ACAIT2020
Intended journal: *'Int. J. Bio-Inspired Computation'*, *'CAAI Transactions on Intelligence Technology'* and *'Complexity'*
The first author is a student.

A Spatial Attentive and Temporal Dilated (SATD) GCN for Skeleton-Based Action Recognition

Jiaxu Zhang

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University,
Wuhan 430079, China
E-mail: z.jiaxu@foxmail.com

Gaoxiang Ye

State Grid Wuhan Power Supply Company,
Wuhan 430014, China
E-mail: ygxflyhigh@163.com

Zhigang Tu*

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University,
Wuhan 430079, China
E-mail: tuzhigang@whu.edu.cn
*Corresponding author

Yongtao Qin

Shenzhen Infinova Ltd. Company,
Shenzhen 518100, China.

Jinlu Zhang

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University,
Wuhan 430079, China
E-mail: jinlu.zhang@mail.sdu.edu.cn

Xiangjian Liu

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University,
Wuhan 430079, China
E-mail: liuxj96@whu.edu.cn

Shixu Luo

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University,
Wuhan 430079, China
E-mail: shixuluo@whu.edu.cn

Abstract: Current studies have shown that the spatial-temporal graph convolutional network (ST-GCN) is effective for skeleton-based action recognition. However, for the existed ST-GCN based methods, their temporal kernel size is usually fixed over all layers, which making them cannot fully exploit the temporal dependency between discontinuous frames and different sequence length. In addition, most of these methods use average pooling to obtain global graph feature from vertex features, resulting in a loss of much fine-grained information for action classification. To address these issues, in this work, we propose a novel spatial attentive and temporal dilated graph convolutional network (SATD-GCN). It contains two important components, i.e., a spatial attention pooling module (SAP) and a temporal dilated graph convolution module (TDGC). Specifically, the SAP module can select the human body joints which are beneficial for action recognition by a self-attention mechanism, and alleviate the influence of data redundancy and noise. The TDGC module can effectively extract the temporal features of different time scales, which is useful to improve the temporal perception field and enhance the robustness of the model to different motion speed and sequence length. Importantly, both the SAP module and the TDGC module can be easily integrated into the ST-GCN based models, and significantly improve their performance. Extensive experiments on two large-scale datasets, i.e., NTU-RGB+D and Kinetics-Skeleton, demonstrate that our method achieves the state-of-the-art performance for skeleton-based action recognition.

Keywords: Skeleton-based action recognition; Graph convolutional network; Spatial attention pooling; Temporal dilated convolution

Reference to this paper should be made as follows: Yousefi, B. and Loo, C.K. (2016) ‘Slow feature action prototypes effect assessment in mechanism for recognition of biological movement ventral stream’, *International Journal of Bio-Inspired Computation*, 8(6), pp.410–424.

Biographical notes: Jiaxu Zhang is a research student in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University in Wuhan, China. His research interests include deep learning and computer vision.

Zhigang Tu received the master’s degree in image processing from the School of Electronic Information, Wuhan University, China, and the Ph.D. degree in computer science from Utrecht University, The Netherlands, in 2015. From 2015 to 2016, he was a Post-Doctoral Researcher with Arizona State University, USA. From 2016 to 2018, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently a Professor with 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, machine learning, and specially for motion estimation, object segmentation, object tracking, action recognition and localization, and anomaly detection.

Jinlu Zhang is a research student in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University in Wuhan, China. His research interests include deep learning and human pose estimation.

Xiangjian Liu is a graduate student in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University in Wuhan, China. His research interests include deep learning and computer vision.

Shixu Luo is a research student in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University in Wuhan, China. His research interests include deep learning and computer vision.

1 Introduction

Human action recognition, which has a wide range of applications in intelligent video surveillance, human-machine interaction, medical service, etc. is still a challenging and unsolved problem (Tu et al., 2019; Simonyan et al., 2014; Tu et al., 2018; Baranwal et al., 2017; Yousefi et al., 2016). Human action recognition based on the RGB appearance is usually easily affected by the complex background, illumination change, occlusion and other factors. In recent years, more and more research has been focused on the skeleton-based action recognition since it is robust to against changes in motion speeds, body scales, camera viewpoints and interference of backgrounds. Moreover, increasingly human skeleton data is collected by depth cameras and human pose estimation algorithms (Cao et al., 2017; Chen et al., 2018), which provides sufficient data for skeleton-based action recognition research and application. The skeleton data represents the human action as a sequence of 2D or 3D coordinates of the major body joints, so it is crucial to extract discriminative features in both spatial and temporal domain for action recognition.

The earliest attempts of skeleton-based action recognition treat all the body joints in sequence as a feature vector, and use a classifier such as SVM to classify the feature vector (Vemulapalli et al., 2014). These methods rarely explore the spatial and temporal dependencies of the skeleton sequence and cannot capture the fine-grained information of human action. Due to the rapid progress of deep learning, models based on convolutional neural networks (CNN) and recurrent neural networks (RNN) have become the mainstream, which normally regard the coordinates of human joints as pseudo-images or vector sequences (Soo Kim et al., 2017; Ke et al., 2017; Song et al., 2017; Cao et al., 2018). Although these methods have the ability to exploit spatial-temporal information, they are only suitable for dealing with the regular data in Euclidean space, and are not suitable for handling the graph data in non-Euclidean space. The skeleton is naturally structured as a graph in a non-Euclidean space with the characteristic that the joints as vertexes and their natural connections in the human body as edges.

In order to better leverage the skeleton data in the non-Euclidean space, Yan et al. firstly applied the graph convolutional network (GCN) to model the skeleton-based action recognition (Yan et al., 2018). They proposed a spatial-temporal graph convolutional network (ST-GCN), which constructs a spatial graph based on the natural connections of joints in the human body and adds the temporal edges between corresponding joints in consecutive frames. ST-GCN can aggregate the information of graph vertexes in both spatial and temporal domain to obtain a discriminative feature representation of vertexes, and then uses an average pooling layer on both spatial and temporal domain to get the feature of spatial-temporal graph for action classification. However, there are two disadvantages of ST-GCN: (1) ST-GCN only considers the temporal dependency between adjacent frames on the time sequence, it cannot fully exploit the temporal dependency between frames with multi-scale time span. Besides, the pose variation between adjacent frames is small, which usually cannot reflect the motion information of human action. (2) ST-GCN simply uses average pooling to obtain the global graph feature representation from vertex feature representations, without paying attention to key joints and key frames in the skeleton sequence, thus losing a lot of fine-grained information for action classification. For example, we should pay more attention to the long-term variations of the human hands and upper limbs for the actions "reading" and "writing". Because in the process of reading or writing, human body is mainly moving with its upper limbs, and the movement is very slow. In contrast, for "running" and "hopping", we should pay more attention to the instantaneous movement of human lower limbs. In other words, for different actions, different parts of the human body have different degrees of importance, and their movement speed is also very different. Therefore, how to fully exploit the attentive multi-scale spatial-temporal dependencies of human body joints is one of the crucial problems in skeleton-based action recognition.

