Motif-GCNs with Local and Non-Local Temporal Blocks for Skeleton-Based Action Recognition

Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, Shihong Xia, Yong-Jin Liu

Abstract—Recent works have achieved remarkable performance for action recognition with human skeletal data by utilizing graph convolutional models. Existing models mainly focus on developing graph convolutional operations to encode structural properties of the skeletal graph, whose topology is manually predefined and fixed over all action samples. Some recent works further take sample-dependent relationships among joints into consideration. However, the complex relationships between arbitrary pairwise joints are difficult to learn and the temporal features between frames are not fully exploited by simply using traditional convolutions with small local kernels. In this paper, we propose a motif-based graph convolution method, which makes use of sample-dependent latent relations among non-physically connected joints to impose a high-order locality and assigns different semantic roles to physical neighbors of a joint to encode hierarchical structures. Furthermore, we propose a sparsity-promoting loss function to learn a sparse motif adjacency matrix for latent dependencies in non-physical connections. For extracting effective temporal information, we propose an efficient local temporal block. It adopts partial dense connections to reuse temporal features in local time windows, and enrich a variety of information flow by gradient combination. In addition, we introduce a non-local temporal block to capture global dependencies among frames. Our model can capture local and non-local relationships both spatially and temporally, by integrating the local and non-local temporal blocks into the sparse motif-based graph convolutional networks (SMotif-GCNs). Comprehensive experiments on four large-scale datasets show that our model outperforms the state-of-the-art methods. Our code is publicly available at https://github.com/wenyh1616/SAMotif-GCN.

Index Terms—Action Recognition, Graph Convolutional Neural Networks, Spatio-Temporal Attention, Non-Local Block, Skeleton Sequence.

1 Introduction

With widespread applications like human-computer interaction, medical monitoring, and video surveillance, researchers have been paying more attention to action recognition. Meanwhile, skeleton sequences have become widely and cheaply available due to the development of real-time human pose estimation technologies [1], [2]. Compared to video modalities such as raw RGB or RGB-D data, human skeletons are more compact and thus can significantly improve computational efficiency for classifying human actions. Human skeletons also provide relatively high-level structural information, leading to better action recognition performance, especially for scenes with complicated background [3], [4], [5]. Many existing deep-learning based methods structurize a skeleton sequence by a time series of 2D or 3D joint coordinates or pseudo-images. These data are then sent into recurrent neural networks (RNNs) [5], [6], [7], [8], [9], [10], [11], [12] or convolutional neural networks (CNNs) [4], [10], [13], [14] to capture both intra-frame features and temporal dependencies among frames. However, a more natural way to represent a skeleton is a graph where human body joints and bones are treated as nodes and edges, respectively. Moreover, the topology of the skeleton can be fully exploited in such a graph representation.

Graph convolutional neural networks (GCNs), which generalize CNNs from regular grid-shaped maps to unordered graphs of arbitrary structures have gained increasing attention (e.g., [15], [16], [17], [18], [19], [20]). The GCNs generally stack layers of graph convolution (GC) operations and temporal modeling operations (e.g., traditional convolutions in the temporal domain) to generate deep spatio-temporal features, which are fed into a classifier for action recognition. Existing researches mainly focus on developing new GC operations to capture skeleton structures [18], [19], [21], [22]. By considering the skeletal graphs modeled by GC operations, these GCN methods can be broadly classified into two classes: sample-independent and sample-dependent.

The sample-independent methods construct GCNs to encode skeletal graphs [15], [16], [17], [20], whose topology structures are fixed over all action samples. The sample-dependent methods take the characteristics of different actions into consideration by using a unique graph structure for each sample action [18], [19], [23]. The detailed summary of these two classes are presented in Sec. 2.1. Although the sample-dependent methods [18], [19] take latent and characteristic relationships among joints into consideration, as shown in Fig. 1, so far the state-of-the-art (SOTA) performance of sample-independent class (i.e., Shift-GCN [20]) is better than that of sample-dependent class (i.e., AS-GCN [19]). The reason is possibly that it is difficult to extract

- *Yu-Hui Wen and Yong-Jin Liu are with the BNRist, Department of Computer Science and Technology, Tsinghua University. E-mail: {wenyh1616,liuyongjin}@tsinghua.edu.cn*
- *Lin Gao and Shihong Xia are with Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences and also with University of Chinese Academy of Sciences. Email: {gaolin,xsh}@ict.ac.cn*
- *Hongbo Fu is with the School of Creative Media, City University of Hong Kong. Email: hongbofu@cityu.edu.hk*
- *Fang-Lue Zhang is with the School of Engineering and Computer Science, Victoria University of Wellington. Email: fanglue.zhang@ecs.vuw.ac.nz*
- *Y.J Liu and L. Gao are the corresponding authors.*

proper sample-dependent relationships among all pairwise joints, which have a relatively high capacity involving $V \times V$ variables for the skeleton of V joints. Moreover, the methods [18], [19] encode temporal features by simply using traditional convolutions, in which the temporal information in local windows are not fully exploited and global temporal dependencies among frames are not considered.

In this paper, we propose a novel method to effectively improve the performance of the sample-dependent class, which outperforms the SOTA performance of sample-independent class. In particular, we propose a novel sparse motif-based graph convolution (SMotif-GC) that imposes sparse sample-dependent relationships between physically disconnected joints, instead of any pairwise joints [18], [19]. Moreover, we propose a novel local temporal block (LTB) to extract temporal features with partial dense connections by reusing feature maps in local time windows. What's more, it can reduce redundant information by truncating the input feature maps, and enrich a variety of gradient propagation by combining partial feature maps from the input to the output. Essentially, SMotif-GC layers capture local and global relationships between joints for modeling spatial information. To further capture the whole range of temporal dependencies by the attention mechanism [24], we integrate a non-local temporal block (NLTB) into our network. This NLTB leads to an effective representation of the skeleton sequence by constructing global relationships between arbitrary pair of frames.

This paper is an extension of our preliminary conference version [23]. We have made the following improvements: 1) We define a sparse motif adjacency matrix, which can be learned during training to better capture skeletal features with motif-based graph convolutions. 2) We propose a novel local temporal block with partial dense connections based on the variable temporal dense block in the conference version, to improve the encoding ability of local temporal features. It can reduce the large amount of duplicated gradient propagation and enrich a variety of information flow in the variable temporal dense block [23]. 3) We make a more thorough evaluation on four large-scale action recognition datasets (instead of two datasets in [23]).

To sum up, the major contributions of our deep learning architecture for skeleton-based human action recognition include: 1) We propose sparse motif-based graph convolutions, which simultaneously encode hierarchical spatial information and high-order sample-dependent locality. Specifically, we introduce a sparse sample-dependent adjacency matrix for capturing useful latent dependencies among joints. 2) We propose local temporal blocks with partial dense connections for encoding richer local temporal information of skeleton sequences by reusing feature maps and enriching a variety of information flow. 3) We propose to use a non-local temporal block for constructing global dependencies between frames to get a more effective representation of the skeleton sequence. The local temporal blocks are combined with the non-local temporal block to extract temporal features with respect to local and global dependencies between frames. Our extensive evaluation on four large-scale action recognition datasets shows that our model outperforms the state-of-the-art methods.

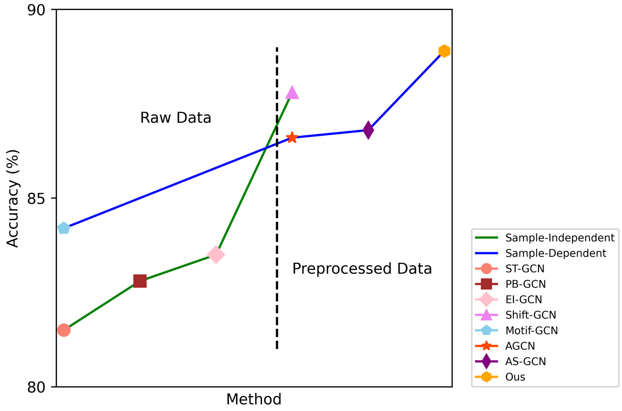


Fig. 1. Comparisons of sample-independent and sample-dependent GCNs for action recognition on NTU-RGB+D (X-Sub evaluation), in terms of accuracy. Our model proposed in this paper achieves the best performance among all the sample-independent and sample-dependent methods. Specifically, it achieves higher accuracy by a relative large margin (88.9% *v.s.* 86.6%) than the previous state-of-the-art method AS-GCN in the same category. The previous GCN-based methods (on the left side of the dashed line) use raw data of joint coordinates as input, while the recent methods (on the right side of the dashed line) use preprocessed joint data. More details about the data preprocessing and comparison results can be found in Sec. 4.6.

2 Related Work

In recent years, skeleton-based action recognition has drawn more and more attention in the industry, and been explored in different aspects for capturing spatial and temporal patterns of skeleton sequences [4], [5], [8], [13], [14], [25]. In this section, we review GCNs and recent approaches which use GCNs for modeling skeletal graphs.

2.1 Graph Convolution

GCNs aim to generalize CNNs to irregular graph structures. According to the principles of defining convolutional operations over graphs, GCNs are generally classified into spatial-domain and spectral-domain approaches. The spatial-domain approaches have to preprocess data and define graph convolutions on the nodes and their neighbors directly [26], [27], [28]. In contrast, the spectral-domain approaches define graph convolutions in the frequency domain [29], [30], [31]. Based on a Chebyshev expansion of graph Laplacian, an improved spectral filtering is proposed [30]. It removes intense computation and yields spatially localized filters for graph convolutional layers. Furthermore, Uni-GCN [31] simplifies the spectral filtering to operate on 1-ring neighborhood, in which the neighboring nodes have a uniform weight of importance. It can be applied to encode skeleton structures.

