The Visual Computer (2023) 39:4501–4512
https://doi.org/10.1007/s00371-022-02603-1

ORIGINAL ARTICLE

Graph-aware transformer for skeleton-based action recognition

Jiaxu Zhang[1] · Wei Xie[2] · Chao Wang[1] · Ruide Tu[3] · Zhigang Tu[1]

Accepted: 14 June 2022 / Published online: 26 July 2022

Abstract
Recently, graph convolutional networks (GCNs) play a critical role in skeleton-based human action recognition. However, most GCN-based methods still have two main limitations: (1) The semantic-level adjacency matrix of the skeleton graph is difficult to be manually defined, which restricts the perception field of GCN and limits its ability to extract the spatial–temporal features. (2) The velocity information of human body joints cannot be efficiently used and fully exploited by GCN, because GCN does not represent the correlation between the velocity vectors explicitly. To address these issues, we propose a graph-aware transformer (GAT), which can make full use of the velocity information and learn discriminative spatial–temporal motion features from the sequence of the skeleton graphs in a data-driven way. Besides, similar to the GCN-based model, our GAT also considers the prior structures of the human body including the link-aware structure and the part-aware structure. Extensive experiments on three large-scale datasets, i.e., NTU-RGB+D 60, NTU-RGB+D 120, and Kinetics-Skeleton, demonstrated that the proposed GAT obtains significant improvement compared to the GCN-based baseline for skeleton action recognition.

Keywords Skeleton action recognition · Visual transformer · Graph-aware transformer · Velocity information of human body joints · Graph neural network

1 Introduction

Human action recognition has become one of the most important tasks in the computer vision field as it has a wide range of applications in intelligent video surveillance [7,67], human–machine interaction [32,36], medical service [53], etc. However, the accuracy of video-based action recognition is limited by the quality of the video and a large amount of video data is not easy to store and transmit. In contrast to video-based human action recognition, skeleton-based human action recognition has attracted great attention due to its robustness against changes in body scales, camera viewpoints, and interference of backgrounds. At the same time, more and more human skeleton data are generated by depth cameras (e.g., Microsoft Kinect) and pose estimation algorithms [5,64], which provides a lot of available data for deep models. The storage and transmission efficiency of skeleton data is also high. In addition, the CNN-based methods have achieved good performance in processing RGB video in Euclidean space for action recognition [12,20,53]. However, unlike RGB video, the skeleton data is a kind of graph structure data, which is in non-Euclidean space. In other words, there is a correlation map between each joint of the human skeleton, but the number of relation joints for each joint is uncertain, and there is no clear order between relation joints. Therefore, how to effectively extract the spatial–temporal information in non-Euclidean space has become a key problem for skeleton-based action recognition, which is the core topic of this work.

Naturally, the skeleton data represents the human action as a sequence of 2D or 3D coordinates of the main body joints, and these joints are connected according to the phys-

✉ Chao Wang
c.wang@whu.edu.cn

✉ Ruide Tu
turuide@mails.ccnu.edu.cn

Jiaxu Zhang
zjiaxu@whu.edu.cn

Wei Xie
XW@mail.ccnu.edu.cn

Zhigang Tu
zhigangtu@whu.edu.cn

[1] State Key Laboratory of Information Engineering in Surveying, Wuhan University, Wuhan 430072, Hubei, China

[2] School of Computer, Central China Normal University, Wuhan 430079, Hubei, China

[3] School Of Information Management, Central China Normal University, Wuhan 430079, Hubei, China

ical structure of human body to construct a skeleton graph [51]. Therefore, the skeleton data expresses human action by the motion information of the skeleton graph, such as the movement speed, the positional relationship, and the angle of the joints. The traditional deep-model-based methods convert the skeleton data into Euclidean space according to predefined rules and then employ CNN or RNN to learn deep features of the skeleton sequence [4,25]. Currently, the GCN-based strategy has become the mainstream to handle the problem of skeleton-based human action recognition [10,21,41,46,62,69,71,73], as it can alternately perform convolutional operation on spatial and temporal edges to jointly learn the spatial–temporal information of the skeleton graph sequence. Compared with CNN and RNN, GCN has the advantage in processing the graph structure data with the result that it can maintain the non-Euclidean characteristics of the skeleton graph. However, as a core problem of GCN, the prior adjacency matrix guides the information aggregation in the non-Euclidean space and limits its perception field, which is difficult to be manually defined. There are many works focused on designing suitable adjacency matrix for skeleton-based action recognition [41,45,62], but these incremental modules increase the complexity of the model and are not easy to follow. In this work, we abandon the traditional GCN-based model and exploit an exquisite transformer-based model to capture features from the skeleton sequence, which is fully data-driven and without any complex incremental modules. Similar with GCN, the transformer-based model can also process the graph structure data in non-Euclidean space instead of converting the skeleton data to the Euclidean space in a manually defined way (like CNN or RNN). The transformer achieves remarkable performance in the field of natural language processing [54] and some basic visual tasks [6,17], but there are few researches that discussed the applicability of the transformer in the field of skeleton-based action recognition. On the basis of keeping the elegant structure of transformer, to make the transformer more suitable for processing the skeleton data of the graph structure, we carefully design a graph-aware transformer (GAT) without increasing any trainable parameters. The GAT takes two important structures of the human body (i.e., the link graph structure and the part graph structure) as prior masks and uses a improved multi-head attention mechanism to extract deep feature of the skeleton sequence.

A recent work 2s-AGCN [45], which utilizes the first-order spatial difference of the skeleton data (i.e., bone vectors) to construct a two-stream GCN, significantly enhances the GCN-based models. However, few GCN-based works can effectively use the first-order temporal difference information (i.e., the velocity vector) and they cannot fully exploit this velocity feature. Worse still, simply adopting the multi-stream strategy to utilize the velocity information can multiply the number of parameters and computation costs of the model. To address this issue, in the GAT, we combine the multi-head attention mechanism with the first-order temporal difference of the skeleton data and propose a velocity-driven correlation, which can make full use of the correlation of joint velocities. Thus, our GAT can learn both position-driven attention maps and velocity-driven attention maps to capture motion features effectively. Notably, unlike the 2s-AGCN algorithm, we do not utilize additional streams but merge the joint positions and the velocity features into one stream. In this way, we can make full use of the velocity vectors without increasing the model size.

In summary, the main contributions of this paper are threefold: (i) A graph-aware transformer (GAT) is carefully designed to extract the spatial–temporal information of the skeleton sequence, which takes two important human body structures as prior masks and uses an improved multi-head attention mechanism to achieve data-driven feature extraction. (ii) The first-order temporal difference of the skeleton sequence is fully utilized by combining the velocity vectors with the guidance of the multi-head attention mechanism, which can effectively learn a velocity-driven attention map to extract motion features. (iii) Extensive experiments on three large-scale datasets demonstrated that our GAT obtains remarkable performance for skeleton-based action recognition and significantly outperforms the GCN baseline.

