Multilevel Position-aware Attention Enhanced Network for Skeleton-Based Action Recognition

Dandan Zhang 107552204117@stu.xju.edu.cn
Sicong Zhan zhansicong@stu.xju.edu.cn
School of Computer Science and Technology, Xinjiang University, Urumqi, China

Jia Wang * jw1024@xju.edu.cn
School of Computer Science and Technology, Xinjiang University, Urumqi, China
Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Effectively capturing the spatiotemporal dependencies between joints is crucial for skeleton-based action recognition. However, existing methods do not consider the sparsity of skeleton data, which hinders the accurate capture of complex posture information and subtle action variations. Moreover, the locality of temporal features requires the model to focus on certain key features. Yet, most methods overlook the impact of temporal redundancy on feature focus, resulting in ineffective capture of significant temporal features. To address the issue of skeleton sparsity, we propose a Multilevel Position-aware Attention module (MPA) that explicitly leverages the relative positional information of the input data to enrich spatial information. To achieve a more effective focus on local temporal features, we develop a Multi-scale Temporal Excitation module (MTE). By scaling temporal features, the MTE module elevates the prominence of salient features and facilitates the capture of multi-scale features. Furthermore, we propose a Part Partition Encoding module (PPE) to aggregate joint data into part data, thereby providing the model with high-level information carried by the interactions between body parts. The MPA, MTE, and PPE are integrated into a unified framework called MPAE-Net. Extensive experimental results demonstrate that the MPAE-Net achieves state-of-the-art performance on two large-scale datasets, NTU RGB+D and NTU RGB+D 120.

Keywords: Skeleton-Based Action Recognition, Multilevel Position-aware Attention, Multi-scale Temporal Excitation, Part Partition Encoding

1. Introduction

Action recognition is a critical area in computer vision with extensive applications in various real-world scenarios, such as video surveillance Xin et al. (2023), autonomous driving Peng et al. (2024), and human-computer interaction Liu et al. (2022). Based on the types of data used, action recognition can be primarily classified into RGB image-based methods and skeleton-based methods. Early studies extensively utilized RGB image data to explore contextual action information in videos, which is easily affected by factors such as body scale variations, cluttered backgrounds, and changes in viewpoints. Recently, with the

*

development of depth sensors Zhang (2012) and human pose estimation algorithms Zheng et al. (2023), skeleton-based action recognition has developed rapidly. Compared to RGB image data, skeleton data consists only of a limited set of 2D/3D coordinates, which is lower-dimensional, more compact, and easier to process. Given these advantages, this paper focuses on skeleton-based action recognition in videos.

Skeleton-based action recognition aims to accurately classify different actions by analyzing the motion trajectories and posture changes of human skeleton. Recently, deep learning has been widely explored to model the skeleton sequence. Some studies adopt Recurrent Neural Networks (RNNs) Liu et al. (2016) and Convolutional Neural Networks (CNNs) Ke et al. (2018) to convert skeleton data into vector sequences or pseudo-images. However, they did not explicitly explore the topological information of the human skeleton, and therefore could only capture limited features. As the human skeleton can naturally be represented as graph structures, where nodes represent joints and edges represent the links between joints. Graph Convolutional Networks (GCNs) can directly model skeleton data as spatiotemporal graphs to explore the topological information. For this reason, GCN-based methods Yan et al. (2018); Shi et al. (2019); Cheng et al. (2021); Wen et al. (2022) have gained widespread attention. However, GCN-based methods rely on the human topological structure to capture joint features, which limits their ability to capture the relationships between distant joints. Meanwhile, Transformer-based methods Shi et al. (2020); Plizzari et al. (2021); Qiu et al. (2023); Wu et al. (2023) do not rely on the human topological structure but utilize the attention mechanism to construct global joint connections, thereby capturing their relationships. However, skeleton data is composed of only a limited number of coordinates in space, and this sparsity brings some new challenges for spatial modeling of skeleton data, such as the lack of rich contextual information and the weaker ability to resist interference. Moreover, considering the locality of temporal features, specific moments (such as the takeoff and landing moments in jumping actions) are crucial for accurately recognizing certain actions. However, the relatively slow changes in human actions over time result in redundant information in the temporal dimension, which makes it difficult for the model to effectively focus on key temporal features. Meanwhile, most studies Yan et al. (2018); Plizzari et al. (2021); Qiu et al. (2023) only utilize joint data, neglecting part data that provides overall action information. For example, a jumping action involves the coordinated movement of arms and legs rather than the precise positioning of the finger and toe joints.

