Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Discriminative Micro-Action-Aware Joint Amplifier for Skeleton-Based Action Recognition

Wenming Cao
Shenzhen University
Gai Wang
Shenzhen University
Xinpeng Yin
2110436215@email.szu.edu.cn

Shenzhen University

Research Article

Keywords: Action Recognition, Spherical Harmonic, Fine-grained Action Recognition, Graph Convolutional Network

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10356230/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Discriminative Micro-Action-Aware Joint Amplifier for Skeleton-Based Action Recognition

Wenming Cao[1], Gai Wang[1], Xinpeng Yin[1*]

[1]State Key Laboratory of Radio Frequency Heterogeneous Integration, Shenzhen University, Yuehai Street, Shenzhen, 518060, Guangdong, China.

*Corresponding author(s). E-mail(s): 2110436215@email.szu.edu.cn;
Contributing authors: wmcao@szu.edu.cn;
2400042078@mails.szu.edu.cn;

Abstract

Skeleton-based human action recognition has achieved significant progress due to its robustness against appearance variations and environmental disturbances. However, fine-grained action recognition remains challenging because discriminative information is often contained in subtle movements of a few joints and can be easily overwhelmed by dominant motion patterns during feature learning. Existing methods mainly improve feature representation through spatio-temporal modeling, attention mechanisms, and adaptive graph learning, but they rarely explore frequency-domain enhancement of local geometric motion patterns. In this paper, we propose a Discriminative Micro-Action-Aware Joint Amplifier (DMJA) for skeleton-based action recognition. Different from existing spherical harmonic-based methods that mainly focus on rotation robustness or geometric representation, DMJA treats spherical harmonic decomposition as a discriminative feature enhancement strategy for micro-action modeling. Specifically, a discriminative joint selection module is first introduced to identify informative joints with fine-grained motion responses. Then, the local geometric relationships among the selected joints are projected into the spherical harmonic domain, where different frequency components are decoupled and adaptively enhanced to highlight subtle directional variations of key joints. By integrating the enhanced micro-action features with conventional skeleton representations, DMJA learns more discriminative action representations. Extensive experiments on multiple benchmark skeleton-based action recognition datasets demonstrate the effectiveness of the proposed method.

Keywords: Action Recognition; Spherical Harmonic; Fine-grained Action Recognition; Graph Convolutional Network

1 Introduction

Human action recognition (HAR) is an important research topic in computer vision. It has wide applications in intelligent surveillance [1], human-computer interaction [2], rehabilitation assessment, and sports analysis. Skeleton data provide only the spatial coordinates of human joints. They are less affected by illumination changes, background clutter, and viewpoint variations than RGB videos [3]. Skeleton data also have lower dimensionality and higher robustness. With the development of pose estimation methods, such as Kinect and OpenPose, skeleton-based action recognition has attracted increasing attention in the pattern recognition community [4, 5].

Fine-grained skeleton-based action recognition is challenging because different action categories often differ only in subtle movements of a few joints [6, 7]. In contrast, joints with large motion amplitudes usually produce stronger feature responses and dominate the feature learning process. As a result, the discriminative information contained in micro-actions is easily suppressed. Recent studies have improved discriminative feature learning by introducing attention mechanisms [8], adaptive graph learning [9], and contrastive learning [10]. These methods enable the network to focus on informative spatial regions and temporal segments, thereby improving recognition performance. However, most existing methods enhance features directly in the spatial feature space. Their performance still relies on the original spatial representation. They also lack effective modeling of local geometric directional variations between joints. Consequently, they cannot sufficiently amplify the discriminative information contained in subtle motions.

Human actions involve not only joint position variations but also rich local directional changes and multi-scale motion patterns [11]. From the perspective of signal processing, subtle motions are generally correspond to relatively higher-frequency motion variations [12]. Existing skeleton representations mainly use joint coordinates, bone vectors, and motion differences to describe human motions [13]. Although these representations effectively capture spatial locations and motion trends, they do not explicitly model motion components at different frequencies. Therefore, learning only in the spatial-temporal domain is insufficient to highlight the discriminative characteristics of subtle motions. Previous studies have shown that spherical harmonic representations can not only provide rotation-invariant or rotation-equivariant geometric features but also characterize local geometric structures at different frequencies. This property enables a richer frequency-domain representation for complex 3D signals. However, existing studies mainly employ spherical harmonics to improve rotational robustness and geometric representation [14, 15]. They rarely investigate the contribution of different spectral components to action discrimination. In skeleton-based action recognition, spherical harmonics are mainly used to describe motion directions or geometric structures. Their frequency decomposition capability has not

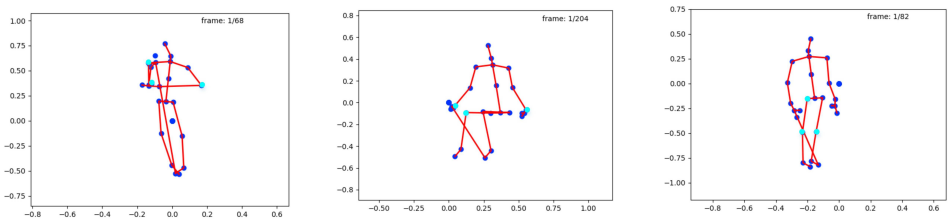


Fig. 1: The row highlights the discriminative micro-motion joints automatically selected. The selected joints mainly concentrate on body parts with subtle but informative movements, providing reliable cues for subsequent local spherical harmonic enhancement.

been fully exploited for discriminative micro-action enhancement. In particular, existing methods lack a dedicated frequency-domain enhancement mechanism for subtle directional changes of local key joints. Consequently, these discriminative cues are easily overwhelmed by dominant motion patterns during deep feature propagation.

To address these limitations, we propose Discriminative Micro-Action-Aware Joint Amplifier for Skeleton-Based Action Recognition(DMJA)in Fig. 1. Unlike previous methods that mainly employ spherical harmonics to obtain rotation-robust representations, our method treats spherical harmonics as a discriminative feature enhancement tool. Specifically, the local key joint motions are first projected into the spherical harmonic domain. The spherical harmonic coefficients are then used to decouple motion components at different frequencies. Next, the high-frequency spectral components associated with subtle motions are adaptively enhanced to amplify fine-grained directional variations of key joints and improve their discriminative capability. Since spherical harmonics preserve local geometric continuity while providing multi-frequency representations, the proposed LSHA effectively strengthens micro-action features and improves the performance of fine-grained skeleton-based action recognition.

