Structure-Feature Fusion Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition

ZHITAO ZHANG\textsuperscript{1}, ZHENGYOU WANG\textsuperscript{1,2}, SHANNA ZHUANG\textsuperscript{1,2}, AND FUYU HUANG\textsuperscript{3}

\textsuperscript{1}School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
\textsuperscript{2}Hebei Key Laboratory for Electromagnetic Environmental Effects and Information Processing, Shijiazhuang 050043, China
\textsuperscript{3}Department of Electronic and Optical Engineering, Army Engineering University, Shijiazhuang 050003, China

Corresponding authors: Zhengyou Wang (zhengyouwang@stdu.edu.cn) and Shanna Zhuang (zsn19850419@163.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61801507 and Grant 61972267, in part by the National Defense Research Project under Grant 20182B040023, in part by the Nature Science Foundation of Hebei Province under Grant F2019210306, in part by the Key Science and Technology Projects of Hebei Education Department under Grant ZD2017247, and in part by the Postgraduate Innovation Funding Project of Hebei Province under Grant YC2020071.

ABSTRACT Human skeleton contains intuitive information of actions and has high robustness in dynamic environment. Therefore, it has been widely studied in action recognition tasks. Most existing methods of skeleton recognition are based on graph convolutional networks (GCNs), which extract the topological structure of graphs to describe the dependencies between joints. However, the GCNs pay excessive attention to skeleton structure and neglect the feature information of skeleton joints. Accordingly, how to fuse the feature of both skeleton structure and joints is a problem to be solved. In addition, non-linear temporal convolutional network (TCN), which has more robustness and learning capability, is rarely investigated in existing methods. With the comprehensive consideration of the dependence between structure and feature on graphs, we propose a novel structure-feature fusion adaptive GCN (SFAGCN) for skeleton based action recognition. The topological structure of the skeleton graph and the feature of the joints can be fused by the decoupled spatiotemporal correlation in our model effectively. The relevance of spatiotemporal data can be preserved well by the fusion strategy, with data integrity ensured. Moreover, Gated TCN is used to extract temporal feature, improving the network performance further. We choose two-stream adaptive GCNs and shift-GCN as the baseline. To demonstrate the effectiveness of our methods, extensive experiments are implemented on the three large-scale datasets, namely, NTU-RGBD 60, NTU-RGBD 120 and Kinetics-Skeleton 400. The accuracy of top-1 on above datasets are improved by more than 0.6% on average, where the performance of SFAGCN exceeds the state-of-the-art methods.

INDEX TERMS Action recognition, structure-feature, adaptive fusion, graph convolution networks.

I. INTRODUCTION
Human action recognition is widely applied in video surveillance and human-computer interaction. Existing recognition methods are usually based on skeleton data, which are represented as sequences or two-dimensional grids. Recently, a significant improvement in recognition accuracy has been achieved by constructing the skeleton as a graph network. Compared with traditional video action recognition, skeleton-based human action recognition employs a smaller amount of data, while high-level semantic information can be provided. At the same time, the robustness against changes in human scale and complicated background interference can also be improved [1]–[8].

Previous methods based on skeleton data simply employ joints coordinate feature streams for action recognition, ignoring the associations between different skeleton joints [2], [9], [10]. With the development of deep learning, researchers manually construct skeleton data into a series of coordinate vectors or pseudo-image features [11]–[15],...
The illustration of extraction and fusion of nodes structure and feature information (a) GCN extracts structure information, (b) MLP extracts joints feature information, (c) Weighted feature map after concatenation.

which are fed into recurrent neural networks (RNNs), convolution neural networks (CNNs) or graph convolutional networks (GCNs). Recently, Yan et al. [8] apply GCNs to extract spatial features of skeleton data, and then add temporal edges between the corresponding joints in consecutive frames by temporal convolutional networks (TCNs). They further build the final spatiotemporal graph convolutional network (ST-GCN). Since the natural connection of skeleton joints is more suitable for extracting features in non-Euclidean domains, the effect of this network is far superior to previous methods. Afterwards, many improvements of ST-GCN have been proposed [1], [3], [4], [16]–[19], most of which modify the spatial module of GCNs to enhance the expressiveness of the model.