To address this issue, a novel spatial attentive and temporal dilated graph convolutional network (SATD-GCN) is proposed in this work. Specifically, in the spatial domain, we propose a spatial attention pooling (SAP) module, which uses the self-attention mechanism to pick important vertexes and remove unimportant vertexes in the graph. In this way, it carries out an spatial attention pooling in the process of spatial graph convolution, which avoids the loss of fine-gained information and reduces the impact of noise caused by average pooling. It should be noted that although unimportant vertexes are removed but their useful information is preserved, because before pooling, their useful features have been aggregated on other vertexes by the spatial graph convolution. In the temporal domain, to give the network multi-scale temporal perception field, we propose a temporal dilated graph convolution (TDGC) module. Similar to the dilated convolution, TDGC extracts the non-adjacent graph sequence with multi-scale interval to expand the temporal receptive field. Both the SAP module and TDGC module can be easily embedded into the spatial-temporal graph convolution models, and significantly improves the performance of them. (see Section 5). Although the latest research on skeleton-based action recognition also uses spatial-temporal attention mechanism to refine features (Si et al., 2019; Huang et al., 2020), in contrast, the purposed SAP module not only refine features, but also reduces the number of graph vertexes properly and alleviates the influence of data redundancy and noise. Moreover, following the work of 2s-AGCN (Shi et al., 2019), we also use the lengths and directions of bones as the second-order information to construct a two-stream (i.e. joint stream and bone stream) SATD-GCN to boost the precision.

The main contribution of this work lies in three folds:

- A spatial attention pooling module is designed to adaptively capture important vertexes and remove unimportant vertexes in the graph, which is effective to reduce the number of graph vertexes and enhance the extraction of discriminative vertex features.
- A temporal dilated graph convolution module is exploited to expand the receptive field of temporal graph convolution, which can adapt to different speed of joint movement in different action and learn temporal features from subtle motions to large-scale motions hierarchically.
- A two-stream spatial attentive and temporal dilated graph convolutional network is constructed by combining the SAP module and the TDGC module, which outperforms the state-of-the-art methods for skeleton-based action recognition.

2 Related work

2.1 Skeleton-based action recognition

Conventional methods for skeleton-based action recognition usually design handcrafted features, i.e., relative positions of joints (Vemulapalli et al., 2014) or rotations, translations between body parts (Vemulapalli et al., 2016), etc., to represent human motion. However, these methods cannot effectively extract the spatial-temporal correlation of skeleton sequence in a wide range, thus the performance of these handcrafted-feature-based methods is unsatisfied. With the collection of skeleton data becomes easy and the development of deep learning technology, using the deep networks for data-driven feature learning has become the mainstream for skeleton-based action recognition. Shahroudy et al. (2016) treat 3D coordinates of all joints of human body in time sequence as a vector sequence and then use RNN to extract temporal information. Similar to (Shahroudy et al., 2016), there are many RNN-based methods has beend proposed and have obtained good results (Song et al., 2017; Cao et al., 2018; Du et al., 2015; Liu et al., 2016; Zhang et al., 2017). However, in these RNN-based methods, the graph structure of human body joints is directly regarded as vectors, leading to the spatial structure information of human body has been ignored. To solve this problem, CNN-based methods have been studied to model the skeleton data as a pseudo-image on the manually designed transformation rules (Soo Kim et al., 2017; Ke et al., 2017; Liu et al., 2017a; Liu et al., 2017b; Li et al., 2017b; Li et al., 2017a), which do not directly process the graph structure skeleton data in the non-Euclidean space. CNN-based methods increase a large amount of redundant computation, and cannot fully exploit the human body structure.

Recently, GCN-based methods promote the performance of the skeleton-based action recognition to a higher level (Yan et al., 2018; Shi et al., 2019; Si et al., 2018; Thakkar et al., 2018; Li et al., 2019b; Zhao et al., 2019; Wen et al., 2019), which construct a skeleton graph whose vertices are joints and edges are bones, and apply GCN to extract correlated features. The existed GCN-based methods can be roughly divided into two categories. The first type of approach leverages GCN to extract spatial correlation of skeleton graph, and then uses RNN to capture the temporal correlation (Huang et al., 2020; Zhao et al., 2019). The second type of approach uses spatial-temporal GCN to process the graph sequence directly (Yan et al., 2018; Shi et al., 2019; Li et al., 2019b; Wen et al., 2019), which can adapt to the non-Euclidean space data in time sequence very well, and achieves the state-of-the-art performance. Yan et al. (2018) firstly proposed the ST-GCN, in which each ST-GCN layer constructs the spatial characteristics with a graph convolutional operation, and models the temporal dynamic with a temporal convolutional operation. Li et al. (2019b) introduced an encoder-decoder structure to capture richer joint correlations and action-specific latent vertexes dependencies. Wen et al. (2019) explored a motif-based graph convolution to encode the hierarchical spatial structure, and applied a variable temporal dense block to exploit local temporal information over different ranges of human skeleton sequences. Although these studies optimize the extraction of spatial-temporal features of the skeleton graph sequence, the pooling method they used in both the spatial domain and the temporal domain is simple, which cannot effectively retain important features, and is also vulnerable to noise. In addition, these methods do not have the multi-scale temporal perception field, therefore they are unable to deal with the different length of graph sequence and the different speed of human body movement well. Following GCN-based methods, our model combines TDGC module and SAP module proposed in this work to extract spatial-temporal features more effectively.

2.2 Graph convolutional network

In the real world, many data are in the irregular non-Euclidean space such as the molecular structure (Zitnik et al., 2017), transportation network (Zhou et al., 2020), social network (Tang et al., 2009), and skeleton graph (Yan et al., 2018). Therefore, how to improve the feature extraction ability of the deep model in the non-Euclidean space is a hot research topic recently. Scarselli et al. (2008) first proposed a graph neural network (GNN) that can handle the graph structure data. GNN is a trainable model which is able to aggregate vertex information in terms of the manually designed rules in the graph structure. Defferrard et al. (2016) used the Fourier transform of graph structure data to expand the convolution operation into the non-Euclidean space, and propose the graph convolutional network (GCN) for graph classification. Kipf et al. (2016) applied the GCN for semi-supervised learning, and verified the validity of GCN. However, these convolution methods operate the graph structure data in the spectral domain, so the computational speed is very inefficient. Later, Monti et al. (2017) modified the spectral domain GCN to construct a more effective spatial domain GCN, which directly operates on the graph vertexes and avoids the complex steps e.g., the Fourier transform and the Chebyshev polynomial approximation. Our work also use GCN to model the skeleton-based action recognition problem, and we follow the work of ST-GCN (Yan et al., 2018) to extract the feature of human body joints from both spatial dimension and temporal dimension.

3 Background

In this section, we introduce the basic background knowledge of this work.

3.1 Notations

In this paper, we use $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the skeleton graph, where $\mathcal{V}$ is the set of n body joints and $\mathcal{E}$ is a set of m bones. We consider the adjacent matrix of the skeleton graph as $A \in \{0,1\}^{n\times n}$, where $A_{i,j} = 1$ if the i-th and the j-th joints are connected and 0 otherwise. Let $D \in \mathbb{R}^{n\times n}$ be the diagonal degree matrix, where $D_{i,i} = \sum_j A_{i,j}$. Following the work of ST-GCN (Yan et al., 2018), we divide one root vertex and its one-order neighbors into three sets, including (1) the root vertex itself, (2) the centripetal group, which is closer to the body barycenter than the root, and (3) the centrifugal group, which is farther away to the body barycenter than the root. In this way, A is accordingly classified to be A^{root}, $A^{centripetal}$ and $A^{centrifugal}$, which can better express the structural information of the skeleton graph. We denote the partition group set as $P = \{root, centripetal, centrifugal\}$ and $\sum_{p\in P} A^p = A$. Let $X \in \mathbb{R}^{n\times 3\times T}$ be the 3D joint positions across T frames. Let $X_t = X_{:,:,t} \in \mathbb{R}^{n\times 3}$ be the 3D joint positions at the t-th frame, which slices the t-th frame in the last dimension of X. $X_t^i = X_{i,:,t} \in \mathbb{R}^3$ be the positions of the i-th joint at the t-th frame.

3.2 Spatial-temporal GCN

ST-GCN (Yan et al., 2018) consists of a series of ST-GCN blocks. Each block contains a spatial GCN layer followed by a temporal GCN layer, which can extract spatial and temporal features alternatively. In the spatial dimension, the convolution operation on the skeleton graph is:

$$X_{out} = \sum\nolimits_{p\in P} \widetilde{A^p} X_{in} W^p. \tag{1}$$

Where $X_{in} \in \mathbb{R}^{n\times d_{in}}$ and $X_{out} \in \mathbb{R}^{n\times d_{out}}$ are the input and the output features of all joints in one frame respectively, and d_{in} and d_{out} is the channel dimension of them. $\widetilde{A^p} = {D^p}^{-\frac{1}{2}} A^p {D^p}^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}$ is the normalized adjacent matrix of each partition. $W^p \in \mathbb{R}^{d_{in}\times d_{out}}$ is the trainable weights for each partition in spatial GCN. In ST-GCN, the adjacent matrix A is manually defined according to the physical structure of human body, which cannot adaptively represent the interdependence of different parts of human body in different actions. Following the work of AGCN (Shi et al., 2019), we reconstruct Eq.1 as following:

$$X_{out} = \sum\nolimits_{p\in P} \left(\widetilde{A^p} + B^p + C^p\right) X_{in} W^p. \tag{2}$$

Where $B^p \in \mathbb{R}^{n\times n}$ is a trainable adjacent matrix that can be optimized together with the other parameters in the training process. There are no constraints on the value of B^p, which means that the graph is completely learned according to the training data. $C^p \in \mathbb{R}^{n\times n}$ is a vertex-dependent adjacent matrix which can determine whether there is a connection between two vertexes and how strong the connection is. We calculate C^p as follows:

$$C^p = softmax\left(\left(X_{in} W_\phi^p\right)\left({W_\theta^p}^T X_{in}^T\right)\right). \tag{3}$$

Where $W_\phi \in \mathbb{R}^{d_{in}\times n}$ and $W_\theta \in \mathbb{R}^{d_{in}\times n}$ are the trainable parameters of the embedding functions. The $softmax(\cdot)$ function operates on each row of the matrix.

For the temporal dimension, since the corresponding vertexes in continuous graph frames are the linear structure, it is straightforward to perform the temporal graph convolution similar to the classical convolution operation. Specifically, we perform a 2D convolution on the output feature map calculated by the spatial convolution with a $K_t \times 1$ kernel, where K_t is the kernel size of the temporal dimension.

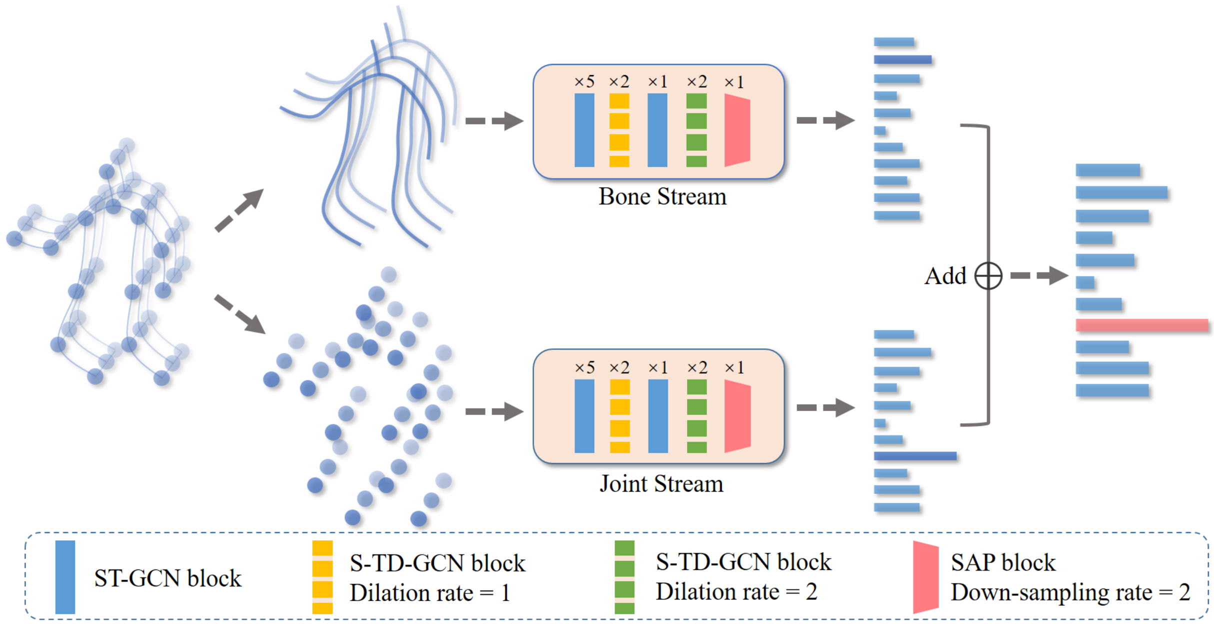
4 Spatial attentive and temporal dilated GCN