Considering the importance of the aforementioned issues, we construct a novel Multi-level Position-aware Attention Enhanced Network (MPAE-Net) for skeleton-based action recognition, which consists of three components: Part Partition Encoding (PPE), Multi-level Position-aware Attention (MPA), and Multi-scale Temporal Excitation (MTE). PPE is used to generate part data. A simple method for adding part data is to segment the joints by regions and then to directly connect these joints from different regions Jia et al. (2024). In contrast, PPE extends the channel information of joints to enrich features and then aggregates the information within parts. This approach reduces noise and highlights the overall characteristics of the part data. Therefore, part data and joint data together construct multilevel input for the model. To overcome the impact of skeleton sparsity on the spatial modeling of skeleton sequences, we propose the MPA module. Specifically, MPA improves the attention mechanism in the vanilla Transformer Vaswani et al. (2017) by ex-

plicitly introducing the relative positional information of the input data. By calculating the joint projections on their relative positions, MPA captures the features at specific locations. This helps the model understand how joint features are influenced by positional information. Additionally, considering the different importance of data and positional information, MPA employs a gating mechanism to dynamically control their fusion. Moreover, to achieve more effective focus on important local temporal features, we propose the MTE module. Inspired by the squeeze-and-excitation (SENet) Hu et al. (2018), MTE calculates the excitation weights of temporal features and uses these weights to dynamically recalibrate skeleton features, thereby amplifying the significance of key features. Then, by employing convolutional operations within the MTE, we capture multi-scale temporal features. In summary, the main innovations and contributions of this work are as follows:

- We propose the PPE module to generate part data to aggregate information from multiple joints and provide overall action features. Part data and joint data together provide the model with both holistic and detailed features.
- We propose the MPA and MTE modules to capture spatiotemporal features. MPA explicitly incorporates positional information to overcome the impact of skeleton sparsity. MTE dynamically recalibrates temporal features and achieves effective focus on key features.
- Extensive experiments on two large datasets, NTU RGB+D 60 and NTU RGB+D 120, validate the effectiveness of each proposed module and demonstrate the competitiveness of the MPAE-Net.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed modules in the MPAE-Net. Section 4 presents a comprehensive analysis of the experimental results. Section 5 summarizes the paper.

2. Related Work

2.1. Skeleton-based action recognition

Skeleton-based action recognition mainly utilizes deep learning methods to extract features from skeleton data and then identify the actions. Deep learning methods can be classified into four main types based on the employed architectural frameworks. RNN-based methods Liu et al. (2016); Li et al. (2018b) primarily deal with the temporal relationships within skeleton data. Liu et al. (2016) introduced trust gates to selectively focus on the important joints in the skeleton sequence. Meanwhile, CNN-based methods Ke et al. (2018); Li et al. (2018a) primarily utilize local convolution operations to capture skeleton features. Li et al. (2018a) used joint positions and motion information to construct a co-occurrence matrix and extracted local features of this matrix. However, these earlier methods do not consider the physical links between different joints, resulting in their limited ability to capture certain features. Therefore, GCN-based methods Yan et al. (2018); Shi et al. (2019); Cheng et al. (2020); Ye et al. (2020); Liu et al. (2020); Chen et al. (2021); Cheng et al. (2021); Wen et al. (2022) model human skeleton as spatiotemporal graphs and use a pipeline of GCNs and temporal convolutional networks (TCNs) to capture joint features. Yan et al. (2018)

proposed ST-GCN, which initially used GCNs to explore the relationships of joints. Shi et al. (2019) addressed the static topology in the ST-GCN by proposing an adaptive graph convolution to dynamically adjust the topology of graphs. Cheng et al. (2020) proposed Shift-GCN, which introduced shift operations in convolution to capture the relationships between joints. Subsequently, Cheng et al. (2021) proposed ShiftGCN++, a lightweight network that combines shift graph operations with lightweight point-wise convolution for efficient skeleton action recognition. Although GCN-based methods excel in this task, they have a limited ability to capture distant joint connections. In contrast, Transformer-based methods Shi et al. (2020); Plizzari et al. (2021); Qiu et al. (2023); Wu et al. (2023) utilize the attention mechanism to construct global joint correlation. Shi et al. (2020) decoupled the spatial and temporal dimensions, thereby separately capturing joint features. Plizzari et al. (2021) combined GCNs and spatiotemporal attention to extract skeleton features. Qiu et al. (2023) proposed STSA-Net, which divides action sequences into multiple spatiotemporal segments and extracted both intra-segment and inter-segment features. The aforementioned Transformer-based methods effectively capture global joint dependencies, but they do not consider the impact of skeleton sparsity on skeleton sequence modeling. Simultaneously, due to the locality of temporal features, utilizing the attention mechanism in the temporal dimension may not be optimal; instead, it is necessary to focus on temporally local important features.

2.2. Self-attention mechanism

The self-attention mechanism captures long-range dependencies in sequences by computing weighted sums of different parts. Bahdanau et al. (2014) initially proposed the self-attention mechanism to dynamically focus on important parts of the input sequences during translation tasks. Vaswani et al. (2017) proposed the Transformer, which features a self-attention mechanism to establish global dependencies between inputs and outputs. Due to its superior performance, the self-attention mechanism has been widely adopted in various computer vision tasks, such as image classification Zhang et al. (2023); Roy et al. (2023), object detection Wang et al. (2023a); Gu et al. (2023), and action recognition Ahn et al. (2023); Xu et al. (2023). Zhang et al. (2023) combined multi-attention to address the issues of information redundancy and inaccurate feature extraction in hyper-spectral images. Wang et al. (2023a) used the attention mechanism to enhance the weighting of adjacency relationships in point cloud data for 3D object detection. Xu et al. (2023) improved channel attention by introducing Discrete Cosine Transform(DCT) and proposed a novel multi-frequency channel attention framework for action recognition. In the domain of skeleton-based action recognition, Plizzari et al. (2021) proposed ST-TR that combined GCNs and the attention mechanism to process spatial and temporal information. Wu et al. (2023) utilized the attention mechanism to capture the relationships between joints and body parts, therefore providing more interpretable representations for different skeleton action sequences. The aforementioned Transformer-based methods merely utilized the attention mechanism to extract spatiotemporal skeleton features but did not improve the mechanism itself. In contrast, we improved this by explicitly leveraging the positional information of the input data in MPA to overcome the impact of skeleton sparsity.