2 Related works

2.1 Skeleton-based action recognition

Conventional skeleton-based action recognition methods usually employ handcrafted features [1,55,56] or utilize RNNs [4,23,32,49,76], CNNs [25,28,29,35] to learn features of the skeleton sequence. Vemulapalli *et al*. [55] designed rolling maps to represent the relative 3D rotations between various body parts, which is a key motion feature of skeleton sequence. Liu *et al*. [32] extended the RNN-based methods to spatial–temporal domains to analyze the action-related information. Zhu *et al*. [80] proposed a cuboid CNN to fully exploit the local movements of human joints in skeleton actions. These methods cannot effectively extract the spatial–temporal correlation from the joints of skeleton graph and also cannot fully exploit the human body structure. Yan *et al*. firstly proposed a GCN-based method ST-GCN [62], which significantly boosts the performance of skeleton-based action recognition. Then later, GCN-based methods have become the mainstream. Based on ST-GCN, many variants have been explored [10,30,46,69,71,73,76], which typically introduce some incremental modules, e.g., the attention module [10], the context-aware module [73], and the semantics-guided module [71], to enhance the network capacity. Shi *et al*. [45] took the first-order spatial difference of joints (i.e., bone vec-

tors) as a second stream and designed a two-stream adaptive GCN. Wen *et al*. [59] introduced a motif-based graph convolution to encode the hierarchical spatial structure. Zhang *et al*. [69] explored a spatial attentive and temporal dilated GCN to extract the features of skeleton sequences with different spatial attention weights and temporal scales. Peng *et al*. [41] turned to neural architecture search (NAS) and proposed the first automatically designed GCN, which can further strengthen the representation ability of the adjacency matrix in GCN. Different from the above works, to completely overcome the limitation of the adjacency matrix, we innovatively propose a transformer-based model, i.e., GAT to replace GCN as the backbone to extract features of the skeleton sequence. GAT not only retains the adjacency matrixes of the skeleton graph as the prior knowledge, but also has a correlation-driven global perception field and a more powerful capability to learn spatial–temporal deep features. Recently, Chiara et al. [42] also proposed to use the transformer to process the skeleton data, but they still employed the GCN layers when extracting the low-level features and they also ignored the velocity features. In contrast, with the help of the graph-aware masks, our GAT model can effectively extract the low-level features and the high-level features with rich velocity information. More importantly, we do not integrate the GCN-based model and any other incremental modules with the transformer, so our GAT model is more concise and easy to implement.

2.2 Visual transformer

Transformers, which have been widely used in natural language processing (NLP) tasks, are the models that rely on the multi-head self-attention mechanism to draw global correlations from the input features. Vaswani *et al*. [54] first proposed transformer based on multi-head attention mechanism for machine translation task. Devlin *et al*. [16] introduced a new language representation model called BERT (Bidirectional Encoder Representations from Transformers), which pre-trains a transformer on unlabeled text to let the model learn the context of each word. Inspired by the major success of transformer architectures in the field of NLP, recently, using transformer in vision tasks becomes the trend, e.g., object detection [6,13,14,52], image enhancement [8,63], image segmentation [8,58], image generation [38], video processing [65,79], and 3D point cloud processing [75]. For image classification, Dosovitskiy *et al*. proposed a vision transformer (ViT) [17], which divides an image into 16×16 patches and feeds these patches into a standard transformer, obtains remarkable performance. Wu *et al*. represented images as semantic visual tokens and ran transformer to densely model token relationships [60]. For object detection, Carion *et al*. [6] combined the transformer framework with the CNN network and proposed a simple and fully end-to-end object detector named DETR. Zhu *et al*. [13] proposed Deformable DETR, which has become a popular method that significantly improves the detection performance. For video processing, Zhou *et al*. designed an end-to-end transformer model to encode the video into appropriate representations [79]. Zeng *et al*. simultaneously fill missing regions in all input video frames by a self-attention module for video inpainting. For skeleton-based action recognition, Chiara *et al*. [42] proposed to improve GCN by combining spatial and temporal attention modules to explore the spatial–temporal correlation of the skeleton graph sequences. These works demonstrated that transformers have strong visual feature extraction capability and tremendous potential compared with CNNs. However, in the field of skeleton-based action recognition, there are few studies discuss the applicability of the transformer in extracting low-level and high-level skeleton features. Moreover, how to effectively use the transformer to process graph structure data is also a meaningful topic.

3 Background

3.1 Problem formulation

In this paper, we use $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the skeleton graph, where $\mathcal{V}$ means the vertexes and $\mathcal{E}$ denotes the edges. For the joint graph, $\mathcal{V}$ is the set of n joints and $\mathcal{E}$ is the set of m bones. For the bone graph, on the contrary, $\mathcal{V}$ is the set of m bones, and $\mathcal{E}$ is the set of n joints. We consider the adjacency matrix of the skeleton graph as $A \in \{0, 1\}^{n \times n}$, where $A_{i,j} = 1$ if the ith and the jth vertexes are connected, and $A_{i,j} = 0$ otherwise. The initial position feature of the skeleton joints is their 3D (or 2D) coordinates. By taking the first-order spatial difference of the joints, we can get the representation of the bones, which is a sequence of 3D (or 2D) vectors. Let $X_j \in \mathbb{R}^{n \times 3 \times T}$ be the 3D joint positions across T frames and $X_b \in \mathbb{R}^{n \times 3 \times T}$ be the 3D bone vectors.

3.2 Multi-head self-attention

The self-attention function proposed in [54] can be described as mapping a query and a set of key-value pairs to an output, where the query (Q), key (K), value (V), and output are all feature vectors. The output is computed as a weighted sum of V, where the weight assigned to each V is computed by a correlation function of Q with the corresponding K. In practice, the self-attention function is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where d_k is the dimension of K. Integrating multiple self-attention head, the multi-head self-attention can be formulated as:

$$Multihead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O, \quad (2)$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$. The projections are parameter matrices $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W^O \in \mathbb{R}^{hd_v\times d}$. Multi-head self-attention allows the model to jointly attend to information from different representation sub-spaces.