Although ST-GCN has been proved to be significantly superior to traditional CNN in skeleton data, there are still two disadvantages of ST-GCN. First, both CNN and GCN only focus on the local characteristics of nodes, and lack the ability of global inference [20]. There are some variants of ST-GCN, such as two-stream adaptive graph neural networks (2s-AGCN) [4] and directed graph neural networks (DGNN) [3], which can enhance the global inference ability of joint points by changing the spatial reasoning structure. For instance, non-local model [21] imposes global dependence, highlighted by [20] and [22]. But the performance may be degraded by excessive global inference. Specifically, clamping enhances non-local inference, but it is still easy to bring about misclassification due to the weights of non-connect structure information. Second, despite the fact that graph convolution can well indicate the rich structural information of nodes and edges, the effect of node feature is always ignored. The feature of joints contains the position information of human skeleton, which is an essential evaluation criterion for action recognition. Therefore, it is of great importance for a critical network to extract this feature. Multi-layer perceptron (MLP) is composed of simple interconnected neurons or nodes, as shown in Fig.1(b), with parameters globally shared. It can completely extract the features of joints. However, there are too many parameters in MLP, which would lead to overfit, resulting in a poor network generalization ability [20].

To solve above problems, a novel structure-feature fusion network is proposed in this work, which adaptively fuse skeleton structure and feature of joints. In the proposed model, the fusion of structure and feature not only rely on simple hyperparameter, but also comprehensively consider the influence of spatiotemporal on action recognition. Inspired by Selective Kernel Networks (SKNet) [23], which utilizes Squeeze-and-Excitation Networks (SENet) [24] to select the channel weights of different kernels, we employ SENet to decouple spatial and temporal information on the feature level and the structural level respectively to obtain two-stream attention weights. The skeleton structure and joints feature extracted by GCN and MLP, respectively, are fused with above attention weights as shown in Fig.1. Structure-Feature Fusion Adaptive Graph Convolutional Networks (SFAGCN) decouples the structure and feature information in spatiotemporal graph. The decoupled data is then fused to ensure the validity of the data. The data-driven method increases the flexibility of spatiotemporal structure-feature fusion, and retains the characteristics of original skeleton data to the greatest extent.

A further cause for concern, in spatiotemporal graph modeling, the ability to learn about time dependence is the criterion for evaluating the model. Existing methods for processing time dependence are usually divided into two classes. The first class makes use of RNN to obtain time dependence by capturing long-distance sequences. But its complex iterative propagation would lead to gradient explosion or disappearance easily [25], [26]. The second is built on CNN, which takes the advantages of parallel computing and stable gradient. Therefore, it has been extensively used in spatiotemporal modeling in TCN. However, with the increase of convolution neural network layers, the model parameters and complexity would also grow rapidly. The model may over-fitting and performance may decrease, which will affect the generalization ability of the system. Compared with convolution network, there are a lot of units in RNN, such as input gate, forget gate and output gate. These units bring more flexible connections for neurons. Motivated by WaveNet [27], [28], the presented model in this paper embeds the gate TCN module in ST-GCN. Gated TCN considers connections between temporal convolutional units and enhances the temporal nonlocality of the network. In ablation experiments, the effectiveness has been verified on the NTU-RGBD 60 dataset. Top-1 accuracy is improved by at least 0.4% on average (IV-C2).

The structure-feature fusion block can be applied to all ST-GCN-based models. To verify the superiority of the proposed model SFAGCN, extensive experiments are performed on NTU-RGBD 60 [11] and NTU-RGBD 120 [29]. With 2s-AGCN and Shift-GCN as baseline, the top-1 recognition accuracy is improved by at least 0.5% on both datasets.

Overall, the main contributions of this work are as follows: (1) We propose a novel, adaptive structure and feature fusion framework, which can well inherit the advantages of all the works in the domain of spatiotemporal graphical and neural networks. (2) The proposed SFAGCN model outperforms state-of-the-art methods on two widely used datasets.

The main contributions and advantages of this work are highlighted by [20] and [22].
of MLP and GCN. Moreover, adaptive fusion can also decouple spatiotemporal information and retain original features.

(2) WaveNet is pruned for temporal skeleton graph modeling. Gated TCN is applied to time-dependence extraction to improve the performance of the model.

(3) The SFAGCN block proposed in this paper can be widely used in ST-GCN-based baseline. The performances on three large-scale datasets exceed the state-of-the-art.

II. RELATED WORK

This section provides a brief overview of the previous literature for action recognition models based on skeleton, as well as the methods of structure-feature fusion.

A. SKELETON-BASED ACTION RECOGNITION

Traditional methods of skeleton-based action recognition mainly focus on hand-crafted features. Dynamic models are employed to analyze action features, such as a rotation and translation of skeleton joints encoded by Lie Group [10], and motion trajectory description of temporal characteristics [9].

