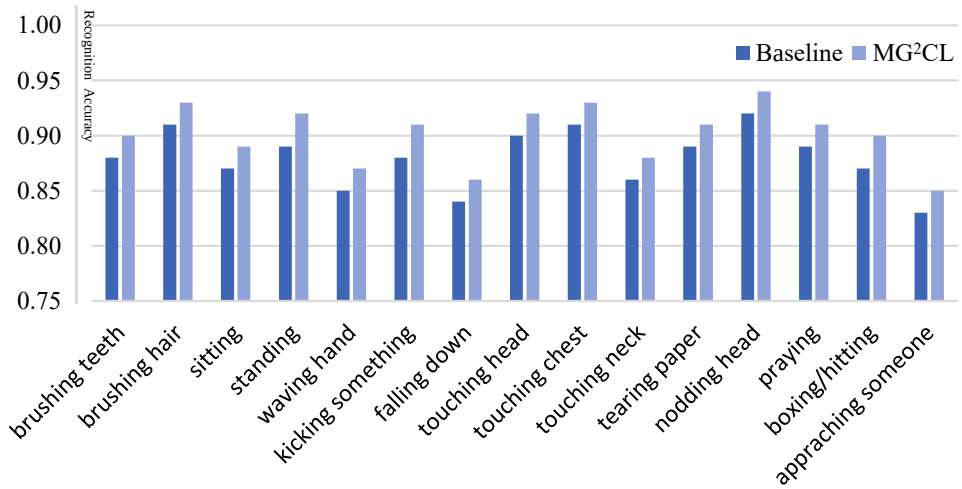

Figure 6. Comparison of recognition accuracy between the baseline model and MG^{2}CL on locally similar actions.

and "brushing teeth", engage similar body regions, particularly the arms. However, the nature of their execution differs: "brushing hair" involves elbow-driven wrist movements, while "brushing teeth" emphasizes wrist and fingertip actions.

The visualization of these 15 actions demonstrates that integrating the multi-granularity contrastive learning framework significantly reduces the misclassification of such similar actions, thus improving the model's precision and robustness. This showcases the power of our method in resolving ambiguities between actions that share local similarities, leading to more accurate and reliable action recognition.

5. Conclusions

In this work, we introduced a novel contrastive learning framework, MG^{2}CL, specifically designed to overcome key limitations in skeleton-based action recognition. By incorporating a multi-granularity graph structure with learned relational modeling, our method captures relationships between both distant and semantically related joints and improves representation learning. This innovative structure not only improves performance but also demonstrates strong transferability, making it compatible with a wide range of GCN-based methods. Furthermore, our spatio-temporal masking strategy generates diverse and informative skeleton samples, augmenting the dataset and improving the model's generalization ability. The integration of a semantic-level memory bank further refines the framework's capacity to distinguish subtle differences between actions, significantly reducing misclassifications in cases of locally similar movements. Experimental results across three widely used benchmark datasets demonstrate the effectiveness and versatility of MG^{2}CL. The promising outcomes affirm that our method offers substantial improvements in both action recognition accuracy and model efficiency, while its performance may still depend on skeleton quality and action context. In our future work, we aim to extend the application of the multi-granularity graph structure to broader skeletal self-supervised tasks and downstream applications.

Author Contributions

Z.W.: conceptualization, methodology, writing—original draft preparation; P.W.: software, validation; X.C.: visualization, investigation; T.W.: writing—reviewing and editing; F.L.: data curation, formal analysis; S.J.: supervision, project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62406095, 62306100, and 62406096), the Natural Science Foundation of Anhui Province (Grant No. 2308085MF213), the Open Fund of the State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology (Grant No. OEIAPT202503), the Hefei Key Technology R&D "Champion-Based Selection" Project (Grant No. 2024SGJ010), and the Natural Science Research Project of the Anhui Educational Committee (Grant No. 2025AHGXZK40026).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets used in this study are publicly available benchmark datasets, including NTU RGB+D 60, NTU RGB+D 120, and Northwestern-UCLA. The datasets can be obtained from their respective official sources as cited in the manuscript.

Acknowledgments

The authors acknowledge the support of the AI General Computing Platform of Hefei University.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

We used ChatGPT during the preparation of this manuscript to assist with English language polishing and expression refinement, aiming to bring the writing closer to the academic writing conventions of native English speakers. In accordance with the journal's author guidelines on the use of artificial intelligence and AI-assisted technologies, we hereby declare that after using the above tool, the authors have reviewed and edited all content as

necessary, and take full responsibility for the final content of this manuscript and any subsequent publication. The authors confirm that the use of AI tools was strictly limited to auxiliary language expression and formatting, and did not involve any research design, data collection, data analysis, result calculation, or conclusion derivation. All substantive academic contributions (including core ideas, experimental data, logical reasoning, and final conclusions) were independently completed by the author team.

