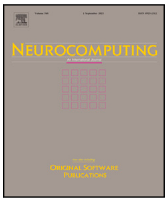
Contents lists available at ScienceDirect

Neurocomputing

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



PRG-Net: Point Relationship-Guided Network for 3D human action recognition

Yao Du [a], Zhenjie Hou [a,*], En Lin [b], Xing Li [c], Jiuzhen Liang [a], Xinwen Zhou [a]

[a] School of Computer Science and Artificial Intelligence, Aliyun School of Big Data, School of Software, Changzhou University, Changzhou, 213164, China
[b] Goldcard Smart Group Co., Ltd, Zhejiang, 310018, China
[c] College of Information Science and Technology, College of Artificial intelligence, Nanjing Forestry University, Nanjing, 210098, China

ARTICLE INFO

Communicated by J. Cheng

Keywords:
3D action recognition
Point cloud sequences
Inter-region relations
Graph aggregation

ABSTRACT

Point clouds contain rich spatial information, providing important supplementary clues for human action recognition. Recent methods for action recognition based on point cloud sequences primarily rely on complex spatiotemporal local encoding. However, these methods often utilize max-pooling operations to select features when extracting local features, restricting feature updates to local neighborhoods and failing to fully exploit the relationships between regions. Moreover, cross-frame encoding can also lead to the loss of spatiotemporal information. In this study, we propose PRG-Net, a Point Relation Guided Network, to further improve the learning of spatiotemporal features in point clouds. First, we designed two core modules: the Spatial Feature Aggregation (SFA) and the Spatial Feature Descriptor (SFD) modules. The SFA module expands the spatial structure between regions using dynamic aggregation techniques, while the SFD module guides the region aggregation process by Attention-Weighted Descriptors. They enhance the modeling of human spatial structure by expanding the relationships between regions. Second, we introduce inter-frame motion encoding techniques that can obtain the final spatiotemporal representation of the human body through the aggregation of cross-frame vectors, without relying on complex spatiotemporal local encoding. We evaluate PRG-Net on publicly available human action recognition datasets, including NTU RGB+D 60, NTU RGB+D 120, UTD-MHAD, and MSR Action 3D. Experimental results demonstrate that our method outperforms state-of-the-art point-based 3D action recognition methods significantly. Furthermore, we conduct extended experiments on the SHREC 2017 dataset for gesture recognition, and the results show that our method maintains competitive performance on that dataset as well.

1. Introduction

Vision-based human action recognition is a well-studied yet challenging problem in the field of computer vision, with significant potential applications in human–computer interaction, such as automatic driving or augmented reality.

The majority of approaches for action recognition primarily rely on video data [1–3] or skeleton data [4–6]. However, recent studies have explored the utilization of sequences of 3D point clouds as an alternative input modality and have highlighted several advantageous properties of point clouds [7–11]. Most importantly, point clouds accurately represent the 3D geometric structure of the entire visible body surface, rendering them indispensable for achieving accurate action recognition. Furthermore, point clouds exhibit robustness against varying camera perspectives and are less affected by issues such as illumination and background clutter. Collectively, these advantages position point clouds as a promising avenue for advancing the field of human action recognition. For these reasons, our work specifically focuses on the task of human action recognition using point cloud sequences.

The unordered nature and lack of explicit spatial structure in point cloud sequences pose significant challenges in accurately modeling their spatial structures. We have observed that existing point cloud-based action recognition methods [7–9] typically rely on constructing complex spatiotemporal local encoding to handle dynamic point clouds. These methods often construct local regions in the spatiotemporal domain to extract spatiotemporal features. However, since the local regions are typically computed independently, and the local features are only aggregated through max-pooling, the relationships between regions have not been fully utilized. Nevertheless, the relationships between regions also provide valuable information for 3D action understanding, such as the similarity, proximity, and symmetry of different body parts. By considering the relationships between regions, we can

[*] Corresponding author.
E-mail address: houzj@cczu.edu.cn (Z. Hou).

https://doi.org/10.1016/j.neucom.2025.130015
Received 15 July 2024; Received in revised form 9 January 2025; Accepted 10 March 2025
Available online 17 March 2025
0925-2312/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

gain a more comprehensive understanding of the action, rather than being limited to the aggregation of local features.

In this paper, we argue that considering the relationships between human body regions can enhance the representation for 3D human action recognition. To this end, we propose an end-to-end framework, called the Point Relation Guided Network (PRG-Net), to explore the relationships between different parts of the human body. PRG-Net consists of two core modules: the Spatial Feature Aggregation (SFA) module and the Spatial Feature Descriptor (SFD) module. For the Spatial Feature Aggregation module, it dynamically partitions the point cloud into regions and performs grouping and aggregation along the channel dimension to establish relationships between these regions. The Spatial Feature Descriptor integrates local information into the channels using a multi-local max-pooling mechanism, and guides the region aggregation process through a grouped activation operation. To comprehensively evaluate the inter-relationships, we perform region shuffling and execute the Spatial Feature Aggregation module twice. To capture temporal information, we designed an inter-frame motion encoding technique that integrates temporal position encoding into the feature vector sequence, enabling joint learning of temporal and spatial features. By performing max pooling operations within different temporal intervals and concatenating the spatiotemporal features from each interval, we obtain a comprehensive spatiotemporal feature representation. We conduct extensive experiments on five public datasets, including MSR Action 3D, NTU RGB+D 60, NTU RGB+D 120, UTD-MHAD, and SHREC'17, to validate the effectiveness of our method.

In summary, the main contributions of our work are as follows:

(1) We propose the PRG-Net network for 3D human action recognition. PRG-Net can model the inter-relationships between human body regions, while achieving cross-frame vector aggregation without additional local encoding, providing a richer feature representation for 3D human action recognition.

(2) We design the Spatial Feature Descriptor (SFD) and Spatial Feature Aggregation (SFA) modules. SFA utilizes dynamic aggregation to optimize the spatial structure between regions, while SFD focuses on the description of local features, guiding the aggregation process. These two modules work collaboratively to enhance the spatial structure modeling of human actions.

(3) We have achieved state-of-the-art accuracy on the point-based action recognition benchmarks of MSR Action 3D and UTD-MHAD datasets, as demonstrated through extensive experiments.

2. Related work

2.1. Deep learning on point set

Due to the unordered and irregular nature of point cloud data, traditional convolutional networks face challenges when directly processing such data. Early studies primarily employed multi-view projections [12,13] and voxelization techniques [14,15] to extract point cloud features, which led to the loss of three-dimensional information and increased computational resources. Recent research has focused on directly processing point cloud sequences, with Qi et al. [16] introducing the landmark PointNet. This algorithm takes point cloud sequences as input and effectively addresses the impact of point cloud unorderedness on results by introducing symmetric functions. However, since PointNet only considers the feature information of individual points, it has limitations in capturing local structures. Subsequently, Qi et al. [17] further developed PointNet++ by incorporating a hierarchical feature extraction mechanism to effectively capture local geometric structure features. The success of PointNet++ laid the foundation for the development of many subsequent networks. DGCNN [18] constructs the neighborhood relationships between points based on their distances in feature space and generates features for central points through the aggregation of paired features. Recent works have begun to explore the application of attention mechanisms in learning point features. Yang et al. [19] proposed a point attention transformer that utilizes parameter-efficient group shuffle attention to learn relationships between points. Lin et al. [20] introduced a new mechanism to enhance point cloud features by learning the best single attention point (LAP) for each input point. Lu et al. [21] proposed 3DGTN, a novel point cloud representation network known as the 3D Dual Self-attention Global Local Transformer Network (3DGTN), which effectively fuses global and local features through dual attention. The above research has significantly advanced the field of 3D point cloud analysis.

