Scale-Aware Graph Convolutional Network with Part-Level Refinement for Skeleton-Based Human Action Recognition

Chang Li, Yingchi Mao, Qian Huang, Xiaowei Zhu, Jie Wu, *Fellow, IEEE*

***Abstract*—Graph Convolutional Networks (GCNs) have been widely used in skeleton-based human action recognition and have achieved promising results. However, current GCN-based methods are limited by their inability to refine semantic-guided joint relations and perform adaptive multi-scale analysis. These limitations impair their performance, particularly for analogical actions involving the interaction of the same body parts (e.g., drinking water and eating) as well as deficient actions with limited spatial-temporal information (e.g., subtle action writing and transient action sneezing). To solve these problems, we propose Part-level Refined Spatial Graph Convolution (PR-SGC) and Scale-aware Temporal Graph Convolution (Sa-TGC) for optimal action representation. The PR-SGC divides the skeleton into body parts and embeds this high-level semantics to refine the physical adjacency matrix. The Sa-TGC leverages the dynamic scale-aware mechanism to extract context-dependent multi-scale features. On this basis, we develop a novel Scale-aware Graph Convolutional Network with Part-level Refinement (SaPR-GCN), which is on par with state-of-the-art benchmarks on NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA datasets.**

***Index Terms*—Action Recognition, Graph Convolutional Networks, Spatiotemporal Modeling, Multi-scale Analysis.**

I. Introduction

HUMAN action recognition is a fundamental topic in computer vision with broad applications, including video surveillance [1] and human-computer interaction [2]. Human actions can be described by multimodal data, such as RGB videos, depth videos, and skeleton sequences. Due to the compactness of representation and robustness to environmental variations, skeleton-based human action recognition has attracted increasing attention.

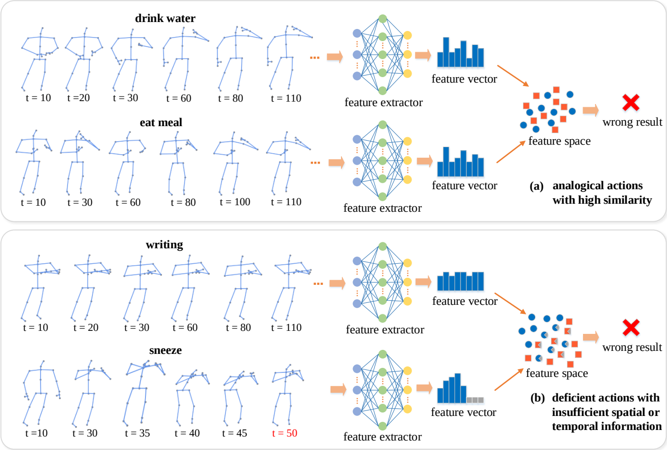


Fig. 1. Examples of analogical and deficient actions. Analogical actions denote similar actions involving interactions between the same body parts, e.g., drinking water and eating meals cover the resemble interaction trajectories of hands and head. The extracted features are closely distributed in feature space. Besides, deficient actions refer to the actions that provide insufficient dynamic information, including continuous actions with subtle movements, e.g., writing, and transient actions with short durations, e.g., sneezing. The above skeleton sequences are derived from NTU RGB+D 60 dataset.

In essence, skeleton sequences consist of isomorphic spatial-temporal graphs, where bones are considered spatial edges while the identical joints between two adjacent frames are connected as temporal edges [3]. Therefore, many researchers have employed Graph Convolutional Networks (GCNs) in skeleton-based action recognition and achieved promising results. However, existing GCN-based methods still have challenges in recognizing some specific categories of actions, which we defined as analogical actions and deficient actions, as shown in Fig. 1. Specifically, analogical actions denote the action pairs with high similarity which often involve interactions between the same body parts. As shown in Fig. 1(a), drinking water and eating meals cover the resemble interaction trajectories of hands and head. The extracted features are closely distributed in high-dimensional space, and thus their inter-class difference is too small to distinguish. Besides, deficient actions refer to the actions that provide insufficient dynamic information for recognition, including continuous actions with subtle movements, e.g., writing, and transient actions with short durations, e.g., sneezing. As shown in Fig. 1(b), writing only has slight spatial variation of hands, and sneezing involves wide range changes in multiple parts but has a shorter time interval. The essential nature of these actions