With the development of deep learning, the feature extraction methods of skeleton joints based on CNN have gradually emerged. Li et al. [30] designed the automatic rearrangement and selection module of important skeleton joints. Liu et al. [31] proposed the Synthesized CNN to eliminate the effect of view variations on spatiotemporal locations of skeleton joints. Li et al. [32] proposed an end-to-end co-occurrence feature learning framework to learn hierarchical co-occurrence features from skeleton sequences. Meanwhile, Recurrent Neural Networks (RNNs) has gradually attracted the attention of researchers in the field of action recognition, since RNN has been widely used in natural language processing (NLP) and proved its effectiveness.

The skeleton sequences are usually fed into multiple RNN modules along the temporal dimension, with each joint corresponding to a recurrent network block. Song et al. [6] proposed a noval spatiotemporal attention long short-term memory (LSTM) network, which selectively focused on the differentiated joints of the skeleton in each frame, and poured attention into the output of multi-frames in different degrees. Li et al. [33] categorized action classes with multiple RNNs in a treelike hierarchy.

Over the last several years, GCNs have been utilized in the field of skeleton-based action recognition. For human skeleton, joints and bones are strongly coupled. It naturally conforms to the topological structure of the graph. Skeleton data is often expressed as a spatiotemporal graph modeling \( G = (V, E) \), where the joints correspond to the vertex \( V \) of the graph, and the skeleton is represented as edge \( E \). Therefore, GCNs are applied to feature extraction to obtain the structural properties of graphs in non-Euclidean domain, which can not be realized by RNNs or CNNs. Yan et al. [8] proposed a spatiotemporal GCN (ST-GCN). It is the first time to employ GCN to extract the spatial features of skeleton sequences. But ST-GCN only extracts the skeleton features of natural connection, and ignores the influence of unnatural connection features on action recognition. Shi et al. [4] proposed a 2s-AGCN, which used non-local block to extract the association between joints, with bones viewed as the basis for action recognition. The skeleton data was also represented as a directed acyclic graph based on the kinematic dependency between the joints and bones. However, the computational complexity of the above methods is too high, and the receptive fields of both spatial graph and temporal graph are not flexible enough. To solve above two problems, Cheng et al. [18] proposed a shift-GCN network, which greatly reduces the computational complexity of the model and provides a flexible receptive field for the model.

B. STRUCTURE-FEATURE CONNECTION MECHANISM

The graph contains a multitude of nodes and edges, which include rich structure and feature information. Although GCN is suitable for obtaining graph structure information, it ignores node characteristic information to a certain extent. On the other side, MLP has strong global reasoning ability, due to the ability of adaptive aggregating feature information of node features. However, it cannot extract the structural information of the graph. Wang et al. [20] used MLP and CNN to extract global and local feature information, respectively. Fusing above two stream sequences, the multi-modal characteristics of the data are obtained. Zhang et al. [34] considered both graph structure and feature information as the basis of graph pooling node selection. In order to make the evaluation criteria of the skeleton graph more objective and robust, we propose a feature fusion method with adaptive structure, which comprehensively considers the global and local characteristics of the data, as well as the relationship between spatial and temporal features. The method proposed in this paper is more accurate and effective.

III. METHODS

In this section, we firstly formulate the problems to be resolved in this paper. And then we introduce two important parts of the framework, which are SFAGCN and Gated TCN. At last, we illustrate the overall architecture of the framework.

A. PRELIMINARIES

A human skeleton is represented by \( G = (V, E) \) where \( V = \{v_1, v_2, \ldots, v_N\} \) is the set of \( N \) graph joints, and \( E \) is the set of edges captured by an adjacency matrix \( A \in \mathbb{R}^{N \times N} \). If an edge directs from \( v_i \) to \( v_j \), \( A_{ij} = 1 \), otherwise 0. Take the example of 3D vector, each joint vector is represented by \( v_i = (x_i, y_i, z_i) \), with the edge expressed as \( e_i = (v_i, v_j) \). Considering the set of action frames, graph sequences are represented as \( S = \{G_1, G_2, \ldots, G_T\} \), where \( T \) is the total number of frames. The spatial joints set of all frame sequence can be expressed as tensor \( X \in \mathbb{R}^{T \times N \times C} \), where \( N \) is the number of joints, and \( C \) denotes the number of channels at the \( t \)th frame. GCNs is an operation of processing node information in non-Euclidean domain. Both the feature \( X \) and
the adjacency matrix $A$ are fed in graph to be embeded. GCN updates regular space features hierarchically as follows,

$$X_i^{l+1} = \sigma(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}X_i^{l} \Theta^{l}),$$ \hspace{1cm} (1)

where $\tilde{A} = A + I$ is the adjacency matrix in which self-loops are utilized to maintain the identity. $D$ is the opposite angle matrix of $\tilde{A}$. $\sigma(\cdot)$ is the activation function. The structure information is fully described in the graph structure, but the feature of $X$ is ignored due to the structure feature extraction, leading to insufficient global feature extraction ability.

