Learning Comprehensive Spatiotemporal Representations for Skeleton-based Action Recognition

Fang Ren, Chao Tang, Anyang Tong, Wenjian Wang

***Abstract*—Currently, Graph Convolutional Network is extensively employed for skeleton-based action recognition and has achieved appreciable performance. However, there are certain unsolved issues with Graph Convolutional Network-based approaches. The existing methods extract the spatiotemporal features of the skeleton sequence by stacking a series of spatiotemporal graph convolution modules using a single modeling method. The network of above structures is difficult to comprehensively characterize the spatiotemporal representations of the actions, which is potentially discriminatory and is not conducive for fully mining the spatiotemporal features with the coupling. Moreover, the extraction of multi-scale spatiotemporal features of the skeleton sequence requires heavy computing resources and the aggregation of extracted features is insufficient. To address the above shortcomings, we propose the Dual-path Spatial Temporal Graph Convolutional Network (DST-GCN), which builds a homogeneous parallel dual-channel network. This network employs different spatiotemporal modeling methods to extract the synergistic and complementary action features to learn comprehensive spatiotemporal representations. Under this framework, first, this paper proposes a Weighted Adaptive Graph Convolutional Network (WA-GCN) which can provide weights to joints according to their importance of action recognition to learn the spatial topology of the human skeleton efficiently and flexibly. Second, we design an Enhanced Multi-Scale Temporal Convolutional Network (EMS-TCN) to ensure the extraction of multi-scale temporal features, while strengthening the long-term and short-term action representations of actions in the temporal dimension, with respect to both the global and local aspects. Finally, a Standard Euclidean Distance based Pairwise Gaussian Loss (SED-PGL) is presented. The loss function excludes the differences between the dimensions of sample features, and takes into account the inter-class separation and intra-class compactness of the samples, which helps the model to distinguish similar actions. Experimental results on the public datasets NTU RGB+D, NTU RGB+D 120, and Kinetics-Skeleton have achieved the state-of-the-art performance in the industry, which also verifies the effectiveness of the proposed method.**

***Index Terms*—Skeleton-based action recognition, spatiotemporal modeling, adaptive graph convolutional network, multi-scale temporal feature aggregation, loss function.**

Fang Ren and Chao Tang are with the School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China. (E-mail: renfang@stu.hfuu.edu.cn; tangchao@hfuu.edu.cn).

Anyang Tong is with the Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China. (E-mail: tonganyang@stu.hfuu.edu.cn).

Wenjian Wang is with the School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China. (E-mail: wjwang@sxu.edu.cn).

Corresponding author: Chao Tang.

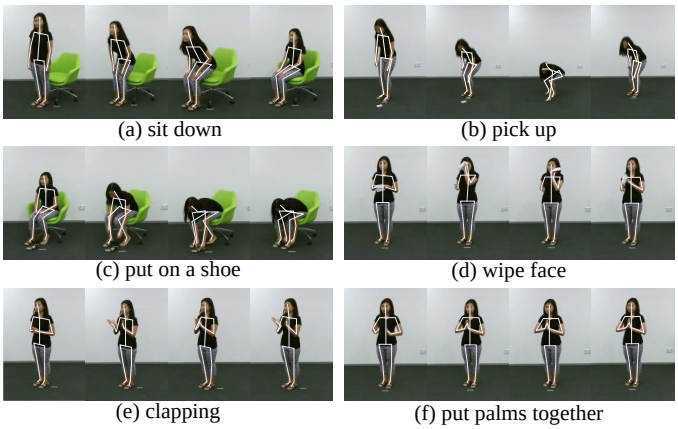


Fig. 1. Demonstration of typical skeleton-based actions. The motivations for our proposed method are demonstrated through a specific analysis of the characteristics of skeleton actions in the temporal and spatial dimensions.

I. Introduction

HUMAN action recognition (HAR) is an important research field in computer vision and has multiple applications such as video surveillance [1], human-computer interaction [2], and motion analysis [3]. Skeleton-based action recognition is a significant branch in the field of HAR. Compared with RGB data, skeleton data contains rich and compact human structure information and has strong adaptability to dynamic environments and complex backgrounds. Therefore, it can extract more robust action features to efficiently complete action recognition task.

Early skeleton-based action recognition relied on manual feature extraction. Owing to the development of deep learning, researchers first considered using classical network models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to extract the spatiotemporal features of skeleton sequence. However, considering that the skeleton has non-Euclidean data, classical CNN-based and RNN-based methods cannot explore the topological information of the human skeleton, thus making the research unsustainable [4]. Skeleton spatiotemporal graphs have semantic in temporal dimension, so certain researchers have attempted to employ Transformer to extract the spatiotemporal features of the skeleton sequence [5]. Furthermore, owing to the advantages of Graph Convolutional Network (GCN) [6] in spatial modeling, skeleton-based action recognition has

entered a new stage [7].

The human skeleton graph has a natural topology and a certain action of the human body constitutes a set of skeleton spatiotemporal graphs. The GCN-based method can spatially model the topology of skeleton and extend it to the temporal domain, intending to extract the spatiotemporal features of the skeleton sequence. Currently, the existing skeleton-based action recognition methods based on GCN has the following problems. First, the previous work [8], [9] employed a single spatiotemporal modeling method and stacked a series of feature extraction modules to obtain spatiotemporal features. The spatial modeling followed by temporal modeling will naturally enhance the spatial characteristics and weaken temporal characteristics. The above spatiotemporal modeling order has potential discrimination, which is not conducive for fully mining the coupled spatiotemporal features. Further, it is difficult to comprehensively learn the spatiotemporal representations of the skeleton sequence. Second, for obtaining the multi-scale spatiotemporal features, previous works [10], [11] taxed the computing resources and ignored the aggregation of multi-scale spatiotemporal features. Therefore, the characteristics of skeleton-based action in both the spatial and temporal dimensions are significant for the action recognition task.

The skeleton sequence is the trajectory of joints within a certain period of temporal and spatial domains, and thus, it is necessary to pay attention to the temporal and spatial characteristics contained in it. The NTU RGB+D dataset contains certain common actions in daily life, such as "drink water", "brush hair", and other actions. Taking the two actions of "sit down" and "pick up" as examples, the representations in the spatial and temporal dimensions demonstrate the inter-dependent spatiotemporal characteristics, as shown in Fig. 1(a) and (b). We consider that the various spatiotemporal modeling orders can fully consider the coupling of spatiotemporal features, which affects the focus of the model on the spatial and temporal dimensions. Take "put on a shoe" and "wipe face" as examples. They are the interactions between distant joints, but there is no physical connection between these joints in the human skeleton, as shown in Fig. 1(c) and (d). Therefore, a model is needed to create node connections that do not exist in reality, so as to obtain the long-distance spatial dependence of joints. We take "clapping" and "put palms together" as examples. The action "clapping" is a dynamic hand movement, but "put palms together" is a continuous static action. They have certain similarities in the spatial dimension, though the information in the temporal dimension is quite different, as shown in Fig. 1(e) and (f). As we can see, some actions take on different characteristics in different length of time segments which can be used to effectively identify the actions.

By analyzing the action patterns, in order to solve the above two problems, we propose a novel Dual-path Spatial Temporal Graph Convolutional Network (DST-GCN) framework, which contains Spatial Temporal Modeling Path (STM-Path) and Temporal Spatial Modeling Path (TSM-Path), for learning synergistic and complementary spatiotemporal features of the skeleton sequence. The STM-Path consists of the Spatial Temporal Convolutional Block (STCB), which employs an order of spatial modeling followed by temporal modeling. On the contrary, Temporal Spatial Convolutional Block (TSCB) in TSM-Path employs the order of temporal modeling followed by spatial modeling. The overall framework is displayed in Fig. 2. Considering that temporal and spatial features are coupled, DST-GCN is employed to design two spatiotemporal modeling methods to eliminate potential discrimination, aiming to provide complementary spatiotemporal feature representations for the same action. According to the spatial characteristics, we propose a Weighted Adaptive Graph Convolutional Network (WA-GCN) to learn the spatial dependence of the non-adjacent joints at a small computational cost. Further, we obtain different weights for joints with different partition strategies to reflect their contribution to action recognition. For the temporal characteristics, we propose a Enhanced Multi-Scale Temporal Convolutional Network (EMS-TCN). This network extracts multi-scale temporal features and strengthens the long-term and short-term representations of the actions in the temporal dimension from both the global and local aspects. Besides, to better distinguish similar actions, we design the Standard Euclidean Distance based Pairwise Gaussian Loss (SED-PGL) to train the proposed method for obtaining better feature space. The contributions of our work can be summarized as follows.

- We propose DST-GCN framework which contains two spatiotemporal modeling methods and incorporate multi data streams and score fusion strategies for learning comprehensive spatiotemporal representations of the skeleton sequence.
- We propose WA-GCN to learn the spatial dependence of the non-adjacent joints efficiently and flexibly. WA-GCN is a data-driven network that adaptively learns the topology of the skeleton and the importance of node partitioning strategies.
- We propose EMS-TCN to learn discriminative temporal features which takes into account both the global and local information of actions in the feature aggregation operation.
- We design the SED-PGL to train the model. This loss function controls the inter-class separability and intra-class compactness, which is beneficial for distinguishing similar actions.

The rest of the paper is organized as follows. In Section II, we review the development of the skeleton-based action recognition methods and the spatiotemporal modeling methods. The Section III introduces the relevant background knowledge for this paper, including GCN and multi-scale temporal features. The Section IV introduces the methods proposed in this paper, including DST-GCN, WA-GCN, EMS-TCN, and SED-PGL. The Section V shows the relevant experimental results and experimental analysis, which proves the effectiveness of the proposed method. Finally, we summarize our work and prospective studies in Section VI.

II. Related Work

This section contains two parts. First, we will review the development of various methods in the field of skeleton-based action recognition. Then, we will discuss the influence of different spatiotemporal modeling methods on the skeleton-based action recognition methods.