3. Method

The overall architecture of MPAE-Net is shown in Fig. 1. The entire processing steps are as follows. First, the raw skeleton data was input into the MPAE-Net, which employed the PPE module to generate part data. Part data and joint data provide overall and detailed action information, respectively. Then, the MPA module was employed to overcome the impact of skeleton sparsity by explicitly incorporating positional information, thereby capturing richer joint and part spatial features. After fusion, these features are fed into the MTE module, which effectively focuses on temporally local features. It calculates the importance weights of these features, thereby amplifying their salience and capturing multi-scale temporal features. Features obtained after spatiotemporal processing are then subjected to the global pooling and the fully connected layers for action prediction.

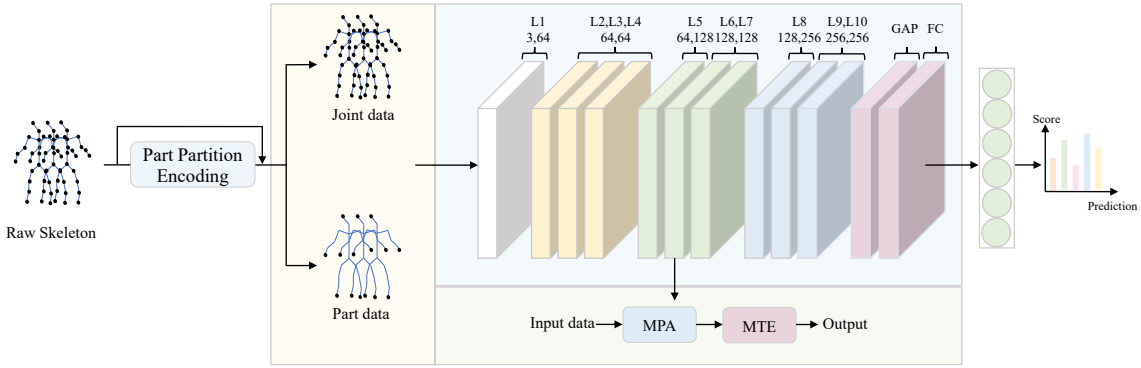

Figure 1: Overview of the MLPA-Net. MLPA-Net consists of the PPE module and 10 spatiotemporal feature extraction layers. Each of these layer includes the MPA and MTE modules.

3.1. Part Partition Encoding

To generate part-level data and provide with comprehensive information, we propose the PPE module, as illustrated on the left of Fig. 2. This module expands joint dimensions to obtain richer features and then aggregates them into corresponding body parts.

The PPE module initially divides joint data into limbs and torso segments based on the predefined patterns. The joint data $\boldsymbol{X}_J \in \mathbb{R}^{C\times T\times V}$, which contains T frames, each with V joints across C channel dimensions. The feature expansion layer $F_C(.)$ is utilized to expand the channel dimension of $\boldsymbol{X}_J$, thereby enhancing the expressive ability. $F_C(.)$ comprises two convolutional layers, each followed by the BatchNorm and the Leaky ReLU functions. Then, $\boldsymbol{X}_J$ is divided into M parts to obtain $\boldsymbol{x}_i \in \mathbb{R}^{C_1\times T\times V_i}, i = 1, 2, \ldots, M$.

We assume that there are at most V_1 joints in $\boldsymbol{x}_i$. For situations where V_i is less than V_1, the missing joints are filled with zeros. All joint features in $\boldsymbol{x}_i$ are average pooled to aggregate overall features, as shown in Eq. 1.

$$\hat{\boldsymbol{x}}_i = F_{avg}(\boldsymbol{x}_i) \tag{1}$$

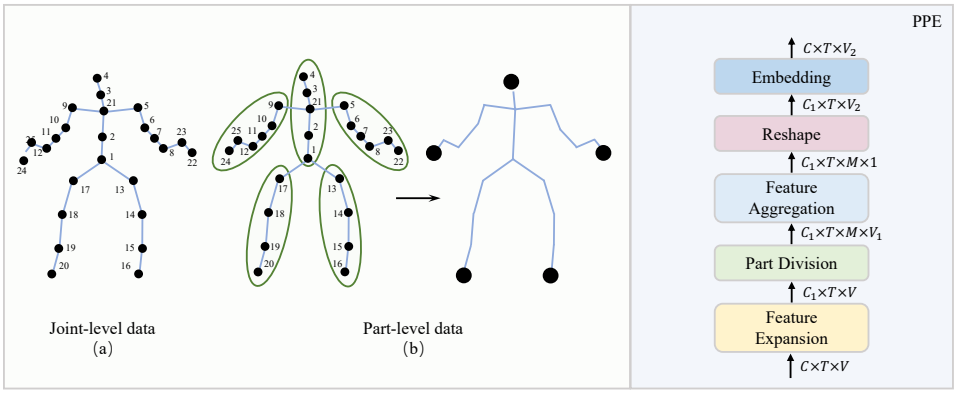

Figure 2: Illustration of the multi-level data and the PPE module. Joint data and part data are presented on the left, while the process of the PPE module is detailed on the right.