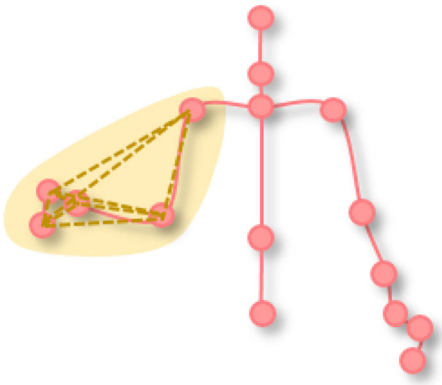

Fig. 2 Illustration of the Link Graph (red) and the Part Graph (yellow)

4 Method

4.1 Graph-aware self-attention

The GCN-based models have achieved remarkable performance in skeleton-based action recognition. An important reason is that GCN can fully explore the physical structure of the human body through the prior adjacency matrix. Peng *et al.* [41] have proved that the first-order adjacency matrix of the human body plays a key role in extracting the low-level features of the skeleton sequence. While to extract high-level features, the higher-order adjacency matrix is more important. Although existing studies have designed a variety of learnable adjacency matrices to improve the performance of GCN [41,45,62], higher-order adjacency matrices are more difficult to design and their robustness is poor. Different from GCN, the transformer employs the attention map (see Eq. 1) to aggregate information. This process is completely data-driven and does not require any prior knowledge. Therefore, the transformer is suitable for extracting high-level motion features from the skeleton sequence. To fully exploit the potential ability of the transformer in extracting high-level features, and enable the transformer to have the same ability as GCN to use the prior knowledge of the skeleton graph to extract low-level features, in this work, we propose a graph-aware self-attention module. The structure of the graph-aware self-attention module is shown in Fig. 1. We add a temporal convolutional network (TCN) layer [62] at the end of the module and employ a residual connection [22]. More importantly, based on the multi-head self-attention mechanism, we propose a multi-head graph-aware self-attention, whose detailed structure can be referred to the right part of Fig. 1. We use a variety of prior graph-aware (G-A) masks to constrain the attention map (QK^T) in this module. Specifically, let W^A and M^i, respectively, denote the attention map and the graph-aware mask, the graph-aware attention map can be calculated as $W^A \odot M^i$, where $\odot$ represents the element-wise multiplication. Next, we will introduce how to construct the G-A masks M.

There are two kinds of natural connecting structures among the joints of the human body. The first is the link graph, which connects all the joints of the human body according to the body's physical structure (see Figs. 2 and 3 Link Graph). During the movement of the human body, the physical adjacent joints have a significant influence on each other. We use the same representation rule of the adjacency matrix to represent this link-graph mask M^L. That is $M^L_{i,j} = 1$ if the ith and the jth vertexes are connected, and 0 otherwise. The second kind of natural connecting structure is the part graph, which connects part of the joints according to the part-based body structure (see Figs. 2 and 3 Part Graph). Different parts of the human body show different motion characteristics during moving. We divide the human body into 5 parts, namely

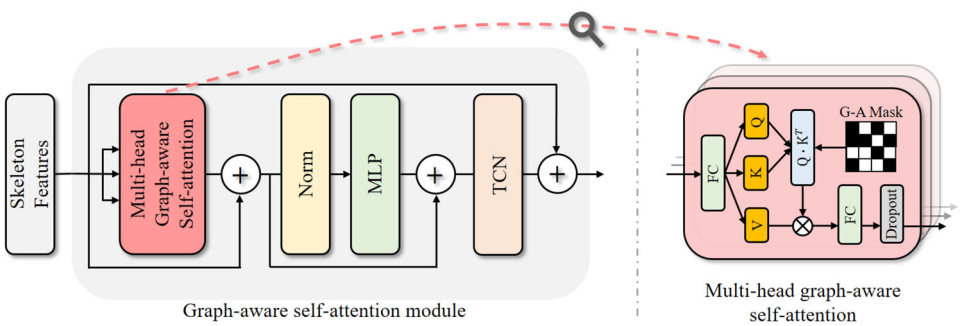

Fig. 1 The architecture of the graph-aware self-attention module

upper left limb, upper right limb, lower left limb, lower right limb, and torso. The joints contained in one part are regarded as a fully connected graph. We use the part-graph mask M^P to represent this partial connecting structure. By combining the attention head with these two kinds of masks, we get the link-aware attention head and the part-aware attention head, respectively. The link-aware attention head focuses on extracting the motion information based on the physical structure of the human body, while the part-aware attention head is more interested in extracting the motion correlation of human body parts. To avoid losing the global information, we also use the free attention head without any masks. By fusing the above three kinds of attention heads, we construct a graph-aware transformer (GAT), which can make full use of the prior graph structure and effectively extract discriminative motion features of the skeleton graph.

4.2 Graph-aware transformer

The architecture of the deigned graph-aware transformer (GAT) is shown in Fig. 3. Firstly, we use a normalization layer (Norm) and a single-head self-attention layer (SSA) to preprocess the input skeleton data. The dimension of the input skeleton data is $\mathbb{R}^{n\times 3\times T}$. After being preprocessed, the output feature dimension changes to $\mathbb{R}^{n\times 54\times T}$, which is used as the input of the GAT. The GAT consists of two parts. One part has 6 layers, the other part has 3 layers. Each layer is a multi-head attention module with h attention heads. The multi-head attention operation can be formulated as Eq. 2. In the first part, we use three kinds of different attention heads, namely the link-aware attention head (LG-A), the part-aware attention head (PG-A), and the free attention head (without the graph-aware mask, Free-A). The LG-A is the self-attention head with the link-graph-aware mask. The PG-A is the self-attention head with the part-graph-aware mask. At this stage, the model can extract rich low-level features with the help of the prior knowledge provided by a variety of attention heads. The output feature dimension of the first part is $\mathbb{R}^{n\times 192\times T/2}$. In the second part, we only use the free attention head to extract high-level features, which is completely data-driven, and there is no prior graph structure to limit its perception field. The output feature dimension of the second part is $\mathbb{R}^{n\times 276\times T/4}$. Finally, the classifier, which consists of a global average pooling layer (Avg-Pooling), a fully connected layer (FC), and a softmax function, is used to classify the human action of the skeleton sequence based on the features extracted by the GAT.

4.3 Velocity-driven correlation

Referring to Eq. 1, in theory, the self-attention mechanism is driven by the correlation between joints. If the product QK^T of the two joints is larger, the correlation is stronger, and the attention mapping is more obvious. Intuitively, in the process of human motion, joints with similar velocities have a strong correlation. For example, in the running period, there is a significant correlation between the speed of the hands and feet. Another example is clapping hands, where the correlated speed and direction of the two hands are the key features to distinguish this action. Based on the above analysis, combined with the characteristic of the self-attention mechanism, we use the first-order temporal difference (velocity vector) of the joint coordinates to enhance the feature of the joints, which is formulated as:

$$S_j = X_j^{:,:,1:T} - X_j^{:,:,0:T-1}, \quad X_j^{'} = Concat(X_j, S_j), \tag{3}$$

where S_j is the velocity vector of the joints, $X_j^{'}$ is the enhanced feature. To keep the dimension of S_j match the dimension of X_j, we fill $S_j \in \mathbb{R}^{n\times 3\times T-1}$ with 0 to make it satisfy $S_j \in \mathbb{R}^{n\times 3\times T}$. Among the GCN-based models, there was no work to use this velocity vector to enhance the feature of the joints. The reason is that GCN cannot fully exploit the correlation between joint velocities, so the performance of the model does not significantly improved after adding this new feature. In contrast, our model is based on the transformer, which can effectively use the correlation between joint velocities and learn velocity-driven attention map to extract motion features. Furthermore, unlike the existing researches that use an additional stream to process the first-order spatial difference, we directly concatenate the velocity vector of the joint with the initial position vector and use a single stream network to process this enhanced feature, so that there is almost no increase in computation cost. The dimension of the enhanced input skeleton data is $\mathbb{R}^{n\times 6\times T}$.