2.2. Dynamic point cloud modeling

Our research focuses on 3D human action recognition based on point cloud sequences. A key issue that needs to be addressed for dynamic point cloud recognition algorithms is how to effectively extract spatiotemporal information. This is because, although point cloud sequences exhibit irregularities in the spatial dimension, they maintain a certain regularity and validity in the temporal dimension. Therefore, accurately encoding the spatial structural information of the human body and effectively modeling its spatiotemporal dynamics are critical issues to consider. In the area of point cloud sequence modeling, MeteorNet [7] extended 3D points to 4D points, incorporating the temporal dimension into PointNet++ for processing. However, merely adding the temporal dimension to PointNet++ does not directly establish dependencies between frames. H et al. [22] proposed PointRNN, a point-based recurrent neural network that extends the grouping operation from the spatial domain to the temporal domain, aiming to capture spatiotemporal information in point cloud sequences. Y et al. [23] introduced PointLSTM, a novel LSTM unit designed for unordered point clouds, which aims to capture long-term relationships among points. Fan et al. [8] proposed PSTNet, which decouples the spatiotemporal dimensions of point clouds and introduces PointTube to capture the spatiotemporal structure of point cloud sequences. Subsequently, to avoid point tracking, Fan et al. [9] introduced the Point 4D Transformer, which effectively captures spatiotemporal dependencies in point cloud data through a self-attention mechanism. Li et al. [11] proposed SequentialPointNet, which abstracts point clouds as a sequence of hyperpoints and designs a Hyperpoint-Mixer module to model their spatiotemporal structure, significantly simplifying encoding complexity and improving computational efficiency. Recent methods [24–26] have begun to explore the possibilities of applying self-supervised or semi-supervised techniques in dynamic 4D point cloud modeling. In summary, to effectively extract spatiotemporal information from point cloud sequences, researchers have proposed various innovative methods, all of which have contributed to the advancement of 3D human action recognition technology to varying degrees.

2.3. Spatiotemporal modeling in action recognition

In the early stages of video-driven human action recognition research, 3D convolutional neural networks were primarily used to extract spatiotemporal features of the human body. Later, to further enhance spatiotemporal modeling capabilities, some researchers [1] explored dual-stream architectures to independently capture temporal and spatial information, which were then used for classification. Additionally, the introduction of attention mechanisms enabled networks to focus on key visual information critical for action recognition [27,28]. With the decreasing costs of depth cameras like Kinect, 3D action recognition has gradually become a hot research area. Some researchers have used skeletal sequences to encode the local structure of the human body, with Yan et al. [5] being the first to propose ST-GCN, aimed at capturing spatial configurations and temporal dynamics in skeleton data. This led to significant interest in using GCNs for skeleton-based action recognition, resulting in a large body of

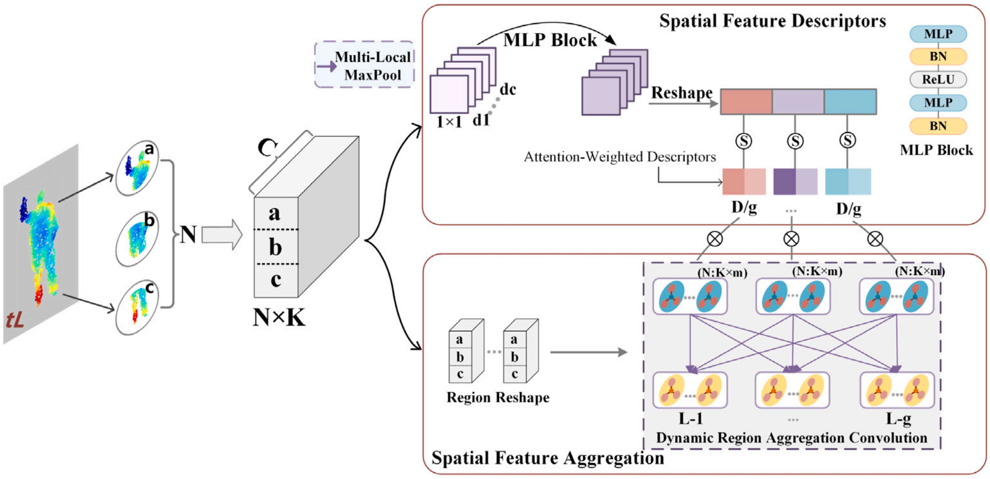


Fig. 1. Spatial aggregation process.

work [29,30]. As low-cost point cloud acquisition technologies matured, human action recognition based on point cloud data became a research hotspot due to its robustness in representing object contours and handling changes in viewpoints. MeteorNet [7] was the first to explore dynamic raw point cloud sequences, enhancing the processing of spatiotemporal information by extending 3D point cloud data into 4D. Subsequently, PSTNet [8] modeled the spatiotemporal information of raw point cloud sequences by hierarchically decomposing spatial and temporal dimensions. Recently, researchers have begun to employ feature fusion methods to enhance the expressive power of human spatiotemporal features. Kim et al. [31] proposed a new 3D deformable transformer that adaptively processes spatiotemporal features and jointly learns video and skeletal features through a cross-modal learning strategy, thereby enhancing action recognition capabilities. Jathushan [32] introduced a Lagrangian action recognition model (LART) that integrates 3D pose and contextualized appearance information, improving action recognition accuracy through the analysis of crowd trajectories.

3. Method

In this section, we introduce our approach to human action recognition based on point clouds. We present our PRG module and the overall architecture.

3.1. PRG module

The proposed PRG module consists of two key modules. The first component is the Spatial Feature Descriptor (SFD) module, which aims to extract discriminative descriptors of the human body structure. This module guides the region aggregation process by focusing on the description of local features, thereby more accurately capturing the structural information of the human body. The second component is the Spatial Feature Aggregation (SFA) module, which focuses on modeling the relationships between different body regions. Through dynamic aggregation techniques, the SFA module expands the spatial structure between regions, further enhancing the modeling capability of the human body's spatial relationships.

The input of our methods is a sequence of 3D point clouds, denoted as $\left\{P^{(1)}, P^{(2)}, \dots, P^{(T)}\right\}$. Here, $P^{(t)} = \left\{p_i^{(t)}\right\}_{i=1}^{N}$ represents a point cloud of N points at time t, each point having associated point-wise features $\mathcal{F} = \left\{F_{Pi}\right\}_{i=1}^{N}$. where $F_{Pi} \in \mathbb{R}^{1\times C}$ denotes the C-dimensional feature of the point p_i.

We first construct local associations for the input point cloud sequence. Inspired by edge convolution [18], we compute the difference between the central point feature F_{Pi} and the features of its jth nearest neighbor F_{Pj} to obtain the local edge features within the local coordinate system. For each point F_{Pi}, a feature vector is generated by combining the edge features and point features together, as described by Eqs. (1) and (2):

$$E_p^t = p_i^t, E_s^t = p_i^t - p_j^t \tag{1}$$

$$F^t_{\{i=1,\dots,N\}} = \left[E_p^t, E_s^t\right] \tag{2}$$

where E_s^t represents the edge feature vector, and E_p^t denotes the feature of the central point. t represents the local structural encoding of the tth frame. N is the number of central points, and there are K nearest neighbor points in total.

3.1.1. Spatial feature descriptors

To efficiently achieve spatial feature aggregation, we first extract the most representative parts from the features of spatial points, constructing initial human spatial feature descriptors. Based on this foundation, we group them and dynamically allocate weights to different groups. This design ensures a reasonable balance of contributions from each feature group during the subsequent spatial aggregation stage, significantly optimizing the overall effectiveness of feature aggregation.

Specifically, we first use Multi-local Max Pooling operations to select key local human body features and generate human spatial descriptors. This process is achieved by choosing the edges with the highest responses within N×K local groups. As shown below:

$$d^t = \text{Max}\left\{F^t \mid \left(P^t_{\{i=1,\dots,N\}}, P^t_j\right) \in E^t_{N\times K}\right\} \tag{3}$$

$$D = \left\{d_1, d_2, \dots, d_c\right\} \tag{4}$$

where $E_{N\times K}$ is the set of N×K edges originating from N points P_1, $P_2, \dots, P_i$, and d^t represents the updated feature vector for each channel, where c denotes the length of the feature channel. Then, we obtain these feature descriptors D.