**B. RELATION TO PRIOR WORKS**

Traditional methods usually model skeleton data as vector sequences, pseudo images and graphics, which are processed by RNNs, CNNs and GCNs, respectively. RNNs are highly capable of handling and mining serialized data for correlation. However, RNNs have limited ability to process the spatial structure of the same frame. CNNs model the spatial features of skeleton data as pseudo-images, which is conducive to the extraction of spatial features. Nevertheless, the temporal relevance of sequential data is ignored by CNNs, with poor performance in structured data extracting. Yan et al. [8] considered the kinematic dependence between joints and skeletons, and proposed ST-GCN network according to the physical structure of the human body. The model is also methodologically flawed, in the processing of skeleton joints. And the adjacency matrix based on prior knowledge suffers bias for the association problem. Thus the relationship of unconnected joints is difficult to model. To solve this problem, 2s-AGCN [4] using non-local block to increase the association capability between any joints unconnected has been proposed. However, non-local block not only employs a large amount of calculation, but also is easy to overfit. Multi-scale graph 3D (MS-G3D) [35] facilitated direct information streams across spatiotemporal level. Therefore, it removes the redundant dependencies between node features from different neighborhoods. MLP adopts linear combination of different joint sets, which has strong global reasoning capability. Through the down-sampling operation, the problem of a large amount of calculations can be solved in non local (NL) block. We suggest that different evaluation methods should be considered comprehensively for the above networks, so as to increase the diversity of evaluation and enhance the objectivity of joint selection.

**GCN:** Human skeleton graph naturally contains rich structural information. Therefore, it is effective to extract the structure information of skeleton joints by graph convolution. The expression is as follows,

$$S_1 = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}XW)$$ \hspace{1cm} (2)

where $S_1$ is the joints structure characteristics calculated by GCN.

**MLP:** Although joints contain rich node feature information, GCNs often ignore it. So we employ MLP to extract node features. It has strong global inference ability, and can aggregate feature information adaptively. In order to reduce calculation costs, the down sampling operation has been adopted,

$$S_2 = \sigma(fc(ReLU(fc(GAP(X)))))$$ \hspace{1cm} (3)

where $X$ is the node characteristic matrix. $ReLU$ and $\sigma$ are both activation functions. $GAP$ denotes global average pooling, $fc$ represents full connection layer, as shown in Fig.2.

In summary, we choose GCN and MLP to extract skeleton structure and joint features, respectively, which can combine the advantages of above methods.

**C. ADAPTIVE STRUCTURE AND FEATURE FUSION BLOCK**

In this paper, we propose SFAGCN to make full use of features of the structure and characteristics on spatialtemporal level. Through the analysis in Sec.III-B, three adaptive selection strategies are designed in this section, as shown in Fig.3.

The first strategy of spatial structure feature fusion, as shown in Fig.3(a), employs a single hyperparameter $\alpha$ to fuse the results of two node evaluation methods according to,

$$S_{fusion} = \alpha S_1 + (1-\alpha)S_2$$ \hspace{1cm} (4)

Hyperparameters can be pre-defined or learned. To facilitate observation, we select custom hyperparameters. Results in Sec.IV-C3 would verify the superiority of the model through experiments.

Compared with the scalar variable of the first strategy, the second strategy, as show in Fig.3(b), sets the trainable parameters to adaptively determine the perceptual connection of each joint. The strategy depends on the connection of samples and enhances the attention of spatiotemporal joints. The fusion is shown as follows

$$S_{fusion} = w_{x,z}f(w_{x,1}[S_1 : S_2])$$ \hspace{1cm} (5)

where $[\cdot]$ is a concatenation operation. $w_{x,z}$ is a linear transformation parameter. $f(\cdot)$ denotes a nonlinear activation function. The size of the parameter matrices is the same as that of input matrices. Although this strategy considers dependencies of joints and improves the perception ability of joints,
FIGURE 3. Illustration of three fusion strategies. (a) represents a simple fusion strategy, which combines the Structure and Features extracted with GCN and MLP by super parameter $\alpha$ and $(1-\alpha)$ to extract the spatial structure feature information. (b) denotes multi hyper parameter fusion method, which is more complex and has more parameters, and is prone to over fitting. (c) denotes an adaptive fusion method, which compresses and extracts multi-channel data through SEblock to obtain the structure and feature weight. The weight is fused with the extracted features to obtain the spatial structure feature information.

