Neurocomputing 518 (2023) 30–38

Contents lists available at ScienceDirect

Neurocomputing

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

Spatio-temporal segments attention for skeleton-based action recognition

Helei Qiu, Biao Hou *, Bo Ren, Xiaohua Zhang

School of Artificial Intelligence, Xidian University, Xi'an 710071, Shaanxi, China

ARTICLE INFO

Article history:
Received 30 July 2022
Revised 21 October 2022
Accepted 30 October 2022
Available online 5 November 2022
Communicated by Zidong Wang

Keywords:
Action recognition
Skeleton
Self-attention
Spatio-temporal joints
Feature aggregation

ABSTRACT

Capturing the dependencies between joints is critical in skeleton-based action recognition. However, the existing methods cannot effectively capture the correlation of different joints between frames, which is very useful since different body parts (such as the arms and legs in "long jump") between adjacent frames move together. Focus on this issue, a novel spatio-temporal segments attention method is proposed. The skeleton sequence is divided into several segments, and several consecutive frames contained in each segment are encoded. And then an intra-segment self-attention module is proposed to capture the relationship of different joints in consecutive frames. In addition, an inter-segment action attention module is introduced to capture the relationship between segments to enhance the ability to distinguish similar actions. Compared with the state-of-the-art methods, our method achieves better performance on two large-scale datasets.

1. Introduction

Recently, skeleton-based action recognition has attracted substantial attention for its compact skeleton data representation. Unlike RGB representation, skeleton data only contains pose information, which is not affected by the changes in background, illumination, viewpoint and so on. In addition, the development of cost-effective depth cameras and human pose estimation methods makes it easier to obtain human skeleton information.

The raw skeleton data is usually converted into a graph or point sequence, and then input into a deep network (such as graph neural network (GCN) and Transformer [1]) for feature extraction. The GCN-based methods [2–6] rely on the inherent graph topology of the human body, which greatly improves the performance of skeleton-based action recognition. Further, some methods [7,8,3] add connections between local spatio-temporal joints without human topology to capture the co-occurrence information of joints. [7] builds multi-level local spatio-temporal maps on sequences, and gradually generates higher-level feature maps. [8] adds connections between adjacent nodes between frames, and extracts additional features based on the extended time graphs. [3] uses dense cross spatio-temporal edges as skip connections to directly disseminate information on spatio-temporal graphs. The Transformer-based methods [9–11] model the relationship between all joints in a sequence without relying on the human structure. [9] regards spatio-temporal skeleton data as a single sequence to capture the related information of all joints. [10] introduces self-attention into graph convolution, and uses them to model the correlation of intra-frame and inter-frame joints respectively. Similarly, [11] uses pure self-attention to capture the relationship between temporal joints and the relationship of spatial joints respectively.

It is observed that different joints in several consecutive frames are related. For example, in the action "long jump", the arms of the previous frame are related to the legs of the next frame since the joints of these parts move together in cases of motion. Therefore, it is very useful to extract the related features of different joints between adjacent frames. However, the above methods still cannot effectively capture this correlation. In the GCN-based methods, [7] constructs spatio-temporal graphs only propagate information on the joints within a frame and the same joints between frames, and do not pay attention to the relationship between different joints between frames. [8] only focuses on the relationship between local joints between frames. Although the receptive field of graph convolution can be increased by the higher power of the adjacency matrix, it will bring the problem of biased node weight, which makes remote modeling inefficient. To solve this problem, [3] eliminates redundant dependencies between neighborhoods that are closer and farther away through the proposed neighborhood de-entanglement method. In the Transformer-based methods, [9] utilize self-attention to calculate the relationship of all

* Corresponding author.
E-mail address: avcodec@hotmail.com (B. Hou).

https://doi.org/10.1016/j.neucom.2022.10.084
0925-2312/

joints in the sequence at the same time will significantly increase the computational cost, and the correlation of different joints between frames far apart is low. Similar to [7,10,11] does not pay attention to the correlation of different joints between several consecutive frames (Fig. 1a), and the extracted related motion features are too simple.

Based on the above observations, we construct a novel spatio-temporal segments attention network (STSA-Net) for skeleton-based action recognition. Specifically, a skeleton sequence is divided into several non-overlapping segments, and each segment contains several consecutive frames. Because different joints in several consecutive frames are related, each segment is flattened into a short sequence, and then an intra-segment self-attention (ISSA) module is proposed to extract the related information of joints in each segment simply and effectively as shown in Fig. 1b. This module not only captures the correlation of different joints between consecutive frames but also hardly increases the computational cost. Because the size of the time dimension is greatly reduced. In addition, a segment can be regarded as a sub-action, and a series of sub-actions form a complete action. Based on this, an inter-segment action attention (ISAA) module is proposed to aggregate these sub-actions to capture the key information between segments, and this module will help to distinguish similar actions. Finally, multi-modal data is used to further improve performance.

The main contributions of this work are as follows:s

- A spatio-temporal segments encoding strategy is proposed to explicitly flatten the joints in several consecutive frames so that our model can capture the related information of different joints between frames.
- An intra-segment self-attention is proposed to capture the related information of different joints in consecutive frames, and an inter-segment action attention module is used to integrate all segments.
- Ablation experiments verify the effectiveness of each component of our model, and the performance of our method exceeds the existing state-of-the-art methods on two challenging benchmarks.

2. Related Work

In this section, the related work is introduced, including the skeleton-based action recognition methods, the self-attention mechanism and its application in this field, and spatio-temporal joint context-aware methods.

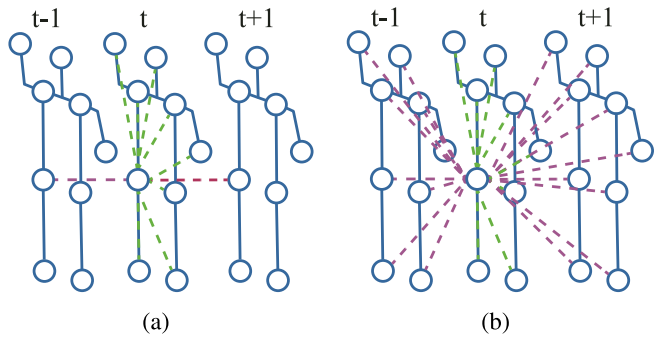

Fig. 1. Two spatio-temporal self-attention schemes. (a): This scheme only establishes the relationship between intra-frame joints and the same joints between inter-frames. (b): This scheme captures the relationship of all joints in several consecutive frames at the same time.