Experimental results on the NTU RGB+D 60 [16], NTU RGB+D 120 [17], and NW-UCLA [18] datasets demonstrate the effectiveness of the proposed method. The proposed method achieves better performance in fine-grained action recognition. Visualization results further show that the dynamic joint selection module can accurately identify discriminative key joints. They also show that the spherical harmonic direction encoding can enhance the representation of local directional variations. As a result, the recognition accuracy of micro-actions is improved. The main contributions of this work are summarized as follows:

- A discriminative micro-action amplification framework based on spherical harmonic direction enhancement is proposed for skeleton-based action recognition. The proposed framework focuses on discriminative micro-action joint mining and directional feature enhancement. It effectively models subtle local motion differences and provides a new solution for fine-grained skeleton-based action recognition.

- A dynamic discriminative joint selection strategy is proposed. This strategy jointly analyzes motion responses and temporal variation characteristics of joints. It adaptively selects key joints with high discriminative capability and suppresses the influence of static joints and redundant motions. As a result, the network can focus on discriminative micro-action regions and achieve more effective feature representation and better recognition performance.
- Extensive experiments are conducted on the NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA datasets. The experimental results demonstrate that the proposed method consistently improves the performance of existing skeleton-based action recognition models.

2 The related work

2.1 Skeleton-Based Action Recognition

In recent years, skeleton-based action recognition has achieved remarkable progress. Existing studies have improved recognition performance from multiple perspectives [11, 19, 20] , including CNNs, RNNs, GCNs, and Transformers. Among them, graph convolutional networks (GCNs), represented by ST-GCN [21], can effectively model spatial topological relationships among human joints and temporal dynamics of actions. Based on this framework, methods such as 2s-AGCN [22], CTR-GCN [13], and MS-GCN [23] further enhance skeleton feature representation through adaptive graph learning, multi-scale feature aggregation, and topology refinement [24–26]. In addition, several studies introduce attention mechanisms or discriminative feature learning strategies. These methods enable the network to focus on more informative spatial regions [27] and temporal segments [9], thereby improving action recognition performance.

Although the above methods achieve promising recognition performance, their primary focus is on global motion modeling and graph topology optimization. Their ability to capture discriminative micro-action information in fine-grained action recognition remains limited [21]. When different action categories differ only in subtle movements of a few joints, joints with large motion amplitudes usually generate stronger feature responses. As a result, they tend to dominate the feature learning process, while discriminative micro-action cues may be overlooked. In addition, most existing joint modeling methods rely on feature response intensity or attention weights to identify important joints. They lack an explicit mechanism for selecting micro-action joints and may not accurately localize truly discriminative local motion regions [9].

On the other hand, existing skeleton-based action recognition methods commonly use joint coordinates, bone vectors, and motion differences to describe human motion states [28] . These representations can effectively characterize joint position variations and motion trends. However, their ability to model relative directional relationships among local joints is limited. As a result, subtle directional variations in fine-grained actions may not be sufficiently represented.Spherical harmonics have demonstrated strong capability in directional feature representation for three-dimensional shape analysis, point cloud processing, and spherical signal modeling. Nevertheless, their

application to skeleton-based action recognition remains relatively limited. In particular, the use of spherical harmonics for modeling local directional relationships among key joints has not been fully explored.

Therefore, adaptively identifying discriminative micro-action joints from complex skeleton sequences and enhancing subtle motion representations with more effective directional descriptors remain challenging problems in skeleton-based action recognition. To address these issues, we propose a Discriminative Micro-Action Joint Amplifier with Spherical Harmonic Direction Enhancement. The proposed framework dynamically selects key micro-action joints and employs spherical harmonics to encode local directional relationships in the frequency domain. In this way, fine-grained action differences can be modeled more effectively.

2.2 Spherical Harmonics (SH)

To further improve the discriminability of micro-motion features, recent studies have proposed various feature enhancement strategies. Chang *et al.* [7] proposed the Fine-grained Contrastive Enhancement module. FCE enlarges the feature distance between different action classes through fine-grained contrastive learning. As a result, it enhances the discrimination of subtle motion patterns. However, these methods mainly rely on contrastive constraints in the feature space and provide limited modeling of local geometric structures.

Recently, spherical harmonics (SH) have been widely adopted in 3D vision because of their powerful spectral representation and rotation modeling capability [29–32]. SPHARM-Net [15] introduced constrained spherical convolution filters with infinitely many spectral components and developed an end-to-end framework without data augmentation. The proposed filters encode spherical signals in the spectral domain without explicitly expanding all spherical harmonic bases. This design enables effective modeling of multi-frequency information. Beyond medical shape analysis, spherical harmonics have also been applied to point cloud analysis, 3D object recognition, and spherical signal processing. Their main advantage is the ability to learn rotation-invariant or rotation-equivariant representations [33, 34]. Meanwhile, Fang *et al.* [35] showed that conventional convolutional neural networks still achieve higher efficiency and better performance in several non-spherical vision tasks. This observation indicates that spherical harmonics are better used as a complementary geometric spectral representation rather than a replacement for conventional convolutions.

In skeleton-based action recognition, Ref. [36] first introduced spherical harmonics into skeleton representation. It maps joint motions onto a spherical space and characterizes motion directions with spherical harmonic spectra. This provides a new spectral representation for skeleton dynamics. However, existing methods mainly employ spherical harmonics to model geometric structures and motion directions. Their primary objective is to improve rotation robustness and geometric representation [14, 37]. They rarely exploit spherical harmonic spectra to enhance the discriminability of micro-motion features. In particular, the high-frequency components associated with subtle joint motions have not been fully explored.

Motivated by these observations, we introduce spherical harmonics as a micro-motion feature amplifier for skeleton-based action recognition. Different from previous

methods that focus on rotation-invariant representations, our method exploits the intrinsic multi-frequency decomposition capability of spherical harmonics. Specifically, local joint motions are projected into the spherical spectral domain. The high-frequency spectral components are then enhanced to amplify subtle motion variations and discriminative joint features. Moreover, spherical harmonics decouple motion information at different frequencies while preserving local geometric continuity. This property further strengthens micro-motion representations during deep feature propagation and ultimately improves fine-grained action recognition performance.

3 Method

3.1 Theoretical Background

Spherical Harmonics (SH) [38] are a set of complete orthonormal basis functions defined on the unit sphere. They correspond to the angular solutions of the Laplace equation in spherical coordinates. By projecting spherical signals onto orthogonal frequency bases, SH provides an efficient frequency-domain representation of geometric structures. Therefore, SH has been widely applied in three-dimensional shape analysis, point cloud processing, pose estimation, and geometric deep learning.

For a point $\mathbf{p} = (x, y, z)$ in Cartesian coordinates, its spherical coordinate representation is defined as

$$r = \sqrt{x^2 + y^2 + z^2}, \tag{1}$$

$$\theta = \arccos\left(\frac{z}{r}\right), \tag{2}$$

$$\phi = \operatorname{atan2}(y, x), \tag{3}$$

where r denotes the radial distance, θ denotes the colatitude, and ϕ denotes the azimuth angle. For signals defined on the unit sphere, $r = 1$, and the spherical signal can be represented as $f(\theta, \phi)$.