As presented in Sec. 1, for the application of skeleton-based action recognition, two classes of GCN-based methods exist: sample-independent and sample-dependent. The methods in the sample-independent class consider structural features of the skeleton, whose graph topology is fixed over all action samples. For example, spatio-temporal GCN (ST-GCN) [15] applies graph convolutions on physically connected neighbors of each joint [31] to encode spatial features of the skeletal data in each frame. They also propose a partitioning strategy by dividing neighbors of a joint

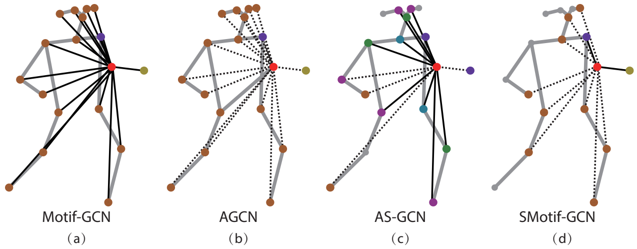


Fig. 2. Skeletal graphs modeled by four sample-dependent GCNs. These GCN methods construct dependencies between the current joint (shown in red color) and its related joints in different partitions (shown in different colors). The dependencies are predefined (shown as black solid lines) or learned from training (shown as black dotted lines). The joints and the physical connections that are not considered for the graph convolution operation on the current joint are shown in grey. (a) Motif-GCN [23] uses predefined relationships between the current joint and its related joints in different partitions. (b) AGCN [18] learns sample-dependent relationships between all pairwise joints. (c) AS-GCN [19] adopts predefined relationships between the current joint and its 1-hop to 4-hop neighbors (shown in purple, blue, green and magenta), and learns sample-dependent links between arbitrary pairwise joints. (d) Our proposed SMotif-GCN learns sparse sample-dependent relationships between the current joint and its non-physically connected joints. The physically connected dependencies are predefined on 1-ring neighborhood. Skeletal graphs modeled by different sample-independent GCNs are illustrated in Fig. A3 of the appendix.

into different subsets. Extrinsic-intrinsic GCN (EI-GCN) [16] utilizes both the extrinsic and intrinsic relationships among joints in a L-hop neighborhood [30]. Shift-GCN [20] adopts shift operations to gather information from all the other joints to the current joint to enlarge the receptive field of the Shift-GC operation to the whole skeleton. These methods have shown encouraging performance for action recognition by modeling skeletal structures with GCNs (Fig. 1). However, these methods neglect the sample-dependent features for action recognition. Based on our investigation in daily life, pairwise relationships between joints may have different degrees of correlations in different action samples of the same class. Thus, it is important to study how to classify human actions with respect to unique relationships among joints for each action sample.

In the second class of sample-dependent GCNs, our Motif-GCN model proposed in the conference version [23] is the first GCN-based method, which constructs sample-dependent relationships among non-physically connected joints for action recognition. As shown in Fig. 2(a), this model considers different roles of physically connected joints (parent and child joints of the current joint are shown in purple and dark yellow, respectively) to encode structural features. As shown in Fig. 2(b), Adaptive GCN (AGCN) learns an adaptive graph structure based on the attention mechanism [32], [33], by calculating the similarity between all pairwise joints with a normalized embedded Gaussian function. As shown in Fig. 2(c), Actional-Structural GCN (AS-GCN) constructs structural links similar to EI-GCN [16], and uses an encoder to learn actional links between any pairwise joints. However, AS-GCN needs a pretraining process to train the encoder to get sparse actional links.

Different from the aforementioned sample-dependent methods, our SMotif-GCN model proposed in this paper considers sample-dependent relationships among joints with prior knowledge of human motions. Moreover, we introduce an SMAM, in which the weights of relationships between pairwise disconnected joints are learned and updated during training, as shown in Fig. 2(d). A new sparsity-promoting loss function is also proposed to make the non-zero values in the SMAM to be as sparse as possible.

2.2 Temporal Information Modeling

RNNs or Long Short-Term Memory (LSTM) networks have been widely used for learning temporal information of long-term skeleton sequences [5], [6], [7], [8], [9], [10], [11], [12]. However, convolutional operations obtain promising performance for the task because of efficient parallelization learning in the temporal domain of long sequences [4], [10], [13], [14], [15]. Specifically, Kim and Reiter [4] construct a CNN with residual connections to recognize human actions represented by sequences of joint coordinates. Liu *et al*. [13] manually transform skeleton sequences into a series of color images, which are fed to CNNs for classifying action categories. Yan *et al*. [15] construct a spatio-temporal graph with physical connections of joints and temporal edges between corresponding nodes in consecutive frames to represent a skeleton sequence. And, they extend graph convolution to learn spatio-temporal features. Since the temporal connections between joints in consecutive frames are regular, performing the graph convolution in the temporal domain is similar to the classical convolution. Following Yan *et al*. [15], there are many other works [17], [18], [19], [22] that perform convolutional operations on skeleton sequences for extracting temporal information.

Instead of directly defining a graph convolution on a predefined spatio-temporal structure [15], we use separate spatial and temporal sub-modules. It is more flexible to design the structure of each sub-module for extracting spatial and temporal features, separately. In addition, it is easier to learn with spatio-temporal decomposition methods [34]. For each frame in a sequence, we feed it into our proposed graph convolution, and then concatenate the output in the time axis to obtain a 3D tensor. Instead of using traditional convolutions [4], [10], [13], [14], [15], we propose a novel local temporal block to process the 3D tensor with densely connected layers to reuse local temporal features.

To draw global dependencies, self-attention mechanisms have been proposed and successfully applied to various applications, such as graph processing [33] and visual recognition [24], [35]. Wang *et al*. [24] formalize the self-attention as non-local operations, which can model pixel-level pairwise relationships among video frames. Hu et al. [36], [37] utilize a 1D non-local module to extract non-local features in the spatial domain, followed by another 1D non-local module to extract temporal features. Finally, a 2D non-local module is used to encode spatio-temporal features. The spatial and temporal features are encoded only with the non-local operations, while in our work, we encode spatial and temporal features with separated sub-modules, in which the temporal sub-module extracts local and global temporal features by an integration of local and non-local temporal blocks.

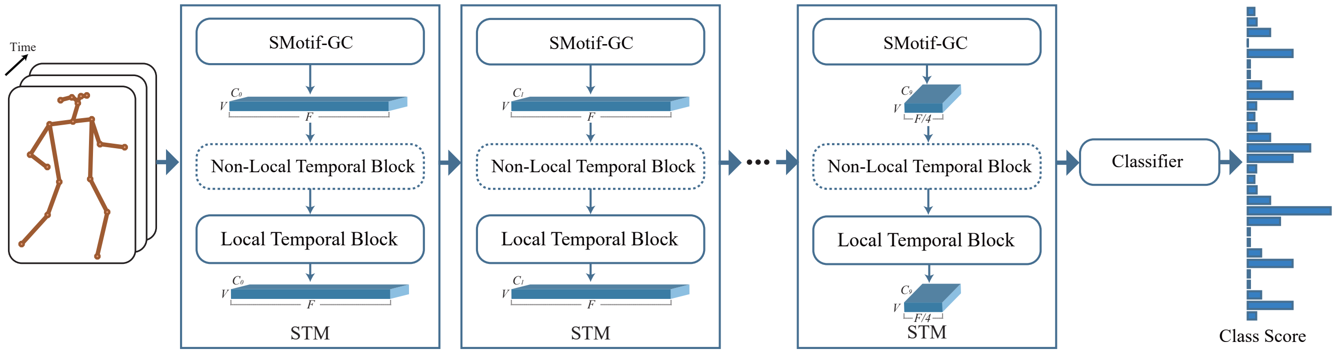


Fig. 3. Our model takes a sequence (F frames) of skeletons (V joints) as input and uses multiple layers of spatio-temporal modules (STMs) to generate higher-level feature maps. There are nine layers of STMs with an initial spatio-temporal head unit. C_i ($i = 0, 1, ..., 9$) represents the number of channels. The structure of the initial head unit is the same as the STM except that no residual connection is used in the head unit. The non-local temporal block is only used in the 7-th STM, so it is shown in a dotted line block. Then, the feature maps are fed into a Softmax classifier to get the final class score for recognizing actions in 30 classes (e.g., Kinetic-M dataset). More details of the overall architecture and the structure of the STM can be found in Sec. 4.2.

3 Methodology

3.1 Overview

Our model takes as input a skeleton sequence extracted from depth data [1] or videos [2] by applying pose estimation algorithms. Each frame in such a sequence has a set of 2D or 3D joint coordinates to represent a skeleton. To capture the topological structure, the human skeleton is modeled by a graph with joints as nodes and bones as edges.

As shown in Fig. 3, given a sequence of skeletal graphs, we construct a new network architecture, which stacks multiple layers of spatio-temporal modules (STMs) to extract features to feed into a classifier to predict the class score. Each STM contains separate sub-modules to tackle spatial and temporal information. Specifically, we propose a motif-based graph convolution (Motif-GC) to learn the spatial structure of skeletal data at each frame (Sec. 3.3). We further propose to learn a sparse motif adjacency matrix (SMAM) (Sec. 3.4), in which the relationships between non-physically connected joints can be updated through layers. The Motif-GC operation is updated to the sparse motif-based graph convolution (SMotif-GC) by introducing the SMAM. For modeling temporal information, a novel local temporal block (LTB) is proposed to capture richer temporal features by using partial dense connections (Sec. 3.5) than traditional temporal convolutions in local time windows. Furthermore, we capture long-range dependencies in the temporal domain by using a non-local temporal block (NLTB) with the attention mechanism (Sec. 3.6).

Below we first briefly review the traditional graph convolutional methods for modeling human skeletons and then introduce our SMotif-GCN with local and non-local temporal blocks.

3.2 Preliminary on Graph Convolution for Skeleton

A human skeleton can be represented by hinged joints and bones, which inherently lie in a graph structure with joints as nodes and bones as edges. Traditional methods construct a fixed graph $G(X, \mathcal{A})$ for each skeleton [15], [16], [17]. Here, $X \in \mathbb{R}^{V \times D}$ represents a node set of V joints with D-dimension coordinates. $\mathcal{A} \in \mathbb{R}^{V \times V}$ is an adjacency matrix for joints in X and is generally defined as:

$$a_{i,j} = \begin{cases} 0 & \text{if joints } i \text{ and } j \text{ are physically disconnected} \\ 1 & \text{if joints } i \text{ and } j \text{ are physically connected} \end{cases} \tag{1}$$

where $a_{i,j}$ is an element at row i and column j in the adjacency matrix $\mathcal{A}$. The graph convolution network for the skeletal graph $G(X, \mathcal{A})$ is named as Uni-GCN, which defines a uniform importance for physically connected neighboring joints of the current joint. Specifically, each joint is physically connected to itself. Mathematically, the graph convolutions can be conducted using the spectral graph theory [31], as follows:

$$Z = \mathcal{D}^{-1/2} \mathcal{A} \mathcal{D}^{-1/2} X W. \tag{2}$$

Here, $\mathcal{D}_{i,i} = \sum_{j=1}^{V} \mathcal{A}_{i,j} \in \mathbb{R}^{V \times V}$ represents the diagonal matrix of $\mathcal{A}$. $W \in \mathbb{R}^{D \times C_{out}}$ is a matrix of trainable weights, with C_{out} representing the number of the output channels.