2.1. Skeleton-Based Action Recognition

Skeleton-based action recognition has been widely studied for decades. Previously, skeleton-based motion modeling methods mainly used 3D information of joints to design handcrafted features [12,13]. With the breakthrough of high-performance computing, deep learning shows excellent ability to extract features. At present, the deep learning methods for skeleton-based action recognition are mainly divided into the following four categories: (1) The RNN-based methods [14–16] is based on the natural time properties of the skeleton sequence, and then modeled by Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), etc.; (2) The CNN-based methods [17–19] usually convert the skeleton sequence into the pseudo-images using specific transformation rules, and model it with efficient image classification networks. Considering the advantages of both, the CNN-based methods are usually combined with the RNN-based methods to model skeleton information. (3) The GCN-based [7,20,2,3,21] methods utilize the natural topology of skeleton data in space to encode the skeleton into spatio-temporal graph and use graph convolution network to model skeleton information. (4) The Transformer-based [9–11] methods utilize self-attention to model the joints of the skeleton sequence. Most methods usually calculate the correlation of joints in the space and time domains respectively to reduce the computational cost.

2.2. Self-Attention Mechanism

Recently, Transformer [1] has become the leading model in natural language processing. The self-attention mechanism is the important component of the Transformer, which learns the relationship between each element in the long-term sequence in parallel, and solves the problem that LSTM and RNN networks cannot effectively model long-term sequence and process data recursively. Due to the advantages of self-attention, which has also been introduced into computer vision tasks such as image classification and recognition [22,23], object detection [24,25] and action recognition [26,10,27,28]. [22] combined CNN and self-attention to model the local and global dependencies for image classification. [26] used self-attention to learn spatio-temporal features from a sequence of frame-level patches for video action recognition. [28] constructed a shifted chunk Transformer to learn hierarchical spatio-temporal features from a local tiny patch to a global video clip. For the application of Transformer in computer vision, please refer to the survey [29]. For skeleton-based action recognition tasks, [11] uses two self-attention modules to capture spatial and temporal joint correlations respectively. [10] combines self-attention and graph convolution to capture intra-frame and inter-frame joint correlations respectively. Unlike [11,10], we utilize an intra-segment self-attention module to capture the correlation of different joints in consecutive frames and use an action attention module to aggregate the information between segments.

2.3. Context-Aware Methods

The context-aware methods [2,30,31,8,32,3] are designed to extract the features of spatio-temporal non-local joints since they are also related (such as clapping and typing request the cooperation of both hands). [2] superimposes an adaptive matrix on the fixed adjacency matrix to learn the non-local relationship between joints and alleviate the limitation caused by the fixed graph topology. However, the method only connects the same joints between frames, and cannot model the relationship of different joints between frames. Based on the LSTM network, [30] uses the proposed global context memory cell to selectively focus on informative joints in each frame. In addition, a recurrent attention

mechanism is introduced to identify the informative joints, and thus achieve a better attention representation. [31] proposed a context-aware graph convolution model, which uses three different functions of inner product, bi-linear form and trainable relevance score to calculate the correlation between joints, and then embeds it into the graph convolution to enrich the local response of each body joint by using the information of all other joints. [8] focuses on adding connections on adjacent vertices between frames and extracting additional features based on the extended temporal graph. [3] uses the proposed graph convolution operator to spread information on the spatio-temporal graph by using dense cross spatio-temporal edges as jump connections and separates the importance of nodes in different neighborhoods to achieve effective remote modeling. In our work, self-attention is used to capture the co-occurrence information of spatio-temporal joints in consecutive frames, and the action attention module is used to focus on the key information between segments.

3. Method

In this section, the overall architecture of the proposed method first is summarized. In the following, the spatio-temporal segments encoding strategy is introduced. Then the ISSA module for modeling the correlation of different joints in consecutive frames is described in detail. Finally, the ISAA module for aggregating inter-segment action information is introduced.

3.1. Overall Architecture

The overall architecture of our model is shown in Fig. 2. The input is a skeleton sequence containing V_0 joints and T_0 frames, which is first passed to the segments encoding module to get formatted data. And then the encoded data is input into the STSA block to capture the related information of spatio-temporal joints. The proposed model contains a total of L STSA blocks, and each block is composed of ISSA and ISAA. The ISSA module is used to model the relationship of different joints in consecutive frames and the ISAA module is used to aggregate inter-segment information. Finally, the obtained features are input into a global average pooling layer and a fully connected layer to obtain classification scores.

3.2. Segments Encoding

To model the relationship between different joints in several consecutive frames, we propose a strategy to encode these joints. The encoding process of spatio-temporal segments is shown in the left block of Fig. 2. Firstly, the raw skeleton sequence $\mathbf{X}_0 \in \mathbb{R}^{C_0 \times T_0 \times V_0}$ is fed to a feature mapping layer to expand the input channel to a set number C_1. The feature mapping layer is implemented by one convolution layer with BatchNorm and Leaky ReLU functions. Subsequently, the skeleton sequence is divided into T non-overlapping segments:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T], \mathbf{x}_i \in \mathbb{R}^{C_1 \times n \times V_0} \tag{1}$$

where n denotes the number of frames contained in a segment. In the following, each segment is flattened:

$$\mathbf{X} \in \mathbb{R}^{C_1 \times T \times n \times V_0} \rightarrow \mathbb{R}^{C_1 \times T \times V} \tag{2}$$

where $T = T_0/n, V = n \times V_0$. Finally, $\mathbf{X}$ is fed to a spatio-temporal segment encoding layer implemented by one convolution layer with Leaky ReLU function to get the final segments encoding $\mathbf{X} \in \mathbb{R}^{C \times T \times V}$.