We input these feature descriptors into a shared Multi-Layer Perceptron (MLP) for processing. These descriptors $D \in \mathbb{R}^{1\times C}$ are subsequently divided into g groups, and the features within each group are weighted using the sigmoid function. This enables assigning different weights to different feature groups during subsequent spatial aggregation to balance their contributions. The weighted spatial feature descriptors are represented as follows:

$$\tilde{d^i} = f\left(d^i\right) \in \mathbb{R}^{1\times m} \tag{5}$$

where $i \in \{1, 2, \dots, g\}$, where $m = \frac{c}{g}$ represents the number of descriptors per group, f denotes the sigmoid function.

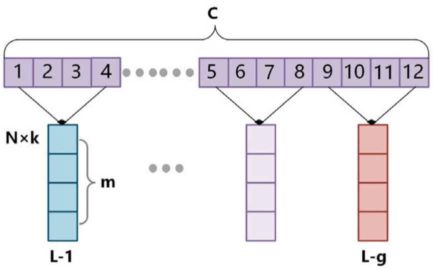


Fig. 2. Region shape.

3.1.2. Spatial feature aggregation

We have observed that in existing methods [8–10], human body local regions are usually computed independently, which can hinder the learning of internal relationships within the human body. Given the diversity of human features, we propose performing aggregation over body regions to enhance the spatial structural representation of the human body.

To capture inter-body relations, we first adaptively decompose the point cloud into a set of local regions, each local region is constructed using a dynamic graph. To be more precise, we select N points as centroids and obtain N local regions, then we employ a reshaping function $f_{rs} : \mathbb{R}^{k\times c} \longrightarrow \mathbb{R}^{k\times m\times g}$(where $m \times g = c$) to initially divide the local region into groups along the channel dimension and there are total of g groups $\left[L_1, L_2, \ldots, L_g\right] \in \mathbb{R}^{n\times k\times m}$, which corresponds to the number of descriptor groups. The simplified visualization of the reshape function is shown in Fig. 2.

Dynamic region Aggregation convolution. As illustrated in Fig. 1, we introduce a dynamic region aggregation convolution (DRAC) to perform channel-wise aggregation among these regions to establish their relationships. Specifically, we multiply the weighted descriptors with their corresponding local regions to balance the contributions of each group during the aggregation operation, represented as follows:

$$W_{\text{agg}} = \left[W_1 \times \tilde{d}^1, W_2 \times \tilde{d}^2, \ldots, W_g \times \tilde{d}^g\right] \in \mathbb{R}^{g\times 1} \tag{6}$$

where W_{agg} denotes the aggregation weight, we multiply each descriptor $\tilde{d}^i$ with its corresponding local region to obtain the weight of aggregation for different local regions.

Hence, the proposed aggregation operation is formulated as follows:

$$L_{\text{agg }1} = \sigma\left(\left[L^1, L^2, \ldots, L^g\right] \cdot \mathrm{W}_{\text{agg}}\right) \tag{7}$$

The output of the DRAC combines information from different groups, thereby leveraging hierarchical representations with distinct local receptive fields. As a result, the DRAC can be implemented using 3D convolutions with a kernel size of 1 × 1 × 1, and its output $L_{\text{agg }1}$ dimensions are (n, k, m, g), where n, k and g, are number of regions, neighbors and groups, and m is number of channels per each group.

Region Shuffle. Inspired by ShuffleNet [33], we propose a region shuffling method to enhance the spatial diversity of the human body. In particular, we transpose the into and subsequently flatten it, which serves as the input to the secondary aggregation operation(which is same as the DRAC) (see Fig. 3).

3.2. Network architecture details

The overall framework is depicted in Fig. 4. We employ a sequential structure to process spatial and temporal information. Specifically, for spatial structure learning, we utilize specially designed point cloud networks DGCNN [18] to encode the spatial structure of the human body. Subsequently, we introduce the SFD and SFA modules to learn richer spatial feature representations by considering the relationships between regions. For temporal information learning, we design the Interframe motion encoding method to effectively aggregate spatiotemporal features.

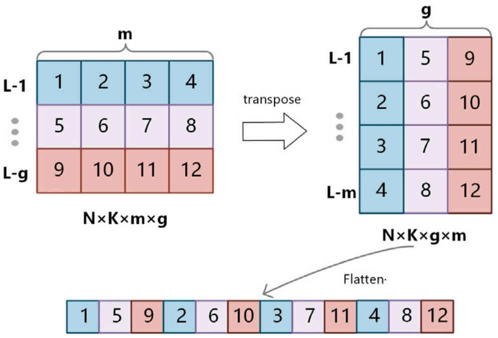


Fig. 3. Region shuffle.

3.2.1. Motion encoding

The temporal encoding module is shown in Fig. 5. We first perform max pooling on each frame to aggregate information across N dimensions, thereby extracting the key spatial features of the human body in each frame (denoted as $F_i = \left\{P_1, P_2, \ldots, P_m\right\}_{i=1,2\ldots t}$).

Secondly, we introduced positional encoding to explicitly inject temporal positional information into each frame's features through temporal displacement vectors. This helps the model understand each frame's relative position in time, enhancing its ability to perceive sequential information. Specifically, we use sinusoidal and cosine functions of different frequencies to encode positional information for each timestep's feature vector. Additionally, we apply an MLP module to perform a nonlinear transformation on the features enriched with temporal information, learning the joint relationships between temporal and spatial features. The entire process is represented as follows:

$$p_t^{2i} = \sin\left(t/10000^{2i/dm}\right) \tag{8}$$

$$p_t^{2i+1} = \cos\left(t/10000^{2i/dm}\right) \tag{9}$$

$$\hat{F}_i^t = MLP\left(F_i^t + \vec{P}_i^t\right) \tag{10}$$

where d_m represents the number of channels in the feature vector, t represents the time position, and i represents the dimension position. $\hat{F}_i^t$ represents the feature vector after injecting the position encoding.

Subsequently, we introduced max pooling along the temporal dimension to aggregate features across all frames, extracting global motion patterns and obtaining the global feature F1. Meanwhile, to capture fine-grained human motion characteristics, we adopted a Hierarchical Max Pooling approach, performing feature aggregation within T/2 temporal partitions to extract motion features F2 and F3 from local temporal windows. Finally, we fused F1, F2, and F3 to form the final representation of human actions. For classification, we applied a fully connected layer followed by softmax normalization to produce a C-dimensional vector representing class probabilities.

4. Experiments

In this section, we perform action recognition on two small-scale datasets, MSR Action3D and UTD-MHAD, as well as two large-scale datasets, NTU RGB+D 120 and NTU RGB+D 60. Additionally, we perform gesture recognition on the SHREC'17 dataset.

4.1. Datasets

NTU RGB+D 120 [34] is one of the largest and most challenging datasets for 3D action recognition, containing 114,480 RGB-D samples captured using Microsoft Kinect v2. The dataset provides a wealth of multimodal data, including RGB videos, depth maps, 3D skeleton data, and infrared videos, covering 120 action classes. It features two

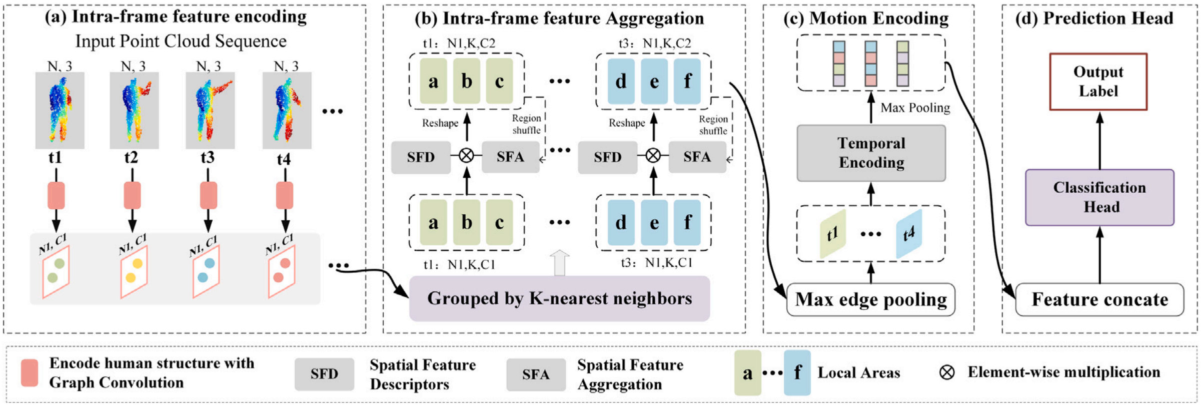

(a) Intra-frame feature encoding
Input Point Cloud Sequence
N, 3
t1
t2
t3
t4
N1, C1
(b) Intra-frame feature Aggregation
t1: N1,K,C2
t3: N1,K,C2
a b c
d e f
Reshape
Region shuffle
SFD
SFA
t1: N1,K,C1
t3: N1,K,C1
Grouped by K-nearest neighbors
(c) Motion Encoding
Max Pooling
Temporal Encoding
t1
t4
Max edge pooling
(d) Prediction Head
Output Label
Classification Head
Feature concate
Encode human structure with Graph Convolution
SFD Spatial Feature Descriptors
SFA Spatial Feature Aggregation
a ··· f Local Areas
⊗ Element-wise multiplication