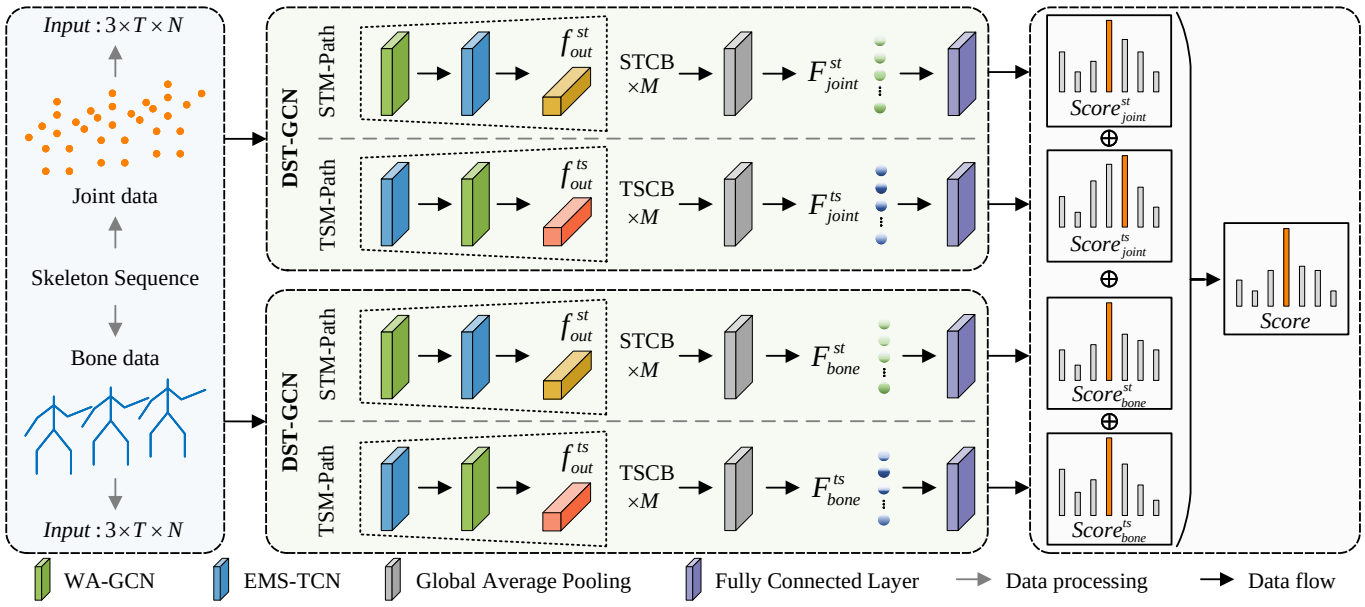


Fig. 2. The overall structure of our proposed method. We employ two data streams of skeleton sequence to train the DST-GCN and the predictions from each path are fused by score fusion strategy to obtain the final prediction. DST-GCN contains STM-Path and TSM-Path and each path consists of M convolutional blocks. DST-GCN are detailed in Section IV to learn comprehensive spatiotemporal representations of the skeleton sequence. WA-GCN and EMS-TCN are detailed in Section IV for spatial modeling and temporal modeling, respectively. $\oplus$ denotes element wise summation.

A. Skeleton-based Action Recognition

The early task of human action recognition has been based on RGB data [12], [13], [14], [15]. With the appearance of Kinect camera [16] and attitude estimation algorithm [17], skeleton-based action recognition has been studied by several researchers. At first, the researchers have manually designed the relevant features of the human skeleton and used machine learning algorithms to identify the actions [18], [19], [20]. However, by employing the method based on manual design, it is difficult to extract advanced features and cannot be applied to large datasets, which severely limits the recognition performance and efficiency. Owing to the interesting results of deep learning network model on RGB data, several researchers have also applied the classic deep learning network model, viz., RNN and CNN, to the task of skeleton-based action recognition.

RNNs and their variants, Long Short-Term Memory Network (LSTM), have natural advantages in modeling the temporal series data. For the skeleton-based action recognition, researchers focused on the co-occurrence features of joints to extract the spatiotemporal features [21], [22], [23]. Furthermore, feature enhancement [24], [25], [26] and special network structure [27], [28], [29], [30] have been employed to improve the performance of the network. Unlike RNNs, CNNs have the ability to extract the advanced features and with efficiency in spatial modeling. Smoe researchers have favored 2D CNN [31], [32], [33] owing to its advantages of simple network structure and fast running speed. Furthermore, 3D CNN [34], [35] have been employed in the skeleton-based action recognition for exploring the spatial and temporal features in the skeleton sequence. However, RNNs and CNNs are both insufficient for the extraction of spatial topology information. Different from above methods based on RNN or CNN, our proposed method based on GCN which has the advantages in spatial modeling leading to higher recognition performance.

Owing to the explosive development of Vision Transformer in the field of computer vision [36], certain researchers attempted using the Transformer to extract the spatiotemporal features of skeleton sequence. The skeleton spatial temporal graph of a skeleton sequence is a set of skeleton data on a temporal dimension and has certain semantics. Zhang et al. [37] have proposed Spatial-Temporal Specialized Transformer, who designed the Spatial Transformer Block and the Directional Temporal Transformer Block models in spatial and temporal dimensions to extract the spatiotemporal features of skeleton sequence. Pang et al. [38] have proposed the Interaction Graph Transformer, which models the human body parts of interactions for action recognition. Transformer has the advantages of extracting global information, and the dynamic attention mechanism included is also conducive for extracting discriminating features. Currently, the Transformer-based skeleton human action recognition method is in the early stages of development. This type of method shows tremendous development potential. But there are still certain problems such as large calculation amount and slow training speed. The DTS-GCN method proposed in this paper focuses on the spatiotemporal modeling of the skeleton sequence and obtains an excellent recognition performance with less parameters.

GCN has a natural advantage while processing a graph structure data like the human skeleton. Yan et al [8] have proposed the Spatial-Temporal Graph Convolutional Network (ST-GCN). They have first defined the skeleton spatiotemporal graph for the skeleton-based action recognition task, and then extracted the spatiotemporal features of the skeleton sequence by stacking a series of spatiotemporal graph convolutional

blocks. Finally, it has employed a SoftMax classifier for the category prediction. Unlike the ST-GCN, where the local spatial features of joints are extracted, Li et al [39] have proposed the Actional-Structural Graph Convolution Network (AS-GCN). The AS-GCN captures the specific joint dependencies, through actions, and multi-neighborhood joint dependencies, by extending the existing skeleton graph for structural connections. Pre-defined topology graphs of the human skeleton do not reflect the invisible connections between the joints. Therefore, the researchers attempted mining the connections between non-adjacent joints [10] and construct an adaptive graph structure [9], [40], besides training the network model using multiple data streams. The above models focus on the spatial modeling of the skeleton, though the temporal features of the skeleton sequence are not sufficiently extracted. Multi-scale G3D(MS-G3D) [11] and Multi-scale Spatial Temporal Graph Convolutional Network (MST-GCN) [41] attempted in extracting the multi-scale spatiotemporal features of the skeleton sequence. Furthermore, Zhou et al. [42] innovatively applied the contrastive learning to the process of network training, which facilitates the network model to learn the discriminative representations to distinguish similar actions. GCN-based skeleton-based action recognition methods usually employ a single spatiotemporal modeling approach to extract spatiotemporal features. Different from these above GCN-based methods [8], [9], [40], [41], the proposed DST-GCN contains two synergistic and complementary spatiotemporal modeling approaches. In two spatiotemporal modeling approaches, we enhance the spatial feature extraction network and temporal feature extraction network by improving feature aggregation operations. Furthermore, we employ the SED-PGL as the loss function to optimize the feature space which is beneficial to classify the similar actions.

B. Spatiotemporal Modeling

Skeleton-based action can be regarded as the spatiotemporal interaction of joints. The temporal and spatial characteristics that are shown are interrelated, and have a certain coupling. Spatiotemporal modeling of the skeleton sequence is of great significance for the skeleton-based action recognition. Currently, GCN-based methods [8], [9], [11] extract the action features by stacking a series of spatiotemporal feature extraction network modules which have been combined in the order of spatial modeling and temporal modeling to complete the spatiotemporal modeling. The extraction of temporal features in the above-mentioned spatiotemporal extraction module network has been based on the extracted spatial features. Such a single spatiotemporal modeling method is discriminatory against the extraction of temporal features. However, the temporal and spatial features are coupled and interdependent. A single spatiotemporal modeling approach is not sufficient for fully exploring the action features. Based on this research motivation, we designed two spatiotemporal modeling approaches in proposed DST-GCN.

Considering the spatiotemporal modeling, View adaptive neural networks [43] have been designed as a two-stream structure (referred to as VA-fusion), using RNN stream and CNN stream, to extract the temporal and spatial features, respectively. Finally, they use the score fusion to predict action categories. In order to improve the capability of extracting the features from skeleton data, Symmetrical Enhanced Fusion Network (SEFN) [44] have designed a multi-stream framework, viz., spatial stream graph convolution product, temporal stream graph convolution product, and temporal-spatial fusion stream to enhance the expression ability of skeleton-based actions. MS-G3D [11] have proposed a novel spatial-temporal graph convolution operator G3D to obtain the inter-temporal dependence of joints, for extracting the multi-scale spatiotemporal features of skeleton sequence. The above methods design the network structure from the perspective of spatiotemporal modeling. In contrast to these methods on spatiotemporal modeling, the DST-GCN proposed in this paper has the following differences. (1) VA-fusion attempts to extract spatiotemporal features using two different networks and one spatiotemporal modeling approach. However, DST-GCN is a method based on GCN which has two spatiotemporal modeling approaches. Complementary spatiotemporal modeling approaches are beneficial for learning comprehensive spatiotemporal representations. (2) SEFN adopts spatial stream, temporal stream and fusion stream which combines temporal and spatial features to form a three-stream network structure. This method uses four different skeleton data streams to obtain the spatiotemporal features of skeleton sequence but ignores spatiotemporal correlations. Different from SEFN, DST-GCN has designed a heterogeneous parallel dual-path framework for the coupling of spatiotemporal characteristics. Dual-path spatiotemporal framework can provide two cooperative and complementary spatiotemporal feature representations for the same action. Compared with SEFN, DST-GCN has a simpler network model framework and better parallel training ability. (3) MS-G3D employs high-order polynomials of joint adjacency matrix to aggregate the multi-scale spatial information, so as to capture complex inter-temporal node correlation. This method extracts features with a 3D operator which taxes a lot of computing resources. DST-GCN adopts a dual-path parallel network structure to broaden the width of the model and explore the spatial and temporal characteristics of synergy and complementarity. Owing to the residual structure and convolution operation in the network, DST-GCN has achieved excellent recognition performance with fewer parameters.

III. BACKGROUND

In this section, we will introduce the background theory of the method in this paper from two aspects. First, we will introduce GCN for spatial feature extraction in skeleton-based action recognition methods. Then, we will introduce classical methods for extracting temporal features in skeleton sequence.