In this section, we introduce the components of our proposed spatial attentive and temporal dilated graph convolutional network (SATD-GCN) in detail.

4.1 Model architecture

Our model consists of two streams, i.e., a joint stream and a bone stream. The joint stream takes human body joints as graph vertexes and bones as graph edges to construct the skeleton graph sequence, and the initial feature of the vertex is its 3D coordinate corresponding to the human body joint. The bone stream takes human bones as graph vertexes and joints as graph edges, and the initial feature of the bone is the coordinate of the target joint minus the coordinate of the source joint. We define the joint close to the center of gravity of the skeleton is the source joint and the joint far

Figure 1 The overall architecture of the proposed SATD-GCN.



away from the center of gravity is the target joint. For example, given a bone with its source joint $v_1 = (x_1, y_1, z_1)$ and its target joint $v_2 = (x_2, y_2, z_2)$, the initial feature of the bone is calculated as $v_2 - v_1 = (x_2 - x_1, y_2 - y_1, z_2 - z_1)$. The overall architecture of the SATD-GCN is shown in Figure 1. Given a sample, we first calculate the data of bones based on the data of joints. Then, the joint data and bone data are fed into the joint stream and the bone stream, respectively. In the two stream network, we first apply five ST-GCN blocks to extract low-level feature of the vertexes. Then, we apply two spatial and temporal dilated graph convolution blocks (S-TD-GCN) with a dilation rate 1 followed by a ST-GCN block to extract high-level feature of the vertexes. Next, we use two S-TD-GCN blocks with a dilation rate 2 followed by a spatial attention pooling block (SAP) with down-sampling rate 2 to further extract high-level features and capture the important vertexes while remove the unimportant vertexes. Finally, we apply the average pooling to a few of remaining important vertexes on both the spatial and temporal domain. The $softmax$ scores of the two streams are combined to obtain the fused score for the action label prediction.

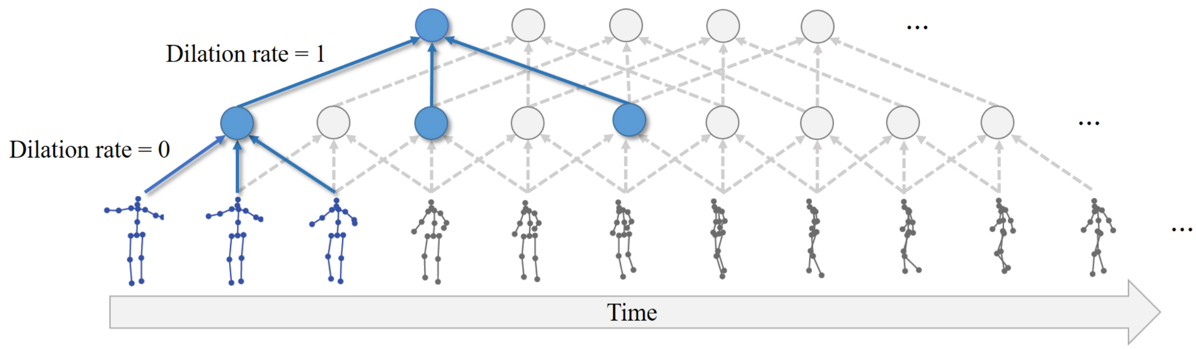
4.2 Temporal dilated graph convolution module

The spatial-temporal GCN first aggregates vertex information in the spatial domain based on the spatial adjacency of the skeleton graph. With the help of multiple adaptive adjacency matrices and the vertex subset partition, ST-GCN can adapt to different spatial correlations of human body joints. However, in the temporal domain, ST-GCN doesn't have the ability to extract multi-scale correlations of non-adjacent frames. This disadvantage makes the ST-GCN cannot adapt to different speed and time span of various human actions perfectly. In order to explore the time sequence information and human motion feature more effectively, we propose a novel temporal dilated graph convolution module (TDGC module). As shown in Figure 2, we use the temporal convolution with continuous kernel to extract low-level features. When to extract high-level features, we let the temporal kernels have gaps and we call the size of this gap is the "dilation rate". By dilated the temporal graph convolution, our model can learn the dependence between non-adjacent frames, and significantly expand the temporal perception field. In addition, by gradually increasing the dilation rate, our model is able to perceive the motion of human body at different time scales. In our SATD-GCN model, as shown in Figure 1, we apply two spatial- temporal dilated-graph convolution (S-TD-GCN) blocks with a dilation rate 1 after five ST-GCN blocks. And then we add two S-TD-GCN blocks with a dilation rate 2 after one ST-GCN block. Experiments show that this kind of structure can extract temporal features from subtle motions to large-scale motions hierarchically.