3.3 Motif-Based Graph Convolution

Different from the existing methods [15], [16], [17], [18], [19], [21], we propose a novel Motif-GC method to extract features from a motif graph defined for skeletal structures. As shown in Fig. 2(d), the graph represents different semantic linkage patterns among physically connected and disconnected joints. Moreover, we define semantic roles of neighbor joints that describe structural similarities. The motif is a well-known concept that defines different patterns of connections in complex social networks [38], [39]. We introduce the motif notation to capture semantic proximity in a skeletal graph based on human motion knowledge (more details of the motif concept can be found in the appendix). The physical connections between joints are intuitive and stable during motion, and hierarchical structures exist among the joints. On the contrary, the non-physically connected relationships are latent and could vary due to motion. The physical and non-physical dependencies between joints should be considered simultaneously but treated differently. Thus, we define two motifs for both the dependencies. In the first motif of physical connections, we define three semantic roles ($K_1 = 3$) for the immediate neighbors of each joint: the joint itself, its parent node, and its child node. By considering different semantic roles of immediate neighbors, joints with physical connections are encoded hierarchically. The second motif defines one semantic role ($K_2 = 1$) for the underlying relationships between physically disconnected joints to impose high-order localities.

Given the input X_t at frame t, the Motif-GC for the motif graph with M motifs ($M = 2$) can be formulated as follows:

$$\hat{X}_t = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \tilde{\mathcal{A}}^{m,k} X_t W^{m,k}, \tag{3}$$

where $X_t \in \mathbb{R}^{V \times C}$ represents V nodes with C-dimension features. Each motif has K_m semantic roles. As illustrated in the last paragraph, $K_{m=1} = 3$ is defined for the first motif, and $K_{m=2} = 1$ is defined for the second motif. The motif adjacency matrix for a specific semantic role k in motif m is $\tilde{\mathcal{A}}^{m,k} \in \mathbb{R}^{V \times V}$, and $W^{m,k} \in \mathbb{R}^{C \times C_{out}}$ is the trainable filter parameter matrix. Similar to existing works [40], [41], we construct the motif adjacency matrices with a function Ψ^m. The weights in each motif adjacency matrix reflect the importance of neighbor joints. Generally, we can define a uniform type of Ψ^m for the two motifs in Eq. (3) as:

$$\tilde{\mathcal{A}}^{m,k} = \Psi^m(\mathcal{A}^{m,k}) = (\mathcal{D}^{m,k})^{-1} \mathcal{A}^{m,k}, \tag{4}$$

where $\mathcal{D}^{m,k} \in \mathbb{R}^{V \times V}$ is a diagonal matrix denoted as $\mathcal{D}^{m,k}_{i,i} = \sum_{j=1}^{V} \mathcal{A}^{m,k}_{i,j}$. For the first motif, $\mathcal{A}^{m=1,k}$ is defined in the same way as in Eq. 1. Additionally, we define a weighted adjacency matrix (WAM) for the latent sample-dependent relationships in the second motif. We notice that larger weights should be assigned to pairs of joints with shorter Euclidean distances. For example, the dependency between both hands is more important for recognizing the action "clapping hands" than the relationships between hands and head. Specifically, the weight of joints i and j in WAM is calculated as:

$$a_{i,j} = max(E) - e_{i,j}, \tag{5}$$

where $e_{i,j}$ is an element at row i and column j in the matrix $E \in \mathbb{R}^{V \times V}$ that represents the distance between every pair of non-physically connected joints. To calculate $e_{i,j}$, we first compute the average skeleton representation $\overline{X} \in \mathbb{R}^{V \times C}$ of the input skeleton sequence $X_{t=1 \ldots F}$ along the time dimension. Then, $e_{i,j} = \|\overline{x}_i - \overline{x}_j\|_2$, where $\|\overline{x}_i - \overline{x}_j\|_2$ denotes the Euclidean distance between non-physically connected joints i and j of $\overline{X}$. The value of $e_{i,j}$ between physically connected joints i and j is zero.

Finally, the output feature maps of different motifs are combined with element-wise sum operation. $\hat{X}_t$ denotes the output of the proposed motif-based graph convolution.

3.4 Sparse Motif-Based Graph Convolution

In Sec. 3.3, the non-physically connected relationships between joints are introduced to encode sample-dependent skeletal structures. The semantic information contained in joints of different layers is updated from lower to higher layers. However, with the fixed adjacency matrix, the relationships cannot be updated across layers of the network. That means a predefined graph structure shared among all layers is not flexible enough to adapt to the variety of the semantic information.

To solve this problem, we impose an SMAM that encodes non-physical connections adaptively into our Motif-GC operation. Fortunately, the motif-based graph convolution proposed in Sec. 3.3 can be easily extended to adopt different types of Ψ^m for the motifs. Different from the original definition in Eq. (4), we define a learnable adjacency matrix for the second motif. In more detail, we can update the elements in the adjacency matrix $\tilde{\mathcal{A}}^{m,l}$ ($m = 2$) across layers and calculate the matrix at layer l as:

$$a^l_{i,j} = max(E^l) - e^l_{i,j}, \tag{6}$$

where $e^l_{i,j}$ is an element at row i and column j in the matrix $E^l \in \mathbb{R}^{V \times V}$ that represents the distance of the latent features between every pair of non-physically connected joints. $e_{i,j}$ is calculated based on the average feature maps $\overline{X}^l \in \mathbb{R}^{V \times C}$ of the input features at the layer l along the time dimension. Then, $e^l_{i,j} = \left\|\overline{x}^l_i - \overline{x}^l_j\right\|_2$ denotes the Euclidean distance of deep features between non-physically connected joints i and j of $\overline{X}^l$. And, $e^l_{i,j}$ for physically connected joints i and j is zero.

Finally, the weighted dependencies between physically disconnected joints are defined as:

$$\tilde{\mathcal{A}}^{m,l} = (\mathcal{D}^{m,l})^{-1} \mathcal{A}^{m,l} \odot \mathcal{M}, \tag{7}$$

where $\mathcal{D}^{m,l}$ is a diagonal matrix. $\mathcal{M} \in \mathbb{R}^{V \times V}$ is a learnable mask to control the sparsity level of the adjacency matrix $\tilde{\mathcal{A}}^{m,l}$ in all layers, and $\odot$ denotes dot production. To impose sparsity to $\tilde{\mathcal{A}}^{m,l}$, we add a regularization term with an L_1 matrix norm $\|\mathcal{M}\|_1 = \sum_{i=1}^{V} \sum_{j=1}^{V} |m_{i,j}|$, with $m_{i,j}$ being an element at row i and column j in the matrix M, to our loss function:

$$\mathcal{L} = loss(\hat{y}, y) + \lambda \|\mathcal{M}\|_1, \tag{8}$$

where $\hat{y} = \text{Softmax}(\text{Pool}(\text{SMotif-GCN}(X, A)))$ is the output of the SMotif-GCN model with the global average pooling operation and the standard Softmax classifier. λ is the weight of the regularization term, and $loss(\hat{y}, y)$ is the softmax loss that measures the difference between the prediction result $\hat{y}$ and the ground-truth label y. Specifically, $\mathcal{A}^{m,l}$ is normalized with the diagonal matrix $\mathcal{D}^{m,l}$ to avoid the magnitude explosion. The mask $\mathcal{M}$ is used after the normalization. It is initialized with all ones to avoid breaking the normalization, and then the sparse-promoting loss function helps to avoid the magnitude explosion during training.

3.5 Local Temporal Block

We concatenate the output feature maps of the Motif-GC operation $\hat{X}_{t=1 \ldots F}$ along the time axis to obtain a 3D tensor $\hat{X} \in \mathbb{R}^{F \times V \times C_{out}}$, which is then sent into a temporal submodule for further information extraction. The traditional methods (e.g., [15], [17], [18], [19], [22]) adopt a classical 2D convolution ($T \times 1$ Conv) for modeling temporal features of $\hat{X}$. We propose a Local Temporal Block (LTB) with dense connections among convolutional layers that encode information in local time windows, inspired by densely concatenated convolutional networks [42]. The dense connections

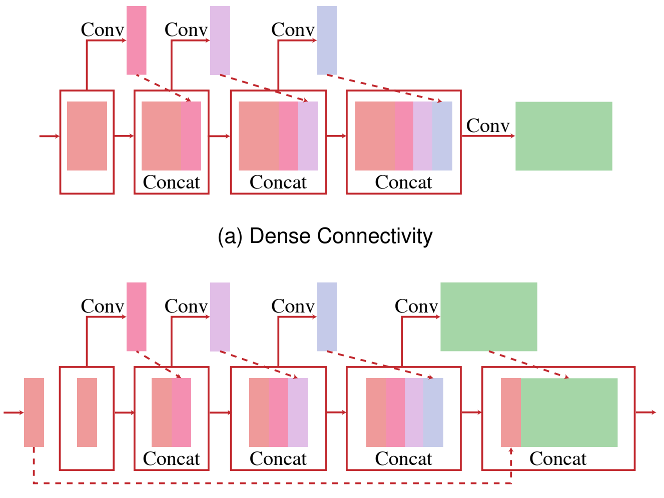


Fig. 4. Illustration of feature map flow in the local temporal block with (a) Dense Connectivity and (b) Partial Dense Connectivity. In (a), each layer takes all preceding feature maps as input. In (b), the feature maps of the first layer are split into two parts with equal channels. One part of feature maps goes though dense connections, while the other part of feature maps is directly concatenated with the output of densely concatenated layers.

are parameter-efficient and enable dense knowledge propagation, because they could reuse and preserve features of different layers. However, the number of connections grows quadratically with the depth of densely connected layers. Instead of introducing dense connections to all layers, we integrate densely connected layers into each STM separately to control the growth of connections. What's more, we propose to use a partial dense connectivity scheme (Fig. 4) to reduce duplicate gradient information and capture richer temporal features by promoting variability of the gradients [43]. The experiments in Sec. 4.5 show the effectiveness of the LTB by introducing the partial dense connectivity. Below, we present how to design the LTB and incorporate it into our model for extracting temporal features.