A. GCN

GCN is a type of Graph Neural Network and is commonly employed to process the graph-structured data, such as molecular structures, social networks, and knowledge graphs. GCN is categorized into spectral graph convolution [45] and spatial graph convolution [46], based on the principles of constructing

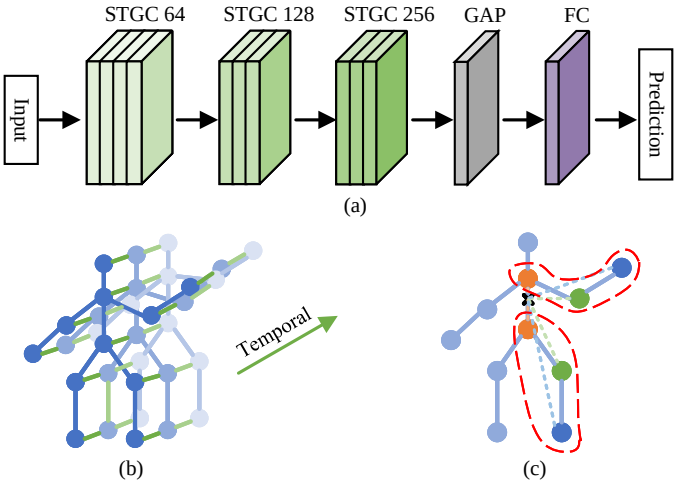


Fig. 3. The background theory of GCN for spatial feature extraction in skeleton-based action recognition methods: (a) denotes serialized spatio-temporal modeling, (b) denotes spatial temporal graph of a skeleton sequence, (c) denotes spatial partitioning strategy.

graph convolution operations. Conventional CNNs have translation invariance for the convolution operations on images, though the number of neighboring joints of each node in the graph is not consistent. This makes it impossible to directly perform the convolution operations on the graph structure data. The spectral graph convolution attempts to transform the spatial domain information of the graph into the spectral domain (frequency domain), which in turn accomplishes the convolution operation. The difference is that the spatial domain graph convolution employs a similar operation on the conventional convolution, thus defining the graph convolution based on the spatial relationship of the joints. The standard convolution operation of CNN on an image is shown as:

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{w=1}^{K} f_{in}(P(x,h,w)) \cdot W(h,w) \tag{1}$$

where $K \times K$ denotes the convolution kernel size, $P(x,h,w)$ enumerates all neighboring pixels in a $h \times w$ region centered on the spatial pixel point x, and $W(h,w)$ denotes a weight vector. Yan et al. [8] proposed ST-GCN, which innovatively applies graph convolution to the skeleton-based action recognition. ST-GCN performs the spatiotemporal modeling of skeleton sequence by sequentially stacking 10 layers of Spatial Temporal Graph Convolution (STGC) as shown in Fig 3(a). Except for the first layer, all network layers contain residual structures, and the numbers in the figure indicate the number of channels in the current network layer.

ST-GCN defines the spatiotemporal graph of the skeleton sequence based on the natural connections of human skeleton joints and the connections of the same joints between neighborhood frames, as displayed in Fig. 3(b). Particularly, the human skeleton and action consists of N joints and T frames, respectively. Then the undirected spatiotemporal graph $G = (V, E)$ is represented by a set V of joints and a set E of edges. $V = (v_{ti} | t = 1, ..., T, i = 1, ..., N)$, where v_{ti} denotes the information of the *i-th* joint in frame *t*. E consists of two subsets, E_S and E_F. $E_S = \{v_{ti}v_{tj} | (i,j) \in H\}$ denotes a natural connection of joints within the same frame, and H denotes a natural connection pair of human joints. $E_F = \{v_{ti}v_{(t+1)i}\}$ denotes a connection of the same joint point between two neighborhood frames.

In addition, this work performs the partitioning for each joint of the human skeleton. The spatial partitioning strategy is displayed in Fig. 3(c). The green node indicates the root node, the orange node indicates the centripetal node, and the blue node indicates the centrifugal node. The partitioning strategy for the neighborhood node v_{tj} of v_{ti} can be expressed as:

$$l_{ti}(v_{tj}) = \begin{cases} 0, if r_j = r_i \\ 1, if r_j < r_i \\ 2, if r_j > r_i \end{cases} \tag{2}$$

where r_i denotes the distance from joint v_{ti} to the center of gravity point of the human body. Therefore, the Spatial Graph Convolution applicable to the human skeleton information has been defined by rewriting the sampling function P and W in Eq. (1). The Eq. (1) is transformed into:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(P(v_{ti}, v_{tj})) \cdot W(l_{ti}(v_{tj})) \tag{3}$$

where Z_{ti} is the normalization term used to balance the contribution of each subset. $B(v_{ti})$ denotes the set of 1-neighborhood joints of v_{ti}. $P(v_{ti}, v_{tj})$ denotes all neighboring nodes v_{tj} of v_{ti}, W denotes the weight function, and l_{ti} denotes the subset labels that are assigned to each v_{ti} 1-neighborhood joints v_{tj} with a fixed subset label that prescribes the order of the convolution operation. The adjacency matrix can represent the joints connection information of the skeleton. The above equation can be converted into Eq. (4) for ST-GCN.

$$f_{out} = \sum_{k}^{K} W_k (f_{in} A_k) \odot M_k \tag{4}$$

W_k is the weight matrix of f_{in}, and A_k is the normalized adjacency matrix of joints. M_k is the matrix of learnable parameters used to represent the importance of each edge, and $\odot$ denotes the dot product, and $K = 3$.

B. Multi-scale Temporal Feature

The skeleton-based action presents the multi-scale characteristics in the temporal dimension, i.e., an action will present multiple characteristics in different temporal segments and the duration of different actions is different. The convolution kernel 9×1 employed by TCN in 2s-AGCN [9] attempts to completely characterize the action in the temporal dimension as displayed in Fig. 4(a). However, this method does not consider the characteristics of actions in different temporal length segments. The MS-TCN in MS-G3D [11] employs the convolution operation of different receptive fields to extract the multi-scale temporal features, and the structure is displayed in Fig. 4(b). MS-TCN has been the most commonly adopted multi-scale temporal feature extraction network module due to its simple structure and excellent performance. Through the analysis of the actual action and previous research methods,

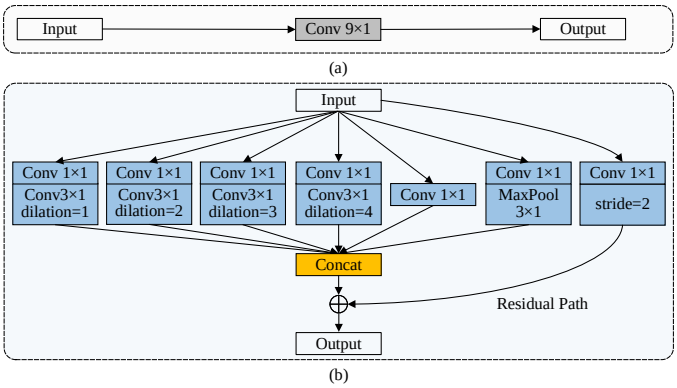


Fig. 4. Development of temporal convolutional network in skeleton-based action recognition methods: (a) denotes Temporal Convolutional Network (TCN) in 2s-AGCN [9], (b) denotes Multi-Scale Temporal Convolutional Network (MS-TCN) in MS-G3D [11].

first, we think that the action characteristics of different temporal scales can successfully describe the performance of different length temporal segments in the action. Second, we think that the long-distance dependence information of the action in the temporal dimension belongs to the global characteristics, whereas the short-distance dependence information belongs to the local characteristics. Global features can provide complete and continuous temporal information, whereas local features can provide efficient and discriminating temporal information. Therefore, it is of tremendous significance to extract the multi-scale temporal features and strengthen them from both the global and local aspects for skeleton-based action recognition.

IV. Method

In this section, we will introduce each network module of this method in detail. First, we will introduce DST-GCN, which is used to learn the comprehensive spatiotemporal representations. Second, we will introduce WA-GCN for the spatial feature extraction. Then, we will introduce EMS-TCN for the temporal feature extraction. Finally, we will introduce the loss function employed in the model training.

A. DST-GCN

Current GCN-based methods extract spatiotemporal features of skeleton sequence by stacking a series of spatiotemporal convolutional blocks of a single spatiotemporal modeling approach, and this type of network structure neglects the coupling of temporal and spatial features [8], [9], [11], [41]. Considering the coupling of temporal and spatial features, and after the analysis of the action patterns and spatiotemporal modeling methods, we innovate the application of two different spatiotemporal modeling approaches to obtain the synergistic and complementary spatiotemporal features.

We propose the Dual-path Spatial Temporal Graph Convolutional Network (DST-GCN) framework, whose structure is shown in Fig. 5(a). DST-GCN uses the Spatial-Temporal Modeling Path (STM-Path) and Temporal-Spatial Modeling Path (TSM-Path) in parallel for spatiotemporal modeling of skeleton sequence. The input of DST-GCN is a multidimensional array, and we take the Joint data stream as an example, i.e., $Input : 3 \times T \times N$, where 3 denotes the three dimensions of the coordinates of the joints. Here, T denotes the duration of the action, and we uniformly expand the action to 300 frames. N denotes the number of joints in the human skeleton, and the value is taken differently in different datasets. STM-Path extracts the spatiotemporal features of skeleton sequence by stacking a series of Spatial Temporal Convolutional Blocks (STCB). Similarly, TSM-Path extracts spatiotemporal features by stacking a series of Temporal Spatial Convolutional Block (TSCB). The parameter list of convolutional block is [in channels, out channels, stride, adjacency matrix, residual connection] and the specific parameters are listed in Table I. Further, the parameters of STM-Path and TSM-Path are kept consistent. STCB and TSCB provide two different spatiotemporal modeling approaches to learn the diverse spatial-temporal information of the skeleton sequence to complete the comprehensive spatial-temporal characterization. The workflow of DST-GCN is given as follows. First, the skeleton data are inputted into the STM-Path and the TSM-Path pathways respectively. Taking STM-Path as an example, one of the STCBs extracts the spatiotemporal features as:

$$f_{out}^{s} \leftarrow WA - GCN(X) \tag{5}$$

$$f_{out}^{t} \leftarrow EMS - TCN(f_{out}^{s}) \tag{6}$$

$$f_{out}^{st} = res(X) \oplus f_{out}^{t} \tag{7}$$

X is the output of the previous layer of the network, which has been passed through the spatial feature extraction network WA-GCN to obtain the spatial feature f_{out}^{s}, which is subsequently fed into the temporal feature extraction network EMS-TCN to obtain f_{out}^{t}. The spatiotemporal feature f_{out}^{st} has been obtained through the residual concatenation operation. Then, after 10 layers of network to extract the features, we have employed Global Average Pooling (GAP) to reduce the dimensionality of the feature vectors, while retaining the key features, and spread them to obtain the spatiotemporal feature F^{st} by focusing on spatial dimensional information with dimension = 192. Finally, Fully Connect Layer (FC) has been employed for the category prediction to obtain the prediction $Score^{st}$. Similar operations are performed in TSM-Path. The skeleton data are used in the TSCB of TSM-Path to extract temporal features f_{out}^{t} using EMS-TCN followed by spatial features f_{out}^{s} using WA-GCN. Subsequently, the spatiotemporal features F^{ts} focusing on the information of temporal dimension is employed to get the prediction $Score^{ts}$. In the end, we combine STM-Path and TSM-Path to form Dual-Path, i.e., we employ a score fusion strategy on the predictions of the two spatiotemporal modeling paths in order to obtain the $Score^{dual}$ of the dual-path spatiotemporal modeling.