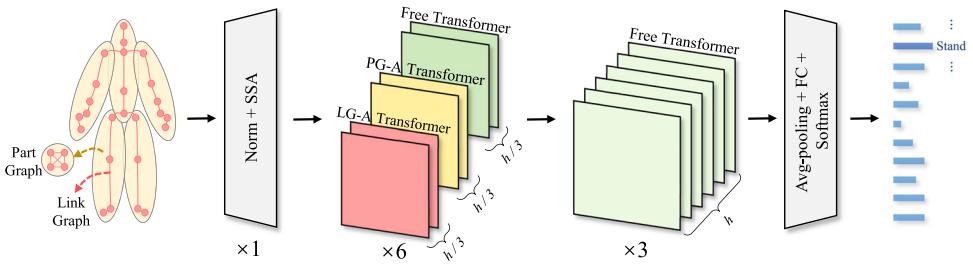

Fig. 3 The architecture of the graph-aware transformer (GAT) model

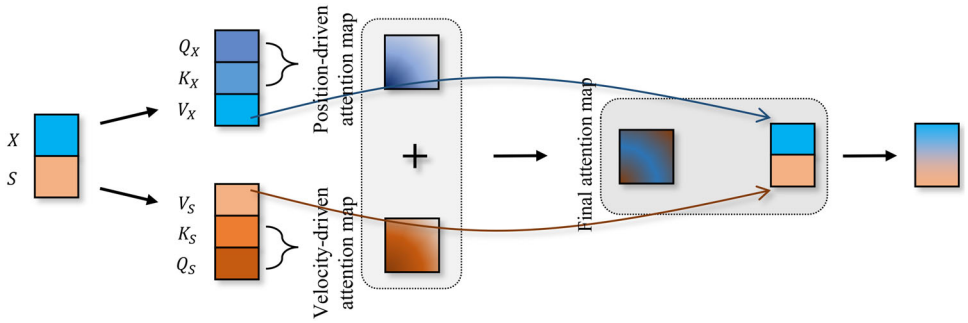

Fig. 4 The flowchart of the velocity-driven correlation mechanism

To further utilize the correlation between joint velocities, we use the position vector and the velocity vector to calculate two different attention maps respectively (*i.e.,* the position-driven attention map and the velocity-driven attention map as shown in Fig. 4). We take the average of the two attention maps as the final attention map. To sum up, our velocity-driven correlation can be formulated as:

$$W^A = \frac{1}{2}(softmax(\frac{Q_X {K_X}^T}{\sqrt{d_k}}) + softmax(\frac{Q_S {K_S}^T}{\sqrt{d_k}})). \quad (4)$$

The graph-aware velocity-driven self-attention function can be formulated as:

$$Attention(X) = (W^A \odot M^i)X^{'}, \quad (5)$$

where W^A and M^i, respectively, denotes the attention map and the graph-aware mask, $X^{'}$ is the velocity enhanced feature.

Our velocity-driven correlation mechanism can fully exploit the speed relationship of the joints and can obtain the attention map containing rich motion information, which helps the model extract the spatial–temporal motion information of the skeleton sequence effectively.

5 Experiments

5.1 Datasets and implementation details

Three popular skeleton action datasets, i.e., NTU RGB+D 60 (NTU60) [43], NTU RGB+D 120 (NTU120), and Kinetics-Skeleton (KS) [62] are selected for our experiments.

5.1.1 NTU-RGB+D 60

NTU60 [43] is a large-scale dataset with annotated 3D joint coordinates of the human body for the task of human action recognition. NTU-RGB+D contains 56,000 action videos with 60 action classes. These videos are in-door-captured from 40 volunteers in different age groups ranging from 10 to 35. For each action, the videos are obtained by 3 cameras from different viewpoints, and the 3D annotations of human body joints are given in the camera coordinate system. Each action video has no more than 2 subjects and there are 25 key joints for each subject in the skeleton sequences. The NTU-RGB+D dataset includes two settings: (1) Cross-Subject (CS) benchmark, which contains 40,320 videos for training and 16,560 videos for testing. In this setting, the training set comes from one subset of 20 subjects and the remaining 19 subjects are used for evaluation; (2) Cross-View (CV) benchmark, which includes 37,920 videos for training and 18,960 videos for testing. In this setting, the training samples come from the camera viewpoints 2 and 3, while the camera viewpoint 1 is used for evaluation. We follow the conventional settings in [43] and report the top-1 accuracy on both benchmarks.

5.1.2 NTU-RGB+D 120

NTU120 [34] is an extension of NTU60, which adds 57367 new skeleton sequences representing 60 new actions, for a total of 113945 videos referring to 120 classes from 106 subjects under 32 camera setups. It includes two settings: (1) cross-subject (X-Sub) benchmark: the 106 subjects are split into training and testing groups. Each group contains 53 subjects. (2) cross-setup (X-Set) benchmark: the training data comes from samples with even setup IDs, and the testing data comes from samples with odd setup IDs.