Subsequently, the aggregated $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{C_1 \times T \times 1}$ from all parts are concatenated, as shown in Eq. 2.

$$\hat{\boldsymbol{X}}_P = Concat(\hat{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, M) \tag{2}$$

Ultimately, the $\hat{\boldsymbol{X}}_P \in \mathbb{R}^{C_1 \times T \times V_2}$ is input to the embedding layer to restore the original channel dimension, resulting in the final obtained part data $\boldsymbol{X}_P \in \mathbb{R}^{C \times T \times V_2}$.

3.2. Multilevel Position-aware Attention Module

To overcome the impact of sparsity on spatial skeleton modeling, we design the MPA module to fully capture the spatial joint and part features, as illustrated on the left side of Fig. 3. For clarity, we describe the entire process using joint data as the processing steps for joints and parts in the MPA module are identical.

The query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ vectors are generated from the joint data $\boldsymbol{X}_J$ through convolutional and vector splitting operations, which are described by the following formulae:

$$\boldsymbol{X}_{conv} = Conv2D(\boldsymbol{X}_J) \in \mathbb{R}^{3\hat{C} \times T \times V} \tag{3}$$

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = Split(\boldsymbol{X}_{conv}) \in \mathbb{R}^{\hat{C} \times T \times V} \tag{4}$$

Here $\hat{C}$ represents the channel dimension of the generated vectors. The vanilla Transformer combines the input with its positional information to generate embedding vectors, and then uses these vectors to obtain $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ vectors. The self-attention mechanism in the vanilla Transformer is expressed as:

$$Y_i = \sum_{j=1}^{N} softmax\left(q_i^T k_j\right) v_j \tag{5}$$

Specifically, q_i, k_j, and v_j are elements in the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ vectors, with i and j denoting the indices of these elemen overlooking the impact of joint positions on skeleton features and, in turn, losing valuable spatial information for accurate action recognition.

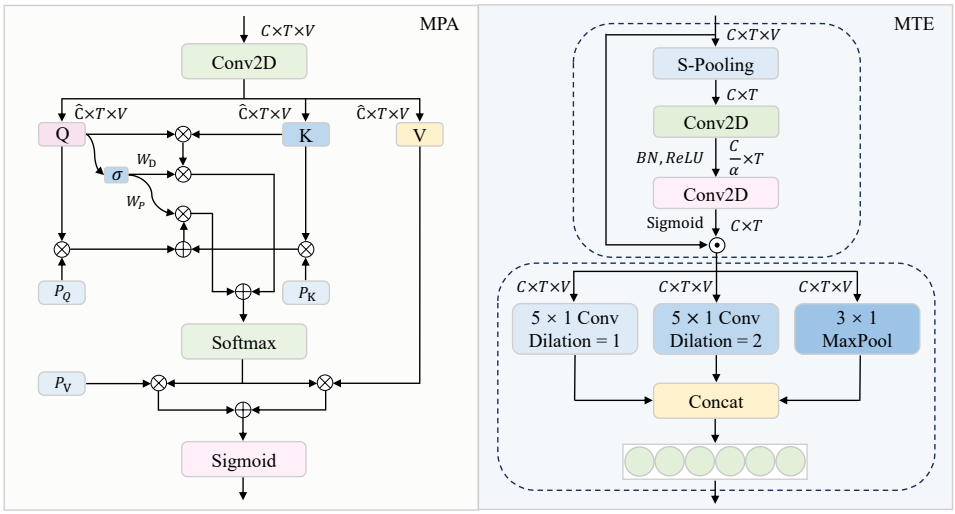

Figure 3: Illustration of the MPA and MTE modules. The left side shows the detailed process of MPA, while the right side displays the implementation of MTE.

Considering the aforementioned reasons, we propose the MPA module, which not only calculates the correlations among these vectors but also explicitly incorporates positional information by computing the dot product with their relative position. Meanwhile, this module employs a gating mechanism to control the fusion of data and positional information, ensuring a balanced integration of both elements. Specifically, the query vector $\boldsymbol{Q}$ generates dynamic weight coefficients through a linear layer to adjust the weights of different information, which allows the MPA module to capture more flexible and rich spatial features of skeleton data. The improved attention mechanism in the MPA module is expressed as:

$$Y_i = \sum_{j=1}^{N} softmax\left(w_i^d \cdot (q_i^T k_j) + w_i^p \cdot (q_i^T p_{ji}^q + k_j^T p_{ji}^k)\right)(v_j + p_{ji}^v) \tag{6}$$

The p_{ji}^q, p_{ji}^k, and p_{ji}^v are elements of $\boldsymbol{P}_Q$, $\boldsymbol{P}_K$, and $\boldsymbol{P}_V$ respectively, which represent the positional information of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ from the position j to i. The w_i^d and w_i^p are the data weight coefficient and positional weight coefficient from $\boldsymbol{W}_D$ and $\boldsymbol{W}_P$, respectively. They are used to balance the fusion of data information and positional information at index i, with $w_i^p = 1 - w_i^d$. Finally, the *sigmoid* is utilized to activate the attention output Y_i, thereby obtaining the output of the MPA module.

Considering the physical connections between joints, we add a topology graph $\boldsymbol{A} \in \mathbb{R}^{V \times V}$ to enhance the processing capability for skeleton data. Moreover, we extend the improved attention mechanism to a multi-head form (empirically set head numbers to 8) to extract diverse features. By concatenating the features of each head, we obtain the final joint spatial feature $\boldsymbol{Y}_J$.

$$\boldsymbol{Y}_J = Concat\left(\boldsymbol{Y}_J^1, \ldots, \boldsymbol{Y}_J^h\right) \in \mathbb{R}^{C \times T \times V} \tag{7}$$

The part spatial features extracted by the MPA module are $\boldsymbol{Y}_P \in \mathbb{R}^{C\times T\times V_2}$. By integrating the joint features with their corresponding part features, the final spatial features $\boldsymbol{Y}_S \in \mathbb{R}^{C\times T\times V}$ are obtained.

3.3. Multi-scale Temporal Excitation Module

Due to the locality of temporal features, specific moments (such as the takeoff and landing moments in jumping actions) are essential for action recognition. Therefore, we design the MTE module, which includes feature excitation and multi-scale convolutional operations to effectively focus on the salient features and capture multi-scale features. Inspired by SENet Hu et al. (2018), the feature excitation operation is performed on the time and channel dimensions as a whole, adjusting the importance of these features. Meanwhile, considering the differences in the duration of various actions, the MTE module utilizes multiple convolutions to extract temporal features.

In order to effectively focus on salient temporal features, the feature excitation operation first performs spatial pooling on $\boldsymbol{Y}_S \in \mathbb{R}^{C\times T\times V}$, as shown in Eq. 8.

$$\hat{\boldsymbol{Y}}_S = F_{Avg}\left(\boldsymbol{Y}_S\right) = \frac{1}{V}\sum_{v=1}^{V}\boldsymbol{Y}_S \tag{8}$$

Then, 2D convolution is used to scale the temporal and channel features of $\hat{\boldsymbol{Y}}_S$, thereby generating the excitation weights, as shown in Eq. 9.

$$\boldsymbol{W}_{CT} = F_2\left(\delta\left(F_1(\hat{\boldsymbol{Y}}_S, \frac{C}{\alpha})\right), C\right) \tag{9}$$

The channel dimensions of F_1 and F_2 are $\frac{C}{\alpha}$ and C, respectively. F_1 is utilized for feature compression, with parameter α determining the degree of compression. F_2 is employed for restoring these features. Batch Norm prevents overfitting, and δ is the ReLU activation function. $\boldsymbol{W}_{CT} \in \mathbb{R}^{C\times T}$ is the generated excitation weights.

Subsequently, we use the *sigmoid* function to nonlinearly scale $\boldsymbol{W}_{CT}$ and ensure the weights are within the range of $[0, 1]$. The scaled excitation weights are then element-wise multiplied with the original features, thereby dynamically adjusting the original features.

$$\hat{\boldsymbol{Y}}_{ST} = \boldsymbol{Y}_S \odot Softmax\left(\boldsymbol{W}_{CT}\right) \tag{10}$$

Inspired by Liu et al. (2020), we employ a simplified MS-TCN to capture temporal features. Dilated convolutions are used to capture cross-step temporal features, while max pooling extracts key temporal features. The outputs of each branch are aggregated to obtain the final spatiotemporal features Y_{ST}.

$$\boldsymbol{Y}_{ST} = concat\left(\phi_1\left(\hat{\boldsymbol{Y}}_{ST}\right), \phi_2\left(\hat{\boldsymbol{Y}}_{ST}\right), \phi_3\left(\hat{\boldsymbol{Y}}_{ST}\right)\right) \tag{11}$$

Specifically, $\phi_1(.)$ and $\phi_2(.)$ represent convolutional layers with different dilation rates, and $\phi_3(.)$ is the max pooling layer. The spatiotemporal features $\boldsymbol{Y}_{ST}$ are then processed through global pooling and fully connected layers to obtain the final action prediction results.

4. Experiments

4.1. Datasets

NTU RGB+D. NTU RGB+D dataset Shahroudy et al. (2016) is a large-scale 3D human action recognition dataset composed of 56,880 skeleton action sequences. It contains 60 action categories performed by 40 volunteers. Each frame contains at most two subjects, with each subject including 25 joints. The dataset has two evaluation benchmarks: **1) Cross-Subject (C-Sub)**: In this setup, both the training and test sets are composed of data from 20 different subjects. It is used to evaluate the generalization ability of the model across different subjects; **2) Cross-View (C-View)**: In this setup, the training and test sets consist of data from different camera views. It is used to assess the robustness of the model across different viewpoints.

NTU RGB+D 120. NTU RGB+D 120 dataset Liu et al. (2019) is an extension of the original NTU RGB+D. It contains 120 action categories performed by 106 volunteers. This dataset comprises a total of 114,480 samples and is divided according to two benchmarks: **1) Cross-Subject (C-Sub)**: In this setup, both the training and test sets consist of data from 53 different subjects; **2) Cross-Setup (C-Set)**: In this setup, the training and test sets respectively contain all samples collected from cameras with even and odd numbers.

4.2. Implementation Details

Following the previous works Plizzari et al. (2021), the Nesterov momentum is set to 0.9 and the weight decay for the SGD optimizer is 0.0004. The batch size is 128, with a total of 65 epochs. The first five epochs are used as a warm-up period. The initial learning rate of the model is 0.1, and it is reduced to 0.1 of its previous value at the 35th and 55th epochs.

4.3. Ablation Studies

4.3.1. Effect of Parameter α

We evaluate the impact of the compression parameter α used in Equation 9 in the MTE. Table 1 shows the best results are achieved when $\alpha = 8$. This indicates that MPAE-Net needs to balance overall and detailed features when compressing temporal and channel features. A too-small compression ratio fails to compress data effectively, whereas a too-large compression ratio may lose crucial detailed features. Both extremes can negatively impact recognition accuracy. Therefore, our model adopts $\alpha = 8$ as the compression ratio.

Table 1: Ablation study of parameter α on the NTU RGB+D using joint modality.

Method	α	C-Sub(%)	C-View(%)
MPAE-Net	4	91.78	96.33
MPAE-Net	8	**91.83**	**96.36**
MPAE-Net	16	91.75	96.31
MPAE-Net	32	91.65	96.18

4.3.2. Ablation Study for MPAE-Net

Table 2 presents the results of the ablation study. Our baseline model is ST-TR Plizzari et al. (2021), which only uses joint data and employs the vanilla attention mechanism in Transformer Vaswani et al. (2017) to extract spatiotemporal features. Methods 1, 2, and 3 indicate the effects of each module. Method 1 shows the effect of the PPE, with improvements of 0.7% and 0.1% over the ST-TR baseline. This suggests that the overall cues provided by part data are crucial for understanding human actions. Method 2 demonstrates the influence of MPA, which explicitly utilizes positional information to capture richer spatial features and resulting in accuracy improvements of 0.6% and 0.1%. Method 3 illustrates the results of MTE and shows an accuracy improvement of 0.4% on the C-Sub benchmark. These results validate the effectiveness of our proposed modules. Methods 4, 5, and 6 showcase the performance enhancements when using any two of the modules. Notably, method 6 employs MPA and MTE for spatiotemporal modeling, which allows for a more equitable comparison with the baseline without the addition of part data. Our method achieves improvements of 1.2% and 0.2% over the baseline. Method 7 shows that MPAE-Net, incorporating all three proposed modules, achieves outstanding performance with accuracies of 91.8% and 96.4% on the C-Sub and C-View benchmarks, respectively.

Table 2: Ablation study of MPAE-Net on the NTU RGB+D using joint modality.

Method	+PPE	+MPA	+MTE	C-Sub(%)	C-View(%)
baseline Plizzari et al. (2021)	-	-	-	89.9	96.1
1	✓			90.6 (↑ 0.7)	96.2 (↑ 0.1)
2		✓		90.5 (↑ 0.6)	96.2 (↑ 0.1)
3			✓	90.3 (↑ 0.4)	96.1
4	✓	✓		91.3 (↑ 1.4)	96.3 (↑ 0.2)
5	✓		✓	91.1 (↑ 1.2)	96.3 (↑ 0.2)
6		✓	✓	91.1 (↑ 1.2)	96.2 (↑ 0.1)
7	✓	✓	✓	**91.8** (↑ 1.9)	**96.4** (↑ 0.3)

4.3.3. Effect of Multi-Modal Data

Most current state-of-the-art methods utilize a multi-stream input. For a fair comparison, we experiment with four different modalities: 1) Js refers to the joint modality; 2) Bs means the bone modality; 3) 2s integrates the joint and bone modalities; 4) 4s combines the joint and bone modalities with their corresponding motion information.

Table 3: Ablation study of the multi-modal data on the NTU RGB+D and the NTU RGB+D 120.

Method	C-Sub 60(%)	C-View 60(%)	C-Sub 120(%)	C-Set 120(%)
Js MPAE-Net	91.8	96.4	87.6	88.8
Bs MPAE-Net	91.2	95.9	87.1	88.3
2s MPAE-Net	92.3	96.6	88.2	89.6
4s MPAE-Net	93.1	96.9	88.7	90.8

4.3.4. Visualization.

To further evaluate MPAE-Net, we visualize and analyze the attention weights of joints and parts in the MPA, as shown in Figure 4. The left part shows the joints of "Jump up" and "Clapping" actions. The darker the color, the more important it is in the action. The middle part shows the attention weights of the joints, with the horizontal and vertical axes representing the joint IDs. For the "Jump up" action, the lower limb joints with IDs 14, 15, 16, 18, 19, and 20, and the upper limb joints with IDs 8, 12, 22, 23, 24, and 25 have darker colors. This indicates the importance of these specific joints and proves the effectiveness of MPA. For the "Clapping" action, the upper limb joints with IDs 8, 12, 22, 23, 24, and 25 have darker colors, which demonstrates that the movement of the upper limb joints is crucial for recognizing this action. The right part of Figure 4 shows the attention weights of part data, with the horizontal and vertical axes representing the part IDs. It can be observed that for the "Jump up" action, the coordination of the upper limb parts with IDs 2 and 3, and the lower limb parts with IDs 4 and 5 is vital for the action. The coordination of the upper limb parts with IDs 2 and 3 is critical for the "Clapping" action. This confirms that part data provides overall action information, while joint data offers detailed interaction information.

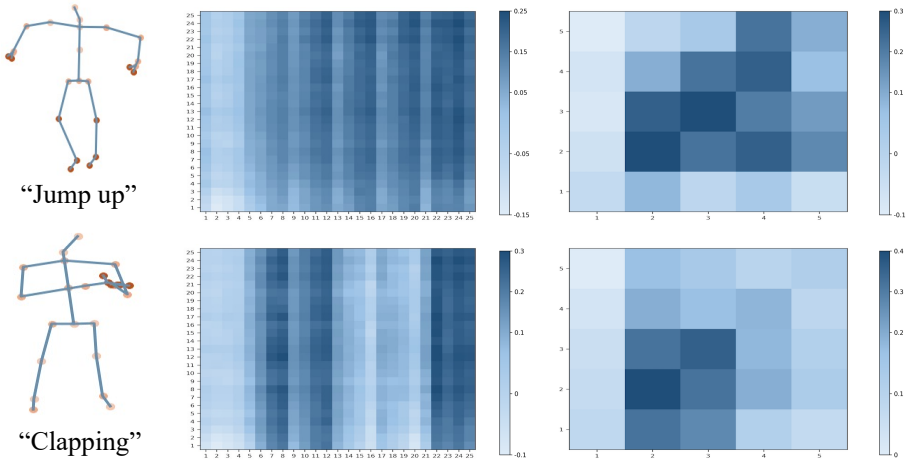

Figure 4: Visualization of the "Jump up" and "Clapping" actions and attention weights of joint and part data in the MPA.

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

We compared our MPAE-Net with state-of-the-art methods on the NTU RGB+D and NTU RGB+D 120 dataset. The results are shown in Table 4. The methods compared include four major categories: RNN-based methods, CNN-based methods, GCN-based methods, and Transformer-based methods.

Compared to RNN and CNN methods Liu et al. (2016); Li et al. (2018b); Ke et al. (2018); Li et al. (2018a), MPAE-Net offers a significant advantage. The main reason is that RNNs and CNNs cannot fully capture the global relationships of joints. In contrast, the

MPA in MPAE-Net employs the improved attention mechanism that is capable of capturing the spatiotemporal connections of all joints. Mainstream GCN methods Yan et al. (2018); Shi et al. (2019); Cheng et al. (2020); Ye et al. (2020); Liu et al. (2020); Chen et al. (2021); Cheng et al. (2021); Wen et al. (2022) utilize the topological structure of human skeleton and have better recognition performance. Compared to GCN-based methods, our MPAE-Net captures features at different levels through a multi-level input. Therefore, MPAE-Net exhibits better performance.

Notably, compared with closely related Transformer-based methods Shi et al. (2020); Plizzari et al. (2021); Wang et al. (2023b); Qiu et al. (2023); Wu et al. (2023), MPAE-Net considers the skeleton sparsity and proposes MPA to capture richer spatial features. Moreover, the MTE takes into account the locality of temporal features and dynamically recalibrates the features, which enhances the ability of the model to focus on salient features. These improvements significantly boost the spatiotemporal modeling capabilities of MPAE-Net. In summary, MPAE-Net achieves competitive results on two datasets, with accuracies of 93.1% and 96.9% on two benchmarks of the NTU RGB+D, and accuracies of 88.7% and 90.8% on two benchmarks of the NTU RGB+D 120.

Table 4: Top-1 accuracy comparison with the SOTA methods on the NTU RGB+D and NTU RGB+D 120 datasets, with the highest performance highlighted in red and the second-highest in blue.