Fig. 4. Framework: We first employ graph convolution techniques to encode the spatial structure of the human body in each frame. Considering that depth sequences contain a significant amount of redundant information when converted into point cloud sequences, we use the Farthest Point Sampling (FPS) method to downsample each frame of the point cloud, reducing the number of points by half, denoted as N_1. Subsequently, these features are input into the Spatial Feature Descriptor (SFD) and Spatial Feature Aggregation (SFA) modules for further processing. Our aggregation process can be performed multiple times. To enhance the diversity of human spatial features, we perform a shuffling operation on the human regions before re-aggregating. Finally, the processed human spatial features are input into Motion Encoding, achieving joint learning of temporal and spatial features, thereby effectively encoding the spatiotemporal characteristics of the human body.

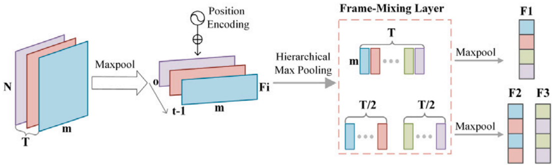


Fig. 5. Temporal encoding.

evaluation protocols: Cross-Subject and Cross-Setup, which are used to test the generalization and robustness of models across different participants and camera settings. The dataset's significant variations in participants and camera viewpoints make it a key benchmark for assessing the performance of 3D action recognition models.

NTU RGB+D 60 [35] is the initial version of the NTU RGB+D 120 dataset, containing 56,880 RGB-D samples across 60 action categories. The dataset was performed by 40 participants and captured using Microsoft Kinect v2 from 80 different viewpoints. It provides a variety of multimodal data, including RGB videos, depth maps, 3D skeleton data, and infrared videos. NTU RGB+D 60 employs two evaluation standards: Cross-View and Cross-Subject. Although its scale is relatively small, its diversity and complexity make NTU RGB+D 60 an important benchmark for evaluating 3D action recognition models.

MSR Action 3D [36] is a classic 3D action recognition dataset, containing 567 samples across 20 action categories, captured using Microsoft Kinect v1. The dataset focuses on motion-intensive actions such as waving, kicking, jumping, and throwing, primarily involving full-body movements. It provides depth images and skeletal sequences. Although relatively small in scale, MSR Action 3D has been widely used as a standard benchmark in the early development of 3D action recognition methods. We utilized all 20 action categories and evenly divided the data by participants, with half used for training and the other half for testing.

UTD-MHAD [37] is a multimodal dataset for 3D action recognition. It was captured using Microsoft Kinect sensors and wearable inertial sensors in an indoor environment, recording 27 actions performed by 8 participants, resulting in a total of 861 action sequences. The dataset provides RGB videos, depth images, skeleton data, and inertial sensor signals. We have fully utilized all 27 actions in the dataset, and the participants have been evenly divided into training and testing groups.

SHREC 2017 [38] is a dataset specifically designed for 3D gesture recognition, containing sequences of 14 unique gestures that can be performed by individual fingers or the entire palm. The dataset includes 28 participants, each performing each gesture 1 to 10 times, resulting in a total of 2800 sequences. We use the official split, dividing the dataset into training and testing sets in a 7:3 ratio, with 1960 training sequences (70%) and 840 testing sequences (30%).

4.2. Experimental setup

Implementation details. In this section, we provide a detailed description of the implementation details of PRG-Net and conduct experiments to validate its effectiveness in action classification tasks. These experiments are performed on the following datasets: NTU RGB+D 60, NTU RGB+D 120, MSR Action 3D, UTD-MHAD, and SHREC'17. PRG-Net first utilizes Graph Convolution [18] to extract body features within a single frame. Then, these features are further expanded through the Spatial Feature Descriptor (SFD) and Spatial Feature Aggregation (SFA) modules. In the SFD module, we employ the k-nearest neighbors (KNN) algorithm to select neighboring nodes, with K set to 20. Next, the Multi-local MaxPooling strategy is applied to generate 128 spatial descriptors. These descriptors are evenly distributed into 64 groups, with each group applying a sigmoid activation function. In the SFA module, we construct 128 groups of local regions and divide these regions along the channel dimension into 64 groups to match the number of descriptor groups. All network parameters of PRG-Net are jointly optimized in an end-to-end manner by minimizing the conventional cross-entropy loss function using the Adam optimizer. During the training process, the initial learning rate is set to 0.001 and decays at a rate of 0.5 every 10 epochs. The network is trained with a batch size of 32 for a total of 150 epochs.
Pre-processing. Due to the large number of points extracted from depth videos and the presence of significant amounts of similar depth information, we follow the approach proposed in [11] to uniformly sample 24 frames along the temporal axis. For each sampled depth image frame, we utilize the publicly available code provided by 3DV-PointNet++ [39] to extract the body outline. Subsequently, the outline is converted into a 3D point cloud, from which 512 points are uniformly sampled. During the training process, we employ the same data augmentation techniques as 3DV-PointNet++, including random rotations around the Y and X axes, data jittering, and random point dropout. These strategies help improve the model's robustness to variations in the input data and enhance its generalization capability.

4.3. Action recognition

Our primary focus is action recognition, which is a fundamental and crucial task in the field of human–computer interaction. We

Table 1
Performance comparison (%) on the MSR Action 3D dataset.

Method/Year	Input	Frames	Accuracy
Klaser et al. (2008) [40]	Depth	18	81.43
Vieira et al. (2012) [41]	Depth	20	78.20
Actionlet(2012) [42]	Skeleton	All	88.21
GFT(2019) [43]	Skeleton	All	74.00
HDM-BG(2019) [44]	Skeleton	All	86.10
DAM(2023) [45]	Skeleton	All	94.07
P4Transformer(2021) [9]	Point	24	90.94
PSTNet(2021) [8]	Point	24	91.20
PSTNet++(2021) [46]	Point	24	92.68
Kinet(2022) [47]	Point	24	93.27
HyperpointNet(2022) [10]	Point	24	91.54
SequentialPointNet(2022) [11]	Point	24	92.64
PST-Transformer(2022) [48]	Point	24	93.73
PSTNet + PointCMP(2023) [25]	Point	24	92.93
CPR(2023) [26]	Point	24	93.03
3DInAction(2024) [49]	Point	24	92.23
PRG-Net(ours)	Point	24	**95.97**

Bold values indicate the best result, while underlined values indicate the second best result.

Table 2
Performance comparison (%) on the UTD-MHAD.

Method/Year	Input	Accuracy
HP-DMM-CNN(2018) [50]	Depth	82.75
Trelinski et al. (2021) [51]	Depth	88.14
DRDIS(2021) [52]	Depth	87.88
LSTM-CNN(2020) [53]	Skeleton	93.20
Gimme signals(2020) [54]	Skeleton	93.33
HDM(2019) [44]	RGB+Skeleton	92.8
Fusion-GCN(2022) [55]	RGB+Skeleton+Inertial	94.42
Multi-stream CNNs(2023) [56]	RGB+Depth+Skeleton+Inertial	**96.95**
SequentialPointNet(2022) [11]	Point	92.31
PointMapNet(2023) [57]	Point	91.61
PRG-Net(ours)	Point	93.95

Bold values indicate the best result, while underlined values indicate the second best result.

first evaluate the effectiveness of our approach on five human action datasets.

4.3.1. Comparison with the state-of-the-art methods

We compared PRG-Net with depth-based, skeleton-based, and point cloud-based methods, and the experimental results are shown in Tables 1–3.

(1) MSR Action 3D dataset. The experimental results of PRG-Net on the MSR Action3D dataset are shown in Table 1. The data indicates that PRG-Net achieved a classification accuracy of 95.97%, which is the best result. Among point-based methods, PRG-Net's accuracy surpassed the second-best PST-Transformer model by 2.24%; it also outperformed the latest 3DInAction-2024 model by 2.94%. The performance advantage of PRG-Net is attributed to our innovative design in extracting spatial features of the human body and capturing spatiotemporal information. The SFA module effectively expands the spatial structure between human body regions through dynamic aggregation techniques, while the SFD module further optimizes the aggregation process of human body regions in the SFA module via attention-weighting, thereby enhancing the overall representation of human body features. Furthermore, the proposed inter-frame motion encoding module avoids constructing spatiotemporal local encodings across frames, effectively preventing the loss of spatiotemporal information of the human body. In contrast, PST-Transformer and CPR utilize Transformer architectures to process spatiotemporal information. However, on a small-scale dataset like MSR Action3D, these models fail to fully realize their potential, resulting in a significant gap when compared to PRG-Net.

(2) UTD-MHAD dataset. To comprehensively evaluate the performance of our network, we conducted comparative experiments on another small dataset, UTD-MHAD. The results, as shown in Table 2, indicate that our method achieved an accuracy of 93.95%, making it the best-performing method among point cloud approaches. Among them, SequentialPointNet, as the first point cloud sequence model applied to the UTD-MHAD dataset, provides an important benchmark for our method, with our model outperforming it by 1.64%. However, compared to the Multi-stream CNNs that utilize multi-modal data, PRG-Net still has room for improvement. Multi-stream CNNs achieved an accuracy of 96.95% by fusing RGB, depth, skeleton, and inertial data, highlighting the potential and value of multi-modal data in action recognition. In our future research, we plan to explore fusion strategies for multi-modal data in order to further enhance the accuracy and robustness of PRG-Net.

(3) NTU RGB+D 60 dataset. To further validate the generalization ability of our method on large-scale datasets, we conducted additional experiments on the NTU 60 dataset and compared our approach with other state-of-the-art methods. As shown in Table 3, PRG-Net achieved accuracies of 91.0% and 97.7% on the Cross-subject and Cross-view metrics, respectively. Although PRG-Net still lags behind some skeleton-based methods in the Cross-subject metric, it has approached their level in the Cross-view metric. Skeleton sequences, by tracking joint movements, can capture finer human motion patterns, thereby better distinguishing subtle differences among individuals performing the same action. These subtle patterns are difficult for point cloud sequences to capture. Compared with point cloud-based methods, PRG-Net demonstrates significant competitiveness on both metrics. Specifically, on the Cross-view metric, PRG-Net is only 0.4% behind the best PSTNet++; on the Cross-subject metric, our method achieves the highest accuracy. In contrast, point cloud data, due to its unordered nature and dynamic changes, typically requires more complex algorithms to address the associated challenges. For example, models such as PST-Transformer, PSTNet++ and CPR leverage the Transformer architecture to efficiently process spatiotemporal information, achieving excellent results in action recognition tasks. However, PRG-Net does not rely on complex spatiotemporal local encoding but instead effectively extracts spatiotemporal features of the human body by expanding human region relationships and inter-frame motion encoding. This simple yet efficient design enables PRG-Net to achieve competitive performance in fierce competition, further validating the effectiveness and reliability of our method.

(4) NTU RGB+D 120 dataset. We conducted a comprehensive comparison of PRG-Net with other advanced methods on the NTU 120 dataset. As shown in Table 3, PRG-Net achieved accuracy rates of 85.4% and 95.6% in the Cross-view and Cross-setup settings, respectively. The performance in the Cross-setup setting is particularly notable, approaching the state-of-the-art level. However, in the Cross-subject setting, PRG-Net still lags behind existing methods. This may be due to the temporal differences in how different individuals perform the same actions. Additionally, while employing a frame-wise parallel spatial structure encoding method improved spatial feature modeling, it also resulted in some loss of temporal information, thereby affecting the precision of the results.

4.4. Model analysis

4.4.1. Ablation experiment of region feature aggregation strategy

In this section, we conducted ablation experiments. First, we employed dynamic graph convolution to encode the spatial structure of each frame of the point cloud. Building on this, to capture the dynamic changes of the human body in the temporal dimension, we combined it with the motion encoding module, forming our initial baseline model, named DGCNN-ME. Based on this model, we further analyzed the contributions of the Spatial Feature Aggregation (SFA) module and the Spatial Feature Descriptor (SFD) module to the overall performance.

Table 3
Performance comparison (%) on the NTU RGB+D 60 and NTU RGB+D 120 dataset, comparing the two evaluation metrics.

Method/Year	Input	NTU60		NTU120	
		Cross-subject	Cross-view	Cross-subject	Cross-setup
Li et al. (2018) [58]	Depth	68.1	83.4	–	–
Wang et al. (2018) [59]	Depth	87.1	84.2	–	–
MVDI(2019) [60]	Depth	84.6	87.3	–	–
MST-GCN(2021) [61]	Skeleton	91.5	96.6	87.5	88.8
Skeletal-GNN(2021) [62]	Skeleton	91.6	96.7	87.5	89.2
Ta-CNN(2022) [63]	Skeleton	90.4	94.8	85.4	86.8
EfficientGCN-B4(2022) [64]	Skeleton	91.7	95.7	88.3	89.1
HD-GCN(2023) [65]	Skeleton	**93.4**	**97.2**	90.1	**91.6**
HetGCN(2023) [66]	Skeleton	90.5	95.3	84.3	85.7
SelfGCN(2024) [67]	Skeleton	93.1	96.6	89.4	91.0
BlockGCN(2024) [68]	Skeleton	93.1	97.0	**90.3**	91.5
3DV-PointNet++(2020) [39]	Point Cloud	88.8	96.3	82.4	93.5
PSTNet(2021) [8]	Point Cloud	90.5	96.5	87.0	93.8
PSTNet++(2021) [46]	Point Cloud	**91.4**	96.7	**88.6**	93.8
P4Transformer(2021) [9]	Point Cloud	90.2	96.4	86.4	93.5
HyperPointNet(2022) [10]	Point Cloud	90.2	97.3	83.2	95.1
SequentialPointNet(2022) [11]	Point Cloud	90.3	97.6	83.5	95.4
PST-Transformer(2022) [48]	Point Cloud	91.0	96.4	87.5	94.0
PointMapNet(2023) [57]	Point Cloud	89.4	96.7	–	–
CPR(2023) [26]	Point Cloud	91.0	96.7	–	–
PRG-Net (ours)	Point Cloud	91.0	**97.7**	85.4	**95.6**

The bold numbers indicate the highest accuracy of the skeleton-based and point cloud methods, while the second highest accuracy of the point cloud method is marked with an underline

Table 4
Performance comparison (%) on the NTU RGB+D 60 dataset, comparing the evaluation metrics CS and CV, i represents the number of executions.

Method	Frame	Cross-view	Cross-subject
DGCNN-ME	20	97.10	90.04
DGCNN-ME with SFA(i)	20	97.34	90.43
DGCNN-ME with SFA(i)+SFD	20	97.72	90.74
DGCNN-ME with SFA(ii)+SFD	20	**97.73**	**91.03**

Bold values highlight the optimal results.

Table 5
Ablation experiment results of SFA and SFD modules on MSR action 3D dataset.

Method	Frame	Accuracy
DGCNN-ME	20	95.60
DGCNN-ME with SFA(i)	20	95.60
DGCNN-ME with SFA(i)+SFD	20	**96.34**
DGCNN-ME with SFA(ii)+SFD	20	95.97

Bold values highlight the optimal results.