4.3 Spatial attention pooling module

In the large scale skeleton datasets i.e. NTU-RGBD or Kinetics-Skeleton, there are some poor information joints, such as the right and left ears can make little contribution to action recognition. On the other side, there are some relatively important joints. For example, most of the actions will have the movement information of the human body's left and right hands or feet. The previous methods usually compress vertex features by means of average pooling in both the temporal domain and the spatial domain (Yan et al., 2018; Shi et al., 2019; Li et al., 2019b), which will inevitably result in the loss of important spatial-temporal information. To solve this problem, we propose a spatial attention

Figure 2 The temporal dilated graph convolution.



pooling module (SAP module). As can be seen from Figure 3, the SAP module uses self-attention mechanism to select the important vertexes in the graph and remove the unimportant vertexes. At the same time, before filtering vertexes, the SAP module also utilizes the attention map to enhance the feature of vertexes (Element-wise multiplication). More specifically, because the SAP module works on the spatial dimension, we first use the temporal average pooling (T-AvgPool) on the skeleton graph sequence to reduce the temporal dimension to one and get a feature map which has $n \times d$ dimension. Furthermore, we use a fully-connected layer (FC) followed by a sigmoid function on the feature map to generate an attention map for each vertex in the graph, which can be interpreted as the relative importance given to vertex in the current graph. The feature of the original graph vertex is multiplied by its attention map to enhance the feature. We rank the attention map from large to small and filter row vectors of adjacency matrix according to the attention map by a down-sampling rate α. In this way, the original $n \times n$ dimensional adjacency matrix becomes $\frac{n}{\alpha} \times n$ dimensional. In SAP module, the physical connection structure of human body no longer exists, so we remove the physical structure adjacency matrix A and the trainable adjacency matrix B, only retain vertex-dependent adjacency matrix C in Eq.2. Finally, we perform ST-GCN operation with the new adjacency matrix on the skeleton graph, so the number of vertexes in the skeleton graph is reduced to $\frac{n}{\alpha}$. It should be noted that we do not directly filter the important vertexes, but indirectly select the vertexes by filtering the adjacency matrix, which can alleviate the non-differentiable problems caused by the selection operation and make the model easy to train. In our SATD-GCN, as shown in Figure 1, we apply one SAP block with the down-sampling rate 2 at the end of the two stream respectively.

5 Experiment

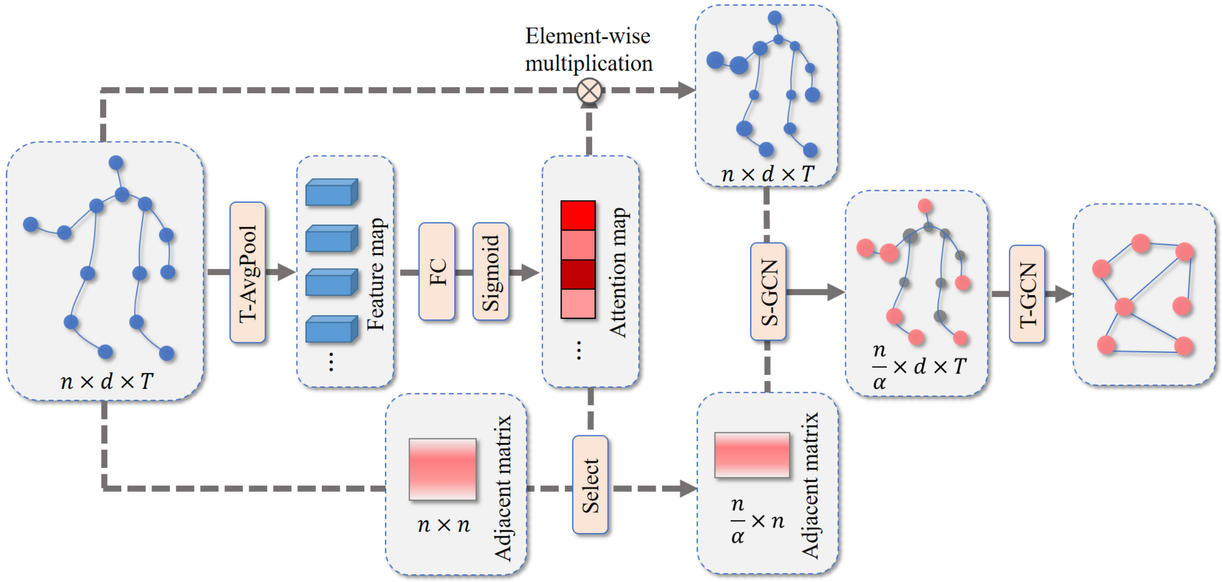
Extensive experiments are conducted and analyzed in this section. Firstly, we introduce two large-scale skeleton datasets, namely NTU-RGB+D (Shahroudy et al., 2016) and Kinetics-Skeleton (Yan et al., 2018). Secondly, our model implementation details and the training details are discussed. Thirdly, we perform an ablation study of each component. Finally, our model is evaluated on these two datasets to compare with the state-of-the-arts.

5.1 Datasets

NTU-RGB+D: NTU-RGB+D a large in-door-captured dataset with annotated 3D joint coordinates for the human action recognition task (Shahroudy et al., 2016). NTU-RGB+D contains 56,000 action videos in 60 action classes, which are captured from 40 volunteers in different age groups ranging from 10 to 35. Each action is obtained by 3 cameras at the same height from different viewpoints, and the provided annotations are given in the camera coordinate system. There are 25 joints for each subject in the skeleton sequences, while each action video has no more than 2 subjects. It includes two settings: (1) Cross-Subject (CS) benchmark, which contains 40,320 videos for training and 16,560 for evaluation. In this setting, the training set comes from one subset of 20 subjects and a model is validated on sequences from the remaining 19 subjects; 2) Cross-View (CV) benchmark, which includes 37,920 videos for training and 18,960 videos for evaluation. The training samples in this setting come from the camera views 2 and 3, and the evaluation samples are all from the camera view 1. We follow the conventional settings and report the top-1 accuracy on both benchmarks.

Kinetics-Skeleton: Kinetics (Kay et al., 2017) consists of 300,000 videos clips in 400 action classes. The video clips of Kinetics are sourced from YouTube and have a great variety, but it only provides raw video clips without skeleton information. Yan et al. (2018) estimate the locations of 18 joints on every frame of the clips using the publicly available OpenPose toolbox and release the Kinetics-Skeleton datasets. In Kinetics-Skeleton, all videos are resized to a resolution of 340×256 and are converted to a frame rate of 30fps. The toolbox generates 2D coordinate and confidence score for totally 18 joints from the resized videos. For the multi-person clips, two people are selected based on the average joint confidence. Each joint is represented as a three-element feature vector that contains the 2D coordinate and confidence

Figure 3 The architecture of the spatial attention pooling module.



score. Following the evaluation method of Yan et al. (2018), we train the models on the training set and report the top-1 and top-5 accuracies on the validation set.