5.1.3 Kinetics-Skeleton

Kinetics [24] consists of 300,000 videos with 400 action classes. The video clips of Kinetics are abundant and various that sourced from YouTube, but it only provides raw videos without skeleton annotation. Yan *et al.* [62] used the OpenPose toolbox to estimate the locations of 18 joints on every frame of the videos and released the Kinetics-Skeleton datasets. In Kinetics-Skeleton, all videos are converted to a frame rate of 30fps and are resized to 340 × 256 resolution. The OpenPose toolbox generates the 2D coordinates and the confidence score for 18 joints of each human body from the

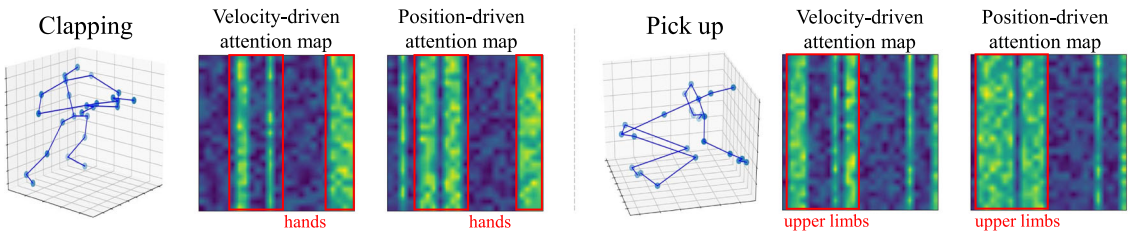

Fig. 5 Visualization of the velocity-driven attention maps and the position-driven attention maps for different human action classes on the NTU60 dataset. The value of the attention map is the average of all the heads' attention maps that without the graph-aware mask operations in the second layer of the model. The brighter area indicates that the weight of the attention map is larger there, which means the correlation between two joints is stronger

processed videos. For the multi-person clips, two major people are selected by calculating the average joint confidence. Each joint is represented by its 2D coordinate and confidence score, which construct a three-element feature vector. Following the evaluation method of Yan *et al*. [62], we train the models on the training set and report the top-1 and top-5 accuracies on the testing set.

We implement our model based on the PyTorch deep learning framework [39]. We apply the stochastic gradient descent (SGD) algorithm with Nesterov momentum (0.9) as the optimizer. We use 4 Nvidia GTX 1080Ti GPUs for the model training, and set the batch size to 48. For the NTU60 and NTU120 datasets, the number of training epoch is set as 60 and the learning rate is set to 0.1. The learning rate decay is set as 0.1 at the 30th epoch, 40th epoch and 50th epoch. For the Kinetics-Skeleton dataset, the number of training epoch is set as 70 and the learning rate is set as 0.1. The learning rate decay is set as 0.1 at the 40th epoch, 50th epoch and 60th epoch.

5.2 Ablation study

We present an ablative analysis on the NTU60 CV benchmark to evaluate the effectiveness of the proposed model. We analyze the effect of the transformer-based model, LG-A mask, PG-A mask, and the velocity-driven correlation. ST-GCN [62] is our baseline.

The results in Table 1 show that compared to the GCN-based baseline, our transformer-based model performs better. The free transformer outperforms the baseline by 2.61% (90.91% vs 88.30%), and the velocity-driven free transformer outperforms the baseline by 4.04% (92.49% vs 88.45%). The results demonstrate that the transformer-based model is significantly stronger than the GCN-based model in terms of extracting spatial–temporal motion features of the skeleton sequence. Besides, the transformer-based model can fully exploit the velocity correlation of human body joints. When the velocity-driven correlation is added, the performance of the free transformer improves by 2.3% (92.49% vs 90.91%), but the GCN-based baseline with velocity enhanced feature only improves by 0.15% (88.45% vs 88.30%). Figure 5 shows the velocity-driven attention maps and the position-driven attention maps for different human action classes on the NTU60 dataset. The value of the attention map is the average of all the heads' attention maps that without the graph-aware mask operations in the second layer of the model. The brighter area indicates that the weight of the attention map is larger there, which means the correlation between the two joints is stronger. We can see that for different actions, the bright area of the velocity-driven attention map is more concentrated than the position-driven attention map. Take the action "clapping" as an example, the activation values of the velocity-driven attention map are concentrated on a few joints of hands with salient motions, which means the velocity-driven attention map can better highlight the significant joints of human action than the position-driven attention map. Therefore, the velocity information can better reflect the motion correlation of the joints during the action procedure. Our GAT, which fuses the velocity-driven correlation mechanism and the position-driven correlation mechanism, can exploit rich motion information from the skeleton sequence.

Extensive experiments are performed to test the impact of the graph-aware masks. Table 1 shows the results of the GAT with LG-A mask, the GAT with PG-A mask, and the GAT with LG-A and PG-A masks. For the GAT with LG-A mask and the GAT with PG-A mask, we make half of their attention heads have masks and the others are free attention heads. For the GAT with LG-A and PG-A masks, we make $h/3$ of its attention heads have LG-A mask, $h/3$ of its attention heads have PG-A mask, and the others are free attention heads (see Fig. 3). The total number of the attention heads h is 6. The LG-A mask can bring 1.29% (93.78% vs 92.49%) improvement to the transformer-based model, and the PG-A mask can bring 0.89% (93.38% vs 92.49%) improvement. The combination of the LG-A mask and the PG-A mask can bring 1.43% (93.92% vs 92.49%) enhancement on top-1 accuracy to the transformer-based model. These experimental results show that the prior graph-aware masks are helpful

Table 1 Comparison of the top-1 and top-5 accuracy on the NTU60 CV benchmark with different model configurations. V-D means velocity-driven. h is the number of the attention heads

Model configs	LG-A	PG-A	V-D	Top-1 (%)	Top-5 (%)
Baseline (ST-GCN)	✓	–	–	88.30	97.10
Baseline	✓	–	✓	88.45	97.82
Baseline	✓	✓	–	88.69	97.53
Free T ($h = 6$)	–	–	–	90.91	98.72
Free T ($h = 6$)	–	–	✓	92.49	98.95
GAT ($h = 6$)	✓	–	✓	93.78	99.14
GAT ($h = 6$)	–	✓	✓	93.38	99.03
GAT ($h = 6$)	✓	✓	✓	**93.92**	**99.19**
GAT ($h = 3$)	✓	✓	✓	92.92	98.89
GAT ($h = 9$)	✓	✓	✓	93.63	99.14
GAT ($h = 6$, all)	✓	✓	✓	93.58	99.09

for the transformer to extract motion details from the physical structure and partial structure of the human body. On the other side, when we use the graph-aware attention heads for all the transformer layers without free attention head, the performance of the model will decrease instead (e.g., see the last row of Table 1, the results are, respectively, 0.34% and 0.10% lower than the GAT ($h = 6$, the 7th row of Table 1) on top-1 and top-5 accuracy). This is because high-level skeleton motion features require the model to have a global perception field, but the graph-aware masks limit the perception field of the model. Therefore, the free attention head without any limitation is necessary for the transformer to extract multi-scale and multi-granularity features. Consequently, in our model, we combine the graph-aware attention heads and the free attention head in the first 6 layers and only use the free attention head in the last 3 layers (see Fig. 3). The influence of the number of the attention heads h in each transformer layer is shown in Table 1. Specifically, compared to the GAT with $h = 3$ and $h = 9$, the GAT with $h = 6$ has the best performance.