The tensor obtained by segments encoding does not contain the order of joints, and the identity of joints cannot be distinguished, which will reduce the performance of action recognition [11]. Con-

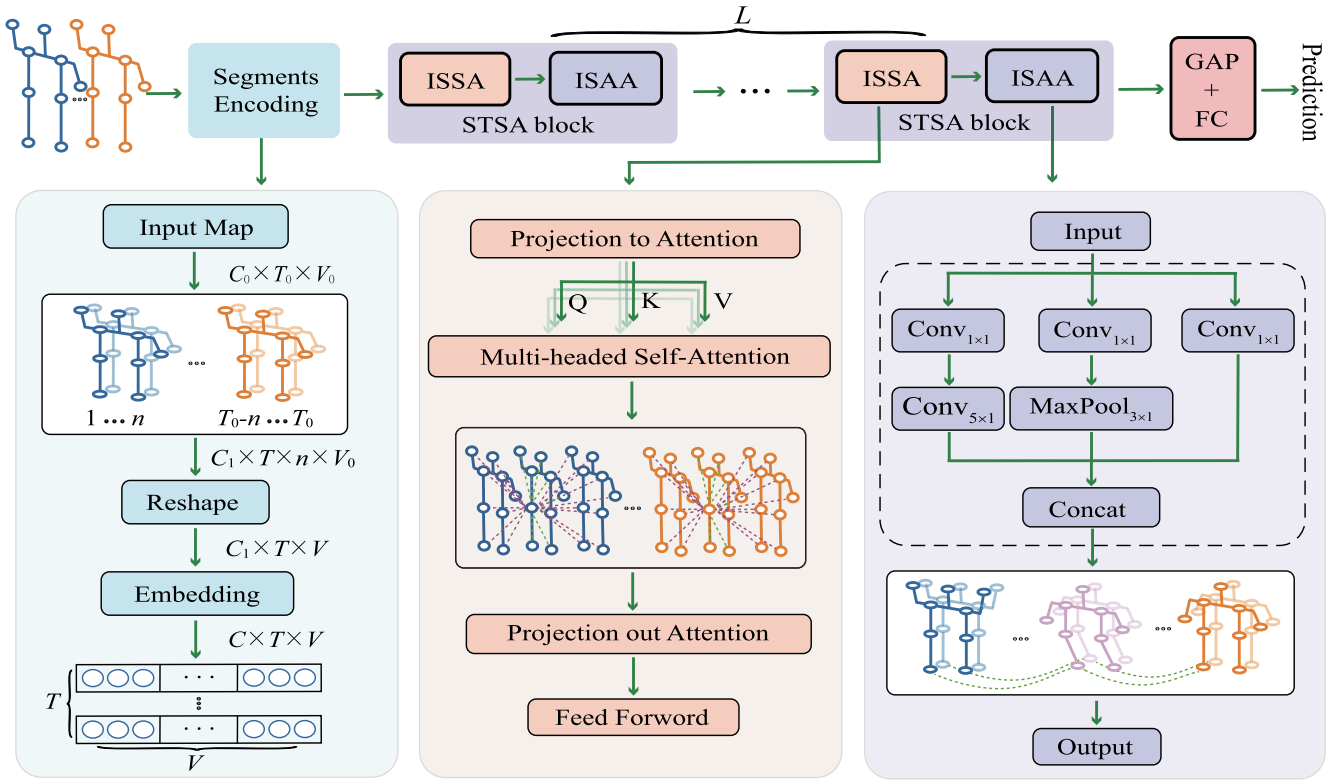

Fig. 2. Illustration of the overall architecture of the proposed method. which consists of three main modules: the spatio-temporal segments encoding, intra-segment self-attention and inter-segment action attention.

sidering this problem, a position encoding module is used to mark each joint. It should be noted that to model the relationship of all joints within a segment, it is necessary to distinguish the same joints in different frames within a segment, so all joints will be marked with different IDs. Following [1], the sine and cosine periodic functions with different frequencies are used to encode relative position information for different joints:

$$\begin{aligned} PE(np, 2i) &= sin\left(np/10000^{2i/C_{in}}\right) \\ PE(np, 2i+1) &= cos\left(np/10000^{2i/C_{in}}\right) \end{aligned} \tag{3}$$

where n denotes the number of frames within a segment, p denotes the position of the joint and i denotes the dimension of the position encoding vector, respectively.

3.3. Intra-Segment Self-Attention

The essence of self-attention can be described as the mapping from a query to a series of key and value pairs. After spatio-temporal segments encoding and positional encoding of the skeleton sequence, the multi-headed self-attention mechanism can be used to model the relationship between input tokens.

Specifically, when calculating self-attention, not only the impact of all other nodes on the node n_i but also the impact of the node n_i on other nodes must be considered. Therefore, the encoded sequence $\mathbf{X}$ is usually projected into the query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ by a convolution layer:

$$\begin{aligned} \widehat{\mathbf{X}} &= Conv_{2D(1\times1)}(\mathbf{X}) \in \mathbb{R}^{3\widehat{C}\times T\times V} \\ \mathbf{Q}, \mathbf{K}, \mathbf{V} &= Split\left(\widehat{\mathbf{X}}\right) \in \mathbb{R}^{\widehat{C}\times T\times V} \end{aligned} \tag{4}$$

where $\mathbf{Q}, \mathbf{K}$ and $\mathbf{V}$ are obtained by splitting the channel of $\widehat{\mathbf{X}}$. Then, the weights can be obtained by calculating the similarity between the query $\mathbf{Q}$ and the transpose of the key $\mathbf{K}$. Like the standard Transformer, the dot-product is simply used as the similarity function. Subsequently, the Tanh function is utilized to normalize the obtained weights. Considering the fixed relationship of human joints, an optimizable correlation matrix $\mathbf{A} \in \mathbb{R}^{T\times V\times V}$ is added to the dot product attention map to learn the inherent topological structure of the human body. At the same time, an optimizable parameter α is used to balance the strength of the attention map. Then, the final attention weight is multiplied by the corresponding value $\mathbf{V}$ to obtain the final attention:

$$\mathbf{X}_{dot} = Tanh\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right) \in \mathbb{R}^{T\times V\times V} \tag{5}$$

$$\mathbf{X}_{attn} = \mathbf{V}(\alpha \times \mathbf{X}_{dot} + \mathbf{A}) \in \mathbb{R}^{\widehat{C}\times T\times V} \tag{6}$$

where $\sqrt{C}$ is to avoid the excessive inner product to increase gradients stability during training.