5.2 Implementation details

Our SATD-GCN model has a total of 12 blocks. In each block, we add a residual connection (He et al., 2016), which enables the model to learn features more effectively and prevents over-fitting. The output channels for each block are 64, 64, 64, 128, 128, 128, 256, 256, 256, 256 and 256. We set the down-sampling rate $\alpha = 2$ and the temporal kernel size $K_t = 9$. A data BN layer is added at the beginning to normalize the input data. The final output is sent to a softmax classifier to obtain the action prediction.

We implement our SATD-GCN model based on the PyTorch deep learning framework (Paszke et al., 2019). We apply the stochastic gradient descent (SGD) algorithm with Nesterov momentum (0.9) as the optimizer. The weight decay is set to 0.0001. We use a Titan XP GPU for the model training and the batch size is 16.

For the NTU-RGB+D dataset, the max number of frames in each sample is 300. We repeat the samples until it reaches to 300 frames if the samples have frames less than 300. There are at most two human bodies in each sample. If the number of bodies in the sample is less than 2, we pad the second body with 0. The number of training epoch is set as 55 and the learning rate is set as 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, there are 150 frames in each sample and 2 bodies in each frame. We randomly choose 150 frames from the input skeleton sequence and slightly disturb the joint coordinates with randomly chosen rotations and translations for data-augmentation. 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 45th epoch, 55th epoch and 65th epoch.

5.3 Ablation study

We test the effectiveness of the components of our SATD-GCN with the Cross-View benchmark on the NTU-RGB+D dataset. We only test one stream (joint stream) in our model as the Bone Stream can be conducted in the same way. From Table 1, we can see that the original performance of one stream ST-GCN (Yan et al., 2018) on the NTU-RGB+D Cross-View benchmark is 88.3%. By applying the adaptive adjacent matrix and the specially designed data preprocessing methods, (Shi et al., 2019) designed an AGCN, where its performance is improved to 93.4%. We use AGCN as the baseline in the work.

To further boost the performance of ST-GCN, we propose two novel modules to effectively learn the temporal and spatial features in the skeleton data, i.e., the TDGC module and the SAP module. The results in the third row and the fourth row show that that either the TDGC module or the SAP module is beneficial for action recognition. Specifically, compared to the baseline AGCN, the TDGC module boosts the performance by 0.6% (94.0% vs 93.4%) and the SAP module enhances the performance by 0.5% (93.9% vs 93.4%), respectively. When integrating these two modules together, the joint stream SATD-GCN obtains the best performance, which improves the accuracy by 1.0% (94.4% vs 93.4%). Experiments demonstrate that both the TDGC module and the SAP module are effective. They can help the network to learn multi-scale and discriminative spatial-temporal features, so as to improve the accuracy of action classification.

Figure 4 visualizes of the attention map in the SAP module. The skeleton graph are plotted based on the physical connections of the human body. Each circle

Figure 4 Visualization of the attention map in the SAP module. The radius size of the circle represents the weight of the joint.

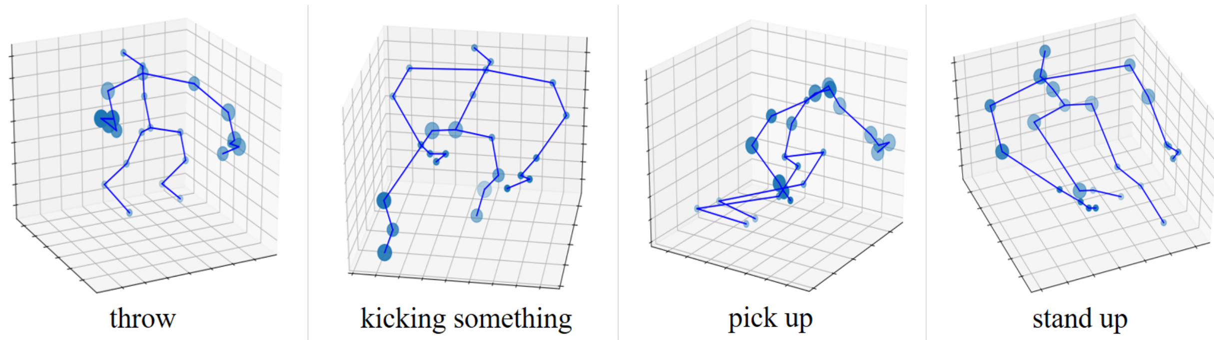


Table 1 Comparisons of the validation accuracy on the NTU-RGB+D Cross-View benchmark with the joint stream SATD-GCN, with or without the TDGC module and the SAP module. w/o means without.

Methods	Accuracy (%)
ST-GCN (Yan et al., 2018)	88.3
AGCN (Shi et al., 2019)	93.4
SATD-GCN (JS) w/o TDGC	93.9
SATD-GCN (JS) w/o SAP	94.0
SATD-GCN (JS)	**94.4**

Table 2 Comparisons of the validation accuracy on the NTU-RGB+D Cross-View benchmark of the joint stream SATD-GCN with different configuration.

Model Configuration	Accuracy (%)
(S-TD-GCN×4, dilation rate=1)	94.1
(S-TD-GCN×2, dilation rate=1)+ (S-TD-GCN×2, dilation rate=2)	**94.4**
(S-TD-GCN×2, dilation rate=2)+ (S-TD-GCN×2, dilation rate=3)	92.8
(SAP×1, down-sampling rate=2)	**94.4**
(SAP×1, down-sampling rate=3)	93.7

represents one joint, and the radius size represents the weight of the joint. It can be seen that for the action "throw", the model pays more attention to the movement of human hands; while for "kicking something", the joints of feet are given the higher weight. For the actions "pick up" and "stand up", the joints of the upper body contain more information, and are selected in the SAP module.

Table 2 shows the effect of different dilation rate in the TDGC module and different down-sampling rate in the SAP module. For the dilation rate, increasing the dilation rate gradually can let the model extracts temporal features from subtle motions to large-scale motions hierarchically and effectively. The joint stream SATD-GCN model, which has two ST-GCN blocks followed by two S-TD-GCN blocks with dilation rate 1 and 2 respectively, obtains the best performance. For the down-sampling rate α, the SAP module with $\alpha = 2$ is the best configuration in our experiment. It should be noted that if the dilation rate and the down-sampling rate are too large, the performance of the model will be damaged. Because the process of padding in the TDGC module and the conduction of deleting vertexes in the SAP module will lead the model to lose some useful information. When the dilation rates of the two S-TD-GCN blocks are set to 2 and 3 respectively, the accuracy is even lower than the baseline (92.8% vs 93.4%). Because the dependence between two frames with too large time span is very weak, and the TDGC module will destroy the fine-grained temporal feature instead.