Furthermore, we have employed two data streams, viz., Joint and Bone, to train the DST-GCN to improve the robustness and generalization of the model. Referring to the definition of joint in Section III, given the source node $v_{ti} = (x_{ti}, y_{ti}, z_{ti})$ and target node $v_{tj} = (x_{tj}, y_{tj}, z_{tj})$ of a skeleton for a frame t, the bone vector has been described as $b_{t,ij} = (x_{tj} - x_{ti}, y_{tj} - y_{ti}, z_{tj} - z_{ti})$, and we similarly constructed the input data in $3 \times T \times N$ dimensions. Different categories of actions exhibit different characteristics in the temporal and spatial

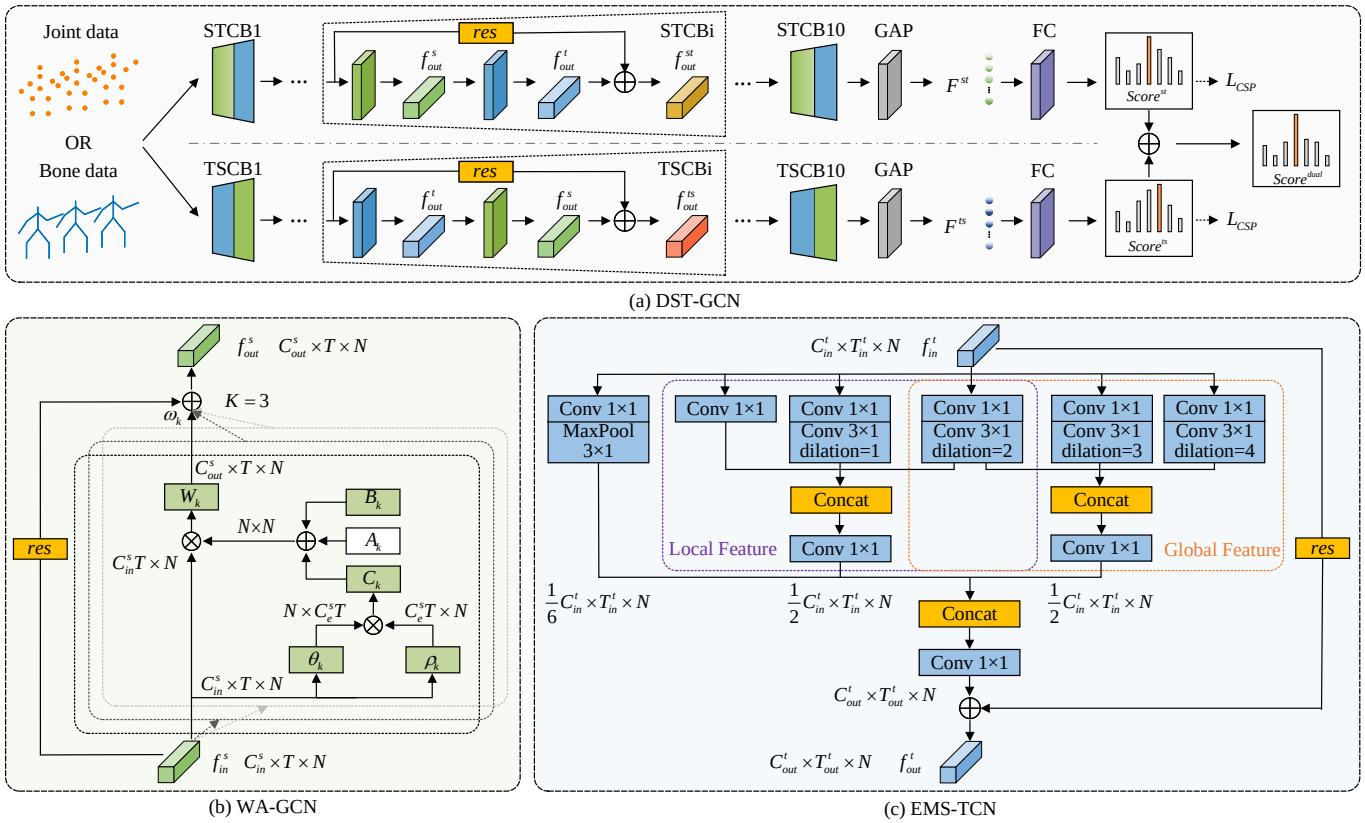


Fig. 5. Detailed network structure of the proposed method: (a) represents the DST-GCN for learning comprehensive spatiotemporal representations of skeleton sequence, (b) represents WA-GCN for spatial modeling, and (c) denotes EMS-TCN for temporal modeling. ⊕ denotes element wise summation. ⊗ denotes matrix multiplication. Concat denotes concatenate by channel dimension.

dimensions, and different spatiotemporal modeling methods focus on different action patterns with different emphases. We use weighted score fusion to predict the final action category to achieve the best recognition rate for the samples in the dual-stream four-pathway framework.

Based on the analysis of skeleton action pattern, DST-GCN is proposed to learn comprehensive representations for skeleton-based action recognition. Two spatiotemporal modeling approaches, viz., STM-Path and TSM-Path, can affect the focus of the DST-GCN on the spatial and temporal dimensions and eliminate potential discrimination. Therefore, DST-GCN can extract the synergistic and complementary action features in skeleton sequence. DST-GCN employs a parallel dual-path framework, which is simple in structure but effective and has good parallel training capability. Two data streams has been employed to train the DST-GCN, which increases the robustness and recognition performance. In addition, DST-GCN achieves sota recognition performance with low complexity and is easy to reproduce and deploy. In summary, the method proposed in this paper has good applicability in practical applications and can also provide a new direction for future research.

B. WA-GCN

The DST-GCN proposed in this paper contains of two spatiotemporal modeling methods, viz., STM-Path and TSM-Path. The two spatiotemporal modeling approaches are composed of STCB and TSCB, and the specific structure is displayed in Fig. 5(a). Obtaining the long-range spatial dependence of joints is the key to action recognition in the spatial dimension. However, the human skeleton belongs to the non-Euclidean data. Therefore, increasing the sensory field in one way, when compared with the Euclidean data, does not accurately procure the spatial features of the joints, instead, it tends to get redundant information and inflates the number of parameters. We argue that the learning the adaptive graph structure of the human skeleton can efficiently and flexibly acquire the spatial feature information of skeleton movements, balancing recognition performance, and consumption.

We propose the Weighted Adaptive Graph Convolutional Network (WA-GCN) as a base module in DST-GCN, and the structure is displayed in Fig. 5(b). WA-GCN is a data-driven network that employs the adjacency matrix of the joints and the weights of the partitioning strategy as part of the network parameters, following the training of the network and updating the parameters. The human skeleton is represented by the adjacency matrix of joints. The adaptive graph creates the connections between joints that are not connected in practice, and establishes dependencies on the non-adjacent joints. Similarly, the joints of different partitioning strategies contribute differently to action recognition. We set the weight parameter for learning the importance and combine it with an adaptive adjacency matrix to increase the flexibility and generalization ability of the model. The dimension of the input

TABLE I
PARAMETERS OF STM-PATH

Layer	Parameter	Layer	Parameter
STCB1	[3,48,1,A,False]	STCB7	[96,96,1,A,Ture]
STCB2	[48,48,1,A,Ture]	STCB8	[96,192,2,A,Ture]
STCB3	[48,48,1,A,Ture]	STCB9	[192,192,1,A,Ture]
STCB4	[48,96,2,A,Ture]	STCB10	[192,192,1,A,Ture]
STCB5	[96,96,1,A,Ture]	GAP	Global Average Pooling
STCB6	[96,96,1,A,Ture]	FC	[192,Action Classes]

feature vector f_{in}^s is $C_{in}^s \times T \times N$. Based on Eq. (4), the process of extracting spatial feature f_{out}^s by WA-GCN is shown as:

$$f_{out}^s = \sum_{k=1}^{K} \omega_k W_k f_{in}^s (A_k + B_k + C_k) \tag{8}$$

where ω weights the node information for different partitioning strategies and be used as part of the network parameters. Convolution operations on 2D images have translation invariance, though each joint has actual semantic properties for the human skeleton. Based on the spatial partitioning strategy in ST-GCN [8], the joints are classified into root, centripetal, and centrifugal nodes. Further, the nodes belonging to different partitions have varied significance for information aggregation. A_k is an $N \times N$ matrix representing the physical connections of joints in reality, where the weights are not updated. This matrix cannot obtain the information between the non-adjacent joints, though it is helpful to improve the stability of the initial network training, and has certain guidance and direction for the training of the network parameters. B_k is also an $N \times N$ matrix whose initial values are all zero. This matrix data has been obtained based on the whole training samples and is the basis for network training. It has a high flexibility for different samples, and can theoretically obtain information between any joints, which enhances the flexibility and generalization ability of the model. The input feature vector f_{in}^s has been projected onto different subspaces using θ_k and ρ_k functions respectively and the correlation C_k between the joints is computed to obtain the information of the key joints in the skeleton, which contribute to the action recognition, according to the strength of the correlation of the joints. The weight of C_k is dependent on the input features of the WA-GCN. The distribution of the input features of the WA-GCNs in the STCB and the TSCB, proposed in this paper, are not identical, which is favorable for obtaining diverse spatial features. The input feature vector f_{in}^s undergoes spatial feature extraction to obtain f_{out}^s, which is then input to the next network module by batch normalization and ReLU activation function, combined with residual connection.

The proposed WA-GCN accounts for the varying importance of joints in different partitioning strategies for action recognition, and sets a learnable weight parameter combining with an adaptive graph. This can efficiently and flexibly learn the spatial topology of the joints in the human skeleton, by taking into account the model complexity and recognition accuracy. The parameter update of the adjacency matrix C_k depends on the features extracted from the shallow network layer. The proposed WA-GCN under the DST-GCN framework can acquire the spatial features more comprehensively.

C. EMS-TCN

A reasonable temporal modeling approach is beneficial for mining the temporal features of the skeleton sequence. In the temporal dimension, we believe that the multi-scale temporal features are conducive for focusing on action features within the different temporal segments. The temporal dimension is a one-dimensional sequence form, which can be considered as Euclidean data with translation invariance. Multi-scale temporal features can effectively identify the differences in the temporal dimension of actions, and thus improve the model generalization ability. Currently available multi-scale temporal convolution methods only capture the multiscale temporal features. However, insufficient attention has been paid to the aggregation of multiscale temporal features, which leads to insufficient differentiation of actions that are similar in the temporal dimension.