Chang Li, Yingchi Mao, Qian Huang, and Xiaowei Zhu is with the College of Computer Science and Software Engineering, Hohai University, Nanjing, Jiangsu, China, 211100. (E-mail: lichang, yingchimao, huangqian, zhuxiaowei@hhu.edu.cn)

Jie Wu is with the Department of Computer and Information Sciences, Temple University, SERC 362, 1925 N. 12th Street, Philadelphia, PA 19122. (E-mail: jiewu@temple.edu)

This work is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. 422003261), the Fundamental Research Funds for the Central Universities (No. B230205027), Key Research and Development Program of China, Yunnan Province (No. 202203AA080009), the Key Research and Development Program of China (No. 2022YFC3005401), the 14th Five-Year Plan for Educational Science of Jiangsu Province (No. D/2021/01/39), the Jiangsu Higher Education Reform Research Project (No. 2021JSJG143) and the 2022 Undergraduate Practice Teaching Reform Research Project of Hohai University.

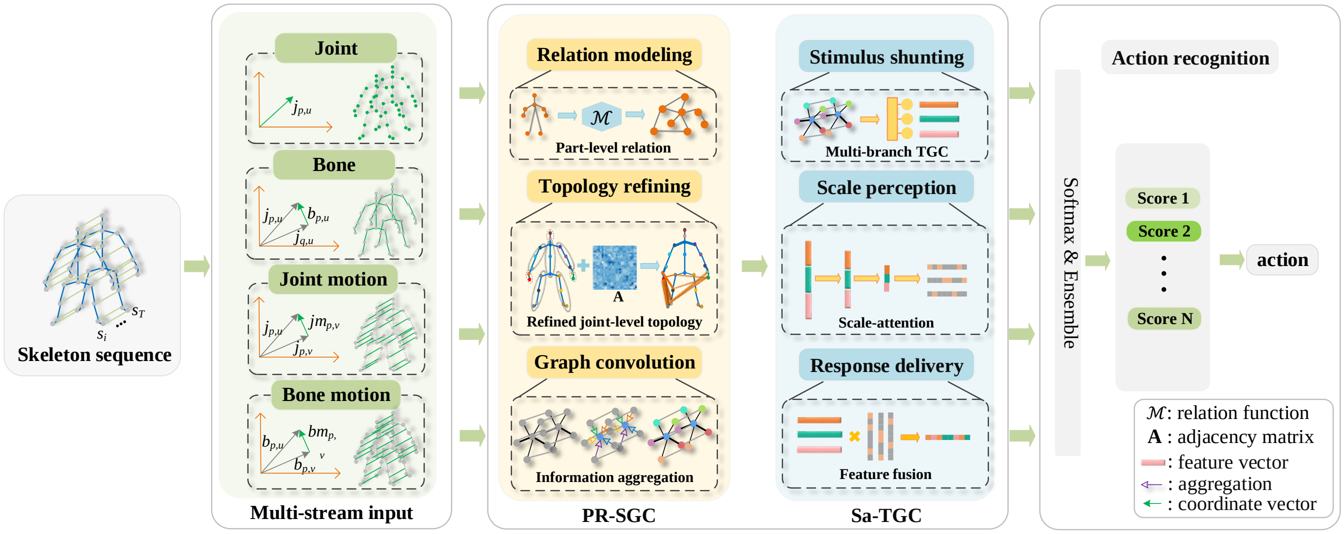


Fig. 2. The pipeline of SaPR-GCN. The data preprocessing module first transforms the input skeleton sequence into multi-stream data. Then, four separate models with identical architecture are trained using joint, bone, joint motion, and bone motion, respectively. The softmax scores of multiple streams are ensembled to obtain the final results. The SaPR-GCN is the stack of PR-SGC and Sa-TGC blocks for effective spatial-temporal representations. Specifically, PR-SGC exploits body part-guided constraints for mining spatial patterns, and Sa-TGC generates multi-scale temporal dynamics in an adaptive context-dependent manner.

requires the model to extract more discriminative features for recognition. After in-depth analysis, we find that the existing GCN-based method is mainly composed of a stack of layers included a Spatial Graph Convolution (SGC) and a Temporal Graph Convolution module (TGC). Based on this, we explore why it is difficult for existing methods to distinguish the above actions from spatial and temporal aspects.