Type	Methods	C-Sub 60(%)	C-View 60(%)	C-Sub 120(%)	C-Set 120(%)	Params (M)	FLOPs (G)
RNN-&CNN-based	ST-LSTM Liu et al. (2016)	69.2	77.7	-	-	-	-
	IndRNN Li et al. (2018b)	81.8	88.0	-	-	-	-
	MTCNN Ke et al. (2018)	81.1	87.4	61.2	63.3	-	-
	HCN Li et al. (2018a)	86.5	91.1	-	-	-	-
GCN-based	ST-GCN Yan et al. (2018)	81.5	88.3	-	-	3.1	16.3
	2S-AGCN Shi et al. (2019)	88.5	95.1	82.9	84.9	6.9	37.3
	4s Shift-GCN Cheng et al. (2020)	90.7	96.5	85.9	87.6	2.8	10.0
	Dynamic GCN Ye et al. (2020)	91.5	96.0	87.3	88.6	14.4	-
	2s MS-G3D Liu et al. (2020)	91.5	96.2	86.9	88.4	2.8	48.8
	4s MST-GCN Chen et al. (2021)	91.5	96.6	87.5	88.5	12.0	-
	4s ShiftGCN++ Cheng et al. (2021)	90.5	96.3	85.6	87.5	-	-
	2s SMotif-GCN+TBs Wen et al. (2022)	90.5	96.1	87.1	87.7	2.0	15.24
Transformer-based	4s DSTA Shi et al. (2020)	91.5	96.4	86.6	89.0	4.1	64.7
	ST-TR Plizzari et al. (2021)	89.9	96.1	82.7	84.7	12.1	259.4
	4s IIP-Transformer Wang et al. (2023b)	92.3	96.4	88.4	89.7	2.9	7.2
	4s STSA-Net Qiu et al. (2023)	92.7	96.7	88.5	90.7	5.8	-
	4s STF-Net Wu et al. (2023)	91.1	96.5	86.5	88.2	6.8	-
Ours	MPAE-Net(Ours)	91.8	96.4	87.6	88.8	2.9	2.2
	2s MPAE-Net(Ours)	92.3	96.6	88.2	89.6	5.8	4.4
	4s MPAE-Net(Ours)	93.1	96.9	88.7	90.8	11.5	8.7

5. Conclusion

In this paper, we propose a novel MPAE-Net for skeleton-based action recognition. MPAE-Net includes three main modules: PPE, MPA, and MTE. The PPE module aggregates joint data into part data to provide overall action features carried by the interactions within body parts. The MPA module addresses the issue of skeleton sparsity by improving the attention mechanism, explicitly utilizing the positional information of skeleton data to capture richer spatial features. Meanwhile, the MTE module excites the temporal features to achieve effective focus on important local features and utilizes multi-scale convolutional operations to extract cross-step temporal features. Extensive ablation experiments demonstrate the effectiveness of each module in MPAE-Net. Furthermore, experimental results on two large

datasets, NTU RGB+D and NTU RGB+D 120, show that the proposed MPAE-Net achieves better performance than existing SOTA methods.

References

Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3330–3339, 2023.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1113–1122, 2021.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 183–192, 2020.

Ke Cheng, Yifan Zhang, Xiangyu He, Jian Cheng, and Hanqing Lu. Extremely lightweight skeleton-based action recognition with shiftgcn++. *IEEE Transactions on Image Processing*, 30:7333–7348, 2021.

Yubin Gu, Honghui Xu, Yueqian Quan, Wanjun Chen, and Jianwei Zheng. Orsi salient object detection via bidimensional attention and full-stage semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Ru Jia, Li Zhao, Rui Yang, Honghong Yang, Xiaojun Wu, Yumei Zhang, Peng Li, and Yuping Su. Hfa-gtnet: Hierarchical fusion adaptive graph transformer network for dance action recognition. *Journal of Visual Communication and Image Representation*, 98: 104038, 2024.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing*, 27(6):2842–2855, 2018.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018a.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466, 2018b.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 816–833. Springer, 2016.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 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.

Xiaokai Liu, Zhaoyang You, Yuxiang He, Sheng Bi, and Jie Wang. Symmetry-driven hyper feature gcn for skeleton-based gait recognition. *Pattern Recognition*, 125:108520, 2022.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M Saquib Sarfraz, Rainer Stiefelhagen, and Alina Roitberg. Navigating open set scenarios for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4487–4496, 2024.

Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III*, pages 694–701. Springer, 2021.

Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal segments attention for skeleton-based action recognition. *Neurocomputing*, 518:30–38, 2023.

Swalpa Kumar Roy, Ankur Deria, Chiranjibi Shah, Juan M Haut, Qian Du, and Antonio Plaza. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang 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*, pages 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian conference on computer vision*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Li Wang, Ziying Song, Xinyu Zhang, Chenfei Wang, Guoxin Zhang, Lei Zhu, Jun Li, and Huaping Liu. Sat-gcn: Self-attention graph convolutional network-based 3d object detection for autonomous driving. *Knowledge-Based Systems*, 259:110080, 2023a.

Qingtian Wang, Shuze Shi, Jiabin He, Jianlin Peng, Tingxi Liu, and Renliang Weng. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition. In *2023 IEEE International Conference on Big Data (BigData)*, pages 936–945. IEEE, 2023b.

Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, Shihong Xia, and Yong-Jin Liu. Motif-gcns with local and non-local temporal blocks for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2009–2023, 2022.

Liyu Wu, Can Zhang, and Yuexian Zou. Spatiotemporal focus for skeleton-based action recognition. *Pattern Recognition*, 136:109231, 2023.

Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 2023.

Shige Xu, Lei Zhang, Yin Tang, Chaolei Han, Hao Wu, and Aiguo Song. Channel attention for sensor-based activity recognition: embedding features into all frequencies in dct domain. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 55–63, 2020.

Bo Zhang, Yaxiong Chen, Yi Rong, Shengwu Xiong, and Xiaoqiang Lu. Matnet: A combining multi-attention and transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023.