We propose the Enhanced Multi-Scale Temporal Convolutional Network (EMS-TCN) as the base module in the DST-GCN and the structure of EMS-TCN is displayed in Fig. 5(c). To extract multi-scale temporal features of the skeleton-based action, we employ the 3×1 convolution kernel with different dilation rates in order to perform feature extraction operations, i.e., to extract multi-scale temporal features in the time dimension using non-symmetric dilated convolution operations. In addition, we strengthen the temporal feature aggregation capability in both local and global aspects. The dimension of the input feature vector f_{in}^t is $C_{in}^t \times T_{in}^t \times N$. First, we utilize non-symmetric dilated convolution for multi-scale temporal feature extraction of the input features. Second, the long-range dependence of the action in the temporal dimension can be viewed as global information and the short-term dependence can be viewed as local information. We aggregate the action features within short and long time segments respectively to obtain local and global features in the time dimension. Finally, in order to fully characterize the temporal features of the actions, we further aggregate the local and global temporal features using concat operation. The Conv 1×1 operation used therein not only captures the information across inter-channel to obtain the corresponding temporal features, but also reduces the model parameters. The input feature vector f_{in}^t has been subjected to the temporal feature extraction to obtain f_{out}^t, which is then input to the next network module by batch normalization and ReLU activation function, and combined with residual linkage.

Compared with the previous work on temporal features, the EMS-GCN proposed in this paper enhances the extraction of multi-scale temporal features from both the global and local aspects. The long-range dependence of actions in the temporal dimension can be regarded as global information, whereas the short-term dependence can be regarded as local information. EMS-GCN takes into account both the global and local information of actions in the feature aggregation operation, which is conducive for obtaining the discriminative temporal features and thus distinguishing the actions that are

similar in the temporal dimension. Based on the specific analysis of action patterns, we obtain the relevant inductive bias and apply it to the deep learning model. This combination of deep learning model and inductive bias [47] is conducive for improving the interpretability and generalization ability of the deep learning models.

D. SED-PGL

Cross Entropy Loss (CEL) is the most common loss function for classification tasks, which can effectively train the network parameters. For a simple binary classification task, the formula for CEL is given as:

$$L_{CEL} = \frac{1}{N_s}\sum_{i} -\left[y_i \cdot log(p_i) + (1-y_i)\cdot log(1-p_i)\right] \quad (9)$$

Where y_i denotes the category label of sample *i*, 1 for the positive category, and 0 for the negative category. p_i denotes the probability that sample *i* is predicted to be in the positive category. Under the multiclassification task, the formula for CEL is updated as:

$$L_{CEL} = -\frac{1}{N_s}\sum_{i}\sum_{c=1}^{Class} y_{ic} log(p_{ic}) \quad (10)$$

where $y_{ic} = 1$ if the true category of sample *i* is *c*, and 0 otherwise. Further, p_{ic} denotes the predicted probability that sample *i* belongs to category *c*. $Class$ denotes the number of categories. Based on the observation of the results obtained from training with CEL, we have found that CEL can effectively differentiate the samples of different classes, but samples of the same class are more dispersed from each other, which can easily lead to misclassification of samples of similar action classes. We propose Standard Euclidean Distance based Pairwise Gaussian Loss (SED-PGL) to solve the problem that CEL only focuses on the inter-class separability and ignores intra-class compactness. Standard Euclidean distance avoids the variability of sample features in different dimensions, hence SED-PGL responds to the correlation between samples by calculating the standard Euclidean distance between the sample features. By embedding it into the Gaussian function, it ensures the intraclass compactness by penalizing samples that are farther away from each other in the same class. Further, we combine CEL and SED-PGL to ensure the inter-class separation of samples from different categories.

The Gaussian function is one of the most common forms of variable distribution in probability theory. We compute the distances between samples and embed them into the Gaussian function as a way of explicitly constructing correlations between samples. However, while using the Euclidean distance to calculate the sample distance, the problem of scale difference in the dimension of the sample features arises. This paper adopts the method of calculating the standard Euclidean distance of the samples, to solve this problem, which avoids the gradient explosion, facilitates the network training, and improves the network performance. The formula for calculating the standard Euclidean distance is given as:

$$d_{i,j} = \sqrt{\sum_{n=1}^{D}\frac{\left(f_n^i - f_n^j\right)^2}{s_n^2+\delta}} \quad (11)$$

where s_n^2 is the variance of the sample and *D* is the dimension of the feature vector. δ is the smoothing factor, which can well avoid the gradient disappearing. According to several comparison tests, we set $\delta = 0.001$. The samples input to the network have been combined in pairs to compute the standard Euclidean distance. f^i and f^j are the feature vectors of the two samples in a set, viz., $i \in \{1, 3, ..., batchsize-1\}$, respectively. Embedding $d_{i,j}$ into the generalized Gaussian function formula is given as:

$$Gaussian(d_{i,j}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d_{i,j}-\mu)^2}{2\sigma^2}} \quad (12)$$

Eq. (12) has been simplified as $Gaussian(d_{i,j}) = e^{-\beta d_{i,j}^2}$ by reducing the variables in it to β. After several comparison experiments, we set $\beta = 0.5$. The integrated probability mapping function, i.e., the formula for the probability of a sample predicting a category, is given as:

$$P(y_{ij}|d_{ij}) = \begin{cases} Gaussian(d_{ij}), y_{ij} = 1 \\ 1 - Gaussian(d_{ij}), y_{ij} = 0 \end{cases} \quad (13)$$

where $y_{ij} = 1$ implies that f^i and f^j belong to the same category, and vice versa is 0. Based on Eq. (10) and bringing in Eq. (13), we obtain the SED-PGL formula as:

$$\begin{aligned} L_{SED-PGL} &= \frac{4}{N_s^2}\sum_{i=1}^{N_s}\sum_{j=i+1}^{N_s} -\log(P(y_{ij}|d_{ij})) \\ &= \frac{4}{N_s^2}\sum_{i=1}^{N_s}\sum_{j=i+1}^{N_s} \left(\beta d_{ij}^2 + (y_{ij}-1)\log(e^{\beta d_{ij}^2}-1)\right) \end{aligned} \quad (14)$$

According to Eq. (14), when two samples belong to the same category and the distance between the samples is large, $L_{SED-PGL}$ applies a penalty, and a larger loss value is computed for controlling the intra-class compactness. Similarly, when two samples do not belong to the same category though the distance between the samples is small, $L_{SED-PGL}$ similarly applies a penalty and a larger loss value has been calculated for controlling the interclass separation. Thus, we combine the cross-entropy loss function and the paired Gaussian loss function based on the standard Euclidean distance as the loss value of the whole network shown as below, and then add a regularity term to control the importance of the two. After many experiments, we set λ= 1.

$$L_{CSP} = L_{CEL} + \lambda L_{SED-PGL} \quad (15)$$

The SED-PGL proposed in this paper calculates the standard Euclidean distance between pairs of samples to reflect the correlation between them, and controls the inter-class separability and intra-class compactness of the samples by penalizing different class samples that are closer, and the same class samples that are farther away. In this paper, SED-PGL and CEL are combined to form L_{CSP} as the loss function of the whole network, which is not only conducive to the stability of the pre-training of the network, but also helps to distinguish the similarity of the action.

V. EXPERIMENT

In this section, we first introduce three skeleton datasets: NTU-RGB+D, NTU-RGB+D 120 and Kinetics-Skeleton. To verify the effectiveness of the proposed method, we will conduct the related ablation experiments on NTU-RGB+D dataset. Besides, we will compare experiment results with state-of-the-art methods on above three large-scale datasets to demonstrate the excellent performance of our method.

A. Datasets

NTU RGB+D [48] contains 60 action classes performed by 40 subjects, with a total of 56,880 video clips. In all the 60 action classes, the first 50 action classes are single-person actions, and the last 10 action classes are double-person interactive actions. 3D coordinates of 25 joints have been collected for each human body. We follow the standard benchmarks, i.e., Cross Subject (X-Sub) and Cross View (X-View). Particularly in X-Sub, 40,320 videos from 20 subjects are employed for training, whereas the remaining 16,560 videos from other 20 subjects will be utilized for testing. In X-View, all 37,920 videos from cameras 2 and 3 are employed for training, and the remaining 18,960 videos from camera 1 are employed for testing.

NTU RGB+D 120 [49] is an expanded version of NTU RGB+D , which contains 120 action classes performed by 106 subjects, with a total of 114,480 video clips. The difference is that this dataset uses the benchmarks based on Cross Subjects (X-Sub) and Cross Settings (X-Set). Particularly in X-Sub, 63,026 videos from 53 agents are employed for training, and the remaining 50,919 videos are employed for testing. In X-Set, there are 32 different camera setting numbers. Further, 54,468 videos with even camera setting numbers are employed for training, and 59,477 videos with odd camera setting numbers will be employed for testing.

Kinetics-Skeleton [50] is a large-scale human action dataset, which contains 300,000 video clips from 400 categories on YouTube. Yan et al. [8] have utilized the Open Pose [51] for the first time to extract 18 joint points of the human skeleton in the video and form a Kinetics-Skeleton dataset. It contains about 240,000 video clips for training and 20,000 video clips for testing. Considering that this dataset has multiple categories and is difficult to identify, we use the accuracy of Top-1 and Top-5 as the evaluation criteria.

B. Training details

Our experiments are carried out on the PyTorch deep learning framework with NVIDIA GeForce RTX4080 GPU. On NTU RGB+D, NTU RGB+D 120 and Kinetics-Skeleton datasets, batch size is set to 16. We employ Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay 1e-4 to train our model for 80 epochs. For NTU RGB+D dataset, the learning rate is set as 0.05 and is divided by 10 at 40-th and 60-th epoch, separately. For NTU RGB+D 120 and Kinetics-Skeleton datasets, the learning rate is set as 0.05 and is divided by 10 at 50-th and 60-th epoch, separately. We follow the data preprocessing methods[1] in MS-G3D and train the network model using the Joint stream and the Bone stream, respectively. We adopt 2s-AGCN[2] as the baseline.

C. Ablation study

We will verify the effectiveness of this method through extensive ablation experiments on NTU RGB+D dataset, including DST-GCN, WA-GCN, EMS-GCN, and SED-PGL.

DST-GCN: We use Joint and Bone data streams to train DST-GCN, and obtain the recognition accuracies of different data streams on different spatiotemporal modeling paths. On the X-View and X-Sub benchmarks, the recognition accuracies obtained by STM-Path, TSM-Path and Dual-Path are presented in the form of histograms as shown in Fig. 6(a) and (b), respectively. The recognition accuracy of the Dual-Path spatiotemporal modeling approach exceeds that of both STM-Path and TSM-Path on two different benchmarks demonstrating the effectiveness of the Dual-Path framework in DST-GCN. Particularly, as shown in Fig. 6(a), STM-Path and TSM-Path achieve the accuracy of 94.83% and 94.95%, respectively, when DST-GCN has been trained on X-View Benchmark using Joint data stream. We have performed score fusion on the predictions of STM-Path and TSM-Path, and the model has achieved the accuracy of 95.83%. Similarly, by employing the Bone data stream to train DST-GCN, the Dual-Path score fusion can effectively improve the recognition accuracy of the model. According to Fig. 6(b), our experiments on X-Sub Benchmark have achieved the same enhancement. Besides, we present the confusion matrices of the DST-GCN, baseline method 2s-AGCN [9] and the difference of these two methods in Fig. 7. The diagonal elements of the confusion matrix represent the accuracy of each category. Our method is generally superior to the baseline method. Specifically, the red dashed box in Fig. 7 (c) shows the difference in recognition accuracy between the two methods demonstrating that the the accuracy improvement is around 4%-12% in some action categories.