Firstly, the SGC is an extension of the convolution on images [4]. The difference is that the image has a regular grid structure, i.e., the relative position of pixels is natural and static, while that in the graph structure is indeterminate. Essentially, the joint relationship in the graph is controlled by the adjacency matrix. Similar to the pixel arrangement is significant to image convolution, the adjacency matrix is critical to graph convolution, which determines the graph structure and feature aggregation during graph convolution. To optimize the SGC, Yan et al. [5] utilized the fixed graph for spatial-temporal modeling, but they failed to capture long-range dependencies, e.g., the information between the two hands is blocked. Shi et al. [3] introduced two supplementary graphs, the structure of which are learned by trainable parameters. Wu et al. [6] constructed multi-directional graphs and global learnable matrices to mine latent joint relationships, including inward, outward, and undirected graphs. Huang et al. [7] proposed a high-resolution skeleton graph (HRG) by creating virtual joints between each pair of physically connected joints and fully connecting every joint as the spatial graph. Despite this, the topology is not optimal and lacks interpretability without physical constraints such as human body parts. Huang et al. [8] devised the part relation block with graph pooling operators to obtain the body parts relationship. However, the pooling function neglected the correlation diversity of joints in the same part. To solve these problems, we propose the Part-level Refined Spatial Graph Convolution (PR-SGC) to mine intrinsic joint relations guided by body part semantics for meticulous topology.

Secondly, as depicted in Fig. 1(b), writing has less spatial variation and requires a more fine-grained feature description while sneezing has a shorter duration and requires capturing a wide range of motion patterns. Therefore, multi-scale analysis is vital to excavate exhaustive information about these actions. Peng et al. [9] applied multi-order polynomials to involve richer connections between multi-distant joins in the spatial dimension, ignoring multi-range temporal dependencies. Chen et al. [10] proposed MST-GCN for multi-scale spatio-temporal modeling. Nevertheless, the temporal scale is limited by the distance factor in the spatial dimension. Liu et al. [11] investigated a unified spatial-temporal operator called MS-G3D with a multi-branch structure. However, the scale is inflexible due to the fixed dilation of temporal convolutions. Due to the lack of an adaptive scale analysis mechanism in the TGC, existing GCNs fail to capture action-specific granularity features, which impair their robustness and generality. Motivated by this, we aim to generate multi-scale spatio-temporal features adaptively and design the Scale-aware Temporal Graph Convolution (Sa-TGC) module.

By coupling the above efforts, we build the Scale-aware Graph Convolutional Network with Part-level Refinement Topology (SaPR-GCN), as shown in Fig. 2, which is competent for recognizing analogical and deficient actions. To sum up, our contributions are as follows:

- We devise PR-SGC that mines intrinsic and meticulous relations between joints guided by body part semantics to extract more fine-grained features.
- We investigate Sa-TGC to enrich the receptive fields of GCN, which generates multi-scale context-dependent features adaptively.
- Taking advantage of PR-SGC and Sa-TGC, we propose a

novel learning framework SaPR-GCN, which can extract more discriminative features, especially for analogical and deficient actions.

- We present some variants of SaPR-GCN, proving that PR-SGC and Sa-TGC can be ported to other GCN-based approachs as separate modules.
- The extensive experimental results and analyses of three public datasets: NTU RGB+D 60, NTU RGB+D 120, and NW-UCLA, show that our model is more interpretable and superior to state-of-the-art methods.

II. Related Work

Topology optimization methods. GCNs have been widely adopted in skeleton-based action recognition to explore a more effective representation of human behaviors. In GCN-based methods, the human skeleton is represented as a graph where the joints are nodes, and the bones are edges. The topology of joints reflected by the adjacency matrix is pivotal in graph convolution, which controls the received fields of graph convolution layers. Yan et al. [5] first constructed the spatial-temporal graph with physical and inter-frame connections between joints for action modeling, laying the foundation for GCN-based methods. Shi et al. [12] designed the directed acyclic GCN, in which incoming and outgoing edges connect vertexes. However, due to the fixed topology, these predefined models lack generality to new samples. Shi et al. [3] proposed a two-stream structure called 2s-AGCN, which embedded the Gaussian function to calculate the similarity of the two vertexes and added learnable parameters to learn the dependency strength between joints. Song et al. [13] focused on the activation degree of joints, which is measured by the class activation maps (CAM). Li et al. [14] introduced the A-link inference module to capture action-specific dependencies for topology optimization. