Received: 30 August 2020 | Revised: 28 November 2020 | Accepted: 21 December 2020

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12012

ORIGINAL RESEARCH PAPER

A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition

Jiaxu Zhang[1] | Gaoxiang Ye[2] | Zhigang Tu[1] | Yongtao Qin[3] | Qianqing Qin[1] | Jinlu Zhang[1] | Jun Liu[4]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

[2]State Grid Wuhan Power Supply Company, Wuhan, China

[3]Shenzhen Infinova Ltd. Company, Shenzhen, China

[4]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore

Correspondence

Zhigang Tu, Luoyu Road 129, Wuhan University, Wuhan, 430079, China.
Email: tuzhigang@whu.edu.cn

Funding information

National Key Research and Development Program of China, Grant/Award Number: 2018YFB1600600

Abstract

Current studies have shown that the spatial-temporal graph convolutional network (ST-GCN) is effective for skeleton-based action recognition. However, for the existing ST-GCN-based methods, their temporal kernel size is usually fixed over all layers, which makes them cannot fully exploit the temporal dependency between discontinuous frames and different sequence lengths. Besides, most of these methods use average pooling to obtain global graph feature from vertex features, resulting in losing much fine-grained information for action classification. To address these issues, in this work, the authors propose a novel spatial attentive and temporal dilated graph convolutional network (SATD-GCN). It contains two important components, that is, 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 alleviates the influence of data redundancy and noise. The TDGC module can effectively extract the temporal features at 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 benchmark datasets, that is, NTU-RGB + D and Kinetics-Skeleton, demonstrate that the authors' method achieves the state-of-the-art performance for skeleton-based action recognition.

1 | INTRODUCTION

Human action recognition, which has a wide range of applications in intelligent video surveillance, human-machine interaction, medical service, and so forth, is still a challenging and unsolved problem [1–4]. 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 have been carried out on skeleton-based action recognition as it is robust 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 [5,6], which provides sufficient data for skeleton-based action recognition study 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 domains 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 [7]. These methods rarely explore the spatial and temporal dependencies of the skeleton sequence and cannot capture the fine-grained information of human action. Owing 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 considers the coordinates of human joints as pseudo-images or vector sequences [8–11]. 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 as edges.

To better leverage the data in the non-Euclidean space, some works process data directly on the graph structure [12,13]. Yan et al. firstly applied the graph convolutional network (GCN) to model the skeleton-based action recognition [14]. 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 domains to obtain a discriminative feature representation of vertexes, and then uses an average pooling layer on both spatial and temporal domains to get the feature of spatial-temporal graph for action classification. Based on the ST-GCN, many variants were explored [17,29,31], which typically introduce some incremental modules, for example the adaptive adjacency matrix [17], the actional-structural module [29], and the variable temporal module [31], to enhance the network capacity. However, there are two drawbacks of the ST-GCN based methods: (1) they only consider the temporal dependency between adjacent frames on the time sequence, causing them cannot fully exploit the temporal dependency between frames in a 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-based methods simply use 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, the human body is mainly moving with its upper limbs in a slow speed. 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 dependency 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 a 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 the average pooling. It should be noted that although unimportant vertexes are removed, 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 a multi-scale interval to expand the temporal receptive field. Both the SAP module and the TDGC module can be easily embedded into the spatial-temporal graph convolution networks, and significantly improves the performance (see Section 5). Although the latest research on skeleton-based action recognition also uses a spatial-temporal attention mechanism to refine the extracted features [15,16], in contrast, the purposed SAP module can not only refine features but also reduce the number of graph vertexes properly and alleviate the influence of data redundancy and noise. Moreover, following the work of 2s-AGCN [17], we also use the length and direction of bones as the second-order information to construct a two-stream (i.e. joint stream and bone stream) SATD-GCN to boost the accuracy.

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 actions and learn temporal features from subtle motion to large-scale motion 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 skeleton-based action recognition methods.