To obtain better performance, the multi-headed self-attention mechanism is usually applied, which allows the model to learn related information in different representation subspaces. Specifically, the self-attention operation is performed on multiple groups of $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ projected by different learnable parameters, and then the multiple groups of attention are concatenated:

$$\mathbf{X}_{ATTN} = Concat\left(\mathbf{X}^1_{attn}, \cdots, \mathbf{X}^h_{attn}\right) \in \mathbb{R}^{H\widehat{C}\times T\times V} \tag{7}$$

In the following, the obtained $\mathbf{X}_{ATTN}$ is projected into an output space by a convolution layer with 1×1 kernel size:

$$\mathbf{X}_{ISSA} = Conv_{2D(1\times1)}(\mathbf{X}_{ATTN}) \in \mathbb{R}^{C\times T\times V} \tag{8}$$

Similar to the transformer, a feed-forward layer implemented by 1×1 2D convolution is added to increase the fitting ability of the network.

3.4. Inter-Segment Action Attention

An action can be regarded as composed of several consecutive sub-actions, such as "long jump" including sub-actions such as "run-up", "take-off" and "landing". In our method, each segment contains a sub-action, which is obtained by modeling several consecutive n frames using ISSA. If the correlation of these sub-actions (such as "run-up", 'take-off" and "landing") is constructed, it will help in the recognition of actions and distinguish similar actions (such as high jump and long jump). Therefore, the ISAA module is proposed to aggregate these sub-actions.

In this work, multi-scale convolution is used to aggregate inter-segment action information, as shown on the right block of Fig. 2. Specifically, 1×1 convolutions are used to reduce the channel dimension of $\mathbf{X}_{ISSA}$ to focus on more effective features, the dimension of each branch is shown in the following equation:

$$\begin{aligned} \mathbf{X}_1 &= \phi_1(\mathbf{X}_{ISSA}) \in \mathbb{R}^{\frac{C}{2}\times T\times V} \\ \mathbf{X}_2 &= \phi_2(\mathbf{X}_{ISSA}) \in \mathbb{R}^{\frac{C}{4}\times T\times V} \\ \mathbf{X}_3 &= \phi_3(\mathbf{X}_{ISSA}) \in \mathbb{R}^{\frac{C}{4}\times T\times V} \end{aligned} \tag{9}$$

Then the obtained features $\mathbf{X}_1$ and $\mathbf{X}_2$ are input into a 5×1 convolution layer with dilation rates 1 and a 3×1 maximum pooling layer respectively, in which the convolution layer is used to extract time information and the maximum pooling layer is used to obtain key features. It should be noted that although the dilation rate of this convolution layer is 1, each segment has n frames, which is equivalent to realizing dilation convolution across n frames, and also avoids the computational overhead caused by a large dilation rate. Unlike [3,33], which use multiple convolutions with different dilation rates, we use the ISSA module to capture the spatio-temporal information in a segment.

Subsequently, $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$ obtained through the convolution layer and the maximum pooling layer and $\mathbf{X}_3$ are concatenated to obtain the inter-segment aggregation feature $\mathbf{X}_{ISAA}$:

$$\mathbf{X}_{ISAA} = Concat\left(\hat{\mathbf{X}}_1 + \hat{\mathbf{X}}_2 + \mathbf{X}_3\right) \in \mathbb{R}^{C\times T\times V} \tag{10}$$

Finally, the residual connections are also used to stabilize network training. It should be noted that all outputs connected to the rest should be regularized.

4. Experiments

In this section, we conducted extensive comparative experiments to evaluate the performance of our method. Firstly, the benchmarks and experimental setup are introduced. In the following, the ablation studies of STSA-Net are carried out on NTU RGB + D [34] skeleton data to evaluate the contribution of each module of our method. Finally, the proposed method is compared with the state-of-the-art methods on NTU RGB + D and NTU RGB + D 120 [35] skeleton data to prove the advantages of STSA-Net, and the corresponding analysis is given.

4.1. Datasets

NTU RGB + D. NTU RGB + D dataset is a large-scale benchmark for 3D human action recognition captured simultaneously using three Microsoft Kinect V2 sensors. The dataset was completed by 40 volunteers and contained 56,000 action sequences in 60 action classes, including 40 daily actions, 9 health-related actions and 11 mutual actions. This experiment only uses the skeleton data con-

taining the three-dimensional positions of 25 body joints per frame. The dataset is divided into training data and testing data by two different standards. The Cross-Subject (CSub) divides the dataset according to the person ID, and training data and the testing data contain 20 subsets respectively. The Cross-View (CView) divides the dataset according to camera ID. The samples collected by cameras 2 and 3 are used for training, and the samples collected by camera 1 are used for testing. It should be noted that the horizontal angles of the three cameras differ by 45° respectively.

NTU RGB + D 120. NTU RGB + D 120 dataset extends NTU RGB + D by adding another 60 classes and another 57,600 samples. The dataset was completed by 106 volunteers and has 114,480 samples and 120 classes in total, including 82 daily actions, 12 health-related actions and 26 mutual actions. Like NTU RGB + D, the dataset is also divided by two different standards. For Cross-Subject (CSub) benchmark, the 106 subjects are split into training and testing groups. Each group consists of 53 subjects. For the Cross-Setup (CSet) benchmark, the samples with even collection setup IDs as the training set and the samples with odd setup IDs as the testing set.

4.2. Experimental Setting

All experiments were performed on 2 GTX 3090 GPUs. All skeleton sequences are padded to 60 frames by replaying the actions. Our model is trained using the SGD optimizer with Nesterov momentum 0.9 and weight decay 0.0004. The training epoch is set to 120, the initial learning rate is 0.1, and it is adjusted to one-tenth at 60 and 90 epochs respectively. The batch size is 64. Each segment contains 3 consecutive frames, that is, $n = 3$. The number of spatio-temporal segments attention blocks is set to 7, and the output channels are 64, 64, 128, 128, 256, 256, and 256, respectively.

4.3. Ablation Studies

The effectiveness of the STSA-Net is investigated on the NTU RGB + D Skeleton dataset. For a fair comparison, other settings are the same except for the explored object.

4.3.1. Ablation Study for STSA-Net

A set of comparative experiments in Table 1 verifies the effectiveness of each component of STSA-Net. SA-TA in the table indicates that self-attention is calculated within and between frames respectively, and n denotes the number of consecutive frames. $n = 1$ refers to modeling only the relationship of intra-frame joints, this scheme is similar to Fig. 1a. $n = 3$ means modeling the relationship of different joints between three consecutive frames at the same time.

When $n = 1$, the SA-TA and STSA-Net only establish the relationship of joints in each frame. Compared with them, the STSA-Net with $n = 3$ has made a significant improvement. This result proves that the ISSA module is very useful, and can capture the relationship between different joints in consecutive frames.

The effect of the ISAA module is also investigated. As shown in Table 1, removing the ISAA module will reduce the performance. The main reason is that this module can effectively model the relationship between segments and capture the key segments in the movement process, which is conducive to distinguishing similar actions to improve performance.

In addition, the accuracy of our model without position encoding is lower than that of the complete model, which shows that position encoding can effectively improve performance. The main reason is that different spatio-temporal joints play different roles in action, and reasonable use of this sequence information will effectively improve performance.

Table 1
Ablation study for STSA-Net on NTU RGB + D Skeleton dataset in the joint modal.

Method	CSub (%)	CView (%)
SA-TA$_{n=1}$	88.6	93.5
STSA-Net$_{n=1}$	89.0	93.8
STSA-Net$_{n=3}$ without PE	88.9	94.0
STSA-Net$_{n=3}$ without ISAA	89.0	94.1
STSA-Net$_{n=3}$	90.3	95.0

4.3.2. Effect of multi-modal Data

Different data patterns have different characteristics. Fusing multi-modal data can significantly improve performance [36]. Like most methods, our model is trained using joint, bone and joint motion modals data respectively, and then averages the reasoning outputs of our models to get the final results. The experimental results in Table 2 verify our viewpoint.

4.3.3. Effect of Parameters n

The effects of the number of consecutive frames n on our model are explored, as shown in Fig. 3. The average frame length of the NTU RGB + D skeleton dataset is about 83 frames, our method sample each skeleton sequence to 60 frames. It can be found that the accuracy of our model is the best when $n = 3$. If n is too small, the relationship between different joints between consecutive frames cannot be effectively captured; If n is too large, the joint relationship between several consecutive frames is too complex, and the correlation between the first and last frames of each part is very low.

4.3.4. Visualization

To further evaluate STSA-Net, we visualize and analyze the attention weight of the ISSA module and the output feature of the ISAA module. The left part of Fig. 4 shows the joints of action

Table 2
Ablation study for the multi-modal data on the NTU RGB + D Skeleton dataset.

Method	CSub (%)	CView (%)
STSA-Net (joint)	90.3	95.0
STSA-Net (bone)	89.7	94.0
STSA-Net (joint motion)	87.5	93.5
Fusion	92.7	96.7

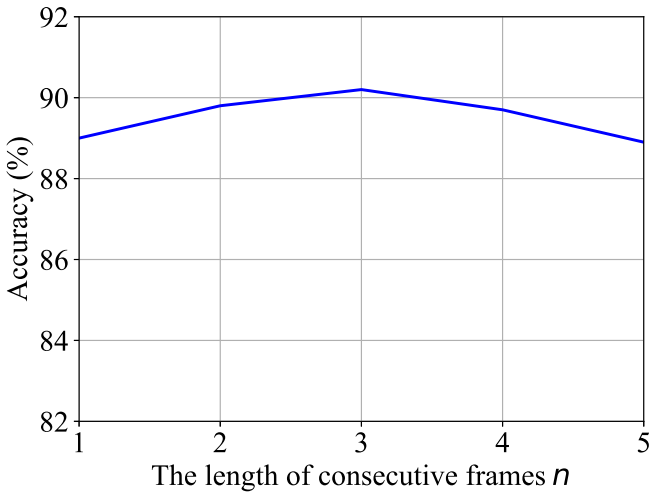

Fig. 3. Effect of the different length n of consecutive frames evaluated on NTU RGB + D skeleton dataset in the joint modal.