The standard complex spherical harmonic basis function is defined as

$$Y_l^m = (\theta, \phi) N_l^m P_l^m(\cos\theta) e^{im\phi}, \tag{4}$$

where l is the degree, m is the order satisfying $-l \leq m \leq l$, $P_l^m(\cdot)$ denotes the associated Legendre polynomial, and N_l^m is the normalization coefficient

$$N_l^m = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}}. \tag{5}$$

Any square-integrable spherical signal can be expanded using spherical harmonic basis functions as

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_l^m Y_l^m(\theta, \phi), \tag{6}$$

where c_l^m denotes the spherical harmonic coefficient. This expansion maps spatial signals into the frequency domain. Low-frequency components mainly describe global structures and overall orientations, whereas high-frequency components characterize local details and subtle variations.

Because spherical harmonic basis functions satisfy the orthogonality condition

$$c_l^m = \int_{S^2} Y_l^m \overline{Y_{l'}^{m'}}, d\Omega = \delta_{ll'}\delta_{mm'}. \tag{7}$$

the spherical harmonic coefficients can be obtained through the Spherical Harmonic Transform (SHT)

$$c_l^m = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi)\overline{Y_l^m(\theta, \phi)} \sin\theta \, d\theta \, d\phi. \tag{8}$$

To characterize the frequency distribution of spherical signals, the energy of the l-th frequency band is defined as

$$E_l = \sum_{m=-l}^{l} |c_l^m|^2. \tag{9}$$

Low-order energy components represent global geometric structures, while high-order components capture local directional details. This property enables multi-scale analysis of three-dimensional geometric signals.

Another important property of SH is rotation equivariance. For a rotation transformation $R \in SO(3)$, the spherical harmonic basis satisfies

$$Y_l^m(R\mathbf{x}) = \sum_{n=-l}^{l} D_{mn}^{(l)}(R) Y_l^n(\mathbf{x}), \tag{10}$$

where $D_{mn}^{(l)}$ denotes the Wigner-D matrix. This property indicates that rotations only induce linear transformations among coefficients within the same degree and do not introduce coupling across different frequency bands. Consequently, spherical harmonic representations exhibit strong robustness to viewpoint and pose variations.

Motivated by these properties, this work employs spherical harmonics to encode local directional relationships among discriminative micro-action joints. By transforming local direction vectors into spherical coordinates and extracting low-order spherical harmonic coefficients, subtle directional variations can be effectively represented, thereby improving fine-grained action recognition performance.

3.2 Overview

To effectively capture subtle and discriminative motion patterns in skeleton sequences, we propose a Discriminative Micro-Motion Joint Amplifier (DMJA). The proposed DMJA consists of three modules, namely Motion Information Extraction (MIE), Subtle Key Joint Selection (SKJS), and Local Spherical Harmonic Amplifier (LSHA). Spherical harmonics (SH) are orthogonal basis functions defined on the unit sphere and provide a powerful tool for directional signal analysis. Specifically, local relative

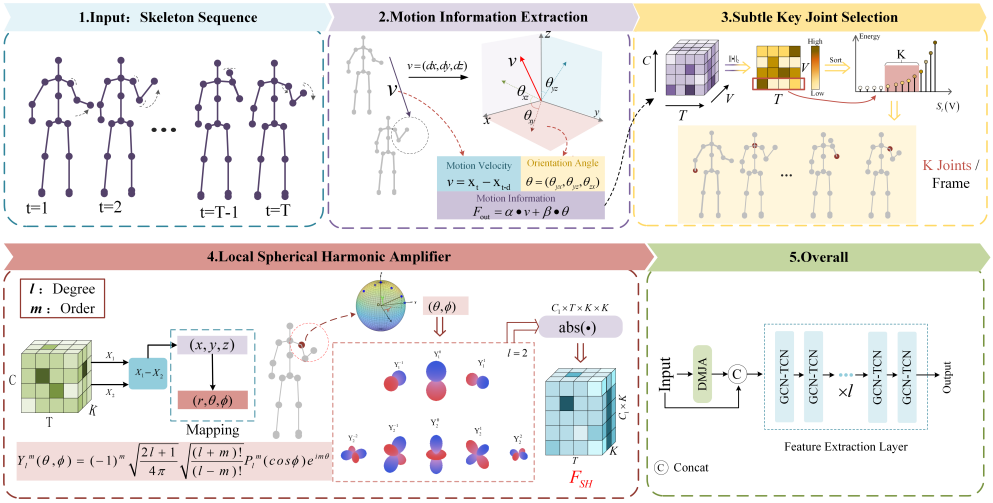


Fig. 2: Overview of the proposed DMJA framework. The light-blue region denotes the input stage, where the motion trends of several representative joints are illustrated for visualization. The purple region corresponds to the Motion Information Extraction (MIE) module, which extracts motion representations from skeleton sequences. The yellow region represents the Subtle Key Joint Selection (SKJS) module, which identifies discriminative micro-motion joints based on motion responses. The orange-red region denotes the Local Spherical Harmonic Amplifier (LSHA) module, which performs local geometric modeling and directional feature enhancement using spherical harmonics. The green region presents the overall processing pipeline. The enhanced features are fused with the original skeleton features and then fed into a multi-scale spatio-temporal graph convolutional network for action classification.

displacement relationships are first constructed among the selected key joints and then transformed into the spherical coordinate system. Subsequently, low-order spherical harmonic basis functions are employed to encode local geometric relationships in the frequency domain. As a result, directional variations between joints are projected into a high-dimensional spherical harmonic feature space.

Due to the high sensitivity of spherical harmonic bases to angular variations, subtle directional changes can generate distinguishable responses across multiple frequency bands. This property enables the proposed method to amplify discriminative micro-motion cues and enhance the representation of fine-grained action patterns. The details of each module and the overall network framework are described in the following subsections.

3.3 Discriminative Micro-Motion Joint Amplifier

Specifically, the Motion Information Extraction (MIE) module first extracts motion representations from skeleton sequences to capture temporal dynamics of joint movements. Based on the extracted motion features, the Subtle Key Joint Selection (SKJS) module adaptively identifies key joints that contain discriminative micro-motion

information according to their motion response strengths. This process suppresses redundant or less informative joints and enables the network to focus on motion regions that are more relevant to action discrimination.

Afterward, the Local Spherical Harmonic Amplifier (LSHA) performs local geometric modeling on the selected key joints using spherical harmonics. By encoding local directional relationships in the frequency domain, LSHA enhances subtle motion patterns and improves their discriminative representation.