References

1. Khaleghi, L.; Sepas-Moghaddam, A.; Marshall, J.; et al. Multiview Video-Based 3-D Hand Pose Estimation. *IEEE Trans. Artif. Intell.* **2023**, *4*, 896–909.
2. Cao, C.; Wang, Y.; Zhang, Y.; et al. Co-Occurrence Matters: Learning Action Relation for Temporal Action Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3327–3339.
3. Vahdani, E.; Tian, Y. Deep Learning-Based Action Detection in Untrimmed Videos: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4302–4320.
4. Sun, Z.; Ke, Q.; Rahmani, H.; et al. Human Action Recognition From Various Data Modalities: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3200–3225.
5. Ahmad, T.; Jin, L.; Zhang, X.; et al. Graph Convolutional Neural Network for Human Action Recognition: A Comprehensive Survey. *IEEE Trans. Artif. Intell.* **2021**, *2*, 128–145.
6. Yan, H.; Liu, Y.; Wei, Y.; et al. SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
7. Do, J.; Kim, M. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
8. Liu, J.; Wang, X.; Wang, C.; et al. Temporal Decoupling Graph Convolutional Network for Skeleton-Based Gesture Recognition. *IEEE Trans. Multimed.* **2024**, *26*, 811–823.
9. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), New Orleans, LA, USA, 2–7 February 2018; pp. 7444–7452.
10. Zhou, M.; Lu, J.; Zhang, G. Multi-Scale Adaptive Convolutional Graph for Multi-Stream Concept Drift. *IEEE Trans. Knowl. Data Eng.* **2026**, *38*, 2982–2994.
11. Zhu, Y.; Shuai, H.; Liu, G.; et al. Multilevel Spatial–Temporal Excited Graph Network for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2023**, *32*, 496–508.
12. Lee, J.; Lee, M.; Lee, D.; et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 10444–10453.
13. Huang, H.; Zhou, H.; Wang, J. Graph Contrastive Learning for Skeleton-based Action Recognition. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
14. Pang, C.; Lu, X.; Lyu, L. Skeleton-based action recognition through contrasting two-stream spatial-temporal networks. *IEEE Trans. Multimed.* **2023**, *25*, 8699–8711.
15. Wu, Z.; Sun, P.; Chen, X.; et al. SelfGCN: Graph Convolution Network with Self-Attention for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2024**, *33*, 4391–4403.
16. Tian, H.; Ma, X.; Li, X.; et al. Skeleton-Based Action Recognition with Select-Assemble-Normalize Graph Convolutional Networks. *IEEE Trans. Multimed.* **2023**, *25*, 8527–8538.
17. Cormier, M.; Schmid, Y.; Beyerer, J. Enhancing Skeleton-Based Action Recognition in Real-World Scenarios through Realistic Data Augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 290–299.
18. Ni, B.; Paramathayalan, V.R.; Moulin, P. Multiple granularity analysis for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 756–763.
19. Chen, T.; Zhou, D.; Wang, J.; et al. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4334–4342.
20. Chen, H.; Shen, Y.; Zhang, Y.; et al. Skeleton-based action recognition through dual-granularity feature fusion with self-adapting graph convolution and multi-scale temporal convolution. *Neurocomputing* **2025**, *639*, 130261.
21. Shahroudy, A.; Liu, J.; Ng, T.; et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1010–1019.
22. Liu, J.; Shahroudy, A.; Perez, M.; et al. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2684–2701.
23. Wang, J.; Nie, X.; Xia, Y.; et al. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656.

24. Liu, Z.; Zhang, H.; Chen, Z.; et al. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Virtual, 14–19 June 2020; pp. 140–149.
25. Chen, Y.; Zhang, Z.; Yuan, C.; et al. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21), Virtual, 11–17 October 2021; pp. 13339–13348.
26. Wu, Z.; Ding, Y.; Wan, L.; et al. Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. *Pattern Recognit.* **2025**, *159*, 111106.
27. Myung, W.; Su, N.; Xue, J.H.; et al. DeGCN: Deformable Graph Convolutional Networks for Skeleton-Based Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 288–300.
28. Huang, L.; Huang, Y.; Ouyang, W.; et al. Part-level graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11045–11052.
29. Trivedi, N.; Sarvadevabhatla, K. PSUMNet: Unified modality part streams are all you need for efficient pose-based action recognition. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 211–227.
30. Li, L.; Wang, M.; Ni, B. 3D human action representation learning via cross-view consistency pursuit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4741–4750.
31. Wu, C.; Wu, J. SCD-Net: Spatiotemporal Clues Disentanglement Network for Self-Supervised Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 5949–5957.
32. Qin, Z.; Liu, Y.; Ji, P.; et al. Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 4783–4797.
33. Xiang, L.; Wang, Z. Joint mixing data augmentation for skeleton-based action recognition. *ACM Trans. Multimed. Comput. Commun. Appl.* **2025**, *21*, 108.
34. Song, Y.F.; Zhang, Z.; Shan, C.; et al. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1474–1488.
35. Shi, L.; Zhang, Y.; Cheng, J.; et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
36. Chen, Z.; Li, S.; Yang, B.; et al. Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI'21), Virtual, 2–9 February 2021; pp. 1113–1122.
37. Chi, H.G.; Ha, M.H.; Chi, S.; et al. InfoGCN: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196.
38. Zhuang, T.; Qin, Z.; Ding, Y.; et al. Temporal Refinement Graph Convolutional Network for Skeleton-Based Action Recognition. *IEEE Trans. Artif. Intell.* **2024**, *5*, 1586–1598.