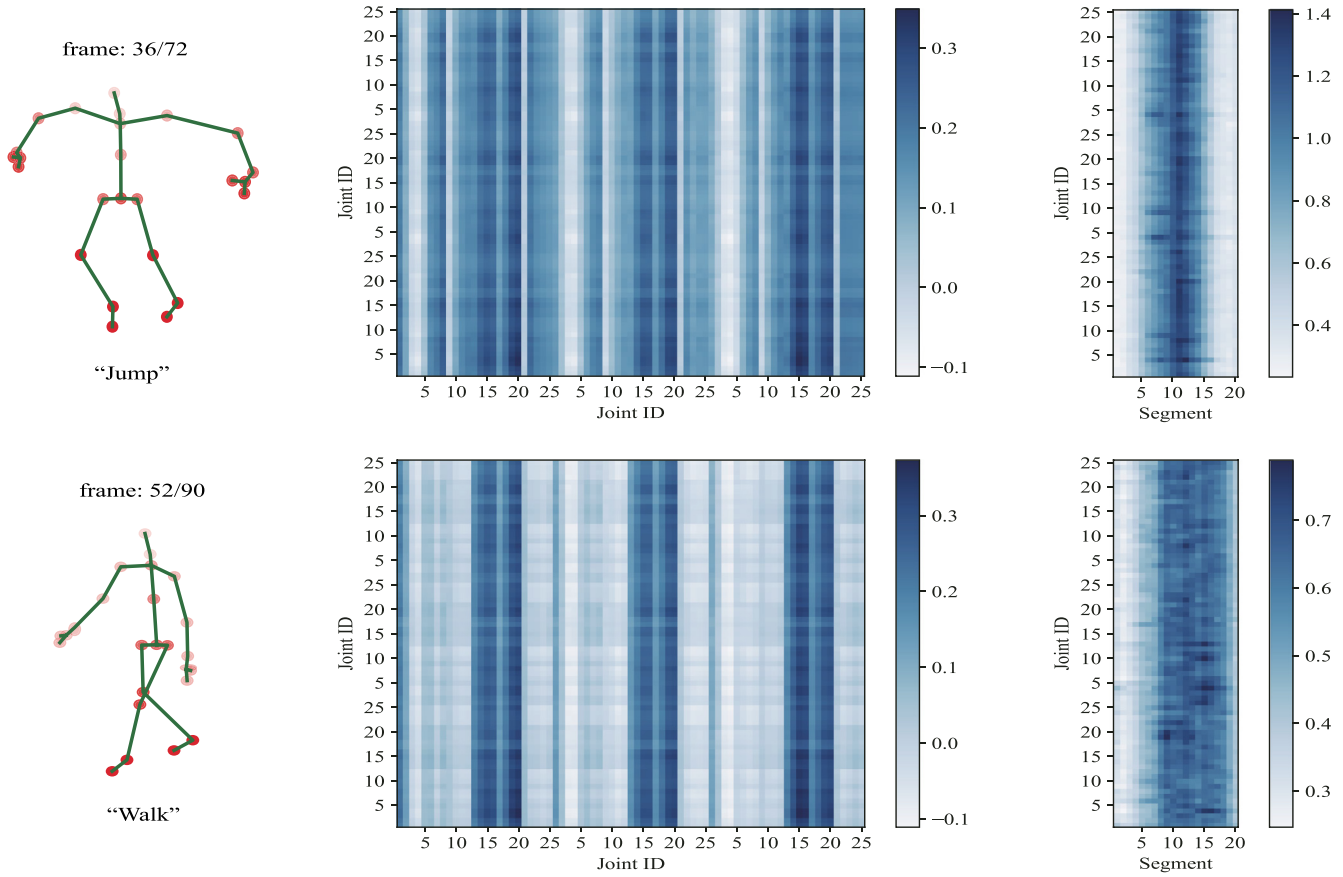

Fig. 4. Visualization of the human skeleton, attention weights of ISSA module and output features of ISAA module.

"jump" and "walk". The darker the color, the more important this joint is in action. The middle part of Fig. 4 shows the attention weights. The horizontal and vertical coordinates represent joint IDs. The darker the color, the greater the correlation. It can be found that in the action "jump", the joint IDS 14, 15, 16, 18, 19, 20 representing the lower limbs and 22, 23, 24, and 25 representing the hands are darker; In the action "walking", the joint IDS 14, 15, 16, 18, 19 and 20 representing the lower limbs and the joints 13, 1 and 7 representing the crotch are darker. These results show that the attention weight of the ISSA module can reasonably capture the correlation of joints in the global intra-segment. The right part of Fig. 4 shows the output features of the ISAA module. The abscissa represents the segment ID, and the darker the color, the more important the segment is. It can be observed that the rising and falling stages of the action "jump" are the key stages, while the stages before the jump and after the fall receive less focus; In the action "walk", the action before starting is almost ignored, and a lot of attention is paid to the action during walking. This shows that the ISAA module can effectively learn sensitive key segment information.

4.3.5. Each class accuracy

Fig. 5 shows the accuracy of each class on the CSub benchmark of the NTU TGB + D 60 dataset in a joint modal. The x-axis represents 60 different action classes. Compared with the baseline ST-GCN, our method has improved in most classes, especially the actions "clapping", "taking a selfie" and "check time (from watch)" to achieve higher accuracy. In addition, it can be found that the accuracy of some actions (such as "reading", "writing" and "typing on a keyboard") is still low. The main reason is that the motion only occurs in a few joints, and the amplitude of action is small, so the algorithm is not enough to model the key joints.

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

The proposed STSA-Net method is compared with the state-of-the-art methods on two datasets: NTU RGB + D, NTU RGB + D 120 Skeleton. Table 3 shows the comparison of Top1 accuracy. The comparison methods include CNN-based, RNN-based, GCN-based and Transformer-based methods.

Compared with the CNN and RNN-based methods [39–41], our proposed STSA-Net has significant advantages. The main reason for the poor performance of methods using CNN or RNN alone is that they cannot make full use of the information of skeleton data. In contrast, the GCN-based methods can effectively use the topology of skeleton data and have better recognition performance. In addition, it is worth mentioning that 3s-CrosSCLR is an unsupervised method, which mines cross-view consistency and achieves competitive results compared with some supervision methods.

Our method is superior to most existing methods on two datasets. Especially, compared with closely related works [3,10,11], STSA-Net has achieved significant improvement. The main reason is that the ISSA module can make full use of the related information of different joints between consecutive frames, and the ISAA module can effectively capture sensitive key segments and model the action information between segments.

The number of parameters of some models is also given. Compared with closely related methods (DSTA-Net, ST-TR), our model has significantly fewer parameters, but the performance of our method is significantly better than theirs.