These three modules jointly form a unified micro-motion enhancement framework. The MIE and SKJS modules provide reliable priors for feature enhancement from the perspectives of motion modeling and key-joint localization, respectively. Guided by these priors, LSHA further amplifies fine-grained motion cues through spherical harmonic direction encoding. Owing to this progressive design, DMJA can effectively discover and enhance discriminative micro-motion information embedded in skeleton sequences, leading to improved fine-grained action recognition performance.

The overall architecture of the proposed framework is illustrated in Fig. 2. The details of each module and the complete processing pipeline are presented in the following subsections.

3.3.1 Motion Information Extraction

To characterize the motion trends of skeleton joints more accurately, we design a Motion Information Extraction (MIE) module. Unlike conventional methods that mainly utilize motion magnitudes, the proposed module further exploits directional information contained in motion vectors. This design enhances the sensitivity of the model to subtle motion variations and improves the representation of fine-grained actions.

Given a skeleton sequence:

$$X \in \mathbb{R}^{C\times T\times V}, \tag{11}$$

where C denotes the number of channels, T denotes the number of frames, and V denotes the number of joints. We first compute the motion vector of each joint between adjacent timestamps. For a temporal scale d, the motion representation is defined as

$$M_t^{(d)} = X_t - X_{t-d}, \tag{12}$$

$$M_t^{(d)} = (\Delta x, \Delta y, \Delta z). \tag{13}$$

To maintain a consistent temporal dimension, zero padding is applied to the first d frames.

To further characterize motion direction information, directional angles of the motion vector are computed on three orthogonal coordinate planes, as illustrated in the second module of Fig. 2:

$$\theta_{yx} = \operatorname{atan2}(\Delta y, \Delta x), \quad \theta_{yz} = \operatorname{atan2}(\Delta y, \Delta z), \quad \theta_{zx} = \operatorname{atan2}(\Delta z, \Delta x). \tag{14}$$

Based on the above directional angles, the motion direction representation is constructed as

$$\Theta = [\theta_{yx}, \theta_{yz}, \theta_{zx}]\,. \tag{15}$$

Compared with motion magnitude alone, directional angles can better describe the variation trends of joint trajectories. Therefore, they are more effective for distinguishing subtle differences among similar actions.

To preserve both motion magnitude and motion direction information, two learnable parameters, α and β, are introduced to adaptively fuse the two representations:

$$F_{\text{out}} = \alpha\Theta + \beta M, \tag{16}$$

where α and β are trainable weighting coefficients that are automatically optimized during network training.

The resulting motion feature contains both motion intensity information and directional variation information. It provides richer motion priors for the subsequent Subtle Key Joint Selection (SKJS) module and Local Spherical Harmonic Amplifier (LSHA).

Fine-grained actions often exhibit similar motion magnitudes but different motion directions. Therefore, the proposed method utilizes motion direction features as an important cue for key-joint selection. This strategy enables more accurate localization of joint regions that contain discriminative micro-motion information.

3.3.2 Subtle Key Joint Selection

The Motion Information Extraction (MIE) module generates motion representations that contain both motion trends and directional variations of skeleton joints. However, not all joints contribute equally to action recognition. Human actions usually exhibit considerable spatial redundancy. Many joints produce irrelevant or repetitive motion information during action execution. In addition, joints with the largest motion amplitudes often correspond to global body movements. In contrast, the differences among fine-grained actions are usually reflected by subtle variations of local joints. Therefore, directly selecting joints with the strongest responses may overlook micro-motion regions that contain discriminative information.

To address this issue, we design a Subtle Key Joint Selection (SKJS) module. The objective of SKJS is to automatically identify key joints that contain discriminative micro-motion cues from all joints in each frame.

First, the response strength of each joint is computed in the feature space:

$$S_{t,v} = \sqrt{\sum_{c=1}^{C} F_{\text{out}}(c,t,v)^2}, \tag{17}$$

where S denotes the joint score matrix. The score reflects the motion activity of each joint at the current timestamp.

Conventional Top-K strategies directly select joints with the highest response values. However, in skeleton-based action recognition, these joints often correspond to large-scale movements, such as arm swinging or body rotation. The extracted information mainly describes global motion patterns rather than subtle action differences.

To prevent dominant joints from controlling the selection process, the joint scores are first sorted in ascending order. Then, two predefined ratio thresholds are employed to remove joints with extremely low responses and joints with excessively high responses. The remaining joints form a candidate set:

$$C = \mathrm{Sort}\left(\mathbf{S}\right)_{[r_l V : \lfloor r_u V \rfloor]}; \tag{18}$$

$$J = \{j_i \mid j_i \in C,\, i = 1, \ldots, K\}, \tag{19}$$

where r_l and r_u denote the lower and upper ratio thresholds, J denotes the set of the selected K subtle key joints.

Through this strategy, the selected joints effectively suppress noise caused by static joints while reducing the dominance of large-amplitude motion joints. Consequently, the model can focus more on fine-grained motion regions that are highly relevant to action discrimination. The selected subtle key joints are then fed into the Local Spherical Harmonic Amplifier (LSHA), providing reliable spatial priors for subsequent discriminative micro-motion feature enhancement.

3.3.3 Local Spherical Harmonic Amplifier

Although the Subtle Key Joint Selection (SKJS) module can identify key joint regions that contain discriminative motion information, the motion amplitudes of these joints are usually small. Consequently, their features can be easily overwhelmed by dominant motion patterns during deep feature propagation. To further enhance the representation of micro-motion information, we propose a Local Spherical Harmonic Amplifier (LSHA). The proposed module constructs high-order geometric relationships among local joints and employs spherical harmonics for frequency-domain representation, thereby enhancing discriminative micro-motion features.

Given the selected key-joint index set J obtained from the SKJS module, the corresponding joint features are first extracted from the original skeleton representation:

$$X_K = \{x_{j_1}, x_{j_2}, \ldots, x_{j_K}\}, \tag{20}$$

where $X_K \in \mathbb{R}^{C \times T \times K}$. Unlike conventional methods that only model the motion of individual joints, our method further models the local geometric relationships among key joints. It computes the relative displacement vectors between key joints and constructs a local joint relationship matrix. This process extends individual joint motion to pairwise geometric relationships. As a result, the local spatial structure of fine-grained actions can be represented more effectively.

$$\Delta x_{ij} = x_i - x_j. \tag{21}$$

To describe the spatial directional relationships among local joints more accurately, the relative displacement vectors are transformed from the Cartesian coordinate system to the spherical coordinate system. The transformation produces the radial distance, azimuth angle, and polar angle. Compared with the Cartesian representation, the spherical representation decouples distance variations from direction variations.