Table 2 Comparison of the top-1 accuracy with the state of the arts on the NTU60 dataset

Methods	Params (M)	CS (%)	CV (%)
HBRNN (2015) [18]	–	59.1	64.0
Deep LSTM (2016) [43]	–	60.7	67.3
ST LSTM (2016) [32]	–	67.2	77.7
TCN (2017) [27]	–	74.3	83.1
Syn CNN (2017) [29]	–	80.0	87.2
CNN+M+T (2017) [28]	–	83.2	89.3
ST-GCN (2018) [62]	6.20	81.5	88.3
GCN+VTDB (2019) [59]	–	84.2	94.2
AS-GCN (2019) [30]	–	86.8	94.2
2s-AGCN (2019) [45]	6.94	88.5	95.1
DGNN (2019) [44]	–	89.9	96.1
CA-GCN (2020) [73]	–	83.5	91.4
SGN (2020) [71]	–	89.0	94.5
Shift-GCN (2020) [11]	–	89.7	96.0
S-TR (2021) [42]	6.14	87.9	94.9
PoseC3D (2021) [19]	–	94.1	97.1
GAT (Ours)	5.86	89.0	95.2

5.3 Comparison with state of the arts

We compare the proposed GAT model with the state-of-the-art skeleton-based action recognition methods on the NTU60, NTU120, and KS datasets. The methods which are selected for comparison include CNN-based methods [2,26,28,29], 3D-CNN-based method [19], RNN-based methods [18,33], GCN-based methods [11,30,44,45,62,73], and transformer-based method [42]. In this experiment, we use a two-stream (joint stream + bone stream) GAT model. The final classification score is the sum of the two-stream scores and the number of parameters is two-stream parameters. The results on the NTU60 dataset are shown in Table 2. Our GAT outperforms the CNN-based methods, RNN-based methods, and GCN-based methods on both the CS and the CV benchmarks, which proved that the transformer has great advantages in dealing with the skeleton data. Our results outperform ST-GCN [62] by 6.9% (95.2% vs 88.3%) on the CV benchmark and 7.5% (89.0% vs 81.5%) on the CS benchmark with less parameters (5.86M vs 6.20M). Figure 6 shows the confusion matrix of our GAT on the NTU60 CV benchmark (left) and the comparison of the classification accuracy with ST-GCN of 60 action categories on the NTU60 CV benchmark (right). The accuracy of the GAT is represented by the red line and the accuracy of the ST-GCN is represented by the blue dotted line. It can be seen that our GAT improves the classification accuracy on all of the action categories, because our velocity-driven correlation can fully exploit the correlation between motion speed of joints, which is helpful for distinguishing the human actions. The results in the red boxes of Fig. 6 also reveal that the main

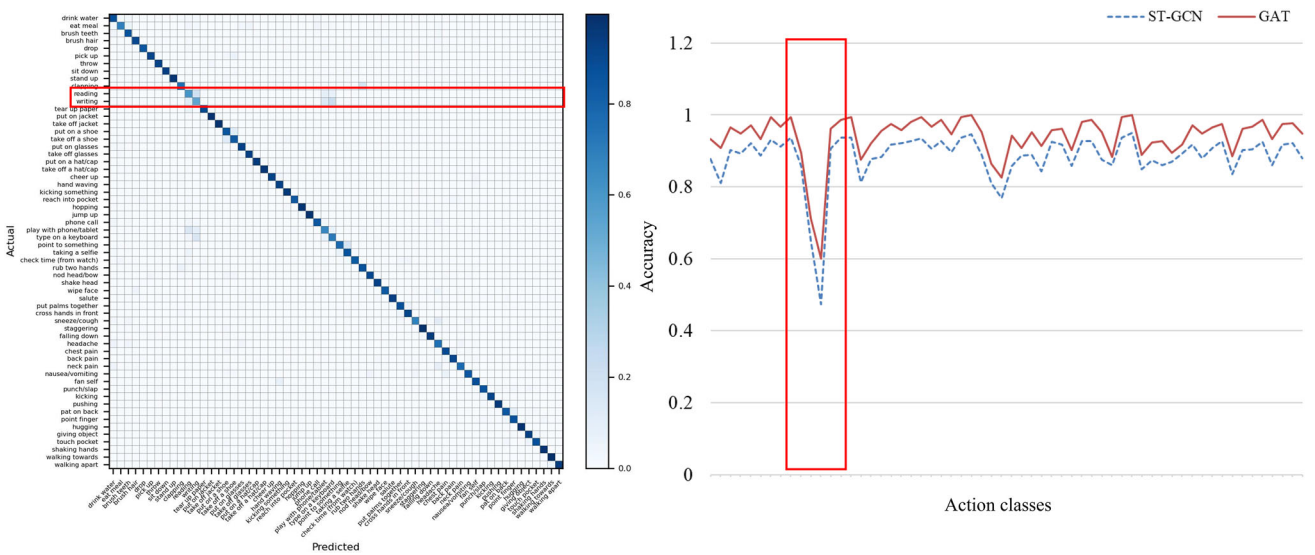

Fig. 6 The confusion matrix on the NTU60 CV benchmark (left). Comparison of the classification accuracy of 60 action categories on the NTU60 CV benchmark (right). The accuracy of the GAT represents as the red line and the accuracy of the ST-GCN represents as the blue dotted line

Table 3 Comparison of the top-1 accuracy with the state of the arts on the NTU120 dataset

Methods	X-Set (%)	X-Sub (%)
ST LSTM (2016) [32]	57.9	55.7
Clips+CNN+MTLN (2017) [25]	62.2	61.8
SkeMotion (2019) [3]	67.7	66.9
TSRJI (2019) [2]	67.9	62.8
ST-GCN (2019) [62]	79.0	77.9
AGCN (2019) [45]	84.9	82.9
Shift-GCN (2020) [11]	86.6	85.3
SGN (2020) [71]	79.2	81.5
S-TR (2021) [42]	83.6	81.0
PoseC3D (2021) [19]	90.3	86.9
GAT (Ours)	86.1	84.0

Table 4 Comparison with the state of the arts on the KS dataset

Methods	Top-1 (%)	Top-5 (%)
PA-LSTM (2016) [43]	16.4	35.3
TCN (2017) [27]	20.3	40.0
ST-GCN (2018) [62]	30.7	52.8
AS-GCN (2019) [30]	34.8	56.5
2s-AGCN (2019) [45]	36.1	58.7
DGNN (2019) [44]	36.9	59.6
CA-GCN (2020) [73]	34.1	56.6
S-TR (2021) [42]	35.4	57.9
GAT (Ours)	35.9	58.9

difficulty for skeleton-based action recognition is to distinguish the confusing human actions, such as "writing" and "type on a keyboard," which needs to be further studied. The results on the NTU-RGB+D 120 dataset are shown in Table 3. Our GAT performs better than the CNN-based methods and RNN-based methods. Compared to the state-of-the-art GCN-based methods Shift-GCN [11], our GAT also obtains competitive results (86.1% vs 86.6% on the X-Set benchmark and 84.0% vs 85.3% on the X-Sub benchmark). Results on the Kinetics-Skeleton dataset are shown in Table 4. Our GAT surpasses the other competitive methods in both top-1 and top-5 accuracy. It demonstrated that our GAT model is more robust to deal with noisy 2D skeleton data in real-world videos. Using only the spatial attention mechanism, the accuracy of our GAT outperforms the S-TR [42] on all the three datasets NTU60, NTU120, and KS.