2 | RELATED WORK

2.1 | Skeleton-based action recognition

Conventional skeleton-based action recognition methods usually extracted handcrafted features, that is, relative positions of joints [7] or rotations, translations between body parts [18], 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. [19] treat

3D coordinates of all joints of the human body in time sequence as a vector sequence and then use RNN to extract the temporal information. Similar to [19], many RNN-based methods have been proposed and good results obtained [10,11,20–22]. 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 the human body is 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 [8,9,23–26], which do not directly process the graph structure skeleton data in the non-Euclidean space and increase a large amount of redundant computation.

Recently, GCN-based methods promote the performance of the skeleton-based action recognition to a higher level [14,17,27–31], which construct a skeleton graph whose vertices are joints and edges are bones and apply the 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 the skeleton graph and then uses RNN to capture the temporal correlation [16,30]. The second type of approach uses spatial-temporal GCN to process the graph sequence directly [14,17,29,31], which can be well adapted to the non-Euclidean space data in time sequence and achieves the state-of-the-art performance. Yan et al. [14] first proposed the ST-GCN, in which each ST-GCN layer constructs the spatial characteristic with a graph convolutional operation and models the temporal dynamic with a temporal convolutional operation. Li et al. [29] introduced an encoder-decoder structure to capture richer joint correlations and action-specific latent vertexes dependencies. Wen et al. [31] 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, resulting in some important features cannot be effectively retained and is vulnerable to noise. Besides, these methods do not have the multi-scale temporal perception field, therefore they are unable to deal with different length of the graph sequence and different speed of the human body movement well. Following the GCN-based methods, our model combines the TDGC module and the SAP module proposed 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 [32], the transportation network [33], the knowledge graph [46], and the skeleton graph [14]. Therefore, how to improve the feature extraction ability of the deep model in the non-Euclidean space is a pressing research topic. Scarselli et al. [34] first proposed a graph neural network (GNN) to handle the graph structure data, as GNN is a trainable model which is able to aggregate the vertex information in terms of the manually designed rules in the graph structure. Defferrard et al. [35] used the Fourier transform of the graph structure data to expand the convolution operation into the non-Euclidean space and proposed a graph convolutional network (GCN) for graph classification. Kipf et al. [36] applied the GCN for semi-supervised learning and verified the validity of GCN. However, these deep learning methods operate the graph structure data in the spectral domain, so the computational speed is inefficient. Monti et al. [37] 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 for example, the Fourier transform and the Chebyshev polynomial approximation. Our work also uses GCN to handle the skeleton-based action recognition, and we follow the work of ST-GCN [14] to extract the feature of human body joints from both the spatial dimension and temporal dimension.

3 | BACKGROUND

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

3.1 | Notations

We use $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ to represent the skeleton graph, where $\mathscr{V}$ is the set of n body joints and $\mathscr{E}$ is the set of m bones. We consider the adjacency matrix of the skeleton graph as $\mathrm{A} \in \{0,1\}^{n\times n}$, where $\mathrm{A}_{i,j} = 1$ 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 $\mathrm{D}_{i,i} = \sum_j \mathrm{A}_{i,j}$. Following the work of ST-GCN [14], we divide one root vertex and its one-order neighbours into three sets, including (1) the root vertex itself, (2) the centripetal group, which is closer to the body barycentre 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 $\mathrm{A}^{\mathrm{root}}$, Acentripetal and $\mathrm{A}^{\mathrm{centrifugal}}$, which can better express the structural information of the skeleton graph. We denote the partition group set as $P = \{\mathrm{root},\ \mathrm{centripetal},\ \mathrm{centrifugal}\}$ and $\sum_{p\in P} \mathrm{A}^p = \mathrm{A}$. Let $\mathscr{X} \in \mathbb{R}^{n\times 3\times T}$ be the 3D joint positions across T frames. Let $\mathrm{X}_t = \mathscr{X}_{:,:,t} \in \mathbb{R}^{n\times 3}$ be the 3D joint positions at the t-th frame, which slices the t-th frame in the last dimension of $\mathscr{X}$. $\mathrm{X}_t^i = \mathscr{X}_{i,:,t} \in \mathbb{R}^3$ be the positions of the i-th joint at the t-th frame.