This property enables a more effective representation of subtle local motion patterns, such as local rotations, swings, and directional deviations in fine-grained actions.

According to Eqs. (1)–(3), the spherical coordinate representation is obtained as $\mathbf{p} = (r, \theta, \phi)$. Compared with the Cartesian coordinate system, the spherical representation describes joint rotations and directional variations more naturally.

On this basis, the angular components of the spherical coordinates are used as the inputs to the spherical harmonic functions. A set of orthogonal spherical harmonic basis functions is employed to perform a frequency-domain expansion of the local geometric relationships. This process produces local motion representations with directional sensitivity and high-order geometric representation capability.Spherical harmonic functions are a set of orthogonal basis functions defined on the unit sphere. They decompose spatial directional information into different frequency components. Given the spherical coordinates (θ, ϕ), the enhanced features $\hat{F}$ can be obtained according to Eqs. (4)(6). Compared with conventional coordinate representations, the spherical harmonic representation maps local geometric relationships into the frequency domain. As a result, subtle directional variations produce stronger responses in high-frequency components.

After the local spherical harmonic transformation, the local geometric relationships among key joints are mapped into the spherical harmonic frequency domain. For a spherical harmonic expansion with the maximum degree l, a total of $l(l+2)$ independent spherical harmonic basis functions are obtained. The outputs of spherical harmonic functions contain both positive and negative responses. However, our method focuses on the response magnitude of local geometric relationships. Following the strategy in [36], the magnitude of the spherical harmonic coefficients is used as the feature representation. This operation removes the influence of sign differences. It also allows the model to focus on the energy distribution of local geometric structures. As a result, the stability of the feature representation is improved. Finally, to facilitate feature aggregation in the subsequent network, the joint dimension and the spherical harmonic frequency dimension are fused. Specifically, the local neighborhood dimension K is flattened and projected to the channel dimension.

$$F_{\text{sh}} = \text{Reshape}\left(\hat{F}\right), F_{\text{sh}} \in \mathbb{R}^{I(I+2)K \times T \times K.} \tag{22}$$

This process encodes the local geometric relationships into multi-frequency spherical harmonic responses. It then reorganizes the channel dimension to obtain a compact representation of the high-order directional information among different joints. As a result, more discriminative fine-grained action features are obtained. These features provide rich local geometric priors for the subsequent action classification.

3.4 Overall

This section presents the overall architecture of the proposed network. As shown in Fig. 2, the network consists of three components: the Discriminative Micro-action Joint Amplifier (DMJA), the feature extraction network, and the classifier.

The input skeleton sequence is first fed into the DMJA module. DMJA consists of three submodules: MIE, SKJS, andLSHA. The module identifies discriminative micro-action regions from the skeleton sequence. It then enhances local geometric variations

using spherical harmonic representations in the frequency domain. As a result, more discriminative fine-grained motion features are obtained.

After the DMJA module, the local spherical harmonic enhanced feature is obtained as

$$F_{\mathrm{SH}} = \mathrm{DMJA}(X). \tag{23}$$

The feature encodes high-order local geometric relationships and multi-frequency directional information among key joints. It enhances the discriminability of subtle motion patterns.

To preserve both the original motion information and the enhanced geometric information, feature fusion is performed by channel concatenation. Specifically, the original skeleton feature and the spherical harmonic feature are concatenated along the channel dimension:

$$F_{\mathrm{in}} = \mathrm{Concat}(X, F_{\mathrm{SH}}). \tag{24}$$

This fusion strategy preserves the global motion information of the original skeleton sequence. It also introduces local geometric priors generated by the DMJA module. Consequently, a more informative action representation is obtained.

The fused feature is then fed into the feature extraction network. CTR-GCN is adopted as the backbone network. Its original ten-layer graph convolution architecture is retained. CTR-GCN dynamically learns the relationships among joints through channel-wise topology refinement. It effectively captures the spatio-temporal dependencies in skeleton sequences.

After the ten-layer CTR-GCN, the high-level action representation is obtained as

$$F_{\mathrm{out}} = \mathrm{CTRGCN}(F_{\mathrm{in}}). \tag{25}$$

Finally, global average pooling and a fully connected classifier are applied to predict the action category:

$$\hat{y} = \mathrm{Classifier}(F_{\mathrm{out}}). \tag{26}$$

It is worth noting that the DMJA module is only introduced at the input stage. It does not modify the architecture of the original CTR-GCN. Therefore, the proposed method enhances the modeling of fine-grained actions and local discriminative motion patterns with only a small computational and parameter overhead.

4 Experiments

4.1 Datasets

NTU RGB+D 60. [16] The NTU RGB+D 60 dataset contains about 60,000 video samples from 60 action classes. Each skeleton consists of 25 joints in each frame. The dataset provides two evaluation protocols. The first is Cross-Subject (X-Sub), where the training and testing sets contain different subjects. The second is Cross-View (X-View), where the training samples are collected from Camera 2 and Camera 3, and the testing samples are collected from Camera 1.

NTU RGB+D 120. [17] NTU RGB+D 120 is an extended version of NTU RGB+D 60. It contains 114,480 video samples from 120 action classes. The dataset is

Table 1: Comparison with state-of-the-art methods on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA datasets.

Method	Year	Mod.	NTU60		NTU120		NW-UCLA
			X-Sub	X-View	X-Sub	X-Set	Top-1
CTR-GCN [13]	2021	4s	92.4	96.8	88.9	91.1	–
MS-G3D++ [20]	2022	2s	92.2	96.6	87.2	89.0	–
HD-GCN [39]	2023	4s	93.0	97.0	89.8	91.2	97.0
BlockGCN [40]	2024	4s	93.1	97.0	90.3	91.5	96.9
DS-GCN [41]	2024	4s	93.1	97.5	89.2	91.1	–
SkateFormer [42]	2024	4s	93.5	97.8	89.4	91.4	98.3
SelfGCN [43]	2024	4s	93.1	96.6	89.4	91.0	96.8
SA-TDGFormer [44]	2025	4s	92.7	96.8	86.9	88.9	–
LA-GCN [45]	2025	4s	93.5	97.2	90.7	91.8	97.6
ProtoGCN [46]	2025	2s	93.0	97.2	89.7	91.2	–
HiOD [47]	2025	4s	93.6	97.4	90.2	91.6	–
SkelFormer [48]	2026	4s	92.8	96.2	89.4	91.0	–
FD-GCN [49]	2026	2s	90.6	96.3	85.5	87.9	95.1
Ours		2s	92.9	96.6	89.2	91.0	97.9

collected from 106 subjects under 32 camera setups. Two evaluation protocols are provided. The first is Cross-Subject (X-Sub), where the 106 subjects are equally divided into training and testing groups. The second is Cross-Setup (X-Set), where samples captured under even-numbered camera setups are used for training, and those captured under odd-numbered camera setups are used for testing.

NW-UCLA. [18] The NW-UCLA dataset contains 1,494 video sequences from 10 action classes. The data are collected by three Microsoft Kinect sensors from different viewpoints. The samples captured by the first two sensors are used for training, while the samples captured by the remaining sensor are used for testing.

4.2 Implementation Details

All experiments are conducted using the PyTorch framework on a single NVIDIA RTX 4090 GPU. The network is trained using stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.004. The model is trained for 75 epochs. A warm-up strategy is adopted during the first five epochs. The initial learning rate is set to 0.1 and is multiplied by 0.1 at the 35th and 55th epochs. For the NTU RGB+D datasets, the mini-batch size is set to 64, and each skeleton sequence is uniformly sampled to 64 frames. For the NW-UCLA dataset, the mini-batch size is set to 16.

4.3 Comparison with State-of-the-Art Methods

Table 1 reports the comparison between the proposed method and recent state-of-the-art methods on the NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA datasets. All comparisons follow the same modality setting. Specifically, methods are compared under identical input modalities, including Joint (J), Joint+Bone (J+B),

and Joint+Bone+Joint Motion+Bone Motion (J+B+JM+BM), to ensure a fair comparison. The Top-1 accuracy is used as the evaluation metric. For NTU RGB+D 60, the Cross-Subject (X-Sub) and Cross-View (X-View) protocols are adopted. For NTU RGB+D 120, the Cross-Subject (X-Sub) and Cross-Setup (X-Set) protocols are used. For NW-UCLA, the standard Top-1 accuracy is reported.

As shown in Table 1, the proposed method achieves competitive performance on all three benchmark datasets. Compared with representative methods, including ST-GCN, 2s-AGCN, MS-G3D, and CTR-GCN, the proposed method obtains higher recognition accuracy under the same modality setting. These results demonstrate the effectiveness of the proposed Discriminative Micro-action Joint Amplifier (DMJA).

Unlike methods that mainly model global joint representations, DMJA first extracts motion information to estimate the motion states of joints. It then identifies discriminative key joints through the Subtle Key Joint Selection module. The Local Spherical Harmonic Amplifier further models the local geometric relationships among the selected joints and enhances subtle directional variations through spherical harmonic frequency-domain representations. As a result, the proposed method learns more discriminative action representations. The experimental results show that DMJA effectively complements the fine-grained motion information in skeleton sequences and improves the discrimination of similar actions. Consequently, consistent performance improvements are achieved on multiple benchmark datasets.

4.4 Ablation Study

Effectiveness of the Subtle Key Joint Selection Strategy. To verify the effectiveness of the proposed subtle key joint selection strategy, we conduct an ablation study on the NTU RGB+D 60 dataset under the Cross-Subject (X-Sub) protocol. In the experiment, eight key joints are selected from different motion amplitude ranges. As shown in Table 2, the highest recognition accuracy of 91.0% is achieved when the selected range is 0.3–0.8. This result improves the accuracy by 0.4% and 0.3% compared with the ranges of 0.7–1 and 0–0.4, respectively. The results indicate that selecting key joints only according to large motion amplitudes cannot produce the optimal action representation. Selecting only joints with very small motion amplitudes also degrades the performance. This is because some subtle joints are sensitive to noise and are weakly related to action semantics. In contrast, the medium motion amplitude range (0.3–0.8) preserves more discriminative micro-motion information while suppressing noisy joints. Therefore, it achieves the best recognition performance.

Effect of the Number of Key Joints. To investigate the influence of the number of key joints, an ablation study is conducted on the NTU RGB+D 60 dataset under the Cross-Subject (X-Sub) protocol. The number of selected key joints K is set to 3, 5, 8, and 10. The results are reported in Table 3. As K increases, the recognition accuracy first improves and then decreases. The model achieves accuracies of 90.7%, 90.8%, 91.0%, and 90.9% when K is set to 3, 5, 8, and 10, respectively. The highest accuracy is obtained at $K = 8$. Although increasing K to 10 introduces more key joints, it slightly degrades the recognition performance and increases the model complexity. These results indicate that an appropriate number of key joints provides a good balance between discriminative information and feature redundancy. When $K = 8$,

Table 2: Ablation study of the proposed Subtle Key Joint Selection strategy on NTU RGB+D 60 (X-Sub).

Motion Magnitude Range	Accuracy (%)
0–0.4	90.7
0.3–0.8	**91.1**
0.7–1.0	90.6

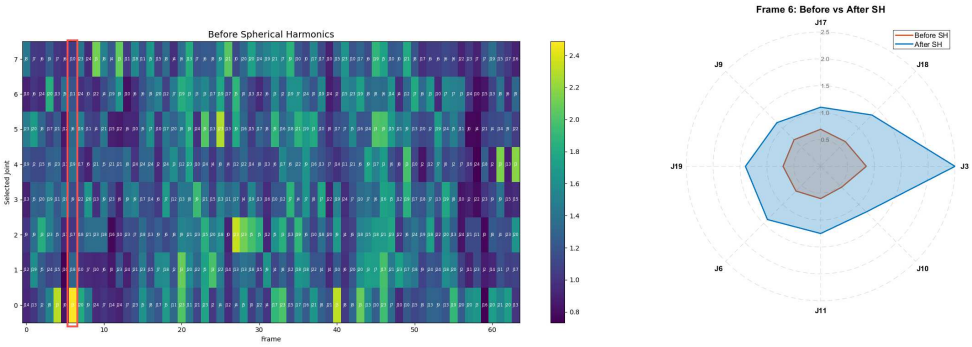
the model preserves sufficient discriminative micro-motion information while reducing the interference from irrelevant joints. Therefore, K is fixed to 8 in all subsequent experiments.

Table 3: Ablation study on the number of selected key joints (K) on NTU RGB+D 60 (X-Sub).

Number of Key Joints (K)	Top-1 Accuracy (%)
1	90.2
3	90.7
5	90.8
8	**91.1**
10	90.9
12	90.8

Feature Visualization. To further verify the effectiveness of the proposed DMJA module, feature visualization is conducted. A skeleton sequence is randomly selected from the training batch. The features before and after the DMJA module are visualized for comparison. The results are shown in Fig. 3. Fig. 3(a) shows the key joints selected in each frame by the proposed Subtle Key Joint Selection module. Fig. 3(b) compares the feature distributions before and after the DMJA module. The orange points represent the original input features, while the blue points represent the enhanced features generated by the DMJA module. It can be observed that the feature responses become more concentrated after the DMJA module. The responses of the discriminative regions are significantly enhanced, while the irrelevant responses are suppressed. These results indicate that the proposed module effectively highlights subtle motion information that is important for action recognition and learns more discriminative feature representations.