5.4 Comparisons to the state-of-the-arts

We compare the proposed SATD-GCN model (two-stream) with the state-of-the-art skeleton-based action recognition methods on both the NTU-RGB+D dataset and the Kinetics-Skeleton dataset. The methods which we selected for comparison include the handcraft-feature-based methods (Vemulapalli et al., 2016; Fernando et al., 2015), the RNN-based methods (Shahroudy et al., 2016; Du et al., 2015; Liu et al., 2016; Song et al., 2017; Zhang et al., 2017; Li et al., 2018a; Li et al., 2018b), the CNN-based methods (Liu et al., 2017a; Soo Kim et al., 2017; Ke et al., 2017; Liu et al., 2017b; Li et al., 2017b; Li et al., 2017a), and the GCN-based methods (Yan et al., 2018; Tang et al., 2018; Li et al., 2019b; Wen et al., 2019; Shi et al., 2019). Results on the NTU-RGB+D dataset are shown in Table 3. Our SATD-GCN outperforms the handcraft-feature-based methods, RNN based methods and CNN based methods for more than $> 4\%$ on both the Cross-Subject and the Cross-View benchmarks, which proved that GCN has great advantages in dealing with skeleton data. Among the GCN-based methods, our SATD-GCN also achieves the state-of-the-art performance. Comparing to ST-GCN (Yan et al., 2018), the improvements of our method reach to 7.8% (89.3% vs 81.5%) and 7.2% (95.5% vs 88.3%) on the Cross-Subject benchmark and the Cross-View benchmark respectively. For the most related work 2s-AGCN (Shi et al., 2019), our results outperform it by 1.1% (89.3% vs 88.2%) on the Cross-Subject benchmark and 0.6% (95.5% vs 94.9%) on the Cross-

Table 3 Comparison of the validation accuracy with state-of-the-art methods on the NTU-RGB+D dataset. X-S means Cross-Subject, X-V means Cross-View.

Methods	X-S (%)	X-V (%)
Lie Group (Vemulapalli et al., 2016)	50.1	82.8
HBRNN (Du et al., 2015)	59.1	64.0
Deep LSTM (Shahroudy et al., 2016)	60.7	67.3
ST-LSTM (Liu et al., 2016)	67.2	77.7
STA-LSTM (Song et al., 2017)	73.4	81.2
VA-LSTM (Zhang et al., 2017)	79.2	87.7
ARRN-LSTM (Li et al., 2018a)	80.7	88.8
Ind-RNN (Li et al., 2018b)	81.8	88.0
Two-Stream 3DCNN (Liu et al., 2017a)	66.8	72.6
TCN (Soo Kim et al., 2017)	74.3	83.1
Clips+CNN+MTLN (Ke et al., 2017)	79.6	84.8
Synthesized CNN (Liu et al., 2017b)	80.0	87.2
CNN+Motion+Trans (Li et al., 2017b)	83.2	89.3
3scale ResNet152 (Li et al., 2017a)	85.0	93.2
ST-GCN (Yan et al., 2018)	81.5	88.3
DPRL+GCNN (Tang et al., 2018)	83.5	89.8
motif-GCNs+VTDB (Li et al., 2019b)	84.2	94.2
AS-GCN (Wen et al., 2019)	86.8	94.2
2s-AGCN (Shi et al., 2019)	88.2	94.9
SATD-GCN (Ours)	**89.3**	**95.5**

Table 4 Comparison of the validation accuracy with state-of-the-art methods on the Kinetics-Skeleton dataset.

Methods	Top-1 (%)	Top-5 (%)
Feature Enc. (Fernando et al., 2015)	14.9	25.8
Deep LSTM (Shahroudy et al., 2016)	16.4	35.3
TCN (Soo Kim et al., 2017)	20.3	40.0
ST-GCN (Yan et al., 2018)	30.7	52.8
AS-GCN (Wen et al., 2019)	34.8	56.5
2s-AGCN (Shi et al., 2019)	35.9	58.6
SATD-GCN (Ours)	**36.6**	**59.8**

View benchmark. It reveals that our SATD-GCN can better classify a variety of human actions by combining the TDGC module and the SAP module.

Table 4 shows the results of the Kinetics-Skeleton dataset, where we compare the proposed SATD-GCN with six state-of-the-art approaches. We can see that our SATD-GCN outperforms the other competitive methods in both Top-1 and Top-5 accuracies. Comparing to ST-GCN (Yan et al., 2018), the improvements of our method reach to 5.9% (36.6% vs 30.7%) and 7.0% (59.8% vs 52.8%) on Top-1 accuracy and Top-5 accuracy respectively. For the most related work 2s-AGCN (Shi et al., 2019), our results outperform it by 0.7% (36.6% vs 35.9%) on Top-1 accuracy and 1.2% (59.8% vs 58.6%) on Top-5 accuracy.

6 Conclusions

In this paper, we propose a novel spatial attentive and temporal dilated graph convolutional network (SATD-GCN), which contains a temporal dilated graph convolution module (TDGC) and a spatial attention pooling module (SAP), for skeleton-based action recognition. The TDGC module can effectively extract the temporal features of different time scales, improve the perception field in the temporal domain, and maintain the robustness of the model to different motion speed and sequence length. The SAP module can select human body joints which are beneficial for action recognition by self-attention mechanism, and alleviate the influence of data redundancy and noise. In addition, both the TDGC module and the SAP module can be easily incorporated into the spatial-temporal graph convolution networks (ST-GCN), and significantly improve the performance of ST-GCN. Due to the contribution of these two modules, our SATD-GCN achieves the state-of-the-art performance on two large-scale action recognition datasets.

References

Baranwal, N., & Nandi, G. C. (2017). 'A mathematical framework for possibility theory-based hidden Markov model', *International Journal of Bio-Inspired Computation*, 10(4), pp.239–247.