3.5.1 *Partial Dense Connectivity*

In densely connected convolutional networks [42], each layer takes the concatenation of the output feature maps from all preceding layers as input as shown in Fig. 4(a). Specifically, the output feature map H_l from layer l is defined by a composite function h_l:

$$H_l = h_l([H_0, H_1, ..., H_{l-1}]) \tag{9}$$

where $[H_0, H_1, ..., H_{l-1}]$ denotes the concatenation of the feature maps from all the early layers. The input H_0 is $\hat{X} \in \mathbb{R}^{F\times V\times C_{out}}$ from the Motif-GC operation. The composite function $h_l(\cdot)$ includes Batch Normalization (BN) [44], rectified linear unit (RELU) [45], and convolution (Conv). Each dense block in our network has three densely connected layers for convolutional operations in the temporal domain. The dense connectivity pattern defined in Eq. (9) has the advantages of reusing features; however, it has a large amount of duplicated gradient information as mentioned in a previous work [43]. To address this issue, we truncate the feature maps that are fed into the temporal block to reduce the redundant gradient information following the previous work [43]. More specifically, as illustrated in Fig. 4(b), the feature maps fed into the first layer are split into two equal parts:

$$H_0 = [H_0', H_0''] . \tag{10}$$

The first part of feature maps H_0' is directly concatenated with the output of the dense layers, and the second part of the feature maps goes through densely connected layers:

$$H_l = h_l([H_0'', H_1, ..., H_{l-1}]) . \tag{11}$$

As the feature maps in H_0' are not updated through the dense layers, the amount of duplicate gradient information in all the following layers is reduced.

3.5.2 *Growth Rate*

We use a hyper-parameter "*growth rate*" (GR) to control the number of output feature maps (denoted as N) of each dense connection [42]. If the composite function h_l is supposed to produce N feature maps, we should get $N_0/2 + N \times (l-1)$ input feature maps for the l-th layer, where N_0 is the number of channels of the first input layer in the partial dense block. GR could be relatively small [42] to reduce the number of filter parameters. We set GR of each dense block according to the number of output channels of each STM. The detailed settings of GR will be evaluated in Sec. 4.5.

3.5.3 *Transition Layer*

We use a transition layer to control the number of the output channels. The transition layer is comprised of a BN and a RELU operation, followed by a Conv operation.

As the input feature maps are split (Eq. (10)), the concatenated feature maps fed into the transition layer are denoted as $[H_0'', H_1, ..., H_l]$. Finally, the output of the transition layer H_{trans} is concatenated with the first part of input feature maps $H_{concat} = [H_0', H_{trans}]$. Our proposed LTB can reduce the amount of duplicate gradient flow as well as enriching a variety of gradient combination (more details can be found in the appendix).

3.6 Non-Local Temporal Block

The Motif-GC method can inherently capture global dependencies of intra-frame joints and adaptively update the dependencies with the SMAM. However, the inter-frame action information is extracted without considering global-range dependencies. In order to model global inter-frame connectivities, we introduce a NLTB that can relate all possible frames of a skeleton sequence in a self-attention mechanism [24]. It is able to enlarge the receptive field from a local time window to the entire sequence with non-local operations.

The non-local operation computes the dependencies at a frame directly by attending at any other frame with possible relationship as:

$$Z_t = \frac{1}{\mathcal{C}(\hat{X})} \sum_{\forall t'} f(\hat{X}_t, \hat{X}_{t'}) g(\hat{X}_{t'}), \tag{12}$$

where $\hat{X}_t$ and $\hat{X}_{t'}$ denote the input of the non-local operation at possible frames t and t', respectively. Z_t is the updated features corresponding to the query frame t, while $g(\cdot)$ is a linear transformation. $\mathcal{C}(\hat{X})$ is defined for normalization with respect to the type of the pairwise function f. There are many choices for the pairwise function f [24]. However, different pairwise functions achieve similar performance [24]. We choose the embedded Gaussian version, because it is the basic version of the non-local operation [24] and can be seen as a generic self-attention form [32]. f is defined as: $f(\hat{X}_t, \hat{X}_{t'}) = e^{\theta(\hat{X}_t)\phi(\hat{X}_{t'})^T}$, where $\theta(\cdot)$ and $\phi(\cdot)$ are learnable embeddings that encode appropriate representations of the input. $\mathcal{C}(\hat{X}) = \sum_{\forall t'} f(\hat{X}_t, \hat{X}_{t'})$ is defined for the embedded Gaussian version, and $\frac{1}{\mathcal{C}(\hat{X})} f(\hat{X}_t, \hat{X}_{t'})$ becomes a nonlinear activation function $Softmax(\theta(\hat{X}_t)\phi(\hat{X}_{t'})^T)$ computed along the dimension t'.

The non-local operation is finally wrapped into the NLTB with a residual connection as:

$$O = W_o Z + \hat{X}, \tag{13}$$

where Z is the concatenation of the output signal of the non-local operation $Z_{t=1\ldots F}$ of all frames, and W_o is a learnable embedding for Z. The residual connection is adopted to eliminate breaking the initial behavior of the model.

4 Experiments

We have evaluated the effectiveness of our proposed model, on four large-scale datasets, namely NTU-RGB+D [6], Kinetics-M [15], [46], NTU-RGB+D-120 [47], and Kinetics [15], [46]. NTU-RGB+D [6] and NTU-RGB+D-120 [47] datasets are captured in a lab environment, which is constrained by experimental setups in the lab. On the contrary, Kinetics-M and Kinetics datasets are captured in the natural environment, which is unconstrained and challenging. By conducting experiments on the four dataset, our model is evaluated in both constrained and unconstrained scenarios. In the ablation studies, we first evaluate our proposed motif-based graph convolutional operations and then focus on analyzing the necessity of temporal modeling modules.

4.1 Evaluation Settings on Datasets

NTU-RGB+D. It is the most widely used dataset with annotated 3D joint coordinates for human action recognition [6]. It contains 56,880 sequences of 60 action classes. These video clips are performed by 40 volunteers in a lab environment, recorded by 3 different camera views simultaneously. Each skeleton sequence has at most 2 subjects and each subject is represented by 25 joints. The authors of this dataset recommend cross-subject (X-Sub) and cross-view (X-View) benchmark evaluations. In the X-Sub benchmark, 40,320 clips from 20 subjects are used for training and 16,560 clips from the remaining 20 subjects are used for testing. In the X-View benchmark, 37,920 clips from camera views 2 and 3 are used for training and the other 18,960 clips from camera view 1 are used for testing. We report our results on both the recommended benchmarks following the conventional settings as suggested in [6].

Kinetics and Kinetics-M. Deepmind Kinetics [46] is a large-scale video dataset containing about 300,000 clips, which cover 400 classes of human actions sourced from YouTube. In the original dataset, each clip lasts around 10 seconds and provides only raw video data without human skeletons. A new dataset that includes skeleton sequences extracted from the video clips has been released for research purposes [15]. To obtain the skeleton data, 2D coordinates of 18 joints are estimated [15] for each person with the real-time Openpose toolbox [2]. For multi-person clips, 2 persons that have higher average joint confidence are chosen. Each action clip is padded to the same length (e.g., 300 frames) by replaying the action from the beginning. Many action classes in the dataset require consideration of relationships between actors and complex scenes for action recognition. Therefore, methods focusing on skeletal data are generally inferior to video-based methods, which can take use of additional information in the background of videos. It has been shown that a subset of 30 action classes strongly related with body motion can make the performance gap smaller [15]. Therefore, we refer to the subset as "*Kinetics-M*" here to evaluate our model with Top-1 and Top-5 classification accuracies as recommended in [46]. The dataset provides 25,000 training clips and a test set of 1,500 clips. Besides, we also evaluate our models on the original Kinetics dataset of skeletal data.

NTU-RGB+D-120. It is an extension of the NTU-RGB+D dataset. It is the largest publicly available dataset for 3D skeleton-based action recognition [47]. The dataset consists of 114,480 samples of 106 volunteers, captured by three cameras from different horizontal angles ($-45°$, $0°$, $+45°$) at the same time. To increase the number of camera viewpoints, the vertical heights of the cameras and their distances to the subjects are changed. Totally, there are 155 different camera viewpoints. The authors of this dataset suggest cross-subject (X-Sub) and cross-setup (X-Set) evaluations. In the X-Sub benchmark, action clips performed by 53 subjects are used for training, and the clips of the remaining 53 subjects are used for testing. In the X-Set benchmark, action clips from 16 camera setups are used for training and the clips from the other 16 setups are used for testing. Each setup has a collection of camera views with different cameras' height and distance to the subjects as illustrated in [47]. We report our results on both the recommended benchmarks.

4.2 Implementation Details

In the overall network architecture, nine layers of STMs are stacked with an initial spatio-temporal head unit, as illustrated in Fig. 5. The structure of the spatio-temporal head unit is the same as the STM except that the residual connection is not used in the head unit. The input and output channel numbers of all STMs are illustrated in Fig. 5. The STM (Fig. 6) extracts spatial and temporal features by the SMotif-GC layer and the LTB, respectively. Before feeding input joint coordinates into our network, we use a BN layer [44] to normalize the data. In the LTB, we set local time window of size 9 for the 3 dense layers. Specifically, a 2D 9×1 convolution is conducted in each dense layer. Additionally, we insert a non-local block at the 7-th STM as "*NL STM*". For down-sampling along the time axis, the strides of temporal convolution layers in the 4-th and 7-th STMs are both set to 2. A global average pooling is performed on the resulting feature maps. Finally, a Softmax classifier is utilized to generate a score for each action category.