Parameter Efficiency. As shown in Table 4, the proposed method requires fewer learnable parameters than the compared methods while achieving superior recognition accuracy. Unlike existing approaches that improve performance by introducing additional network branches or substantially increasing model capacity, the proposed DMJA enhances feature representation through discriminative micro-motion modeling. Specifically, SKJS identifies only a small number of informative joints, and LSHA



(a) (b)

Fig. 3: Visualization of the proposed DMJA module. (a) Selected key joints in the current frame. (b) Joint features before and after the spherical harmonic transformation.

performs local directional enhancement on the selected joints instead of the entire skeleton.

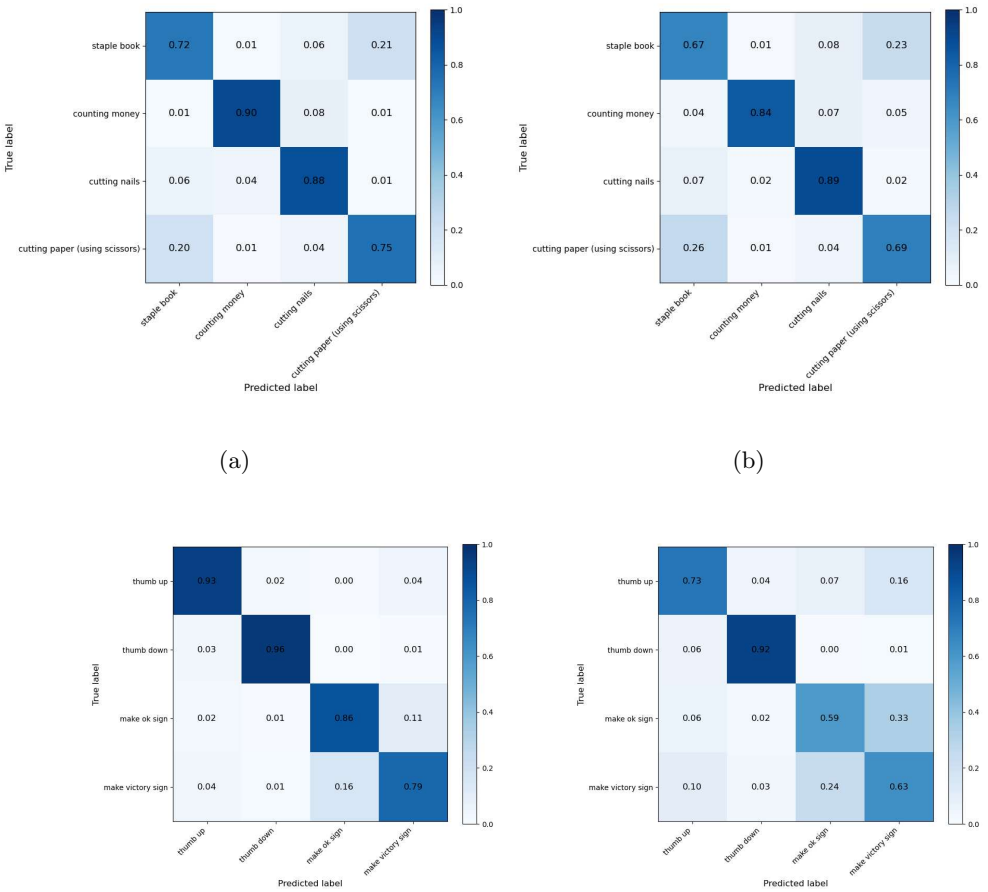
Table 4: Comparison of Model Complexity and Recognition Performance on NTU RGB+D 60 (X-Sub).

Method	Params (M)	Accuracy (%)
ST-GCN	1.22	81.5
CTR-GCN	1.46	89.9
MS-G3D	2.80	89.4
HD-GCN	1.68	90.1
Ours	**1.50**	**91.1**

Visualization of confusion matrix. The overall confusion matrix demonstrates a clearer diagonal distribution, indicating that DMJA achieves accurate recognition across most action categories. Furthermore, the local confusion matrix comparisons in Fig. 4d(a,c) and Fig. 4d(b,d) show the results of DMJA and CTR-GCN, respectively. Compared with CTR-GCN, DMJA demonstrating its superior capability in distinguishing visually similar actions. This improvement benefits from the discriminative micro-motion joint selection and local motion enhancement strategy, which enables DMJA to capture subtle motion variations that are easily overlooked by conventional spatial-temporal modeling methods.

5 Conclusion

We propose a Discriminative Micro-Motion Joint Amplifier for skeleton-based action recognition to enhance subtle discriminative motion information. First, we



(a) (b)

(c) (d)

Fig. 4: (a) and (c) present the local confusion matrices of our proposed model, while (b) and (d) show the corresponding results of the CTR-GCN model. It can be clearly observed that our method achieves better classification performance than CTR-GCN, with fewer misclassified samples in the selected local regions.

design a MIE module to extract joint motion information from multiple temporal scales, providing motion priors for key joint selection. Then, we propose a SKJS module, which adaptively identifies discriminative micro-motion joints according to motion response intensity and reduces noise from irrelevant joints. Finally, we introduce a LSHA module, which maps local directional relationships between key joints into the spherical harmonic domain. By utilizing the multi-frequency representation ability of spherical harmonics, LSHA enhances high-frequency features related to micro-motions and improves the representation of local discriminative information. Notably, our

framework achieves state-of-the-art performance on skeleton-based action recognition benchmarks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant 62571346.

Authorship Contributions

Wenming Cao: Conceptualization, Writing, Supervision - Review and Editing. Gai Wang: Methodology, Data processing, Writing – original draft. Xinpeng Yin: Conceptualization, Methodology, Writing – original draft, Revision.

Declarations

Conflicts of Interest The authors declare no conflicts of interest.

Author declarations and funding statements

This research received funding from the National Natural Science Foundation of China under grant 62571346.