To concretely demonstrate the effectiveness of DST-GCN, we train DST-GCN on the X-View benchmark of NTU-RGB+D dataset using Joint data stream and visualize the recognition accuracies of STM-Path, TSM-Path, and Dual-Path on each category, as shown in Fig. 8. The number of correctly predicted samples in each category is divided by the total number of samples in that category to obtain the recognition accuracy for each category. We can see that Dual-Path has higher recognition rates than STM-Path and TSM-Path for most of the categories. This demonstrates that the two spatio-temporal modeling approaches can extract synergistic and complementary spatio-temporal features. The action recognition rate is further improved after the score fusion of the dual path. As we mentioned in Introduction, ”sit down” and ”pick up” have interdependent spatiotemporal features. Dual-Path improves the recognition accuracy on both categories which shows that skeleton actions do contain temporal and spatial features with coupling. For ”reading”, ”writing” and ”clapping”, traditional spatiotemporal modeling

[1]https://github.com/kenziyuliu/ms-g3d
[2]https://github.com/lshiwjx/2s-AGCN

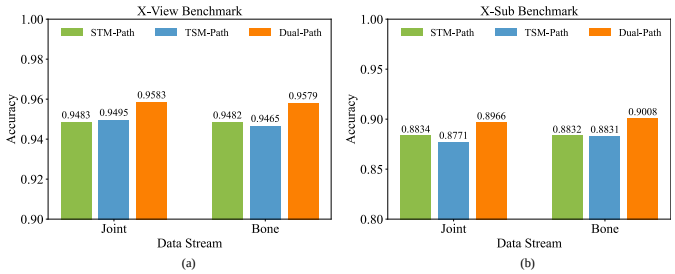


Fig. 6. Accuracy achieved by Joint data stream and Bone data stream on different spatiotemporal modeling paths on X-View and X-Sub benchmarks.

TABLE II
ACCURACY OBTAINED BY DST-GCN ON DIFFERENT DATA STREAMS.

Methods	NTU RGB+D	
	X-View(%)	X-Sub(%)
DST-GCN(Joint)	95.83	89.66
DST-GCN(Bone)	95.79	90.08
DST-GCN(2s)	**96.53**	**91.06**

TABLE III
DST-GCN CONTAINS DIFFERENT TEMPORAL FEATURE EXTRACTION MODULES.

Methods	NTU RGB+D	
	X-View(%)	X-Sub(%)
DST-GCN with TCN	95.84	90.34
DST-GCN with MS-TCN	96.26	90.84
DST-GCN with EMS-TCN	**96.53**	**91.06**

TABLE IV
ACCURACY ACHIEVED BY DIFFERENT LOSS FUNCTIONS ON NTU RGB+D DATASET.

Loss Function	NTU RGB+D	
	X-View(%)	X-Sub(%)
L_{CEL}	96.36	90.72
L_{CSP} with $\delta = 0.01$	96.43	90.84
L_{CSP} with $\delta = 0.001$	**96.53**	**91.06**
L_{CSP} with $\delta = 0.0001$	96.19	90.54

(STM-Path) has potential discrimination. TSM-Path, by virtue of its temporal modeling followed by spatial modeling, pays more attention to the features of the temporal dimension, which complements the features obtained by STM-Path. In summary, STM-path and TSM-path are able to capture focused feature representations of actions in spatial and temporal dimensions, avoiding potential discrimination. Dual-Path can further improves the recognition accuracy.

To show the feature extraction capability of the proposed DST-GCN framework, we visualize the feature space of the vectors F^{st}_{joint}, F^{ts}_{joint}, F^{st}_{bone}, and F^{ts}_{bone}, respectively using the t-SNE tool, as shown in Fig. 9(a). Each dot represents a sample, and different colors represent different categories. The feature vectors extracted by different spatiotemporal modeling paths under the same data stream have different spatial distributions, which essentially reflects the diversity of features acquired by STM-path and TSM-path. The dual-stream four-pathway framework of DST-GCN proposed in this paper demonstrates excellent feature extraction capability, and the spatio-temporal features extracted from different spatiotemporal modeling paths are complementary, which is conducive to learning comprehensive spatiotemporal representations.

According to previous works [9], [40], using multi data streams to train the network can significantly improve the generalization performance of the model. The experimental results of using two data streams, viz., Joint and Bone, to train the network are shown in Table II. On X-View benchmark, the recognition accuracy of the models has reached the accuracy of 95.83% and 95.79%. Further, the recognition accuracy of the two streams network model has reached 96.53%, which exceeds by 0.7% and 0.74% compared to the Joint data stream network and Bone data stream network, respectively. Similarly, on X-Sub Benchmark, the two-stream network model has improved by 1.4% and 0.98% over the Joint data stream network and the Bone data stream network, respectively, to achieve the accuracy of 91.06%. The experimental results show that the strategy of training the network with multiple data streams is also effective for the proposed DST-GCN. The combination of the two-stream strategy and the two spatiotemporal modeling approaches can greatly enhance the ability of the model to extract features and learn spatiotemporal features with comprehensiveness as a way to improve the model generalization ability.

In addition, we also plot a scatterplot showing the relationship between the different methods in terms of accuracy and model complexity, as shown in Fig. 10. For fair comparison, the methods we have shown are all trained using one data stream. DST-GCN is a dual path parallel framework, which broadens the width rather than the depth of the model and increases the ease of model deployment. In addition, the method based on adaptive graph structure can efficiently acquire the spatial information of the human skeleton, and the use of dilated convolution and convolution kernel 1×1 effectively reduces the number of parameters. Compared with the baseline model [9], the proposed DST-GCN improves the recognition accuracy while reducing the spatial complexity of the model. Compared with other methods, our method achieves a balance between performance and complexity.

WA-GCN: We mentioned in the WA-GCN that the topology of the human skeleton is represented by an adjacency matrix of size 25×25. A_k is a predefined skeleton topology with fixed weights and does not follow the iteration of the network to update the parameters. The parameters of B_k are adaptively learned from all the samples, which can create connections between joints that do not exist in reality, and help to explore the hidden spatial information in the human skeleton. The values in B_k follow the iteration of the network to update. Similarly, the parameter training of C_k relies on the sample features extracted from the previous layer of the network, and is sample-adaptive. $A_k + B_k$ form the basis of the human skeleton topology, and C_k is sample adaptive and will dynamically adjust to different samples. In addition, the adaptive weight ω_k, as the network parameter, is conducive to better feature aggregation. On the X-View benchmark of NTU RGB+D dataset, taking STM-Path as an example, we visualize the training process of the skeleton joints of "drink water" as the number of network layers deepens, and the experimental

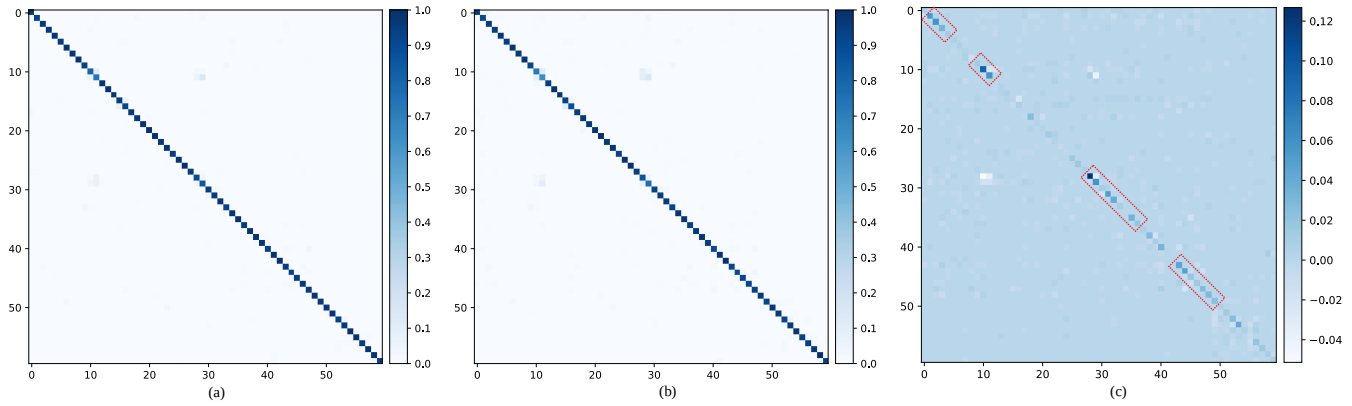


Fig. 7. Visualization of the confusion matrix. (a) denotes the proposed DST-GCN. (b) denotes the baseline 2s-AGCN. (c) denotes the difference of subtracting (a) from (b). Different colors indicate the proportion of each correctly or incorrectly classified category.

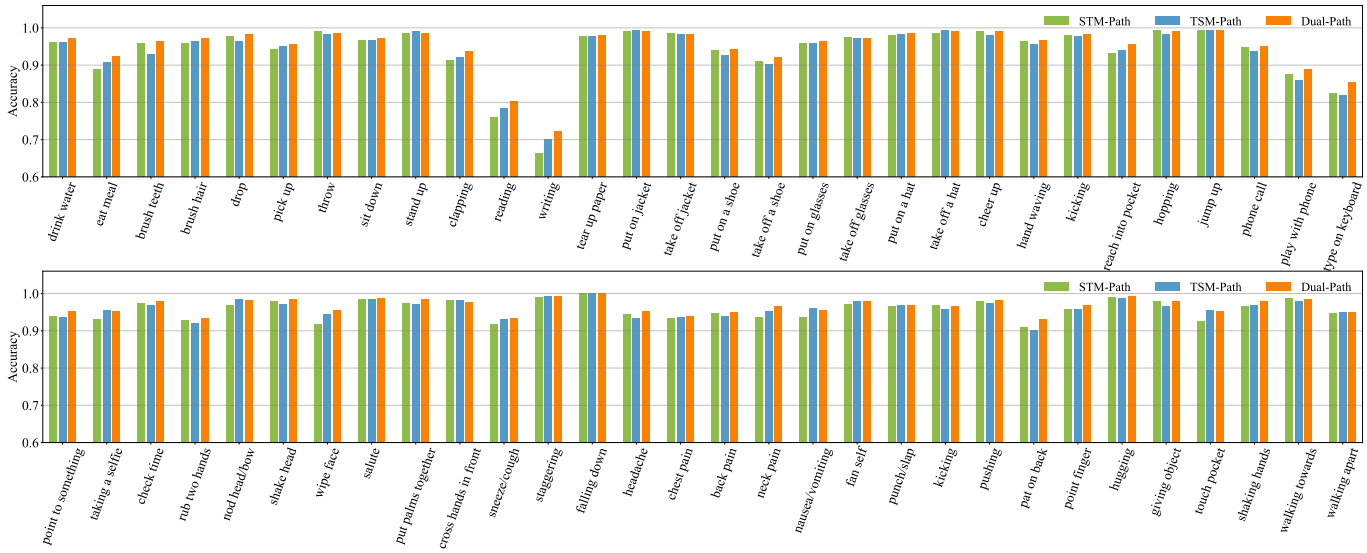


Fig. 8. Comparison of the accuracy achieved by STM-Path, TSM-Path and Dual-Path in NTU RGB+D dataset. The accuracy of each action category in the dataset is presented in two rows.