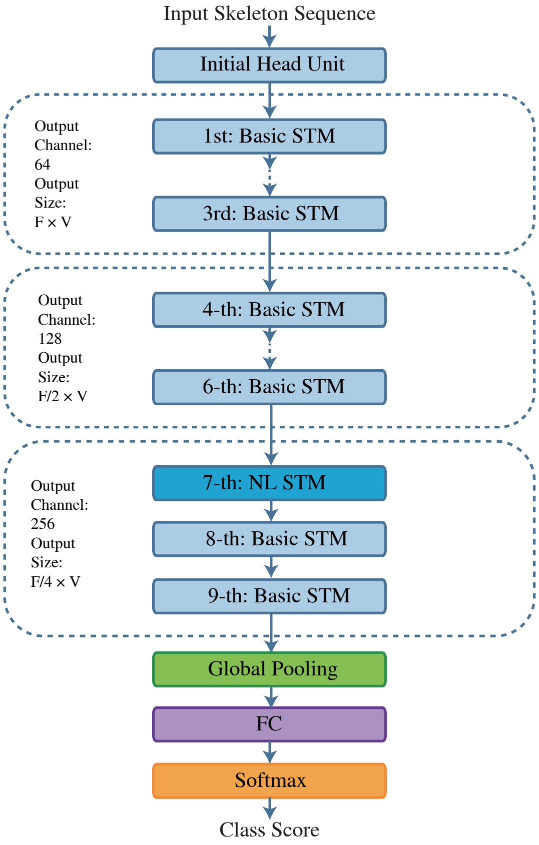


Fig. 5. The architecture of our proposed Motif-GCN, which is composed of multiple spatio-temporal modules (STMs). The non-local spatio-temporal module (NL STM) is used before the last two STMs to further expand the temporal dependency.

We use PyTorch [48] to implement the proposed model and conduct all the experiments with 8 NVIDIA TitanX GPUs. We train the network with stochastic gradient descent with Nesterov momentum (0.9). The weight decay factor is set to 0.0001. In the following subsections, we show that our model achieves the state-of-the-art performance on all the benchmark datasets with the above parameter settings, demonstrating the effectiveness of our model.

We adopt random moving and selecting methods for data augmentation [15] when training our model on the Kinetics-M dataset (with the batch size set to 64) and Kinetics dataset (with the batch size set to 256). The initial learning rate is set to 0.1 and reduced by multiplying by 0.1 at the 50-th, 60-th, and 70-th epochs, and the training process is ended at the 90-th epoch. On the NTU-RGB+D and NTU-RGB+D-120 datasets, we set the batch size to 64 for training. The initial learning rate is 0.1 and reduced by multiplying by 0.1 at the 40-th epoch, and the training process is terminated at the 80-th epoch. In our experiments, all data clips are padded to the same length (i.e., 300 frames) for training and testing in batches. Our system handles the skeletal data of two people in each frame. For the data clips that only have one skeleton, we follow the similar idea in [15], [18] by adding a new skeleton padded with zeros to each frame, so that all the data clips have the same number of skeletons.

4.3 Ablation Studies

NTU-RGB+D is the most widely used dataset for the task of skeleton-based action recognition, and Kinetics-M includes skeleton sequences that strongly related with body motion from Kinetics dataset. NTU-RGB+D-120 and Kinetics datasets can be seen as extensions of the representative NTU-RGB+D and Kinetics-M datasets, whose data are captured in constrained lab and unconstrained natural environments, respectively. Here, we first perform ablation studies and detailed experiments on both the NTU-RGB+D and Kinetics-M datasets to verify the effectiveness of individual critical sub-modules and their combinations for action recognition with noisy joint data. For the NTU-RGB+D dataset, the 3D joint data estimated from depth data is noisy because of occlusion or sensor noise [6], [49]. For the Kinetics dataset, the skeleton sequences are estimated from videos, which are captured in unconstrained and challenging environments, by monocular pose estimation algorithms (e.g. OpenPose). The data is noisy because of the occlusion and motion blur. If there are heavy occlusions, some joints would be missing. In this case, the positions of the missing joints are filled with zeros.

4.3.1 *SMotif-Based Graph Convolution*

In our Motif-GC method, we define two motifs for the physical and non-physical dependencies between joints (see Sec. 3.3). In the motif for physical connections, three semantic roles are assigned to the immediate neighbors of each joint. We also introduce a motif with one semantic role for physically disconnected joints based on human motion knowledge. Here, we compare our proposed network with Motif-GC layers (Motif-GCN) with a baseline method [31] referred to as "Uni-GCN". Uni-GCN defines only one motif and one semantic role for physically connected joints. In Uni-GCN and Motif-GCN, we use traditional convolutions for extracting features in the temporal domain.

The comparison results are summarized in Table 1. On NTU-RGB+D, a large boost in accuracy is observed from 75.8% to 84.0% under X-Sub evaluation. Motif-GCN also outperforms Uni-GCN by a notable margin under X-View evaluation. Consequently, the methods are useful for tackling noisy raw data in the NTU-RGB+D. What's more, the Top-1 recognition accuracy boosts from 76.4% to 83.7% on Kinetics-M. In our Motif-GC method, global dependencies between joints are inherently constructed between joints. This method is robust to handle skeletal data with local noise captured in daily life.

The experiment results suggest that Motif-GCN can effectively extract spatial features for improving action recognition accuracy by considering hierarchical structures and latent high-order relationships. In our previous conference version of Motif-GCN [23], which adds a mask to physical connections as suggested by Yan *et al.* [15], the Top-1 and Top-5 accuracies on Kinetics-M dataset were 82.5% and 95.7%, respectively. Here, we remove the mask in our Motif-GC operations, and obtain better performance (Top-1: 83.7% and Top-5: 95.8%), as shown in Table 1. This comparison shows Motif-GCN can capture effective spatial information by considering high-order relationships between joints. By adding the mask, the relationships between joints may

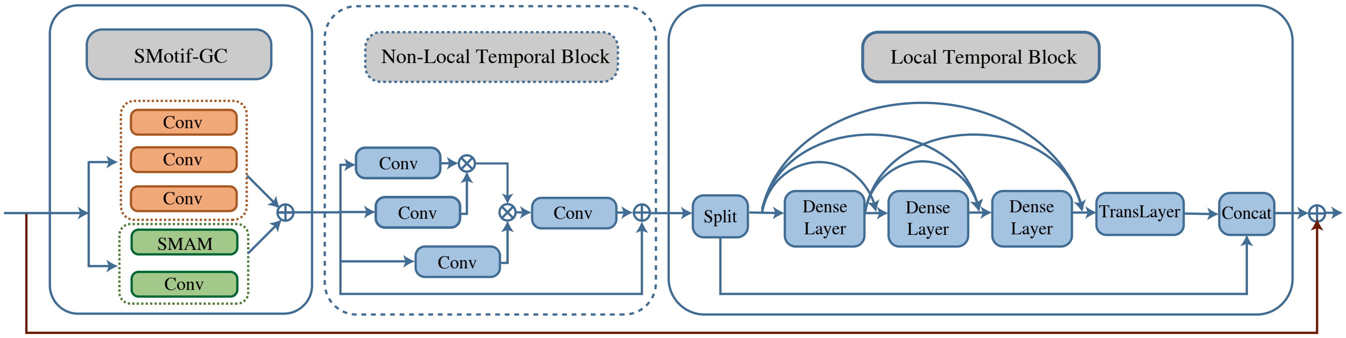


Fig. 6. Architecture of the spatio-temporal module (STM). It contains a sparse motif-based graph convolution (SMotif-GC) sub-module, which adopts a sparse motif adjacency matrix (SMAM), for modeling spatial information. A local temporal block that contains three dense layers and a transition layer (TransLayer) is used to encode features with local time windows. The dense layer is implemented with a combination of BN-RELU-Conv operation. A residual connection (shown with an arrow in dark red) is added to each STM by a shortcut connection from input to output and an element-wise sum operation. A non-local temporal block is used only in the last stage of our network, so it is shown in a dotted line block. ⊕ denotes element-wise sum operation, and ⊗ denotes matrix multiplication.

TABLE 1
Ablation studies on NTU-RGB+D and Kinetics-M datasets. TBs are the combination of LTB and NLTB.

Dataset	NTU-RGB+D		Kinetics-M	
Evaluation	X-Sub	X-View	Top-1	Top-5
Uni-GCN	75.8%	85.2%	76.4%	94.0%
Motif-GCN	84.0%	90.5%	83.7%	95.8%
SMotif-GCN	84.7%	91.4%	84.1%	95.8%
Motif-GCN+LTB	85.4%	92.5%	84.7%	96.5%
Motif-GCN+NLTB	84.8%	92.0%	84.4%	95.9%
Motif-GCN+TBs	86.7%	92.9%	85.0%	96.5%
SMotif-GCN+TBs (Ours)	**87.2%**	**93.6%**	**85.4%**	**96.8%**

become too flexible to learn and encode, leading to a degradation in the performance.

By introducing the SMAM into Motif-GCN as "SMotif-GCN", the accuracies can be further improved on both the NTU-RGB+D and Kinetics-M datasets (Table 1). The experiment results indicate that the learnable SMAM is effective to capture latent dependencies among non-physically connected joints for encoding spatial information.

4.3.2 Local Temporal Block

We evaluate the effectiveness of the LTB for extracting temporal features from local time windows. As shown in Table 1, the accuracies of Motif-GCN+LTB are 85.4% and 92.5% on the NTU-RGB+D dataset under X-Sub and X-View evaluation protocols, respectively. The respective accuracies are higher than those of Motif-GCN (X-Sub: 84.0% and X-View: 90.5%). Moreover, the Top-1 and Top-5 accuracies of Motif-GCN+LTB on the Kinetics-M dataset are 84.7% and 96.5%, which are also higher than 83.7% and 95.8% of the Motif-GCN, respectively. The experiment results show that the LTB can capture local temporal features more effectively, compared to using classical convolutions for modeling temporal information. We will evaluate the LTB further in Sec. 4.5.

4.3.3 Non-Local Temporal Block

We evaluate the effectiveness of utilizing the NLTB for modeling whole-range dependencies. The NLTB can be combined with traditional convolutions in the time axis as "Motif-GCN+NLTB".

Table 1 shows that introducing the NLTB into the temporal sub-module of STM in our Motif-GCN (Fig. 5) brings performance improvement on the NTU-RGB+D and Kinetics-M datasets. These results demonstrate that constructing global dependencies among frames by the NLTB can achieve a more representative feature extraction for action recognition. Not only our Motif-GCN can achieve remarkable gain by considering global relationships in the spatial context, but also the NLTB can further improve the performance by constructing global inter-frame dependencies in the temporal context.

To investigate the effectiveness of the combination of the LTB and the NLTB as temporal blocks (TBs) for modeling temporal information, we analyze the performance of the model "Motif-GCN+TBs". We have evaluated the performance of integrating NLTB into different layers of our network. The proper setting of the NLTB will be given in Sec. 4.5. In Table 1, the results show that this model has better capability for modeling temporal information. Additionally, we integrate the TBs into SMotif-GCN as our full model "SMotif-GCN+TBs (Ours)". Given the high performance of the Motif-GCN+TBs, there is still an improvement in accuracy of SMotif-GCN+TBs by imposing the SMAM.