the spatiotemporal relationship is not analyzed, resulting in less modeling capabilities. The spatial computational complexity of this strategy is $O(TNC)$, which is prone to overfitting. Although the sensing capability of nodes is improved, a large number of parameters are applied. So that the computational complexity is increased and it is easy to over fit.

The third strategy, as show in Fig.3(c) adopts aggregating parameters to reduce the number of parameters. Inspired by SKNet [23], we fuse results from multiple branches(two streams in Fig.3(c)) through element wise multiplication where $F_{sq}$ and $F_{ex}$ are the squeeze and excitation operation, respectively. Human action recognition not only pays attention to the structural information between skeleton joints, but also concerns the spatial position information. Our fusion method is similar to the scale in SE module, where $U_S$ and $U_F$ represent structure and feature information, respectively. Since there is no prior knowledge above the importance of the two kinds of streams, we exploit weighted summation to squeeze and excite fusion data according to,

$\text{attention}_{\text{branches}} = \frac{e^{A_{gcn}Z}}{e^{A_{gcn}Z} + e^{B_{mlp}Z}}$ (6)

where $A$ and $B$ are two branches after compression. This strategy learns the importance of different branches and gains the attention coefficient.

The final fusion vector $S_{fusion}$ is obtained, as follows,

$S_{fusion} = \left( S_1, S_2 \right) \cdot \left( \frac{\text{softmax}(\sigma(f_{ex}(f_{sq}(S_1 : S_2))))}{1 - \text{softmax}(\sigma(f_{ex}(f_{sq}(S_1 : S_2))))} \right)$ (7)

where $(;)$ denotes the cascade of $S_1$ and $S_2$

D. GATED TEMPORAL CONVOLUTION NETWORK

TCNs can reshape a sequence of any length to a uniform length sequence. Compared with RNN, TCN inherits the advantage of stable gradient of CNN, so that over fitting can be avoided to some certain extent. RELU is widely used as the activation function in existing TCN networks. It is partially truncated at $(-\infty, 0)$ so that the range of $(-1, 0)$ is in the off-state. Inspired by WaveNet [27], we employ gate activation unit to optimize TCN. Sigmoid and tanh activation functions are utilized to calculate the result of element multiplication. A residual connection has been added to accelerate convergence and help gradient descent propagate in deeper network model. The gated TCN module fully considers the relationship between the temporal convolution units and enhances the temporal nonlocality of the network. The gated activation unit is listed in the following,

$h = \text{tanh}(\Theta_1 x + b) \odot \sigma(\Theta_2 x + c)$ (8)

where $\odot$ is the hardamard product. $\text{tanh}$ represents the activation function of the output layer, with $\sigma$ the sigmoid function, which controls the ratio of information output to the next layer. The gated TCN module is the temporal part of the SFAGCN block(Sec. III-E), as shown in Fig. 4.

E. ADAPTIVE STRUCTURE AND FEATURE FUSION GRAPH CONVOLUTIONAL BLOCK

The structure of SFAGCN is similar to STGCN. In SFAGCN the spatio module and the temporal module process the feature mapping of $C \times T \times N$ alternately. As shown in Fig.4, a basic block is composed of gate TCN, a structure feature spatio GCN and an adaptive fusion module. In order to stably train and integrate different features, a skip connection is added to each block. Finally, by using one-dimensional convolution transformation, the result and the residual sequence are aligned and fed into the summation block.
FIGURE 4. Illustration of SFAGCN block. The module consists of two parts: Gated TCN and SFA model. The green box represents the input and output stream, the blue box is the basic module, and the yellow circle is the activation function (tanh and sigmoid), ⊕ denotes the elementwise summation, and ⊗ is the matrix multiplication. The residual box (Res) is only needed when \( f_{in} \) is not the same as \( f_{out} \).

FIGURE 5. Illustration of the overall architecture of the SFAGCN. There are a total of \( N=10 \) layers of SFAGCN blocks. The fusion features are finally input into the full convolutional network (FCN) through global average pooling (GAP) and classified by softmax.