3.2 | Spatial-temporal GCN

ST-GCN [14] consists of a series of ST-GCN blocks. Each block contains a spatial GCN layer followed by a temporal

GCN layer, which can extract the spatial and temporal features alternatingly. In the spatial dimension, the convolution operation on the skeleton graph is:

$$\mathrm{X}_{out} = \sum_{p \in P} \widetilde{\mathrm{A}^p}\, \mathrm{X}_{in} \mathrm{W}^p, \tag{1}$$

where $\mathrm{X}_{in} \in \mathbb{R}^{n \times d_{in}}$ and $\mathrm{X}_{out} \in \mathbb{R}^{n \times d_{out}}$ are the input and output features of all joints in one frame respectively, and d_{in} and d_{out} is the channel dimension of them. $\tilde{\mathrm{A}}^p = \mathrm{D}^{p-\frac{1}{2}} \mathrm{A}^p \mathrm{D}^{p-\frac{1}{2}} \in \mathbb{R}^{n \times n}$ is the normalized adjacency matrix for each partition. $\mathrm{W}^p \in \mathbb{R}^{d_{in} \times d_{out}}$ are the trainable weights for each partition in the spatial GCN. In ST-GCN, the adjacency 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. For example, clapping hands, there is a connection between two hands, but they are not physically connected. Follow the work of 2s-AGCN [17], we change Equation 1 to the following form:

$$\mathrm{X}_{out} = \sum_{p \in P} (\widetilde{\mathrm{A}^p} + \mathrm{B}^p + \mathrm{C}^p) \mathrm{X}_{in} \mathrm{W}^p, \tag{2}$$

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

$$\mathrm{C}^p = softmax\left(\left(\mathrm{X}_{in} \mathrm{W}_{\varphi}^{P} \right) \left({\mathrm{W}_{\theta}^{p}}^{T} {\mathrm{X}_{in}}^{T} \right) \right), \tag{3}$$

where $\mathrm{W}_\theta \in \mathbb{R}^{d_{in} \times n}$ and $\mathrm{W}_\varphi \in \mathbb{R}^{d_{in} \times n}$ are the trainable parameters of the embedding functions. *softmax* function operates on each row of the matrix.

For the temporal dimension, since the corresponding vertexes in continuous graph frames are linear structures, it is straightforward to perform the temporal graph convolution similar to the classical convolution operation. Concretely, we perform a 2D convolution on the output feature map calculated by spatial convolution with a $\mathrm{K}_t \times 1$ 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, that is, 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, which closes to the centre of gravity of the skeleton, as the source joint; and define the joint, which is far away from the centre of gravity, as 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 the 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 features 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 the 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 downsampling rate 2 to further extract high-level features and capture the important vertexes while removing the unimportant vertexes. Finally, we apply the average pooling to a few of the remaining important vertexes on both the spatial and temporal domains. The *softmax* scores of the two streams are combined to obtain the final 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 does not have the ability to extract multi-scale correlations of non-adjacent frames. This disadvantage makes the ST-GCN cannot adapt to various human actions which usually with different speed and time span. 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 a 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 using the temporal dilated graph convolution, our model can learn the dependence between non-adjacent frames, and