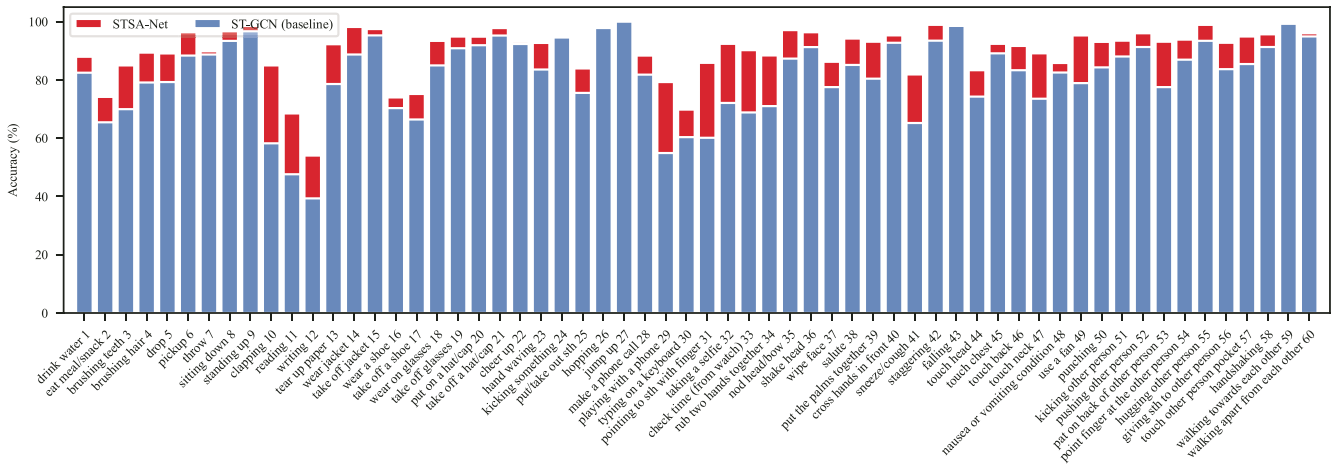

Fig. 5. The accuracy of each class of our STSA-Net and baseline ST-GCN and the comparison between them.

Table 3
Comparison of recognition accuracy (%) and parameters ($\times 10^6$) with state-of-the-art methods on NTU RGB + D (NTU60) and NTU RGB + D 120 (NTU120) Skeleton dataset. The best precision is marked in bold, and the second is underlined.

Methods	Param	NTU60		NTU120	
		CSub	CView	CSub	CSet
3s-CrosSCLR [37]	-	86.2	92.5	80.5	80.4
ST-LSTM [38]	-	69.2	77.7	-	-
MTCNN [39]	-	81.1	87.4	61.2	63.3
IndRNN [40]	-	81.8	88.0	-	-
HCN [41]	-	86.5	91.1	-	-
'TS + SS + PS' Colorization [42]	-	88.0	94.9	-	-
ST-GCN [7]	3.1	81.5	88.3	-	-
2s-AGCN [2]	6.9	88.5	95.1	82.9	84.9
SATD [43]	-	89.3	95.5	-	-
MCC+ 2s-AGCN [44]	-	89.7	96.3	81.3	83.3
DGNN [45]	26.2	89.9	96.1	-	-
Shift-GCN [46]	-	90.7	96.5	85.9	87.6
Dynamic-GCN [47]	14.4	91.5	96.0	85.9	87.6
MS-G3D [3]	6.4	91.5	96.2	86.9	88.4
MST-GCN [48]	12.0	91.5	96.6	87.5	88.8
STST [49]	-	91.9	<u>96.8</u>	-	-
CTR-GCN [33]	5.8	<u>92.4</u>	<u>96.8</u>	<u>88.9</u>	<u>90.6</u>
InfoGCN (4 ensemble) [50]	-	**92.7**	**96.9**	**89.4**	**90.7**
ST-TR [10]	9.8	89.9	96.1	82.7	84.7
DSTA-Net [11]	13.8	91.5	96.4	86.6	89.0
STSA-Net(Ours)	5.8	**92.7**	96.7	88.5	**90.7**

5. Conclusion

A novel spatio-temporal segments attention method for skeleton-based action recognition is proposed. The method consists of three modules: segments encoding, STTA and ISAA, in which segments encoding formats several consecutive frames into sequences, the STTA module is used to effectively capture the relationship of different joints in consecutive frames, and the ISAA module is used to aggregate the action inter-segments. Ablation studies show the effectiveness of the proposed method. On two large-scale datasets NTU RGB + D and NTU RGB + D 120 Skeleton, the proposed STSA-Net achieves better performance than the existing state-of-the-art methods.

CRediT authorship contribution statement

Helei Qiu: Conceptualization, Methodology, Software, Writing – original draft. **Biao Hou:** Funding acquisition, Writing - review & editing. **Bo Ren:** Investigation, Resources, Visualization. **Xiaohua Zhang:** Project administration, Supervision.

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 in part by the Key Scientific Technological Innovation Research Project by Ministry of Education; the National Natural Science Foundation of China under Grant 62171347, 61877066, 61771379, 62001355, 62101405; the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621005; the Key Research and Development Program in Shaanxi Province of China

under Grant 2019ZDLGY03-05 and 2021ZDLGY02-08; the Science and Technology Program in Xi'an of China under Grant XA2020-RGZNTJ-0021; 111 Project.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 6000–6010.

[2] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12018–12027. doi:10.1109/CVPR.2019.01230.

[3] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 140–149. doi:10.1109/CVPR42600.2020.00022.

[4] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021), https://doi.org/10.1109/TPAMI.2021.3053765, 1–1.

[5] X. Hao, J. Li, Y. Guo, T. Jiang, M. Yu, Hypergraph neural network for skeleton-based action recognition, IEEE Transactions on Image Processing 30 (2021) 2263–2275, https://doi.org/10.1109/TIP.2021.3051495.

[6] X. Liu, Y. Li, R. Xia, Adaptive multi-view graph convolutional networks for skeleton-based action recognition, Neurocomputing 444 (2021) 288–300, https://doi.org/10.1016/j.neucom.2020.03.126.

[7] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in, in: Proceedings of the AAAI conference on artificial intelligence, 2018.

[8] Y. Obinata, T. Yamamoto, Temporal extension module for skeleton-based action recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 534–540, https://doi.org/10.1109/ICPR48806.2021.9412113.

[9] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794–7803.

[10] C. Plizzari, M. Cannici, M. Matteucci, Spatial temporal transformer network for skeleton-based action recognition, in: Pattern Recognition. ICPR International Workshops and Challenges, Cham, 2021, pp. 694–701.

[11] L. Shi, Y. Zhang, J. Cheng, H. Lu, Decoupled spatial-temporal attention network for skeleton-based action recognition, in: Asian Conference on Cpmputer Vision, 2020.

[12] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595. doi:10.1109/CVPR.2014.82.

[13] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for rgb-d activity recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (11) (2017) 2186–2200, https://doi.org/10.1109/TPAMI.2016.2640292.