References

[1] Khan, M.A., Javed, K., Khan, S.A., Saba, T., Habib, U., Khan, J.A., Abbasi, A.A.: Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia tools and applications **83**(5), 14885–14911 (2024)

[2] Gammulle, H., Ahmedt-Aristizabal, D., Denman, S., Tychsen-Smith, L., Petersson, L., Fookes, C.: Continuous human action recognition for human-machine interaction: a review. ACM Computing Surveys **55**(13s), 1–38 (2023)

[3] Xing, Y., Zhu, J.: Deep learning-based action recognition with 3D skeleton: a survey. Wiley Online Library (2021)

[4] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp. 1297–1304 (2011). Ieee

[5] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)

[6] Xie, Z., Chen, J., Wang, Y., Xie, B.: Enhanced fine-grained relearning for skeleton-based action recognition: Z. xie et al. The Visual Computer **41**(10), 7983–7995 (2025)

[7] Chang, H., Chen, J., Li, Y., Chen, J., Zhang, X.: Wavelet-decoupling contrastive enhancement network for fine-grained skeleton-based action recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4060–4064 (2024). IEEE

[8] Liu, K., Li, Y., Xu, Y., Liu, S., Liu, S.: Spatial focus attention for fine-grained skeleton-based action tasks. IEEE Signal Processing Letters **29**, 1883–1887 (2022)

[9] Chi, H.-g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20186–20196 (2022)

[10] Bian, X., Chang, D., Yang, Y., He, Z., Liang, K., Ma, Z.: Class-aware contrastive learning for fine-grained skeleton-based action recognition. In: Proceedings of the Asian Conference on Computer Vision, pp. 3638–3654 (2024)

[11] Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3d skeleton-based action recognition using learning method. Cyborg and Bionic Systems **5**, 0100 (2024)

[12] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895 (2020)

[13] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368 (2021)

[14] Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3 d shape descriptors. In: Symposium on Geometry Processing, vol. 6, pp. 156–164 (2003)

[15] Ha, S., Lyu, I.: Spharm-net: Spherical harmonics-based convolution for cortical parcellation. IEEE Transactions on Medical Imaging **41**(10), 2739–2751 (2022)

[16] Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)

[17] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., Kot, A.C.: 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

(2019)

[18] Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.-C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2649–2656 (2014)

[19] Xin, W., Liu, R., Liu, Y., Chen, Y., Yu, W., Miao, Q.: Transformer for skeleton-based action recognition: A review of recent advances. Neurocomputing **537**, 164–186 (2023)

[20] Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2969–2978 (2022)

[21] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[22] Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)

[23] Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1113–1122 (2021)

[24] Song, Y.-F., Zhang, Z., Shan, C., Wang, L.: Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology **31**(5), 1915–1925 (2020)

[25] Duan, H., Wang, J., Chen, K., Lin, D.: Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv preprint arXiv:2210.05895 (2022)

[26] Wu, Z., Ding, Y., Wan, L., Li, T., Nian, F.: Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. Pattern Recognition **159**, 111106 (2025)

[27] Zhang, H., Liu, X., Yu, D., Guan, L., Wang, D., Ma, C., Hu, Z.: Skeleton-based action recognition with multi-stream, multi-scale dilated spatial-temporal graph convolution network. Applied Intelligence **53**(14), 17629–17643 (2023)

[28] Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8 (2019).

IEEE

[29] Barrera, R.G., Estevez, G., Giraldo, J.: Vector spherical harmonics and their application to magnetostatics. European Journal of Physics **6**(4), 287–294 (1985)

[30] Zonca, A., Singer, L., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., Gorski, K.: healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. Journal of Open Source Software **4**(35), 1298 (2019)

[31] Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., Tuia, D.: Geographic location encoding with spherical harmonics and sinusoidal representation networks. In: International Conference on Learning Representations, vol. 2024, pp. 1746–1759 (2024)

[32] De Luca, A., Karayumak, S.C., Leemans, A., Rathi, Y., Swinnen, S., Gooijers, J., Clauwaert, A., Bahr, R., Sandmo, S.B., Sochen, N., *et al.*: Cross-site harmonization of multi-shell diffusion mri measures based on rotational invariant spherical harmonics (rish). NeuroImage **259**, 119439 (2022)

[33] Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning so (3) equivariant representations with spherical cnns. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 52–68 (2018)

[34] Poulenard, A., Rakotosaona, M.-J., Ponty, Y., Ovsjanikov, M.: Effective rotation-invariant point cnn with spherical harmonics kernels. In: 2019 International Conference on 3D Vision (3DV), pp. 47–56 (2019). IEEE

[35] Fang, J., Zhou, D., Song, X., Jin, S., Yang, R., Zhang, L.: Rotpredictor: Unsupervised canonical viewpoint learning for point cloud classification. In: 2020 International Conference on 3D Vision (3DV), pp. 987–996 (2020). IEEE

[36] Prasse, K., Jung, S., Zhou, Y., Keuper, M.: Local spherical harmonics improve skeleton-based hand action recognition. In: DAGM German Conference on Pattern Recognition, pp. 67–82 (2023). Springer

[37] Cohen, T., Weiler, M., Kicanaoglu, B., Welling, M.: Gauge equivariant convolutional networks and the icosahedral cnn. In: International Conference on Machine Learning, pp. 1321–1330 (2019). PMLR

[38] Müller, C.: Spherical Harmonics. Springer, ??? (2006)

[39] Zhou, T., Shen, Y., Chen, K., Cao, Q.: Hierarchical dual graph convolutional network for aspect-based sentiment analysis. Knowledge-Based Systems **276**, 110740 (2023)

[40] Zhou, Y., Yan, X., Cheng, Z.-Q., Yan, Y., Dai, Q., Hua, X.-S.: Blockgcn: Redefine topology awareness for skeleton-based action recognition. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2049–2058 (2024)

[41] Xie, J., Meng, Y., Zhao, Y., Nguyen, A., Yang, X., Zheng, Y.: Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 6225–6233 (2024)

[42] Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: European Conference on Computer Vision, pp. 401–420 (2024). Springer

[43] Wu, Z., Sun, P., Chen, X., Tang, K., Xu, T., Zou, L., Wang, X., Tan, M., Cheng, F., Weise, T.: Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition. IEEE Transactions on Image Processing **33**, 4391–4403 (2024)

[44] Chen, D., Chen, M., Wu, P., Wu, M., Zhang, T., Li, C.: Two-stream spatio-temporal gcn-transformer networks for skeleton-based action recognition. Scientific Reports **15**(1), 4982 (2025)

[45] Xu, H., Gao, Y., Hui, Z., Li, J., Gao, X.: Language knowledge-assisted representation learning for skeleton-based action recognition. IEEE Transactions on Multimedia (2025)

[46] Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 29248–29257 (2025)

[47] Chang, H., Ren, P., Zhang, H., Xie, L., Chen, H., Yin, E.: Hierarchical-aware orthogonal disentanglement framework for fine-grained skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11252–11261 (2025)

[48] Yan, J., Zhang, X., Tan, C., Li, D.: Skelformer: An adaptive hierarchical transformer-based approach on skeleton graphs for human action recognition in video sequences. PloS one **21**(1), 0340390 (2026)

[49] Huo, J., Cai, H., Meng, Q.: Fast distance-enhanced graph convolutional network for skeleton-based action recognition. Pattern Recognition **169**, 111827 (2026)