Cao, C., Lan, C., Zhang, Y., Zeng, W., Lu, H., & Zhang, Y. (2018). 'Skeleton-Based Action Recognition with Gated Convolutional Neural Networks', *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11). pp.3247–3257.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). 'Realtime multi-person 2d pose estimation using part affinity fields', In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7291–7299.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). 'Cascaded pyramid network for multi-person pose estimation', In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.7103–7112.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). 'Convolutional neural networks on graphs with fast localized spectral filtering', In *Advances in neural information processing systems*, pp.3844–3852.

Du, Y., Wang, W., & Wang, L. (2015). 'Hierarchical recurrent neural network for skeleton based action recognition', In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1110–1118.

Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., & Tuytelaars, T. (2015). 'Modeling video evolution for action recognition', In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5378–5387.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). 'Deep residual learning for image recognition', In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770–778.

Huang, J., Xiang, X., Gong, X., & Zhang, B. (2020). 'Long-Short Graph Memory Network for Skeleton-based Action Recognition', In *The IEEE Winter Conference on Applications of Computer Vision*, pp.645–652.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., & Suleyman, M. (2017). 'The kinetics human action video dataset', *arXiv preprint arXiv:1705.06950.*

Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2017). '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.

Kipf, T. N., & Welling, M. (2016). 'A new representation of skeleton sequences for 3d action recognition', *arXiv preprint arXiv:1609.02907.*

Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., & He, M. (2017, July). 'Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN', In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp.601–604.

Li, C., Zhong, Q., Xie, D., & Pu, S. (2017, July). 'Skeleton-based action recognition with convolutional neural networks', In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp.597–600.

Li, L., Gan, Z., Cheng, Y., & Liu, J. (2019). 'Relation-aware graph attention network for visual question answering', In *Proceedings of the IEEE International Conference on Computer Vision*, pp.10313–10322.

Li, L., Zheng, W., Zhang, Z., Huang, Y., & Wang, L. (2018). 'Relation-aware graph attention network for visual question answering', *arXiv preprint arXiv:1805.02556.*

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2019). '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.

Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (2018). 'Independently recurrent neural network (indrnn): Building a longer and deeper rnn', In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.5457–5466.

Liu, H., Tu, J., & Liu, M. (2017). 'Two-stream 3d convolutional neural network for skeleton-based action recognition', *arXiv preprint arXiv:1705.08106.*

Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016, October) 'Spatio-temporal lstm with trust gates for 3d human action recognition', In *European conference on computer vision*, pp.816–833.

Liu, M., Liu, H., & Chen, C. (2017). 'Enhanced skeleton visualization for view invariant human action recognition', *Pattern Recognition*, 68, pp.346–362.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). 'Geometric deep learning on graphs and manifolds using mixture model cnns', In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5115–5124.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Desmaison, A. (2019). 'PyTorch: An Imperative Style, High-Performance Deep Learning Library', In *Advancesin Neural Information Processing Systems*, pp.8024–8035.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). 'The graph neural network model', *IEEE Transactions on Neural Networks*, 20(1), pp.61–80.

Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). '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.

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

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). '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.

Si, C., Jing, Y., Wang, W., Wang, L., & Tan, T. (2018). 'Skeleton-based action recognition with spatial reasoning and temporal stack learning', In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.103–118.

Simonyan, K., & Zisserman, A. (2014). 'Two-stream convolutional networks for action recognition in videos', In *Advances in neural information processing systems*, pp.568–576.

Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017, February). 'An end-to-end spatio-temporal attention model for human action recognition from skeleton data', In *Thirty-first AAAI conference on artificial intelligence*.

Soo Kim, T., & Reiter, A. (2017). 'Interpretable 3d human action analysis with temporal convolutional networks', In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.20–28.

Suhail, M., & Sigal, L. (2019). 'Mixture-Kernel Graph Attention Network for Situation Recognition', In *Proceedings of the IEEE International Conference on Computer Vision*, pp.10363–10372.

Tang, L., & Liu, H. (2009, June). 'Relational learning via latent social dimensions', In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.817–826.

Tang, Y., Tian, Y., Lu, J., Li, P., & Zhou, J. (2018). 'Deep progressive reinforcement learning for skeleton-based action recognition', In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5323–5332.

Thakkar, K., & Narayanan, P. J. (2018). 'Part-based graph convolutional network for action recognition', *arXiv preprint arXiv:1809.04983*.

Tu, Z., Li, H., Zhang, D., Dauwels, J., Li, B., & Yuan, J. (2019). 'Action-stage emphasized spatiotemporal vlad for video action recognition', *IEEE Transactions on Image Processing*, 28(6), pp.2799–2812.

Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B., & Yuan, J. (2018). 'Multi-stream CNN: Learning representations based on human-related regions for action recognition', *Pattern Recognition*, 79, pp.32–43.

Vemulapalli, R., & Chellapa, R. (2016). '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.

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). '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.

Wen, Y. H., Gao, L., Fu, H., Zhang, F. L., & Xia, S. (2019, July). 'Graph CNNs with motif and variable temporal block for skeleton-based action recognition', In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp.8989–8989.

Yan, S., Xiong, Y., & Lin, D. (2018, April). 'Spatial temporal graph convolutional networks for skeleton-based action recognition', In *Thirty-second AAAI conference on artificial intelligence*.

Yousefi, B., & Loo, C. K. (2016). 'Slow feature action prototypes effect assessment in mechanism for recognition of biological movement ventral stream', *International Journal of Bio-Inspired Computation*, 8(6), pp.410–424.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). 'View adaptive recurrent neural networks for high performance human action recognition from skeleton data', In *Proceedings of the IEEE International Conference on Computer Vision*, pp.2117–2126.

Zhao, R., Wang, K., Su, H., & Ji, Q. (2019). 'Bayesian Graph Convolution LSTM for Skeleton Based Action Recognition', In *Proceedings of the IEEE International Conference on Computer Vision*, pp.6882–6892.

Zhou, Z., Wang, Y., Xie, X., Chen, L., & Liu, H. (2020). 'RiskOracle: A Minute-level Citywide Traffic Accident Forecasting Framework', *arXiv preprint arXiv:2003.00819.*.

Zitnik, M., & Leskovec, J. (2017). 'Predicting multicellular function through multi-layer tissue networks', *Bioinformatics*, 33(14), pp.1190–1198.

Corresponding author: Zhigang Tu.
Zhigang Tu is currently a professor at the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University.Wuhan 430079, China.
E-mail: tuzhigang@whu.edu.cn