4.4 Evaluation of SMotif-Based Graph Convolution

As described in Sec. 3.4, our proposed Motif-GC operation is extended by defining a learnable adjacency matrix for the second motif. The elements in the matrix are updated across layers in the self-attention mechanism, with the input $X^l_{t=1...F}$ at the layer l. The weighted dependencies between physically disconnected joints are updated according to Eq. (7).

Fig. 7 illustrates how the adjacency matrices for latent dependencies among non-physically connected joints are updated through layers. Fig. 7(a) and Fig. 7(b) show the adjacency matrices of different layers for action samples "Clapping" and "Check time", respectively. It is obvious that the learned adjacency matrices for different action samples are different. What's more, the matrices can be updated through message passing across different layers. Taking the

"Clapping" sample as an example, the learned adjacency matrices in the lower layers indicate the relationships between both hands are important for recognizing the action. This is in accordance with human motion knowledge for action recognition intuitively. Then, the high-level features strengthen the dependencies between the "Throat" joint and other joints at the 9-th layer. As the dependencies are captured in latent semantics, they are different from intuitions on human action recognition in daily life. The above experiment observations confirm our motivation to learn the adjacency matrix for latent dependencies in Motif-GC methods.

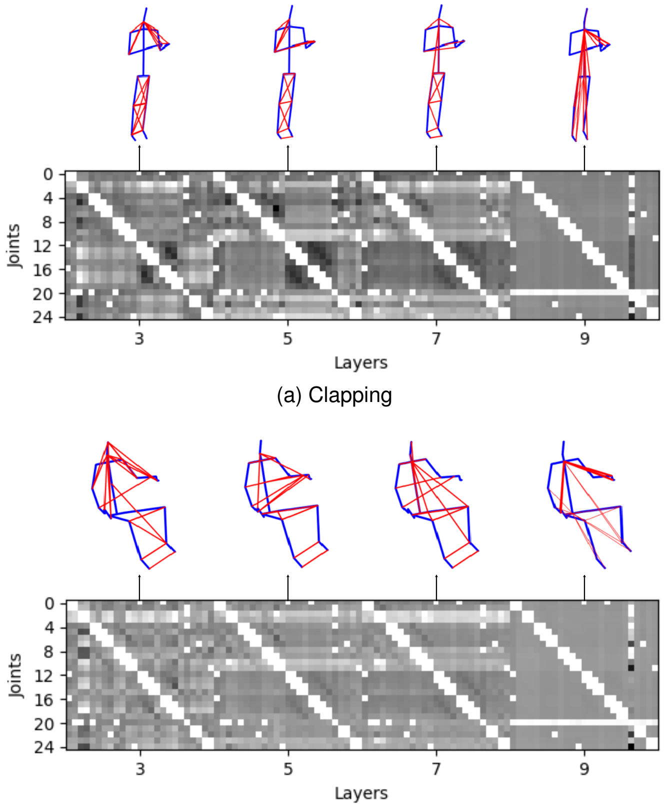
To impose sparsity to the matrix, we add the $L1$ regularization term to our loss function in Eq. (8), where λ trades off the importance of the Softmax loss and the regularization term. Increasing λ promotes the sparsity of the adjacency matrix for non-physically connected relationships among joints. We evaluate the recognition performance with different values of λ using the same test approaches in our experiments. From Fig. 8, we observe that increasing the value of λ to $10e^{-5}$ achieves the best performance on the NTU-RGB+D dataset (X-View). The improvement implies that promoting the sparsity of the adjacency matrix helps capture more useful dependencies among physically disconnected joints. However, we observe that the performance becomes worse when λ increases from $10e^{-5}$ to higher values. If the adjacency matrix for latent dependencies is too sparse, it hinder the Motif-GC operation to capture richer information in the spatial context among joints. We also evaluate the performance of SMotif-GCN on the NTU-RGB+D dataset under X-Sub evaluation. The accuracy is 84.7% with $\lambda = 0$, but it drops to 83.1%, 84.3%, 83.3%, 83.2%, 83.2%, when λ is set to $6e^{-5}$, $8e^{-5}$, $10e^{-5}$, $12e^{-5}$, $14e^{-5}$, respectively.

4.5 Evaluation of Temporal Block

We use the LTB and the NLTB to extract local and global temporal features effectively. Here, we conduct experiments to illustrate how to combine the two temporal blocks to achieve better performance.

The NLTB constructs dependencies between any two frames in a sequence, so the computation complexity is relatively large [24]. In this way, we use only one NLTB in our network. The default setting of the model Motif-GCN+NLTB is to integrate only one NLTB to the 7-th STM.

Firstly, we evaluate the recognition performance of the LTB based on Motif-GCN+NLTB (default setting). As described in Sec. 3.5, the growth rate (GR) is controlled by the transition layer and set according to the number of output channels in the STM (Fig. 5). We denote the number of output channels as OC. The GR value reflects the amount of new information that each densely connected layer contributes to the local temporal features. We thus evaluate the influence of the GR, which is the most critical hyper-parameter governing the LTB. The results are summarized in Table 2. We observe that setting the GR as OC/8 leads to the best performance on the Kinetics-M dataset in terms of both the Top-1 and Top-5 accuracies, and NTU-RGB+D dataset under X-Sub evaluation. Under X-View evaluation (NTU-RGB+D dataset), the highest accuracy is achieved with GR = OC/4. Richer temporal features can be extracted by a higher value of GR. The observation shows that temporal information is more important for recognizing human actions across views than across subjects. The default



(a) Clapping

(b) Check time

Fig. 7. Illustration of the sparse motif adjacency matrices in SMotif-GCN for (a) "Clapping" and (b) "Check time" test action samples from the NTU-RGB+D dataset (X-Sub). The updated adjacency matrices of different layers, and corresponding graph topologies of latent dependencies among joints are shown in different columns. The importance of the latent dependency is denoted by the gray scale (the darker, the more important) of the element in the adjacency matrix. The red lines represent the latent dependencies whose importance values are in the top 20, and the thickness of each line represents the importance of the relationship between pairwise joints.

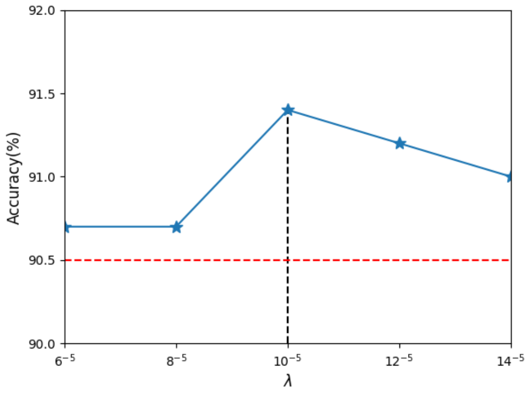


Fig. 8. Performance comparison of SMotif-GCN approach using different settings of λ on the NTU-RGB+D dataset (X-View). The blue line represents experiment results with different λ values used for the loss function in Eq. (8), while the red dashed line indicates the performance of Motif-GC without sparsity.

TABLE 2
Evaluation of different settings of growth rate (GR) for integrating the LTB into Motif-GCN+NLTB model on the NTU-RGB+D and Kinetics-M datasets.

Dataset	NTU-RGB+D		Kinetics-M	
Evaluation	X-Sub	X-View	Top-1	Top-5
Motif-GCN+NLTB	84.8%	92.0%	84.4%	95.8%
+LTB (GR = OC/2)	85.4%	93.1%	84.3%	95.8%
+LTB (GR = OC/4)	86.4%	**93.2%**	84.6%	96.5%
+LTB (GR = OC/8)	**86.7%**	92.9%	**85.0%**	**96.5%**
+LTB (GR = OC/16)	86.6%	92.7%	82.1%	95.8%
+LTB (GR = OC/32)	86.1%	92.2%	82.0%	95.8%

TABLE 3
Evaluation of NLTB integrated into different STMs of Motif-GCN+LTB model on the NTU-RGB+D and Kinetics-M datasets.

Dataset	NTU-RGB+D		Kinetics-M	
Evaluation	X-Sub	X-View	Top-1	Top-5
Motif-GCN+LTB	85.4%	92.5%	84.7%	96.5%
+NLTB (1st STM)	86.4%	92.7%	84.5%	96.3%
+NLTB (4-th STM)	86.2%	92.6%	83.5%	95.6%
+NLTB (7-th STM)	**86.7%**	**92.9%**	**85.0%**	**96.5%**

setting of the LTB is set GR = OC/8. To verify the improvement by adopting the LTB instead of the variable temporal dense block (DB) with the traditional dense connectivity in our conference version [23], we compare the performance of Motif-GCN+LTB (default setting) with Motif-GCN+DB in different settings. The comparison results can be found in the appendix.

Next, we conduct a series of experiments to achieve a good setting of the NLTB based on Motif-GCN+LTB (default setting), and the results are summarized in Table 3. The NLTB can be introduced into our Motif-GCN+LTB flexibly. We observe that adding the NLTB to the 7-th STM achieves the best accuracy on the NTU-RGB+D and Kinetics-M datasets, respectively. It implies that imposing global dependencies between frames with features from higher layers has better representation power than lower layers. On the NTU-RGB+D dataset, the raw joint data are padded with zeros for training and testing in batches. We add masks M_t and M_t' to calculate the dependencies between possible frames t and t', respectively. We set $M_t = 0$ for the frame t whose data is padded, and $M_t = 1$ for the frame whose data is from the captured data. In this way, the similarities between the padded frame and the other frames calculated by the pairwise function are initialized as zeros. The Top-1 accuracy of Motif-GCN+LTB+NLTB (7-th STM) is improved from 85.4% to 86.7% under the X-Sub evaluation and from 92.5% to 92.9% under the X-View evaluation.

On the Kinetics-M dataset, the Top-1 accuracy is improved from 84.7% to 85.0% by introducing the NLTB. If the NLTB is introduced into the 1-st or 4-th STM, the accuracy of the model drops to lower values on the Kinetics-M dataset (Table 3). This indicates that imposing global dependencies among frames of noisy skeleton sequences cannot always bring improvements to the model. The deep feature maps from the 7-th STM have a larger receptive field in the temporal domain, so they are less influenced by local noise and are more proper to be used in the NLTB. It is important to choose proper features to construct global inter-frame dependencies in the self-attention mechanism to improve action recognition accuracy.