[14] S. Wei, Y. Song, Y. Zhang, Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 91–95. doi:10.1109/ICIP.2017.8296249.

[15] C. Si, Y. Jing, W. Wang, L. Wang, T. Tan, Skeleton-based action recognition with spatial reasoning and temporal stack learning, in: Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 106–121.

[16] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, An attention enhanced graph convolutional lstm network for skeleton-based action recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1227–1236. doi:10.1109/CVPR.2019.00132.

[17] 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), 2017, pp. 597–600. doi:10.1109/ICMEW.2017.8026285.

[18] 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, https://doi.org/10.1016/j.neucom.2020.07.068.

[19] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, J. Chen, Memory attention networks for skeleton-based action recognition, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–15, https://doi.org/10.1109/TNNLS.2021.3061115.

[20] W. Peng, J. Shi, T. Varanka, G. Zhao, Rethinking the st-gcns for 3d skeleton-based human action recognition, Neurocomputing 454 (2021) 45–53, https://doi.org/10.1016/j.neucom.2021.05.004.

[21] T. Ahmad, L. Jin, L. Lin, G. Tang, Skeleton-based action recognition using sparse spatio-temporal gcn with edge effective resistance, Neurocomputing 423 (2021) 389–398, https://doi.org/10.1016/j.neucom.2020.10.096.

[22] H. Wu, B. Xiao, N.C.F. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: Introducing convolutions to vision transformers, ArXiv abs/2103.15808 (2021).

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2021).

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision (ECCV 2020), Cham, 2020, pp. 213–229.

[25] M. Zheng, P. Gao, X. Wang, H. Li, H. Dong, End-to-end object detection with adaptive clustering transformer, ArXiv abs/2011.09315 (2020).

[26] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: The Thirty-eighth International Conference on Machine Learning, 2021.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, ArXiv abs/2103.15691 (2021).

[28] X. Zha, W. Zhu, L. Xun, S. Yang, J. Liu, Shifted chunk transformer for spatio-temporal representational learning, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 11384–11396.

[29] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–20, https://doi.org/10.1109/TPAMI.2022.3152247.

[30] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A.C. Kot, Skeleton-based human action recognition with global context-aware attention lstm networks, IEEE Transactions on Image Processing 27 (4) (2018) 1586–1599, https://doi.org/10.1109/TIP.2017.2785279.

[31] X. Zhang, C. Xu, D. Tao, Context aware graph convolution for skeleton-based action recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14321–14330. doi:10.1109/CVPR42600.2020.01434.

[32] J. Li, X. Xie, Z. Zhao, Y. Cao, Q. Pan, G. Shi, Temporal graph modeling for skeleton-based action recognition, ArXiv abs/2012.08804 (2020).

[33] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13339–13348. doi:10.1109/ICCV48922.2021.01311.

[34] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+d: A large scale dataset for 3d human activity analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019. doi:10.1109/CVPR.2016.115.

[35] 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 Transactions on Pattern Analysis and Machine Intelligence 42 (10) (2020) 2684–2701, https://doi.org/10.1109/TPAMI.2019.2916873.

[36] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–20, https://doi.org/10.1109/TPAMI.2022.3183112.

[37] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, W. Zhang, 3d human action representation learning via cross-view consistency pursuit, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4739–4748. doi:10.1109/CVPR46437.2021.00471.

[38] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: Computer Vision – ECCV 2016, Cham, 2016, pp. 816–833.

[39] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, Learning clip representations for skeleton-based 3d action recognition, IEEE Transactions on Image Processing 27 (6) (2018) 2842–2855, https://doi.org/10.1109/TIP.2018.2812099.

[40] S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (indrnn): Building a longer and deeper rnn, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5457–5466. doi:10.1109/CVPR.2018.00572.

[41] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 786–792.

[42] S. Yang, J. Liu, S. Lu, M.H. Er, A.C. Kot, Skeleton cloud colorization for unsupervised 3d action representation learning, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13403–13413. doi:10.1109/ICCV48922.2021.01317.

[43] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, L. Jun, A spatial attentive and temporal dilated(satd)gcn for skeleton-based action recognition, CAAI Transactions on Intelligence Technology 7 (1) (2022) 46–55.

[44] Y. Su, G. Lin, Q. Wu, Self-supervised 3d skeleton action representation learning with motion consistency and continuity, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 13308–13318. doi:10.1109/ICCV48922.2021.01308.

[45] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with directed graph neural networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7904–7913. doi:10.1109/CVPR.2019.00810.

[46] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 180–189. doi:10.1109/CVPR42600.2020.00026.

[47] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, H. Tang, Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020, pp. 55–63. doi:10.1145/3394171.3413941.
[48] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1113–1122.
[49] Y. Zhang, B. Wu, W. Li, L. Duan, C. Gan, Stst: Spatial-temporal specialized transformer for skeleton-based action recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2021, pp. 3229–3237, https://doi.org/10.1145/3474085.3475473.
[50] H.-G. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, Infogcn: Representation learning for human skeleton-based action recognition, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20154–20164. doi:10.1109/CVPR52688.2022.01955.

Helei Qiu received the M.S. degrees in computer science from Dalian University, Dalian, China, in 2019, and is currently studying for a doctorate at the College of Artificial Intelligence, Xidian University, Xi'an, China. His research interest includes computer vision and machine learning, with a focus on object detection and tracking, video action recognition and anomaly detection.

Biao Hou received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2003. Since 2003, he has been with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, where he is currently a Professor. His research interests include compressive sensing and Synthetic Aperture Radar image interpretation.

Bo Ren received the B.S. degree in telecommunications engineering from Northwest University, Xi'an, China, in 2011, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2017. Since 2018, he has been an Assistant Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He is a Visiting Scholar with the GIPSA-lab laboratory (Grenoble Images Parole Signal Automatique), Grenoble, France. His research interests include synthetic aperture radar image interpretation and understanding and sparse representation, and machine learning in remote sensing images.

Xiaohua Zhang received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1997 and 2000, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2004. Since 2005, he has been working in the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. His research interests include multiscale geometric analysis, Synthetic Aperture Radar image processing, compressive Sampling, deep learning and hyperspectral image classification.