Table 6
Ablation experiment results of SFA and SFD modules on UTD-MHAD dataset.

Method	Frame	Accuracy
DGCNN-ME	20	93.49
DGCNN-ME with SFA(i)	20	93.95
DGCNN-ME with SFA(i)+SFD	20	**94.19**
DGCNN-ME with SFA(ii)+SFD	20	93.95

Bold values highlight the optimal results.

Here, SFA(i) denotes the initial aggregation, whereas SFA(ii) signifies the secondary aggregation after region shuffle.

Tables 4–6 show the results of the ablation experiments conducted on the NTU RGB+D 60, MSR Action 3D, and UTD-MHAD datasets. The experimental results demonstrate that the inclusion of the SFA module improves the accuracy for cross-view and cross-subject scenarios on the NTU RGB+D 60 dataset, with an increase from 97.10% to 97.34% and from 90.04% to 90.43%, respectively. This validates the SFA module's ability to provide valuable supplementary information for 3D action understanding tasks. Furthermore, the integration of the SFD module leads to additional performance improvement, yielding accuracies of 97.72% and 90.74%. Additionally, by performing region shuffling and secondary aggregation, the accuracy further improves to 97.73% and 91.03%, demonstrating the complementary advantages of the SFA and SFD modules in optimizing spatial feature representation. For smaller-scale datasets such as MSR Action 3D and UTD-MHAD, experimental results indicate that both the SFA and SFD modules contribute to performance improvements. However, on these datasets, the secondary aggregation after region shuffling did not bring the expected performance boost. We speculate that this might be due to the small scale of the datasets, resulting in limited diversity that the model learns from the secondary aggregation. Nonetheless, the combination of the SFA and SFD modules still demonstrates positive effects on these datasets, further confirming their applicability and effectiveness across different data scales .

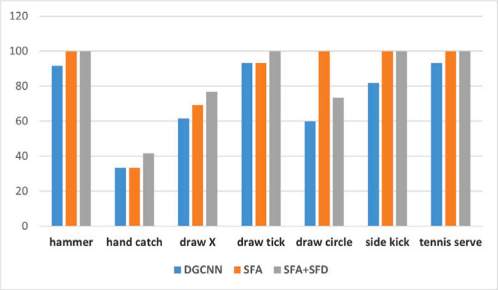


Fig. 6. Impact of SFA and SFD modules on different actions.

4.4.2. Impact of region feature aggregation strategy on different actions

As show in Fig. 6, we further analyzed the impact of our Spatial Feature Aggregation module and Spatial Feature Descriptors on different actions in the MSR Action 3D dataset. The results indicate that for subtle hand-related actions such as "draw circle", "draw X", "draw tick", "hammer", and "hand catch"(see Fig. 7), the SFA and SFD modules significantly improve the accuracy of these actions. Similarly, for large-scale body movements such as "side kick" and "tennis serve", our SFA and SFD modules also contribute to improved recognition rates for these actions. This suggests that our modules play a positive role in different types of actions.

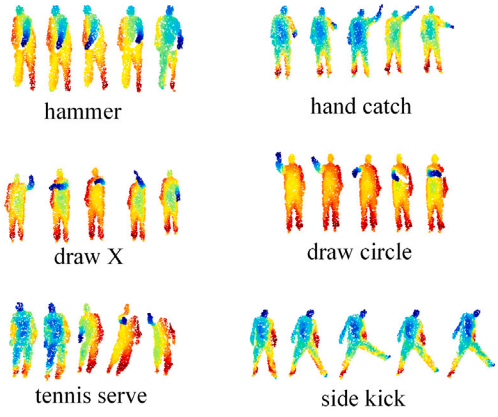


Fig. 7. Visualization of partial actions in MSR Action 3D: To achieve better visual representation, we utilized open3d to assign colors to each point. It is important to note that there is no direct relationship between the assigned colors and the features of the dataset. The original input point cloud only contains positional attributes.

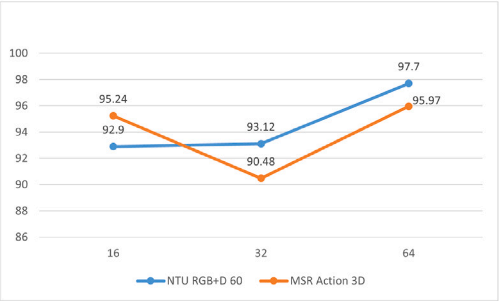


Fig. 8. Accuracy under Different Groupings.

4.4.3. *Impact of different groupings on aggregation effect*

In our study, the SFA and SFD modules enhance the feature diversity of human spatial structures through cross-channel region aggregation. To determine the optimal number of channel groups g, we conducted comparative experiments on the NTU RGB+D 60 and MSR Action 3D datasets, setting the number of groups to 1/2, 1/4, and 1/8 of the input channel number. Considering that our model's input feature dimension is 128, we chose configurations of 64, 32, and 16 as the number of groups. The experimental results, as shown in Fig. 8, indicate that the model achieves the best performance when the number of groups is set to 1/2 of the input channel number for both datasets. By increasing the number of groups, we facilitate cross-channel interactions among more local regions and introduce greater diversity in human features. This interaction helps the model better capture the details and variations of the human spatial structure.

4.4.4. *The effectiveness of region shuffle*

To verify the effectiveness of the region shuffle operation, we conducted ablation experiments on the MSR Action 3D and NTU RGB+D 60 datasets and compared the results with those of the second aggregation method without region shuffle. In the NTU RGB+D 60 dataset, the sequence length was set to 8 frames, while in the MSR Action 3D dataset, the input length was set to 24 frames. The experimental results over the first 50 training epochs were compared, as detailed below.

The Table 7 shows the comparison of the accuracy of the PRG-Net method with and without the region shuffle operation on the MSR Action 3D and NTU RGB+D 60 datasets. The experimental results indicate that the region shuffle operation can enhance the performance of the model. On the MSR Action 3D dataset, the accuracy increased from 94.51% to 95.97%, a gain of 1.46%. Similarly, on the NTU RGB+D 60 dataset, the accuracy improved from 73.83% to 74.42%, an increase of 0.59%. The region shuffle operation, by rearranging local regions of the human body, effectively increases feature diversity, thereby further improving the model's performance in human action recognition tasks.

Table 7
PRG-Net with/without region shuffle on MSR Action 3D and NTU RGB+D 60 Dataset.

Dataset	Method	Accuracy
MSR action 3D	PRG-without shuffle PRG-shuffle	94.51 **95.97**
NTU RGB+D 60	PRG-without shuffle PRG-Shuffle	73.83 **74.42**

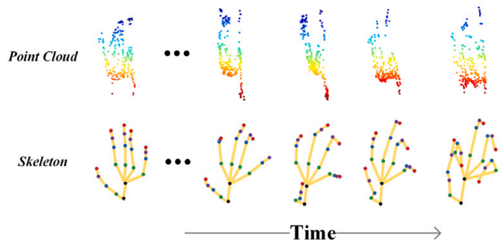


Fig. 9. Visualization of Gesture Recognition: The first row shows point cloud sequences, and the second row shows skeleton sequences.

Table 8
Performance comparison (%) on the SHREC 2017 dataset.