4.6 Comparison with State-of-the-Art Methods

In this section, we present experiments to compare our proposed models with the state-of-the-art methods on different benchmark datasets.

4.6.1 Experiments on NTU-RGB+D Dataset

To demonstrate the effectiveness of our method, we compare it with the following related methods: 1) Early methods that use LSTMs to extract features from skeleton sequences (e.g., [6], [12], [50], [51]); 2) Recent methods that utilize CNNs for temporal modeling of a skeleton sequence represented by a time series of 3D joint coordinates or pseudo-images (e.g., [3], [4], [13], [14], [52]); 3) The most related methods that adopt GCNs to learn the spatial information of human skeletons (e.g., [15], [16], [17], [18], [19], [21], [22], [53], [54]). Among all the above mentioned methods, only TConv [4], ST-GCN [15], EI-GCN [16] and Motif-GCN [23] take the raw (X, Y, Z) values of each skeleton joint as input. The others adopt data augmentation (e.g., [14], [21]) or preprocessing strategies (e.g., [3], [6], [12], [13], [19], [20], [50], [51], [53]). What's more, some models adopt the multi-stream architecture (e.g., [17], [18], [22], [54]).

Without data augmentation or preprocessing strategies, our proposed model SMotif-GCN+TBs achieves higher accuracy than the previous method ST-GCN [17] by a large margin on X-Sub (87.2% *v.s.* 81.5%) and X-View (93.6% *v.s.* 88.3%) benchmarks. Besides, PB-GCN [17] has also reported the action recognition accuracy on raw 3D joint coordinates. Our proposed method outperforms PB-GCN by a large margin on X-Sub (86.9% *v.s.* 82.8%) and X-View (93.6% *v.s.* 90.3%) benchmarks. The results in Table 4 demonstrate the effectiveness of our method in modeling spatial and temporal information for action recognition with noisy skeletal data. Our model can capture local and non-local information spatially and temporally at the same time. Consequently, the model is more effective for encoding noisy data with local disturbance. Moreover, the proposed model SMotif-GCN+TBs also achieves significant improvements over the Motif-GCN method in our conference version [23]. The comparison results show that our Motif-GC method is effective and easy to extend for encoding skeletal data to improve action recognition accuracy.

We perform data preprocessing [6], [18], [22] for a fair comparison with the earlier works that require data preprocessing as well. In more detail, we first normalize each sample, and then translate the normalized sample to the central perspective, as suggested in [6]. The performance of our model SMotif-GCN+TBs with preprocessed joint data, denoted as "SMotif-GCN+TBs (Ours, Preprocessed Joint)", is better than the state-of-the-art method [20] under X-Sub evaluation (88.9% *v.s.* 87.8%). We also use the bone information [18] with each bone represented by a vector pointing from the current joint to its child joint (as shown

TABLE 4
Skeleton-based action recognition performance on the NTU-RGB+D dataset. The accuracies are reported on both the cross-subject (X-Sub) and cross-view (X-View) benchmarks.

Model	X-Sub	X-View
PA-LSTM [6]	60.7%	67.3%
ST-LSTM [12]	69.2%	77.7%
GCA-LSTM [50]	74.4%	82.8%
TConv [4]	74.3%	83.1%
S-TGCN [21]	74.9%	86.3%
ATT-LSTM [51]	76.1%	84.0%
Clips-CNN-MTLN [3]	79.6%	84.8%
VI [13]	80.0%	87.2%
RC-CNN [52]	81.1%	87.4%
ST-GCN [15] (Raw Joint)	81.5%	88.3%
TSI-CNN [14]	82.9%	90.1%
PB-GCN [17] (Raw Joint)	82.8%	90.3%
EI-GCN [16] (Raw Joint)	83.5%	89.8%
Motif-GCN [23] (Raw Joint)	84.2%	90.2%
SR-TSL [53] (Preprocessed Joint)	84.8%	92.4%
AS-GCN [19] (Preprocessed Joint)	86.8%	94.2%
Shift-GCN [20] (Preprocessed Joint)	87.8%	95.1%
PB-GCN [17] (Two streams)	87.5%	93.2%
AGCN [18] (Two streams)	88.5%	95.1%
DAG [22] (Two streams)	89.2%	95.5%
NAS-GCN [54] (Two streams)	89.4%	95.7%
SMotif-GCN+TBs (Ours, Raw Joint)	87.2%	93.6%
SMotif-GCN+TBs (Ours, Preprocessed Joint)	88.9%	95.0%
SMotif-GCN+TBs (Ours, Two Streams)	**90.5%**	**96.1%**

in Fig. 2). The bone data can be combined with joint data as spatial information, whose motion is easily calculated as the difference of spatial coordinates along the temporal dimension [22]. Different streams of spatial and motion information are fed into our model separately. Then, the Softmax scores of different streams are fused to obtain the final score for predicting the action class. With data preprocessing and score fusing, we evaluate the action recognition accuracies of SMotif-GCN+TBs and summarize the results in Table 4. The performance of our model SMotif-GCN+TBs with joint and bone data as two-stream input, denoted as "SMotif-GCN+TBs (Ours, Two streams)" is compared with the state-of-the-art multi-stream models [17], [18], [22], [54]. The results in Table 4 show that our model, with two streams of input, achieves the best performance on the large-scale NTU-RGB+D dataset.

In addition to geometric features (joint and bone positions) of skeletal data, we have also tried to use visual features as additional input in the model. As the video frames from the NTU-RGB+D dataset are captured in the constrained lab environment, the proportion of subjects is similar in different action samples. Thus, we extract latent visual features from local pictures (64×64) centered on joints by the VGG network [55]. Specifically, the latent code (the average of the feature maps in the spatial and temporal domains) from the first layer of the VGG is used. As shown in Table 5, the performance of SMotif-GCN+TBs with visual input is inferior to the performance of the model with geometric input, because skeletal data that can represent high-level structural information is more efficient for action recognition. Given the high performance of the Motif-GCN+TBs with geometric data, there is still an improvement in accuracy of SMotif-GCN+TBs with additional visual data (Table 5).

TABLE 5
The accuracies of our proposed SMotif-GCN+TBs model with different data preprocessing methods, evaluated on the NTU-RGB+D dataset.

Preprocessing Method	X-Sub	X-View
Geometric	90.5%	96.1%
Visual	72.9%	74.8%
Geometric and Visual	**91.7%**	**96.7%**

4.6.2 Experiments on NTU-RGB+D-120 Dataset

On NTU-RGB+D-120 dataset, we compare our proposed model with the state-of-the-art skeleton-based action recognition methods [3], [4], [13], [15], [18], [52]. We follow the standard evaluation protocol in [47], and report both X-Sub and X-Set Top-1 recognition accuracies. As NTU-RGB+D-120 is an extension of the NTU-RGB+D dataset, the hyper parameter settings of LTB and NLTB in our model are set in accordance with the default settings in the NTU-RGB+D dataset.

The NTU-RGB+D-120 dataset contains more action categories, and the skeleton sequences in it are captured in varied environmental setups and performed by diverse human subjects. The high variability in different aspects (i.e., environmental setups and human subjects) makes the action recognition task more challenging. The results in Table 6 show that our model achieves the best performance. Specifically, our full model with raw joint as input can also outperform the state-of-the-art method [20] with preprocessed joint as input. What's more, our SMotif-GCN+TBs with preprocessed joint input achieves higher accuracies than the previous method on X-Sub (83.2% *v.s.* 80.9%) and X-View benchmarks (84.1% *v.s.* 83.2%). The comparison results show the superiority of our model to extract spatial and temporal features from skeleton sequences for action recognition in challenging scenarios.

Similar to the experiments on the NTU-RGB+D dataset (Sec. 4.6.1), we also use additional visual data and conduct experiments on the NTU-RGB+D-120 dataset. As shown in Table 7, the performance of SMotif-GCN+TBs is improved from 87.1% to 88.4% under X-Sub evaluation and from 87.7% to 88.9% under X-Set evaluation.

4.6.3 Experiments on Kinetics-M Dataset

To demonstrate the effectiveness of our model for skeleton-based action recognition in unconstrained natural scenarios, we compare it with the most related methods [4], [15], [18], [19], [23] in terms of Top-1 and Top-5 accuracies on the Kinetics-M dataset as recommended in [46].

TConv [4] uses convolutions for temporal modeling of skeleton sequences represented by clips of time series with 3D joint coordinates. ST-GCN [15] firstly constructs a spatial-temporal graph to represent a skeleton sequence. AGCN [18] proposes to use an adaptive graph structure, and uses a normalized embedded Gaussian function to calculate

TABLE 6
Skeleton-based action recognition performance on the NTU-RGB+D-120 dataset. The accuracies are reported on both the cross-subject (X-Sub) and cross-setup (X-Set) benchmarks.

	X-Sub	X-Set
PA-LSTM [6]	25.5%	26.3%
ST-LSTM [12]	58.2%	60.9%
GCA-LSTM [50]	58.3%	59.2%
Clips-CNN-MTLN [3]	58.4%	57.9%
FSNet [56]	59.9%	62.4%
VI [13]	60.3%	63.2%
ATT-LSTM [51]	61.2%	63.3%
RC-CNNs [52]	62.2%	61.8%
TConv [4] (Raw Joint)	68.2%	67.2%
ST-GCN [15] (Raw Joint)	76.9%	78.5%
Motif-GCN [23] (Raw Joint)	80.2%	81.6%
Shift-GCN [20] (Preprocessed Joint)	80.9%	83.2%
SMotif-GCN+TBs (Ours, Raw Joint)	82.0%	83.1%
SMotif-GCN+TBs (Ours, Preprocessed Joint)	83.2%	84.1%
SMotif-GCN+TBs (Ours, Two streams)	**87.1%**	**87.7%**

TABLE 7
The accuracies of our proposed SMotif-GCN+TBs model with different data preprocessing methods, evaluated on the NTU-RGB+D-120 dataset.

Preprocessing Method	X-Sub	X-Set
Geometric	87.1%	87.7%
Visual	59.5%	59.9%
Geometric and Visual	**88.4%**	**88.9%**

TABLE 8
Skeleton-based action recognition performance on the Kinetics-M dataset in terms of Top-1 and Top-5 accuracies.

	Top-1	Top-5
TConv [4]	70.8%	92.5%
ST-GCN [15]	79.7%	94.2%
AS-GCN [19]	78.8%	94.6%
AGCN [18]	81.2%	95.2%
Motif-GCN [23]	84.2%	96.1%
SMotif-GCN+TBs (Ours, Preprocessed Joint)	85.4%	96.8%
SMotif-GCN+TBs (Ours, Two streams)	**85.7%**	**96.8%**

the similarity of physically connected joints. AS-GCN [19] learns the relationships between arbitrary pairwise joints. The skeletal data in the Kinetics-M dataset is captured in the unconstrained daily-life scenarios, so the data are varied. Consequently, all the related methods adopt preprocessing strategies [15] to normalize the joint data. In addition to joint data, we also use bone data as two streams of input. The results in Table 8 show that our full model SMotif-GCN+TBs with preprocessed joint and two streams of input achieve better performance than all the compared methods.

4.6.4 Experiments on Kinetics Dataset

Similar to the experiments on the Kinetics-M dataset (Sec. 4.6.3), we compare our model with the state-of-the-art methods [4], [6], [15], [18], [19], [22], [57] in terms of Top-1 and Top-5 accuracies for skeleton-based action recognition on the Kinetics dataset. Our model again achieves the best performance, as shown in Table 9. Specifically, our full model with joint data as input can still outperform the state-of-the-art method [22] with two streams data as input. The comparison results show the superiority of our model for action recognition in unconstrained daily-life scenarios.

TABLE 9
Skeleton-based action recognition performance on the Kinetics dataset in terms of Top-1 and Top-5 accuracies.

	Top-1	Top-5
FeatureEnc [57]	14.9%	25.8%
PA-LSTM [6]	16.4%	35.3%
TConv [4]	20.3%	40.0%
ST-GCN [15]	30.7%	52.8%
AS-GCN [19]	34.8%	56.5%
AGCN [18] (Two streams)	36.1%	58.7%
DAG [22] (Two streams)	36.9%	59.6%
SMotif-GCN+TBs (Ours, Preprocessed Joint)	36.8%	59.4%
SMotif-GCN+TBs (Ours, Two streams)	**37.8%**	**60.6%**

There are a great variety (400 categories) of human actions in the Kinetics dataset. Most of the actions require the interactions between the subjects and scenes to be accurately classified. It is challenging to recognize the actions with skeletal data, so the accuracies of our model on this dataset are lower than those on the Kinetics-M dataset.

5 Conclusion

In this paper, we have proposed the Motif-GCN with the SMAM in order to adaptively encode spatial information in human skeletons, both locally and non-locally. The SMAM is introduced to capture richer dependencies between non-physically connected joints and can be updated during the learning process. Besides, we have proposed a novel LTB for modeling temporal information by reusing feature maps of local temporal convolutions with dense connections. We further improve the performance by reducing duplication and enriching variety in gradient propagation of densely concatenated layers. Finally, a NLTB is further used to construct global dependencies, which can effectively extract more representative features to increase recognition accuracy. As a result, our model can effectively encode local and non-local information in the spatial as well as temporal context. We validated our model on four large-scale datasets and our model achieve the best performance on all the datasets.

Many action classes in the Kinetics dataset require recognizing relationships between actors and complex scenes. The model focusing on skeleton-based action recognition is inferior to video-based methods, which can make use of additional information in the background. Moreover, the skeleton sequences in the Kinetics dataset are tracked by monocular pose estimation algorithms (e.g. OpenPose). The occlusion and ambiguity of the 2D skeletal data may also degrade the performance of skeleton-based action recognition methods. However, we believe that the skeletal data can provide complementary information to raw RGB data.

In our future work, we are interested in combining data from other information carriers such as raw RGB to achieve further improvements in action recognition.

Acknowledgments

This work was supported by National Key Research and Development Plan (No. 2021YFF0307702), China Postdoctoral Science Foundation (No.2021M701891), Tsinghua University Initiative Scientific Research Program, National Natural Science Foundation of China (No. 61725204), Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), Youth Innovation Promotion Association CAS, Marsden Fund Council managed by Royal Society of New Zealand (No. MFP-20-VUW-180).

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011, pp. 1297–1304.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 1302–1310.

[3] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *CVPR*, 2017, pp. 4570–4579.

[4] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *CVPRW*, 2017, pp. 1623–1631.

[5] 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 *ICCV*, 2017, pp. 2136–2145.

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb + d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[7] 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," in *AAAI*, 2016, pp. 3697–3704.

[8] 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 *AAAI*, 2017, pp. 4263–4270.

[9] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *WACV*, 2017, pp. 148–157.

[10] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using lstm and cnn," in *ICMEW*, 2017, pp. 585–590.

[11] A. Ben Tanfous, H. Drira, and B. Ben Amor, "Coding kendall's shape trajectories for 3d action recognition," in *CVPR*, 2018, pp. 2840–2849.

[12] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *TPAMI*, pp. 3007–3021, 2018.

[13] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, pp. 346–362, 2017.

[14] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, M. He *et al.*, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *ICMEW*, 2017, pp. 601–604.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition." in *AAAI*, 2018, pp. 7444–7452.

[16] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *CVPR*, 2018, pp. 5323–5332.

[17] K. C. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition." in *BMVC*, 2018, p. 270.

[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12 026–12 035.

[19] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 3595–3603.

[20] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network." in *CVPR*, 2020, pp. 180–189.

[21] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, "Spatio-temporal graph convolution for skeleton based action recognition." in *AAAI*, 2018, pp. 3482–3489.

[22] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *CVPR*, 2019, pp. 7912–7921.

[23] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph cnns with motif and variable temporal block for skeleton-based action recognition." in *AAAI*, 2019, pp. 8989–8996.

[24] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks." in *CVPR*, 2018, pp. 7794–7803.

[25] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*, 2016, pp. 5308–5317.

[26] D. K. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints." in *NeurIPS*, 2015, pp. 2224–2232.

[27] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *NeurIPS*, 2016, pp. 1993–2001.

[28] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017, pp. 1024–1034.

[29] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv*, 2015.

[30] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering." in *NeurIPS*, 2016, pp. 3837–3845.

[31] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR*, 2017.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *ICLR*, 2018, pp. 1–12.

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[35] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification." in *CVPR*, 2017, pp. 6450–6458.

[36] G. Hu, B. Cui, and S. Yu, "Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 1216–1221.

[37] H. Guyue, C. Bo, and Y. Shan, "Joint learning in the spatio-temporal and frequency domains for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2207–2220, 2019.

[38] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, pp. 163–166, 2016.

[39] Y. Fang, W. Lin, V. W. Zheng, M. Wu, K. C.-C. Chang, and X.-L. Li, "Semantic proximity search on graphs with metagraph-based learning," in *ICDE*, 2016, pp. 277–288.

[40] J. B. Lee, R. A. Rossi, X. Kong, S. Kim, E. Koh, and A. Rao, "Graph convolutional networks with motif-based attention." in *CIKM*, 2019, pp. 499–508.

[41] R. A. Rossi, N. K. Ahmed, and E. Koh, "Higher-order network representation learning." in *WWW (Companion Volume)*, 2018, pp. 3–4.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, 2017, pp. 4700–4708.

[43] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn." in *CVPRW*, 2020, pp. 1571–1580.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." in *ICML*, 2015, pp. 448–456.

[45] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." in *AISTATS*, vol. 15, 2011, pp. 315–323.

[46] 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*, 2017.

[47] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "Ntu rgb + d 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, pp. 2684–2701, 2019.

[48] P. Adam, G. Sam, C. Soumith, C. Gregory, Y. Edward, D. Zachary, L. Zeming, D. Alban, A. Luca, and L. Adam, "Automatic differentiation in pytorch," in *NeurIPS Workshop*, 2017.

[49] D. Pagliari and L. Pinto, "Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors," *Sensors*, vol. 15, no. 11, pp. 27 569–27 589, 2015.

[50] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *CVPR*, 2017, pp. 1647–1656.

[51] 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," *TIP*, pp. 1586–1599, 2017.

[52] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *TIP*, pp. 2842–2855, 2018.

[53] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning." in *ECCV*, 2018, pp. 106–121.

[54] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching." in *AAAI*, 2020, pp. 2669–2676.

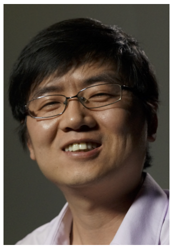
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." in *ICLR*, 2015.

[56] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. K. Chichung, "Skeleton-based online action prediction using scale selection network," *TPAMI*, 2019.

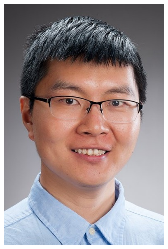
[57] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *CVPR*, 2015, pp. 5378–5387.

Yu-Hui Wen received the bachelor's degree from Harbin Institute of Technology (HIT), and the Ph.D. degree in computer science and technology from University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2020. She is currently a postdoc at Tsinghua University. Her research interests include computer vision, artificial intelligence, virtual reality, and human motion analysis.

Lin Gao received his PhD degree in computer science from Tsinghua University. He is currently an Associate Professor at the Institute of Computing Technology, Chinese Academy of Sciences. He has been awarded the Newton Advanced Fellowship from the Royal Society and the AG young researcher award. His research interests include computer graphics and geometric processing.



Hongbo Fu received the BS degree in information sciences from Peking University, China, in 2002 and the PhD degree in computer science from the Hong Kong University of Science and Technology, in 2007. He is a Professor in the School of Creative Media, City University of Hong Kong. His primary research interests fall in the fields of computer graphics and human computer interaction. He has served as an associate editor of The Visual Computer, Computers & Graphics, and Computer Graphics Forum.



Fang-Lue Zhang is currently a lecturer with Victoria University of Wellington, New Zealand. He received the Bachelors degree from Zhejiang University, Hangzhou, China, in 2009, and the Doctoral degree from Tsinghua University, Beijing, China, in 2015. His research interests include image and video editing, computer vision, and computer graphics. He is a member of IEEE and ACM. He received Victoria Early Career Research Excellence Award in 2019.

Shihong Xia received PhD degree in computer science from University of Chinese Academy of Sciences. He is currently a professor of Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS), and director of the human motion laboratory. His primary research is in the area of computer graphics, virtual reality and Artificial Intelligence.

Yong-Jin Liu is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include computational geometry, computer graphics and computer vision. He is a senior member of the IEEE.