F. OVERALL ARCHITECTURE

The overall architecture (SFAGCN) is shown in Fig.5. We employ continuous skeleton joints or bone vectors between joints as input streams and then feed the streams into the spatiotemporal block (as shown in Fig.4). The extracted features are processed by global average pooling (GAP) and full convolutional network (FCN) successively to obtain the softmax score of each action. The summation of above scores obtain the final prediction. As the skeleton graph is acyclic tree structure, the size of the graph network for joints and bones should be same with each other. Considering that only \( n-1 \) bones can be obtained from \( n \) joint points, we add an empty bone with value 0 to the central joint. The graph network composed of bones is designed according to joints, with an adjacency matrix \( A_{bone} \) similar to that of joints.

IV. EXPERIMENTS

To verify the generalization of the model, this paper conducted experiments on several challenging datasets, that is, NTU-RGBD 60 [11], NTU RGB+D 60 120 [29] and Kinetics 400 [36]. We perform exhaustive ablation experiments on SFAGCN to verify the effectiveness of the proposed model. Then, taking 2s-AGCN [4] and shift-GCN [18] as the baseline, our module is superior to the state-of-the-art methods.

A. DATASETS

NTU-RGBD 60 [11] is a large-scale in-door-captured action recognition dataset, which contains 56000 skeleton sequences of 60 action classes. The clips are performed by 40 volunteers in different age groups ranging from 10 to 35. Each action is captured by 3 different KinectV2 cameras with different views. The dataset provides the 3D coordinates of each frame detected by the kinect depth sensor. Each skeleton graph contains 25 skeleton joints as nodes, and the action contains 1 or 2 subjects. This dataset recommends two benchmarks for classification: (1) Cross-Subject(X-Sub): the benchmark is divided into training set (40320 action segments) and verification set (16560 motion segments). (2) Cross-View(X-View): the training set in the benchmark test contains 37920 action segments. Both the verification dataset and the training dataset adopt different perspectives and contain 18960 action segments. In comparison with the existing methods, top-1 accuracy in two benchmarks is used as the evaluation criterion.

NTU-RGBD 120 [29] extends NTU-RGBD 60 by adding 60 additional action classes, including 57367 skeleton sequences. The action sequences are performed by 106 actors in 32 cameras. Similar to NTU-RGBD 60, it contains the same benchmark, where cross setting means that samples
Kinetics 400 [36] is a large-scale human behavior dataset with a total of 400 classes of actions. There are 306,245 videos in total, each of which has 400-1150 video clips. STGCN [8] estimates the locations of 18 joints on every frame of the clips using OpenPose toolbox. This dataset contains 240,436 training clips and 197,96 test clips, with each skeleton image including 18 body joints. We train the model on the training set, and report the accuracies of top-1 and top-5 on the validation set.

B. TRAINING CONFIGURATIONS

Our experiment is conducted on the Pytorch deep learning framework. The stochastic gradient descent (SGD) with Nesterov momentum is selected as the optimization strategy to back-propagate the gradients. The batch size is 16. Cross-entropy is applied as the loss function on back-propagation. The weight decay is set to 0.001 for final models and is adjusted accordingly during component studies. In order to prove the effectiveness of our model, we adopt the same preprocessing and super parameters in ST-GCN, 2s-AGCN and shift-GCN for fair performance comparison. To improve the performance of our model on NTU-RGBD 60, 120, the training process is increased to 100 epochs. The initial learning rate is set to 0.1 and is divided by 10 at 30th, 60th and 80th epoch, respectively. The model in this paper is implemented on two NVIDIA RTX 2080ti GPUs with CUDA10.0.

For the NTU-RGBD 60 dataset, there are at most two subjects in each clip. If the number of subjects in the clip is less than 2, the second entity will be filled with 0. The maximum number of frames in each video is 300. For videos with less than 300 frames, we repeat sampling until 300 frames are achieved.

C. STRATEGIC ANALYSIS AND ABLEXPERIMENTS

In this section, we verify the effectiveness of SFGCN block, three fusion strategies and Gated TGCN. The experiment of Gated TCN is implemented in non-Digested environment. In the benchmark, top-1 accuracy in action recognition on the NTU-RGBD 60 dataset is evaluated.

1) SFGCN BLOCK

As illustrated in section III-C, there are three types of strategies in fusion adaptive block. We evaluated the first two fusion strategies to verify the effectiveness of our method. Tab.1 demonstrates the fusion top-1 accuracy of strategy 1. Without changing the original parameters, 2s-AGCN [4] is employed as the baseline method. We use a single hyper parameter to fuse the structural features extracted by convolution and the joints features extracted by MLP. With adjusting the fusion parameter $\alpha$, the model obtains a higher recognition accuracy (95.57%) than the original paper (95.10%). By the same token, we also do experiments on the second strategy, and the recognition accuracy is 95.60%.

| Method       | joints | bones | joints&bones |
|--------------|--------|-------|--------------|
| 2s-AGCN      | 93.70  | 94.30 | 95.10        |

Table 1 shows that compared with the original method, the fusion feature recognition rate of strategy 1 is improved, especially when $\alpha = 0.4$, achieving the optimal recognition rate. Compared with strategy 1, the top-1 accuracy of action recognition obtained by strategy 2 is not greatly improved. It can be found from the results that over fitting leads to little improvement in the recognition rate. It is proved that the complexity of this method is too high, and it is easy to appear over fitting. See Sec.III-C for details.

2) GATED TGCN

Existing methods are based on simple TCN to extract temporal dimension information, ignoring different temporal feature correlations. Here, we compare the performance of Gated-TGCN on the NTU-RGBD 60 dataset (joints, skeletons, and two-streams). The accuracy of Gated-TGCN has been improved on joints (Js), bones (Bs) and two stream (2s) data, as illustrated in Sec.III-D. According to Tab.2, we can find that the accuracy on Bs-skeleton data can be improved by 0.75% using the Gated-TGCN, which is higher than that on the other datasets. As Gate TCN focuses more attention on the association of temporal nodes, and the bone data is composed of adjacent joints with a local topology structure, larger receptive field improves the recognition accuracy. When extracting temporal data associations through Gated-TGCN, the skeleton association between different temporal frames is stronger than the joint association. Clearly, Gated-TCN provides better performance in handling with time series combinations.
3) SFAGCN BLOCK

In this paper, three fusion strategies are proposed for structural-feature fusion, with 2s-AGCN and shift-GCN as the baseline models. Strategy 1 and strategy 2 adopt a simple fusion strategy, without considering the spatiotemporal relationship as a reference for the fusion strategy. To address this shortcoming, strategy 3 is proposed to add attention mechanism in the fusion process, extracting the action data. The spatiotemporal parameters of the different branches represent the attention coefficients for structure and feature extraction, respectively. In this paper, SKNet and softmax function for reference allocate multiply parameters to different channel streams. The performance of the proposed SFAGCN on NTU-RGBD 60, 120 and Kinetics-skeleton 400 is improved by 0.7%, 1.5% and 0.3%, respectively, as shown in Tab.3. Compared with NTU-RGBD 60, NTU-RGBD 120 contains more accurate action skeleton data owing to more volunteers and cameras, providing more accurate action skeleton data. Therefore, SFAGCN achieves greater accuracy improvement on NTU-RGBD 120 dataset. The correctness of strategy 3 has been verified in Sec.III-C. Kinetics-skeleton 400 is an action skeleton data extracted from video through openpose [36]. On one hand, the dataset contains more classifications of actions. On the other hand, each frame of the skeleton is only composed of 18 joints and each joint is two-dimensional. Due to the lack of structural-feature information, the improvement of SFAGCN on Kinetics-skeleton 400 dataset is not particularly significant as that on NTU-RGBD 60/120 dataset.

### D. COMPARISON WITH THE STATE-OF-THE-ART

To evaluate the advancement of SFAGCN, we take 2s-AGCN and shift-GCN as baseline. NTU-RGBD 60, NTU-RGBD 120 and Kinetics-Skeleton 400 datasets are used in this experiment. And the corresponding comparison results are shown in Tab.4, Tab.5 and Tab.6 respectively. The handcraft-feature-based [10], RNN-based [6], [11], [33], [37], [38], CNN-based [13], [14], [31], [32] and GCN-based [3], [4], [8], [18], [35] methods are compared with our model. This paper achieves state-of-the-art performance on three datasets. The superior results show the generalization and potential of the proposed approach.

### V. CONCLUSION

In this paper, we proposed SFAGCN for skeleton-based action recognition. This method integrates the structural information of the skeleton data with the position information of the joints. At the same time, the spatiotemporal features are relatively independently integrated to ensure the integrity of the spatiotemporal data. In addition, we applied Gated TCN to increase the connections between neural units and enhance the temporal non-local characteristics of the network. Taking 2s-AGCN and shift-GCN as the baseline, SFAGCN has achieved higher performance on the NTU-RGBD 60 dataset. The experimental results on 3 large-scale datasets demonstrate that our model is effective. In future work, we will study the structure and feature associations between adjacent skeletons at different clips, so as to better model the spatiotemporal information of skeleton actions.
REFERENCES

[1] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Actional-structure graph convolutional networks for skeleton-based action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3595–3603.