Method/Year	Input	Actions	Accuracy
Key frames(2017) [38]	Depth	14	82.90
SoCJ+HoHD+HoWR(2016) [69]	Skeleton	14	88.20
Res-TCN(2018) [70]	Skeleton	14	91.10
STA-Res-TCN(2018) [70]	Skeleton	14	93.60
ST-GCN(2018) [5]	Skeleton	14	92.70
DG-STA(2019) [71]	Skeleton	14	94.40
TCN-Summ(2021) [72]	Skeleton	14	93.57
MS-ISTGCN(2022) [73]	Skeleton	14	96.70
TD-GCN(2023) [74]	Skeleton	14	**97.02**
PointLSTM-Baseline(2019) [23]	Point	14	87.60
PointLSTM(2019) [23]	Point	14	95.90
FPPR-PCD(2022) [75]	Point	14	**96.10**
Kinet(2022) [47]	Point	14	95.20
CPR(2023) [26]	Point	14	93.10
SerialSTTR(2023) [76]	Point	14	93.80
PointDMIG(2024) [77]	Point	14	95.12
PRG-Net(ours)	Point	14	95.36

The bold numbers indicate the highest accuracy of the skeleton-based and point cloud methods

4.5. *Gesture recognition*

Unlike action recognition, gestures are defined by humans, have specific communication purposes, and typically exhibit smaller intra-class variations. Gesture recognition focuses on analyzing hand shapes, poses, and movements, and point cloud data is often preferred as a representation method due to its simplicity and strong spatial positional information. To comprehensively evaluate the generalization ability of our model, we conducted comparative experiments on the SHREC 2017 dataset. We followed the same experimental setup as previous work [23], employing the Feature Point Sampling (FPS) algorithm to sample 256 points from each frame as the data volume for a single-frame point cloud. In the experiment, we selected 1960 action sequences for the training phase of the model, while the remaining data was used for testing (see Fig. 9).

According to the experimental results shown in Table 8, although our method did not achieve the highest level compared to the latest

Table 9
Comparison of parameter, Flops and running time on 3D action recognition accuracy (%). The MSR-Action3D dataset is used. Clip length is 16.

Method	#Parameters (M)	FLOPs (G)	Time (ms)	Accuracy (%)
MeteorNet	17.60	–	109.82	88.21
P4Transformer	44.1	–	187.56	89.56
PSTNet	20.46	–	206.74	89.90
SequentialPointNet	3.72	**2.84**	**9.72**	89.90
PRG-Net(ours)	**3.03**	8.53	27.81	**94.14**

Bold values indicate the best result.

skeleton-based approaches, it demonstrated competitive results when compared to point cloud methods. TD-GCN establishes spatiotemporal dependencies by constructing temporal adjacency matrices, allowing it to capture more nuanced human motion patterns. Since gestures are subtle actions with ambiguous boundaries and significant contextual dependencies, point cloud data often requires more complex methods to overcome the challenges posed by the disorder and dynamic variations of subtle movements. Therefore, further research is needed to explore more effective point cloud processing methods to address these issues. Overall, our results on the SHREC 2017 dataset validate the versatility of our method in various visual sequence learning tasks.

4.6. Memory usage and computational efficiency

In this section, we evaluate the computational efficiency and memory usage of our method, including runtime, number of parameters, and FLOPs. The experiments were conducted on a machine equipped with an Nvidia Quadro RTX 6000 GPU, using the MSR-Action 3D dataset, with comparisons made for sequences of 16 frames in length. The runtime refers to the network's forward inference time for each point cloud sequence. Additionally, we compared our method with several classical approaches. It is important to note that the FLOPs data for MeteorNet, P4Transformer, and PSTNet have not been provided by their authors.

As shown in Table 9, our method achieves a good balance between efficiency and performance. PSTNet and P4Transformer, due to their use of local spatiotemporal encoding, have a larger number of parameters, which in turn increases computation time. In contrast, our designed inter-frame motion encoding effectively captures human movement without relying on the construction of local spatiotemporal regions, significantly reducing computation time. Although our method employs region aggregation techniques to encode the spatial structure of the human body, resulting in slightly higher running time compared to the lightweight network SequentialPointNet, it demonstrates strong competitiveness in terms of accuracy.

5. Conclusion

In this paper, we introduce PRG-Net, an architecture for action recognition based on point cloud sequences. Our method focuses on enhancing the ability of 3D human action recognition by learning the interrelationships among body regions. We propose the Spatial Feature Aggregation (SFA) module and the Spatial Feature Descriptors (SFD) module, which employ dynamic region aggregation to effectively model the spatial structure of human actions. For temporal modeling, we avoid constructing complex local spatio-temporal structures and instead aggregate feature vectors across frames efficiently using a cross-frame max pooling strategy. We conduct extensive experiments on 3D human action recognition datasets, and the results validate the effectiveness of our approach.

CRediT authorship contribution statement

Yao Du: Writing – original draft, Validation, Methodology. **Zhenjie Hou:** Writing – review & editing, Resources. **En Lin:** Resources. **Xing Li:** Validation, Formal analysis. **Jiuzhen Liang:** Investigation. **Xinwen Zhou:** Software, Resources.

Declaration of competing interest

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

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61063021).

Data availability

Data will be made available on request.

References

[1] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[2] M. Majd, R. Safabakhsh, Correlational convolutional LSTM for human action recognition, Neurocomputing 396 (2020) 224–229.
[3] L. Shao, R. Gao, Y. Liu, H. Zhang, Transform based spatio-temporal descriptors for human action recognition, Neurocomputing 74 (6) (2011) 962–973.
[4] C. Li, Q. Zhong, D. Xie, S. Pu, Skeleton-based action recognition with convolutional neural networks, in: 2017 IEEE International Conference on Multimedia & Expo Workshops, ICMEW, IEEE, 2017, pp. 597–600.
[5] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[6] S. Cho, M. Maqbool, F. Liu, H. Foroosh, Self-attention network for skeleton-based human action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 635–644.
[7] X. Liu, M. Yan, J. Bohg, Meteornet: Deep learning on dynamic 3d point cloud sequences, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9246–9255.
[8] H. Fan, X. Yu, Y. Ding, Y. Yang, M. Kankanhalli, Pstnet: Point spatio-temporal convolution on point cloud sequences, 2022, arXiv preprint arXiv:2205.13713.
[9] H. Fan, Y. Yang, M. Kankanhalli, Point 4d transformer networks for spatio-temporal modeling in point cloud videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14204–14213.
[10] X. Li, Q. Huang, T. Yang, Q. Wu, Hyperpointnet for point cloud sequence-based 3D human action recognition, in: 2022 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2022, pp. 1–6.
[11] X. Li, Q. Huang, Z. Wang, Z. Hou, T. Yang, SequentialPointNet: A strong frame-level parallel point cloud sequence network for 3D action recognition, 2021, arXiv preprint arXiv:2111.08492.
[12] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 945–953.
[13] A. Hamdi, S. Giancola, B. Ghanem, Mvtn: Multi-view transformation network for 3d shape recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1–11.
[14] D. Maturana, S. Scherer, Voxnet: A 3d convolutional neural network for real-time object recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2015, pp. 922–928.
[15] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D shapenets: A deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.
[16] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.
[17] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, Adv. Neural Inf. Process. Syst. 30 (2017).
[18] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, J.M. Solomon, Dynamic graph cnn for learning on point clouds, ACM Trans. Graph. (Tog) 38 (5) (2019) 1–12.
[19] J. Yang, Q. Zhang, B. Ni, L. Li, J. Liu, M. Zhou, Q. Tian, Modeling point clouds with self-attention and gumbel subset sampling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3323–3332.
[20] L. Lin, P. Huang, C.-W. Fu, K. Xu, H. Zhang, H. Huang, On learning the right attention point for feature enhancement, Sci. China Inf. Sci. 66 (1) (2023) 112107.

[21] D. Lu, K. Gao, Q. Xie, L. Xu, J. Li, 3DGTN: 3D dual-attention glocal transformer network for point cloud classification and segmentation, IEEE Trans. Geosci. Remote Sens. (2024).

[22] H. Fan, Y. Yang, PointRNN: Point recurrent neural network for moving point cloud processing, 2019, arXiv preprint arXiv:1910.08287.

[23] Y. Min, Y. Zhang, X. Chai, X. Chen, An efficient pointlstm for point clouds based gesture recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5761–5770.

[24] X. Chen, W. Liu, X. Liu, Y. Zhang, J. Han, T. Mei, MAPLE: Masked pseudo-labeling autoencoder for semi-supervised point cloud action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 708–718.

[25] Z. Shen, X. Sheng, L. Wang, Y. Guo, Q. Liu, X. Zhou, Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1212–1222.

[26] X. Sheng, Z. Shen, G. Xiao, Contrastive predictive autoencoders for dynamic point cloud self-supervised learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 9802–9810.

[27] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[29] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.

[30] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Trans. Image Process. 29 (2020) 9532–9545.