6 Conclusions

In this work, we proposed a novel graph-aware transformer (GAT), which can fully utilize the velocity correlation of human joints to extract motion features of the skeleton sequence. The link-aware attention and the part-aware attention are the core modules of the GAT, which are designed by fusing the graph-aware masks with the attention map to effectively make use of the prior skeleton graph structures. Extensive experiments are conducted on three large-scale datasets to evaluate the performance of our method. The results verified that the proposed transformer-based model outperforms the GCN-based baseline by a large margin, and the GAT obtains remarkable performance on extracting

the spatial–temporal deep features for skeleton-based action recognition. In the future, considering the context information and using the attention mechanism to learn the temporal feature more effectively need to be further investigated.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 62106177. It was also supported by the Joint Fund of the Ministry of Education of China under Grant No. 8091B032156. The numerical calculation was supported by the supercomputing system in the Super-computing Center of Wuhan University.

References

1. Agahian, S., Negin, F., Köse, C.: Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. Visual Comp. **35**(4), 519–607 (2019)
2. Caetano, C., Brémond, F., Schwartz, W.R.: Skeleton image representation for 3d action recognition based on tree structure and reference joints. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), IEEE, pp 16–23 (2019a)
3. Caetano, C., Sena, J., Brémond, F., et al.: Skelemotion: a new representation of skeleton joint sequences based on motion information for 3d action recognition. In: 2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS), IEEE, pp 1–8 (2019c)
4. Cao, C., Lan, C., Zhang, Y., et al.: Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans. Circuit Sys. Video Tech. **29**(11), 3247–3257 (2018)
5. Cao, Z., Hidalgo, G., Simon, T., et al.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Patt. Anal. & Mach. Intell. **PP**(99), 1 (2018)
6. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: European Conference on Computer Vision, Springer, Berlin. pp 213–229 (2020b)
7. Chang, Y., Tu, Z., Xie, W., et al.: Clustering driven deep autoencoder for video anomaly detection. In: European conference on computer vision, Springer, Berlin pp 329–345 (2020)
8. Chen, H., Wang, Y., Guo, T., et al.: Pre-trained image processing transformer. In: arXiv preprint arXiv:2012.00364 (2020)
9. Chen, Y., Wang, Z., Peng, Y., et al.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7103–7112 (2018)
10. Cheng, K., Zhang, Y., Cao, C., et al.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: Proceedings of the European conference on computer vision (ECCV) (2020a)
11. Cheng, K., Zhang, Y., He, X., et al.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 183–192 (2020b)
12. Crasto, N., Weinzaepfel, P., Alahari, K., et al.: Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7882–7891 (2019)
13. Dai, Z., Cai, B., Lin, Y., et al.: Deformable transformers for end-to-end object detection. In: arXiv preprint arXiv:2010.04159 (2020a)
14. Dai, Z., Cai, B., Lin, Y., et al.: Up-detr: Unsupervised pre-training for object detection with transformers. In: arXiv preprint arXiv:2011.09094 (2020b)
15. Demisse, G.G., Papadopoulos, K., Aouada, D., et al.: Pose encoding for robust skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 188–194 (2018)
16. Devlin, J., Chang, M.W., Lee, K., et al.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: arXiv preprint arXiv:2010.11929 (2020)
18. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118 (2015)
19. Duan, H., Zhao, Y., Chen, K., et al.: Revisiting skeleton-based action recognition. arXiv preprint arXiv:2104.13586 (2021)
20. Feichtenhofer, C., Fan, H., Malik, J., et al.: Slowfast networks for video recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6202–6211 (2019)
21. Gao, X., Hu, W., Tang, J., et al.: Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of the 27th ACM international conference on multimedia, pp 601–610 (2019)
22. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 770–778 (2016)
23. Hu, Y., Liu, C., Li, Y., et al.: Temporal perceptive network for skeleton-based action recognition. In: BMVC (2017)
24. Kay, W., Carreira, J., Simonyan, K., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
25. Ke, Q., Bennamoun, M., An, S., et al.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3288–3297 (2017)
26. Ke, Q., Bennamoun, M., An, S., et al.: Learning clip representations for skeleton-based 3d action recognition. IEEE Trans. Image Process. **27**(6), 2842–2855 (2018)
27. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), IEEE, pp 1623–1631 (2017)
28. Li, B., Dai, Y., Cheng, X., et al.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: 2017 IEEE International conference on multimedia & expo workshops (ICMEW), IEEE, pp 601–604 (2017a)
29. Li, C., Zhong, Q., Xie, D., et al.: Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International conference on multimedia & Expo Workshops (ICMEW), IEEE, pp 597–600 (2017b)
30. Li, M., Chen, S., Chen, X., et al.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3595–3603 (2019)
31. Li, M., Chen, S., Zhao, Y., et al.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 214–223 (2020)
32. Liu, J., Shahroudy, A., Xu, D., et al.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision, Springer, Berlin pp 816–833 (2016)
33. Liu, J., Wang, G., Hu, P., et al.: Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 1647–1656 (2017a)
34. Liu, J., Shahroudy, A., Perez, M., et al.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. In: CoRR, abs/1905.04757 (2019)

35. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Patt. Recognit. **68**, 346–362 (2017)
36. Ma, C., Wang, A., Chen, G., et al.: Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network. Visual Comp. **34**(6), 1053–1063 (2018)
37. Miyato, T., Si, Maeda, Koyama, M., et al.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Patt. Anal. Mach. Intell. **41**(8), 1979–1993 (2018)
38. Parmar, N., Vaswani, A., Uszkoreit, J., et al.: Image transformer. In: arXiv preprint arXiv:1802.05751 (2020)
39. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems, pp 8026–8037 (2019)
40. Peng, G., Wang, S.: Dual semi-supervised learning for facial action unit recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 8827–8834 (2019)
41. Peng, W., Hong, X., Chen, H., et al.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI conference on artificial intelligence (2020)
42. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. Comp. Vis. Image Understand. **208–209**(103), 219 (2021). https://doi.org/10.1016/j.cviu.2021.103219
43. Shahroudy, A., Liu, J., Ng, T.T., et al.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019 (2016)
44. Shi, L., Zhang, Y., Cheng, J., et al.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7912–7921 (2019)
45. Shi, L., Zhang, Y., Cheng, J., et al.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12,026–12,035 (2019)
46. Si, C., Jing, Y., Wang, W., et al.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the European conference on computer vision (ECCV), pp 103–118 (2018)
47. Si, C., Chen, W., Wang, W., et al.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236 (2019)
48. Si, C., Nie, X., Wang, W., et al.: Adversarial self-supervised learning for semi-supervised 3d action recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 35–51 (2020)
49. Song, S., Lan, C., Xing, J., et al.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-first AAAI conference on artificial intelligence (2017)
50. Song, S., Lan, C., Xing, J., et al.: Spatio-temporal attention-based LSTM networks for 3d action recognition and detection. IEEE Trans. Image Process. **27**(7), 3459–3471 (2018)
51. Straka, M., Hauswiesner, S., Rüther, M., et al.: Skeletal graph based human pose estimation in real-time. In: BMVC, pp 1–12 (2011)
52. Sun, Z., Cao, S., Yang, Y., et al.: Rethinking transformer-based set prediction for object detection. In: arXiv preprint arXiv:2011.10881 (2020)
53. Tu, Z., Xie, W., Qin, Q., et al.: Multi-stream CNN: learning representations based on human-related regions for action recognition. Patt. Recogn. **79**, 32–43 (2018)
54. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008 (2017)
55. Vemulapalli, R., Chellapa, R.: Rolling rotations for recognizing human actions from 3d skeletal data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4471–4479 (2016)
56. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 588–595 (2014)
57. Wang, H., Wang, L.: Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. IEEE Trans. Image Process. **27**(9), 4382–4394 (2018)
58. Wang, Y., Xu, Z., Wang, X., et al.: End-to-end video instance segmentation with transformers. In: arXiv preprint arXiv:2011.14503 (2020)
59. Wen, Y.H., Gao, L., Fu, H., et al.: Graph CNNS with motif and variable temporal block for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8989–8996 (2019)
60. Wu, B., Xu, C., Dai, X., et al.: Visual transformers: token-based image representation and processing for computer vision. In: arXiv preprint arXiv:2006.03677 (2020)
61. Xu, Z., Hu, R., Chen, J., et al.: Semisupervised discriminant multimanifold analysis for action recognition. IEEE Trans. Neur. Netw. Learn Sys. **30**(10), 2951–2962 (2019)
62. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
63. Yang, F., Yang, H., Fu, J., et al.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 5791–5800 (2020)
64. Yuan, X., Kong, L., Feng, D., et al.: Automatic feature point detection and tracking of human actions in time-of-flight videos. IEEE/CAA J. Automat. Sinica. **4**(4), 677–685 (2017). https://doi.org/10.1109/JAS.2017.7510625
65. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: Proceedings of the European conference on computer vision (ECCV), pp 528–543 (2020)
66. Zengeler, N., Kopinski, T., Handmann, U.: Hand gesture recognition in automotive human-machine interaction using depth cameras. Sensors **19**(1), 59 (2019)
67. Zhang, D., He, L., Tu, Z., et al.: Learning motion representation for real-time spatio-temporal action localization. Patt. Recogn. **103**(107), 312 (2020)
68. Zhang, J., Han, Y., Tang, J., et al.: Semi-supervised image-to-video adaptation for video action recognition. IEEE Trans. Cybernet. **47**(4), 960–973 (2016)
69. Zhang, J., Ye, G., Tu, Z., et al.: A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition. CAAI Transactions on intelligence technology pp 1–10 (2021a)
70. Zhang, P., Lan, C., Xing, J., et al.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans. Patt. Anal. Mach. Intell. **41**(8), 1963–1978 (2019)
71. Zhang, P., Lan, C., Zeng, W., et al.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 1112–1121 (2020c)
72. Zhang, X., Xu, C., Tian, X., et al.: Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Trans. Neur. Netw. Learn Sys. **31**(8), 3047–3060 (2019)

73. Zhang, X., Xu, C., Tao, D.: Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14,333–14,342 (2020d)
74. Zhang, X., Li, C., Shi, H., et al.: Adapnet: adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization. IEEE Transactions on Neural Networks and Learning Systems (2020e)
75. Zhao, H., Jiang, L., Jia, J., et al.: Point transformer. In: arXiv preprint arXiv:2012.09164 (2020)
76. Zhao, R., Wang, K., Su, H., et al.: Bayesian graph convolution lstm for skeleton based action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 6882–6892 (2019)
77. Zheng, N., Wen, J., Liu, R., et al.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Thirty-Second AAAI conference on artificial intelligence (2018)
78. Zheng, W., Li, L., Zhang, Z., et al.: Relational network for skeleton-based action recognition. In: 2019 IEEE International conference on multimedia and expo (ICME), pp 826–831 (2019)
79. Zhou, L., Zhou, Y., Corso, J.J., et al.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8739–8748 (2018)
80. Zhu, K., Wang, R., Zhao, Q., et al.: A cuboid CNN model with an attention mechanism for skeleton-based action recognition. IEEE Trans. Multim. **22**(11), 2977–2989 (2019)

Jiaxu Zhang received the B.S. degree from Southeast University, Nanjing, China, in 2020. He is currently working toward the M.S. degree at the LIESMARS (State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing), Wuhan University, China. His research interests include computer vision, computer graphics, and action recognition/retargeting.

Wei Xie received the B.E. degree in electronic information engineering and the Ph.D. degree in communication and information system from Wuhan University, China, in 2004 and 2010, respectively. From 2010 to 2013, he was an Assistant Professor with the Computer School, Wuhan University. He is currently an Associate Professor with the Computer School, Central China Normal University, China. His research interests include motion estimation, super resolution reconstruction, image fusion, and image enhancement.

Chao Wang received the Ph.D. degree in Photogrammetry and remote sensing from Wuhan University, China, in 2009. He is currently a associate professor at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China. His research interests include geographic information science, remote sensing, and land use/cover change.

Ruide Tu received the master's degree from Wuhan Institute of Technology, Wuhan, China, in 2020. He is currently working toward the doctoral in the school of information management of Central China Normal University, China. His research areas include digital government, information management and services, information system analysis, and prediction and so on.

Zhigang Tu started his Master Degree at Wuhan University, China, 2008. In 2015, he received the Ph.D. degree from Utrecht University, the Netherlands. From 2015 to 2016, he was a postdoctoral researcher at Arizona State University, USA. Then from 2016 to 2018, he was a research fellow at Nanyang Technological University, Singapore. He is currently a professor at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote sensing, Wuhan University. His research interests include computer vision, image processing, video analytics, and machine learning. Special for motion estimation, action recognition/localization, human/hand pose estimation, and anomaly event detection.