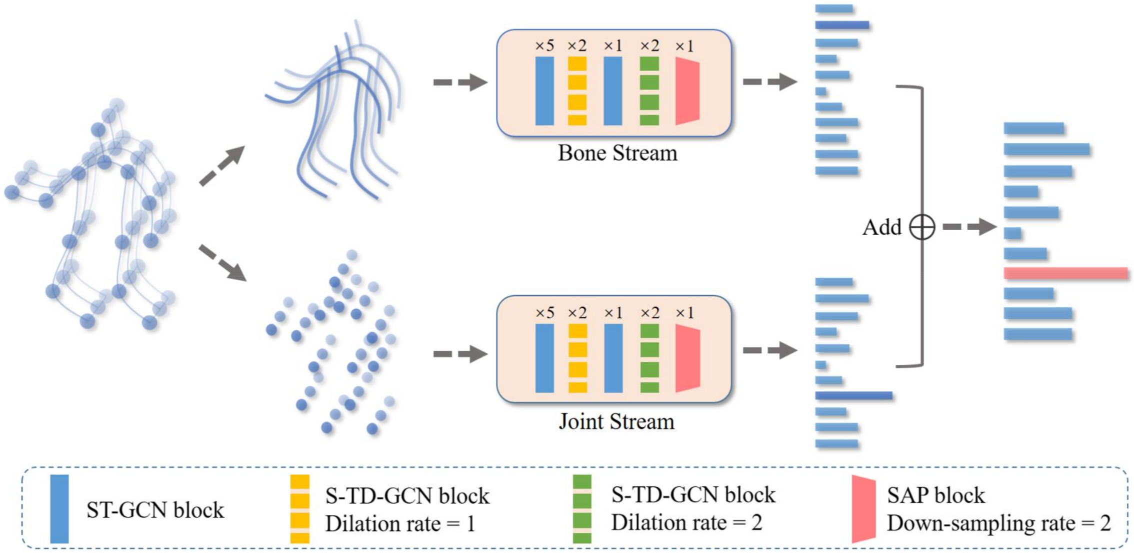

FIGURE 1 The overall architecture of the proposed SATD-GCN

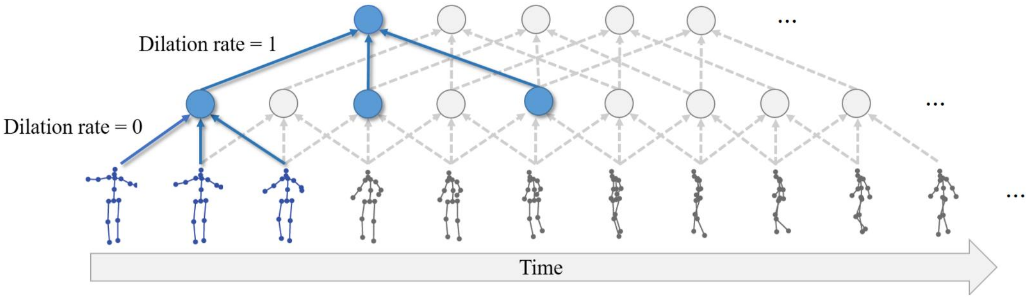

FIGURE 2 The temporal dilated graph convolution

significantly expands 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. It should be noted that the TDGC module does not increase the number of parameters and it can be easily combined with the ST-GCN based model. 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 of 1 after five ST-GCN blocks. And then we add two S-TD-GCN blocks with a dilation rate of 2 after one ST-GCN block. Experiments show that this kind of structure can extract temporal features from subtle motion to large-scale motion hierarchically.

4.3 | Spatial attention pooling module

In the large-scale skeleton datasets that is NTU-RGBD or Kinetics-Skeleton, there are some poor information joints, such as the right and left ears which 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 [14,17,29], which will inevitably be losing important spatial-temporal information. To solve this problem, we propose a spatial attention 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 the adjacency matrix according to the attention map by a down-sampling rate α. In this way, the original $n \times d$ dimensional adjacency matrix

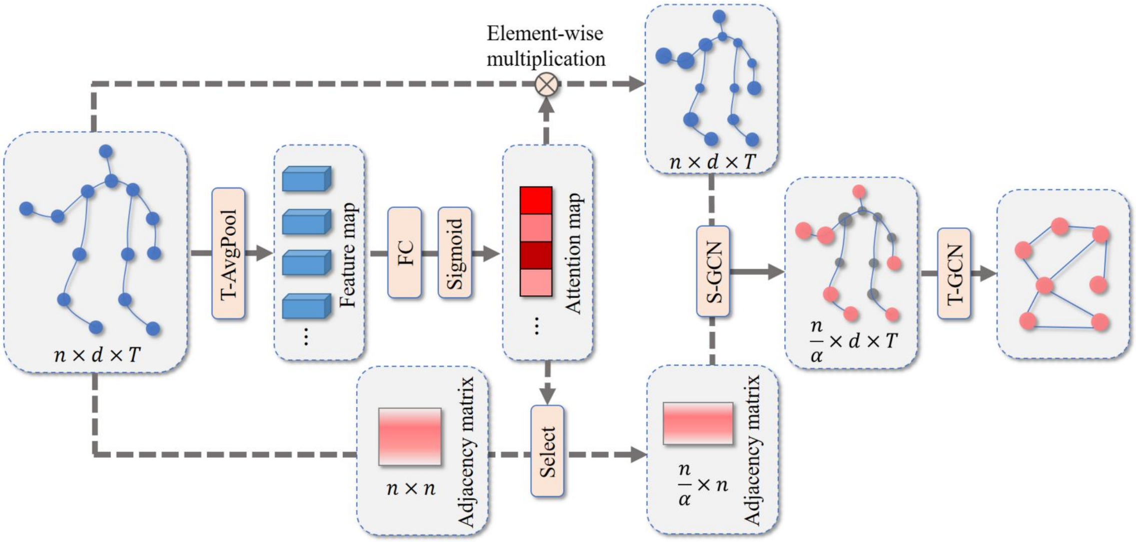

FIGURE 3 The architecture of the spatial attention pooling module

becomes $\frac{n}{\alpha} \times n$ dimensional. In the SAP module, the physical connection structure of the human body no longer exists, so we remove the physical structure adjacency matrix A and the trainable adjacency matrix B, only retain the 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 two at the end of the two streams, respectively.

5 | EXPERIMENT

Extensive experiments are conducted and analysed in this section. Firstly, we introduce two large-scale skeleton datasets, namely NTU-RGB + D [19] and Kinetics-Skeleton [14]. 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 state-of-the-arts.

5.1 | Datasets

5.1.1 | NTU-RGB + D

NTU-RGB + D is a large in-door-captured dataset with annotated 3D joint coordinates for the human action recognition task [19,47]. 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 three 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 two 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 set 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.

5.1.2 | Kinetics-skeleton

Kinetics [38] 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. [14] estimate the locations of 18 joints on every frame of the clips by using the publicly available OpenPose [39] 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 30 fps. 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 score. Following the evaluation method of Yan et al. [14], we train the models on the training set and report the top-1 and top-5 accuracies on the validation set. Because the videos in the Kinetics dataset are captured from the real-world, this experiment can better reflect the performance of the model in real-world situations.

5.2 | Implementation details

Our SATD-GCN model has a total of 12 blocks. In each block, we add a residual connection [40], which enables the model to learn features more effectively and prevents overfitting. 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 [41]. 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 set to 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 two 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.

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

Methods	Accuracy (%)
ST-GCN [14]	88.3
AGCN [17]	93.4
SATD-GCN (JS) w/o TDGC	93.9
SATD-GCN (JS) w/o SAP	94.0
SATD-GCN (JS)	**94.4**

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 [14] on the NTU-RGB + D Cross-View benchmark is 88.3%. By applying the adaptive adjacency matrix and the specially designed data pre-processing methods, Shi et al. designed an AGCN [17], its performance is improved to 93.4%. We use AGCN as the baseline in this 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, that is, the TDGC module and the SAP module. The results in the third row and the fourth row of Table 1 show 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. Besides, compared to AGCN, the increasement of the parameters and training/inference time of our model can be ignored (less than 10%).