[31] S. Kim, D. Ahn, B.C. Ko, Cross-modal learning with 3D deformable attention for action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10265–10275.

[32] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, J. Malik, On the benefits of 3d pose and tracking for human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 640–649.

[33] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[34] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.

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

[36] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 9–14.

[37] C. Chen, R. Jafari, N. Kehtarnavaz, UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: 2015 IEEE International Conference on Image Processing, ICIP, IEEE, 2015, pp. 168–172.

[38] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B.L. Saux, D. Filliat, 3D hand gesture recognition using a depth and skeletal dataset: Shrec'17 track, in: Proceedings of the Workshop on 3D Object Retrieval, 2017, pp. 33–38.

[39] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J.T. Zhou, J. Yuan, 3DV: 3D dynamic voxel for action recognition in depth video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 511–520.

[40] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, 2008, 275–1.

[41] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, M.F. Campos, Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17, Springer, 2012, pp. 252–259.

[42] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.

[43] J.-Y. Kao, A. Ortega, D. Tian, H. Mansour, A. Vetro, Graph based skeleton modeling for human activity analysis, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 2025–2029.

[44] R. Zhao, W. Xu, H. Su, Q. Ji, Bayesian hierarchical dynamic model for human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7733–7742.

[45] F. Ronchetti, F. Quiroga, L. Lanzarini, C. Estrebou, Distribution of action movements (DAM): a descriptor for human action recognition, Front. Comput. Sci. 9 (2015) 956–965.

[46] H. Fan, X. Yu, Y. Yang, M. Kankanhalli, Deep hierarchical representation of point cloud videos via spatio-temporal decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 9918–9930.

[47] J.-X. Zhong, K. Zhou, Q. Hu, B. Wang, N. Trigoni, A. Markham, No pain, big gain: classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8510–8520.

[48] H. Fan, Y. Yang, M. Kankanhalli, Point spatio-temporal transformer networks for point cloud video modeling, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2181–2192.

[49] Y. Ben-Shabat, O. Shrout, S. Gould, 3Dinaction: Understanding human actions in 3d point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19978–19987.

[50] N.E.D. Elmadany, Y. He, L. Guan, Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis, IEEE Trans. Image Process. 27 (11) (2018) 5275–5287.

[51] J. Trelinski, B. Kwolek, CNN-based and DTW features for human activity recognition on depth maps, Neural Comput. Appl. 33 (21) (2021) 14551–14563.

[52] H. Wu, X. Ma, Y. Li, Spatiotemporal multimodal learning with 3D CNNs for video action recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1250–1261.

[53] A. Zhu, Q. Wu, R. Cui, T. Wang, W. Hang, G. Hua, H. Snoussi, Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN, Neurocomputing 414 (2020) 90–100.

[54] R. Memmesheimer, N. Theisen, D. Paulus, Gimme signals: Discriminative signal encoding for multimodal activity recognition, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2020, pp. 10394–10401.

[55] M. Duhme, R. Memmesheimer, D. Paulus, Fusion-gcn: Multimodal action recognition using graph convolutional networks, in: DAGM German conference on pattern recognition, Springer, 2021, pp. 265–281.

[56] F. Khezerlou, A. Baradarani, M.A. Balafar, R.G. Maev, Multi-stream CNNs with orientation-magnitude response maps and weighted inception module for human action recognition, in: 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), IEEE, 2023, pp. 1–5.

[57] X. Li, Q. Huang, Y. Zhang, T. Yang, Z. Wang, Pointmapnet: Point cloud feature map network for 3d human action recognition, Symmetry 15 (2) (2023) 363.

[58] J. Li, Y. Wong, Q. Zhao, M.S. Kankanhalli, Unsupervised learning of view-invariant action representations, Adv. Neural Inf. Process. Syst. 31 (2018).

[59] P. Wang, W. Li, Z. Gao, C. Tang, P.O. Ogunbona, Depth pooling based large-scale 3-d action recognition with convolutional neural networks, IEEE Trans. Multimed. 20 (5) (2018) 1051–1061.

[60] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J.T. Zhou, X. Bai, Action recognition for depth video using multi-view dynamic images, Inform. Sci. 480 (2019) 287–304.

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

[62] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, Q. Xu, Learning skeletal graph neural networks for hard 3d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11436–11445.

[63] K. Xu, F. Ye, Q. Zhong, D. Xie, Topology-aware convolutional neural network for efficient skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2866–2874.

[64] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 1474–1488.

[65] J. Lee, M. Lee, D. Lee, S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10444–10453.

[66] X. Gao, Y. Yang, Y. Wu, S. Du, Learning heterogeneous spatial–temporal context for skeleton-based action recognition, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[67] Z. Wu, P. Sun, X. Chen, K. Tang, T. Xu, L. Zou, X. Wang, M. Tan, F. Cheng, T. Weise, Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition, IEEE Trans. Image Process. (2024).

[68] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, X.-S. Hua, Blockgcn: Redefine topology awareness for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2049–2058.

[69] Q. De Smedt, H. Wannous, J.-P. Vandeborre, Skeleton-based dynamic hand gesture recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1–9.

[70] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, H. Yang, Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.

[71] Y. Chen, L. Zhao, X. Peng, J. Yuan, D.N. Metaxas, Construct dynamic graphs for hand gesture recognition via spatial-temporal attention, 2019, arXiv preprint arXiv:1907.08871.

[72] A. Sabater, I. Alonso, L. Montesano, A.C. Murillo, Domain and view-point agnostic hand action recognition, IEEE Robot. Autom. Lett. 6 (4) (2021) 7823–7830.

[73] J.-H. Song, K. Kong, S.-J. Kang, Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 6227–6239.

[74] J. Liu, X. Wang, C. Wang, Y. Gao, M. Liu, Temporal decoupling graph convolutional network for skeleton-based gesture recognition, IEEE Trans. Multimed. (2023).

[75] A. Bigalke, M.P. Heinrich, Fusing posture and position representations for point cloud-based hand gesture recognition, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 617–626.

[76] S. Zou, J. Zhang, Serial spatial and temporal transformer for point cloud sequences recognition, in: Computer Graphics International Conference, Springer, 2023, pp. 16–27.

[77] Y. Du, Z. Hou, X. Li, J. Liang, K. You, X. Zhou, PointDMIG: a dynamic motion-informed graph neural network for 3D action recognition, Multimedia Syst. 30 (4) (2024) 192.

Yao Du is a current Master's degree student in the Computer Science and Technology program at Changzhou University. His research focus lies in the area of computer vision, with a specific emphasis on human action recognition.

Zhenjie Hou received the Ph.D. degree in mechanical engineering from Inner Mongolia Agricultural University, in 2005. From 1998 to 2010, he was a Professor with the Computer Science Department, Inner Mongolia Agricultural University. In August 2010, he joined Changzhou University. His research interests include signal and image processing, pattern recognition, and computer vision.

En Lin received the B.E. degree in computer science and technology from Zhejiang Gongshang University, Hangzhou, China, in 2018. He is currently engaged in big data analysis and development with Goldcard Smart Group Company Ltd., Hangzhou.

Xing Li received the B.Sc. and M.Sc. degrees in software engineering from Changzhou University, China, in 2016 and 2019, and the Ph.D. degree in Computer Science and Technology from Hohai University, Nanjing, China, in 2023. He is currently an Associate Professor at Nanjing Forestry University. His current research interests include machine learning, computer vision and deep learning, especially human action recognition.

Jiuzhen Liang was born in November 1968. He received the B.S. degree in mathematics from the Daqing Institute of Petroleum, in 1991, the M.S. degree in mathematics from the Harbin Institute Technology, in 1996, and the Ph.D. degree in computer science from Beihang University, in 2001. He is currently with the Department of Computer Science, Changzhou University. He has published 190 articles in journals and proceedings. His research interests include computer vision, image processing, and pattern recognition.

Xinwen Zhou received the Ph.D. degree in Control Science and Engineering from Shanghai Jiao Tong University, China, in 2020. In December 2020, he joined Changzhou University. His research interests include Image processing, artificial intelligence, 3D modeling.