[2] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1110–1118.

[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Skeleton-based action recognition with directed graph neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7912–7921.

[4] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 12026–12035.

[5] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, “Skeleton-based action recognition with spatial reasoning and temporal stack learning,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 103–118.

[6] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in Proc. 31st AAAI Conf. Artif. Intell., 2017, pp. 1–7.

[7] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, “Deep progressive reinforcement learning for skeleton-based action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5323–5332.

[8] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–10.

[9] B. Fernando, E. Gasves, M. José Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 5378–5387.

[10] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 588–595.

[11] A. Shahroury, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1010–1019.

[12] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, “Independently recurrent neural network (IndRNN): Building a longer and deeper RNN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5457–5466.

[13] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “A new representation of skeleton sequences for 3D action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3288–3297.

[14] T. S. Kim and A. Reiter, “Interpretable 3D human action analysis with temporal convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1623–1631.

[15] A. Main de Boissiere and R. Noumeir, “Infrared and 3D skeleton feature fusion for RGB-D action recognition,” 2020, arXiv:2002.12886. [Online]. Available: http://arxiv.org/abs/2002.12886.

[16] B. Li, X. Li, Z. Zhang, and F. Wu, “Spatio-temporal graph routing for skeleton-based action recognition,” in Proc. 33rd AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8561–8568.

[17] T. Ahmad, H. Mao, L. Lin, and G. Tang, “Action recognition using graph wavelet,” in Proc. 2017 IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1444–1452.

[18] L. Zhang, X. Wang, H. Li, G. Zhu, P. Shen, P. Li, X. Lu, S. A. A. Shahs, and M. Bennamoun, “Structure-feature based graph self-adaptive pooling,” in Proc. Web Conf., 2020, pp. 3098–3104.

[19] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 143–152.

[20] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017, arXiv:1705.06950. [Online]. Available: http://arxiv.org/abs/1705.06950.

[21] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive recurrent neural networks for high performance human action recognition from skeleton data,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2117–2126.

[22] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, “Global context-aware attention LSTMs for 3D action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1647–1656.

[23] I. Liu, A. Shahroudy, D. Xu, and G. Wang, “Spatio-temporal LSTMs with trust gates for 3D human action recognition,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 816–833.

[24] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “Learning clip representations for skeleton-based 3D action recognition,” IEEE Trans. Image Process., vol. 27, no. 6, pp. 2842–2855, Jun. 2018.

ZHTAO ZHANG was born in 1995. He received the B.E. degree in information engineering from Shijiazhuang Tiedao University, Shijiazhuang, China, in 2018, where he is currently pursuing the M.S. degree with the School of Information Science and Technology. His research interests include deep learning and computer vision, especially human action analysis.
ZHENGYOU WANG was born in Hunan, China, in 1972. He received the B.E. degree in mechanical and electrical engineering, the M.S. degree in mechatronics, and the Ph.D. degree in power electronics and electric drives from the China University of Mining and Technology (CUMT), Xuzhou, China, in 1994, 1997, and 2002, respectively. From 1997 to 1999, he was an Assistant with CUMT. From 2002 to 2004, he was an Assistant Research Fellow with Zhejiang University. He was an Associate Professor from 2004 to 2010 and a Professor in 2006 with the Jiangxi University of Finance and Economics. He is currently a Professor with the School of Information Science and Technology, Shijiazhuang Tiedao University. His research interests include computer vision, media computing, and intelligent information processing.

SHANNA ZHUANG was born in 1985. She received the B.E. degree in electronic information engineering and the Ph.D. degree in information and communication engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2012, respectively. Since 2012, she has been with the School of Information Science and Technology, Shijiazhuang Tiedao University, where she is currently an Associate Professor. Her research interests include computer vision and adaptive waveform design for cognitive radar.

FUYU HUANG was born in Heibei, China, in 1985. He received the B.S., M.S., and Ph.D. degrees from the Department of Optical and Electronic Engineering, Mechanical Engineering College, Shijiazhuang, China, in 2008, 2010, and 2014, respectively. His research interests include optical signal processing, infrared detection, and optical design.