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

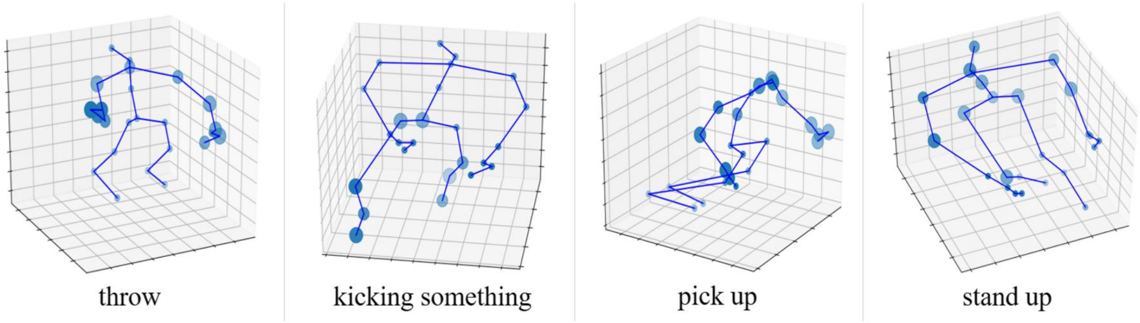

FIGURE 4 Visualization of the attention map in the SAP module. The radius size of the circle represents the weight of the joint

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

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

TABLE 3 Comparison of the validation accuracy with state-of-the-art methods on the NTU-RGB + D dataset

Methods	Cross-Subject (%)	Cross-View (%)
Lie Group [18]	50.1	82.8
HBRNN [20]	59.1	64.0
Deep LSTM [19]	60.7	67.3
ST-LSTM [21]	67.2	77.7
STA-LSTM [10]	73.4	81.2
VA-LSTM [22]	79.2	87.7
ARRN-LSTM [43]	80.7	88.8
Ind-RNN [44]	81.8	88.0
Two-stream 3DCNN [23]	66.8	72.6
TCN [8]	74.3	83.1
Clips + CNN + MTLN [9]	79.6	84.8
Synthesized CNN [24]	80.0	87.2
CNN + Motion + Trans [25]	83.2	89.3
3 scale ResNet152 [26]	85.0	92.3
ST-GCN [14]	81.5	88.3
DPRL + GCNN [45]	83.5	89.8
Motif-GCNs + VTDB [31]	84.2	90.2
AS-GCN [29]	86.8	94.2
2s-AGCN [17]	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. [42]	14.9	25.8
Deep LSTM [19]	16.4	35.3
TCN [8]	20.3	40.0
ST-GCN [14]	30.7	52.8
AS-GCN [29]	34.8	56.5
2s-AGCN [17]	35.9	58.6
SATD-GCN (ours)	**36.6**	**59.8**

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 enables the model extracts temporal features from subtle motion to large-scale motion hierarchically and effectively. The joint stream SATD-GCN model, which has two ST-GCN blocks followed by two S-TD-GCN blocks with dilation rates one and two, respectively, obtains the best performance. For the down-sampling rate α, the SAP module with α=two 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 two and three, 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-art

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 [18,42], the RNN-based methods [10,19–22,43,44], the CNN-based methods [8,9,23–26], and the GCN-based methods [14,17,29,31,44]. 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. Compared to ST-GCN [14], 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 [17], 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-View benchmark. The results reveal 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 compared 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. In contrast to ST-GCN [14], 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 [17], 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 article, we propose a novel SATD-GCN, which contains a TDGC module and an SAP module, for skeleton-based action recognition. The TDGC module can effectively extract the temporal features in different time scales, improve the perception field in the temporal domain, and maintain the robustness 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 ST-GCN, and significantly improve the performance of ST-GCN. Owing to the contribution of these two modules, our SATD-GCN obtains the state-of-the-art performance on two large-scale action recognition benchmark datasets.

ACKNOWLEDGEMENTS

The work is supported by the National Key Research and Development Program of China (No.2018YFB1600600).

ORCID

Jiaxu Zhang https://orcid.org/0000-0002-9551-2708

REFERENCES

1. Tu, Z., et al.: Action-stage emphasized spatiotemporal VLAD for video action recognition. IEEE Trans. Image Process. 28(6), 2799–2812 (2019)
2. Tu, Z., et al.: Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recogn. 79, 32–43 (2018)
3. Sun, Q.R., Wang, W.M., Liu, H.: Study of human action representation in video sequences. CAAI Trans. Intell. Syst. 8, 189–198 (2013)
4. Meixiang, Q., Songhao, P., Guo, L.I.: An overview of visual SLAM. CAAI Trans. Intell. Syst. 11(06), 768–776 (2016)
5. Chen, Y., et al.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7103–7112 (2018)
6. Chen, Y., et al.: So-handnet: self-organizing network for 3d hand pose estimation with semi-supervised learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6961–6970. Long Beach (2019)
7. 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, 588–595. Columbus (2014)
8. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1623–1631. Honolulu (2017)
9. Ke, Q., 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. Honolulu (2017)
10. Song, S., 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. San Francisco (2017)
11. Cao, C., et al.: Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans. Circ. Syst. Video Technol. 29(11), 3247–3257 (2018)
12. Li, Z., et al.: Weakly-supervised semantic guided hashing for social image retrieval. Int. J. Comput. Vis. 128(2), 2265–2278 (2020)
13. Kong, X., Yu, P.S.: Semi-Supervised Local Feature Selection for Data Classification. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM .793. Washington (2010)
14. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans (2018)
15. Si, C., 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. Long Beach (2019)
16. Huang, J., et al.: Long-short graph memory network for skeleton-based action recognition. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 645–652. Snowmass village (2020)
17. Shi, L., 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. 12026–12035. Long Beach (2019)
18. 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. Las Vegas (2016)
19. Shahroudy, A., 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. 1010–1019 (2016)
20. 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. Boston (2015)
21. Liu, J., et al.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision. Springer, Cham, 816–833 (2016)
22. Zhang, P., et al.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2126. Honolulu (2017)
23. Liu, H., Tu, J., Liu, M.: Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)
24. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn. 68, 346–362 (2017)
25. Li, C., et al.: Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, pp. 597–600. Ypsilanti (2017)
26. Li, B., 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. Ypsilanti (2017)
27. Si, C., 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. Munich (2018)

28. Thakkar, K., Narayanan, P.J.: Part-based graph convolutional network for action recognition. arXiv preprint arXiv:1809.04983 (2018)
29. Li, M., 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. Long Beach (2019)
30. Zhao, R., et al.: Bayesian graph convolution LSTM for skeleton based action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 6882–6892. Long Beach (2019)
31. Wen, Y.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. 33, 8989–8996. Honolulu (2019)
32. Zitnik, M., Leskovec, J.: Predicting multicellular function through multi-layer tissue networks. Bioinformatics, 33(14), i190–i198 (2017)
33. Zhou, Z., et al.: Risk oracle: a minute-level citywide traffic accident forecasting framework. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01), 1258–1265. New York (2020)
34. Scarselli, F., et al.: The graph neural network model. IEEE Trans. Neural Netw. 20(1), 61–80 (2008)
35. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Adv. Neural Inf. Process. Syst. 29, 3844–3852 (2016)
36. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
37. Monti, F., et al.: 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. Honolulu (2017)
38. Kay, W, et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
39. Cao, Z., et al.: 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. Honolulu (2017)
40. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. Las Vegas (2016)
41. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037 (2019)
42. Fernando, B., et al.: Modeling video evolution for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5378–5387. Boston (2015)
43. Li, L., et al.: Skeleton-based relational modeling for action recognition. arXiv preprint arXiv:1805.02556 (2018)
44. Li, S., et al.: 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. Salt Lake City (2018)
45. Tang, Y., et al.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5323–5332. Salt Lake City (2018)
46. Peng, Z., et al.: Few-shot image recognition with knowledge transfer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 441–449. Long Beach (2019)
47. Liu, J., et al.: NTU RGB+D 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. 42(10), 2684–2701 (2019)

How to cite this article: Zhang J, Ye G, Tu Z, et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol*. 2021;1–10. https://doi.org/10.1049/cit2.12012