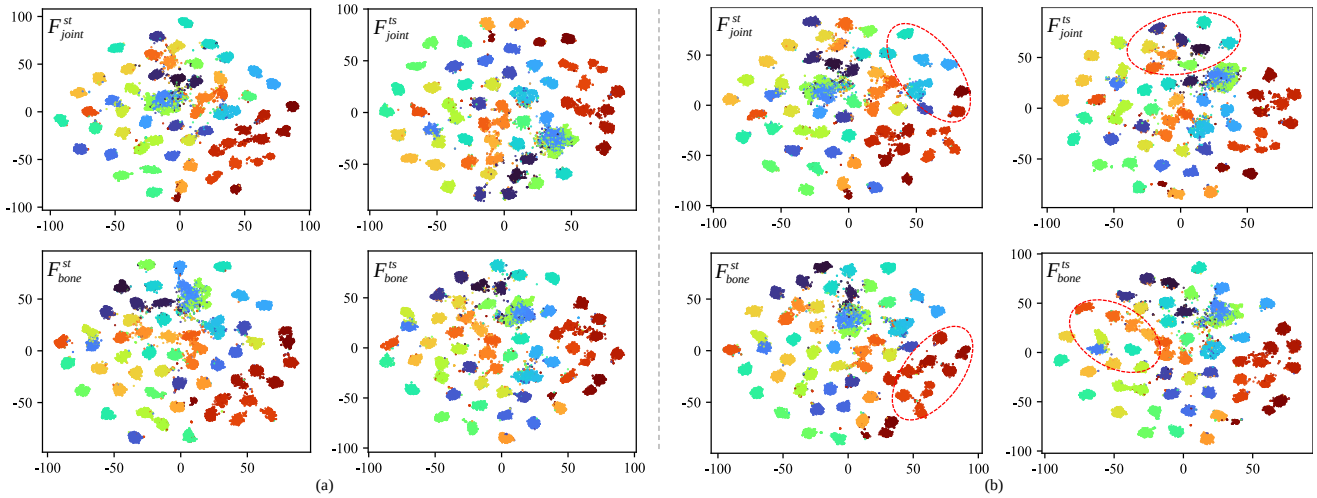


Fig. 9. Feature vectors representation with t-SNE from NTU RGB+D dataset. (a) denotes the representations of F^{st}_{joint}, F^{ts}_{joint}, F^{st}_{bone}, and F^{ts}_{bone} extracted from DST-GCN when L_{CSP} is used as the loss function, respectively. (b) denotes the representations of F^{st}_{joint}, F^{ts}_{joint}, F^{st}_{bone}, and F^{ts}_{bone} extracted from DST-GCN when L_{CEL} is used as the loss function, respectively.

TABLE V
COMPARISONS WITH THE STATE OF-THE-ART METHODS ON THE NTU RGB+D DATASET, NTU RGB+D 120 AND KINECT-SKELETON.

Methods	Source	Parameters (M)	NTU RGB+D		NTU RGB+D 120		Kinetics-Skeleton	
			X-View(%)	X-Sub(%)	X-Set(%)	X-Sub(%)	Top-1(%)	Top-5(%)
ST-GCN [8]	AAAI 2018	3.12	88.3	81.5	-	-	30.7	52.8
2s-AGCN [9]	CVPR 2019	3.47	95.1	88.5	-	-	36.1	58.7
AS-GCN [39]	CVPR 2019	-	94.2	86.8	-	-	34.8	56.5
AGC-LSTM [24]	CVPR 2019	-	95.0	89.2	-	-	-	-
DGNN [10]	CVPR 2019	13.12	96.1	89.9	-	-	36.9	59.6
VA-fusion [43]	TPAMI 2019	-	95.0	89.4	-	-	-	-
MS-G3D [11]	CVPR 2020	3.2	96.2	**91.5**	88.4	86.9	38.0	**60.9**
Shift-GCN [52]	CVPR 2020	-	96.5	90.7	87.6	85.9	-	-
NAS-GCN [53]	AAAI 2020	-	95.7	89.4	-	-	37.1	60.1
MS-AAGCN [40]	TIP 2020	3.78	96.2	90.0	-	-	37.8	**61.0**
Hyper-GNN [54]	TIP 2021	-	95.7	89.5	-	-	37.1	60.0
2s MST-GCN [41]	AAAI 2021	3.0	96.4	91.1	88.3	87.0	37.8	60.3
SEFN [44]	TCSVT 2021	3.5	96.4	90.7	87.8	86.2	**39.3**	**62.1**
CDGC [55]	TCSVT 2022	-	96.5	90.9	87.8	86.3	-	-
FGCN [56]	TIP 2022	-	96.3	90.2	87.4	85.4	-	-
2s Graph2Net [57]	TCSVT 2022	0.9	96.0	90.1	87.6	86.0	36.4	-
4s STF-Net [58]	PR 2023	6.8	96.5	91.1	88.2	86.5	36.1	58.9
SMotif-GCN [59]	TPAMI 2023	-	96.1	90.5	87.7	87.1	37.8	**60.6**
4s-MS&TA-HGCN-FC [60]	TCSVT 2023	-	96.4	90.8	88.4	87.0	**38.7**	**62.3**
EfficientGCN-B2 [61]	TPAMI 2023	-	95.7	**91.4**	87.8	**88.0**	-	-
EfficientGCN-B4 [61]	TPAMI 2023	-	96.1	**92.1**	**88.9**	**88.7**	-	-
LKA-GCN [62]	TVCG 2023	-	96.1	90.7	87.8	86.3	37.8	60.9
DST-GCN (2s)	-	2.36	**96.5**	**91.1**	**88.7**	**87.3**	**38.3**	**60.4**

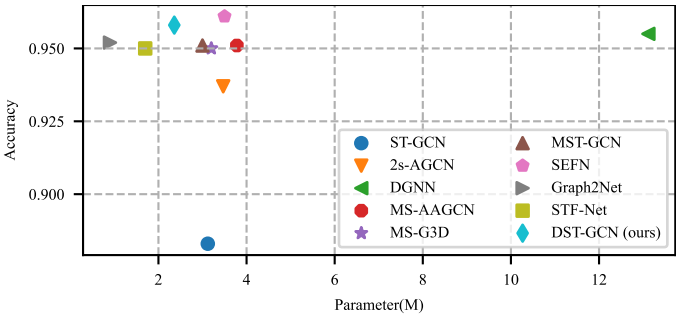


Fig. 10. Parameter vs. accuracy on NTU RGB+D X-View benchmark.

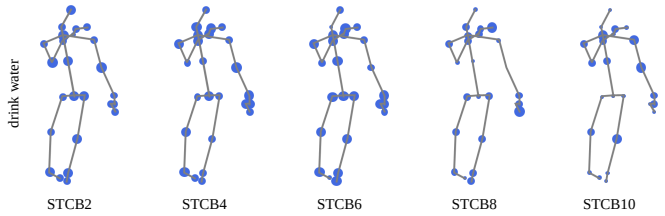


Fig. 11. Visualization of the human skeleton at different network layers.

results are shown in Fig. 11. Each circle represents a joint, and its size indicates its importance in the current action. For the action "drink water", the joints at the shoulder, hand and near the head are of high importance, and the spatial information of the relevant joints helps to recognize the action.

EMS-TCN: To verify the effectiveness of the EMS-TCN proposed in this paper, we replace this module with the TCN module in 2s-AGCN [9] and the MS-TCN module in MS-G3D [11], respectively, while keeping the rest of the parameters unchanged, and then compare the model accuracy. The experimental results are shown in Table III. DST-GCN with the EMS-TCN module achieves an accuracy of 96.53% on X-View Benchmark, which is 0.27% higher than the accuracy achieved by DST-GCN with the MS-TCN module, and 0.69% higher than the accuracy achieved by DST-GCN with the TCN module. Similarly, on X-Sub benchmark, the DST-GCN with the EMS-TCN module has achieved accuracy of 91.06%, which is an improvement of 0.22% and 0.72%, respectively. The TCN module shown in Fig. 4(a) uses only convolution kernel 9×1 and therefore cannot extract multi-scale temporal features. The MS-TCN module shown in Fig. 4(b) can extract multi-scale temporal features, but it is unable to obtain cross-channel information, resulting in a lack of ability to aggregate features. The EMS-TCN extracts multi-scale temporal features and aggregates the extracted multiscale temporal features both globally and locally according to the inductive bias, for strengthening its long-term and short-time action characterization in the temporal dimension, which is conducive for mining the temporal features in the action.

SED-PGL: As shown in Eq. (15), the L_{CSP} proposed in this paper has two parameters, β and δ. Based on a set of comparison experiments, we set $\beta = 0.5$. For the smoothing factoe δ, we design a set of comparison experiments, and the experimental results are shown in Table IV. Meanwhile, we use L_{CEL} alone as the loss function of DST-GCN to compare the recognition accuracy achieved, and the experimental results are also shown in Table IV. On the X-View benchmark of the NTU RGB+D dataset, using L_{CSP} with $\delta = 0.001$ as the loss function, the model achieved an accuracy of 96.53%, which is 0.17% higher than when using L_{CEL} as the loss function. On X-Sub benchmark, using L_{CSP} with $\delta = 0.001$ as the

loss function, the model achieves 91.06% accuracy, which is 0.34% higher than when L_{CEL} is used as the loss function. Furthermore, we observe that the model achieves the best performance when the smoothing factor $\delta = 0.001$ in L_{CSP}. Besides, we train the proposed network with L_{CEL} alone and visualize the four feature vectors, F^{st}_{joint}, F^{ts}_{joint}, F^{st}_{bone}, and F^{ts}_{bone} as shown in Fig. 9(b). Compared with Fig. 9(a), when using L_{CEL} as the loss function, the distance between samples of the same category is large, not compact enough, and the expression of samples with similar actions is not clear enough to differentiate. Details can be seen in the places marked by the red round boxes.

D. Compared with the state-of-the-art

We compare the proposed DST-GCN with several recent state-of-the-art (SOTA) methods to demonstrate the effectiveness for skeleton-based actions recognition. For fair comparison, we follow the data preprocessing method [9], [11] without any other data augmentation. The results achieved by the proposed method on NTU RGB+D, NTU RGB+D 120 and Kinetics-Skeleton datasets are shown in Table V.

First, we compare the results on the NTU RGB+D and NTU RGB+D 120 datasets. For NTU RGB+D dataset, there are some typical methods we should nocite. ST-GCN [8] is the pioneering work which applying GCN in the skeleton-based action recognition. Our DST-GCN surpasses ST-GCN by 8.2% on the X-View benchmark and 9.6% on the X-Sub benchmark. Compared to the baseline 2s-AGCN [9], we improve by 1.4% on the X-View Benchmark and 2.6% on the X-Sub Benchmark. DST-GCN also outperforms other classic GCN-based methods which contains single spatiotemporal modeling approach, such as AS-GCN [39], MS-AAGCN [40] and 2s MST-GCN [41] on both benchmarks with less parameters. VA-fusion [43], SEFN [44], MS-G3D [11] are all designed from the perspective of spatiotemporal modeling. Compared with these methods, the proposed DST-GCN outperforms in terms of model complexity and recognition accuracy, demonstrating the effectiveness of our proposed method. It is important to note that EfficientGCN [61] provides a family baselines. EfficientGCN-Bx is termed with different scaling coefficient "x". EfficientGCN-Bx using three data streams (joint, bone and velocity) has slightly higher recognition accuracy than DST-GCN on X-sub benchmark. In contrast, the proposed DST-GCN achieved the similar performance only using two data stream (joint and bone). Multi-stream data fusion is also one of our future research directions. Similar experimental results appear in the NTU RGB+D 120 X-Sub benchmark. Besides, the proposed DST-GCN also has exciting results on large-scale NTU RGB+D 120 dateset. DST-GCN outperforms most GCN-based methods on the X-Set benchmark and X-Sub benchmark of NTU RGB+D 120. For example, DST-GCN outperforms the current SOTA method, MS-G3D [11], by 0.3% and 0.4% on X-Set Benchmark and X-Sub Benchmark, respectively.

Second, we will compare the results on Kinetics-Skeleton dataset. This dataset is derived from real-world scenarios and has 400 action categories, thus resulting in a low recognition rate. The proposed DST-GCN achieves the Top-1 accuracy of 38.4% and Top-5 accuracy of 60.4%. We need to note that, compared to DST-GCN, SEFN [44] and 4s-MS&TA-HGCN-FC [60] exhibit better recognition performance. Specifically, the above two methods [44], [60] both use four data streams (joint, bone, joint-motion and bone-motion) to train the network. These four skeleton data streams contain richer spatiotemporal information, thus improving the recognition accuracy of the model. On the other hand, improving the recognition accuracy of the model in real complex environments is one of the problems we will try to solve in the future. In summary, all above experimental results verify the effectiveness of our method.

VI. Conclusion

In this paper, first, we have proposed DST-GCN intended for learning the comprehensive spatiotemporal representations of skeleton-based actions, based on the analysis of action patterns and feature extraction networks. DST-GCN designs two feature extraction networks with different spatiotemporal modeling orders to adapt the action patterns exhibited by skeleton-based actions in both the temporal and spatial dimensions. Second, under the framework of DST-GCN, we proposed WA-GCN and EMS-TCN for the spatial and temporal feature extraction, respectively. WA-GCN is a data-driven network that adaptively learns the spatial topology of the human body, establishes the dependencies between different joints, and learns the corresponding weights for different joint partitioning strategies. EMS-TCN focuses on multi-scale temporal features and strengthens the long and short-time action representations of skeleton-based actions in the temporal dimension from both the global and local aspects. Finally, we also designed SED-PGL to train a GCN-based model, by controlling the distance between samples to better distinguish the similar actions. We have proved the effectiveness of the proposed framework and the network modules after a series of ablation experiments. The proposed method outperforms many SOTA works on NTU RGB+D, NTU RGB+D 120 and Kinetics-Skeleton datasets.

For future work, we will explore the following three aspects. First, we will focus on learning robust spatio-temporal features with coupling in skeleton sequence by incorporating contrastive learning and mixed model (GCN and Transformer). Second, for the problem of lack of sufficient labeled data in real-world scenarios, we will explore skeleton-based action recognition under semi-supervised and unsupervised learning frameworks. Third, we are interested to combine skeleton data with RGB data to accomplish multimodal human action recognition which can be a powerful solution to the problems existing in potential practical applications [63], [64], such as object occlusion and multi-person recognition.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62076154, U21A20513), the Anhui Provincial Natural Science Foundation (2008085MF202), the Provincial Graduate Academic Innovation Project (2022xscx145), and the Anhui Province Student Innovation Training Project (S202311059117, 1602575875168014336).

REFERENCES

[1] A. Tong, C. Tang, and W. Wang, "Semi-supervised action recognition from temporal augmentation using curriculum learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1305–1319, 2023.

[2] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1895–1904.

[3] J. Hong, M. Fisher, M. Gharbi, and K. Fatahalian, "Video pose distillation for few-shot, fine-grained sports action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9254–9263.

[4] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2020.

[5] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, and Q. Miao, "Transformer for skeleton-based action recognition: A review of recent advances," *Neurocomputing*, vol. 537, pp. 164–186, 2023.

[6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[7] L. Feng, Y. Zhao, W. Zhao, and J. Tang, "A comparative review of graph convolutional networks for human skeleton-based action recognition," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 4275–4305, 2022.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[10] Shi, Lei and Zhang, Yifan and Cheng, Jian and Lu, Hanqing, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7912–7921.

[11] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 568–576.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[15] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[16] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[18] P. Koniusz, A. Cherian, and F. Porikli, "Tensor representations via kernel linearization for action recognition from 3d skeletons," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 37–53.

[19] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4471–4479.

[20] R. Vemulapalli, F. Arrate, and R. Chellappa, "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*, 2014, pp. 588–595.

[21] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[22] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 816–833.

[23] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2018.

[24] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[25] S. Guan, H. Lu, L. Zhu, and G. Fang, "Afe-cnn: 3d skeleton-based action recognition with action feature enhancement," *Neurocomputing*, vol. 514, pp. 256–267, 2022.

[26] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[27] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer lstm networks," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2330–2343, 2018.

[28] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5457–5466.

[29] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1012–1020.

[30] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 103–118.

[31] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3288–3297.

[32] Ke, Qiuhong and Bennamoun, Mohammed and An, Senjian and Sohel, Ferdous and Boussaid, Farid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2842–2855, 2018.

[33] F. Ren, C. Tang, A. Tong, and W. Wang, "Skeleton-based human action recognition by fusing attention based three-stream convolutional neural network and svm," *Multimedia Tools and Applications*, vol. 83, p. 6273–6295, 2024.

[34] A. Hernandez Ruiz, L. Porzi, S. Rota Bulò, and F. Moreno-Noguer, "3d cnns on distance matrices for human action recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2017, pp. 1087–1095.

[35] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *In 9th International Conference on Learning Representations, ICLR*, 2021.

[37] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "Stst: Spatial-temporal specialized transformer for skeleton-based action recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, p. 3229–3237.

[38] Y. Pang, Q. Ke, H. Rahmani, J. Bailey, and J. Liu, "Igformer: Interaction graph transformer for skeleton-based human interaction recognition," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 605–622.

[39] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[40] Shi, Lei and Zhang, Yifan and Cheng, Jian and Lu, Hanqing, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[41] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1113–1122, 2021.

[42] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 608–10 617.

[43] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

[44] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4394–4408, 2021.

[45] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[46] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and deep locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2014.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[48] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: a large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–19.

[49] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[50] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[51] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[52] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192.

[53] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2669–2676.

[54] X. Hao, J. Li, Y. Guo, T. Jiang, and M. Yu, "Hypergraph neural network for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2263–2275, 2021.

[55] S. Miao, Y. Hou, Z. Gao, M. Xu, and W. Li, "A central difference graph convolutional operator for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4893–4899, 2021.

[56] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank, "Feedback graph convolutional network for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 164–175, 2021.

[57] C. Wu, X.-J. Wu, and J. Kittler, "Graph2net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2120–2132, 2021.

[58] L. Wu, C. Zhang, and Y. Zou, "Spatiotemporal focus for skeleton-based action recognition," *Pattern Recognition*, vol. 136, p. 109231, 2023.

[59] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, S. Xia, and Y.-J. Liu, "Motif-gcns with local and non-local temporal blocks for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2009–2023, 2022.

[60] Z. Huang, Y. Qin, X. Lin, T. Liu, Z. Feng, and Y. Liu, "Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1868–1883, 2022.

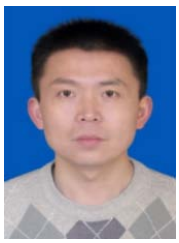
[61] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2022.

[62] Y. Liu, H. Zhang, Y. Li, K. He, and D. Xu, "Skeleton-based human action recognition via large-kernel attention graph convolutional network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2575–2585, 2023.

[63] M. Amsaprabhaa, Y. N. Jane, and H. K. Nehemiah, "Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection," *Expert Systems With Applications*, vol. 212, 2023.

[64] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human interaction learning on 3d skeleton point clouds for video violence recognition," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 74–90.

Fang Ren received his B.S. degree from the School of Information Engineering, Nanhang Jincheng College, China, in 2020. He is currently pursuing a M. E. degree at Hefei University under the supervision of Dr. Chao Tang. His research interests include skeleton-based action recognition, machine learning, and deep learning.



Chao Tang received his M.E. degree from the School of Computer and Information Technology, Shanxi University, in 2009, and his Ph.D. degree from the School of Information Science and Technology, Xiamen University, China, in 2014. In 2016, he was a visiting scholar at Bloomfield College, U.S.A. From 2017 to 2018, he was a visiting scholar at the University of Birmingham, U.K. Since December 2018, he has been an associate professor at the School of Artificial Intelligence and Big Data, Hefei University, China. His research interests include machine learning and computer vision.

Anyang Tong received his M. E. degree from the School of Artificial Intelligence and Big Data, Hefei University, China, in 2023. He is currently pursuing a Ph.D. degree at Hefei University of Technology under the supervision of Dr. Meng Wang. His research interests include semi-supervised action recognition, curriculum learning, and deep learning.

Wenjian Wang received her Ph.D. degree from Institute for Information and System Science, Xi'an Jiaotong University in 2004. Now she is a full-time professor and Ph.D. supervisor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education, Shanxi University. She has published more than 170 academic papers. Her current research interests include machine learning and computational intelligence, etc.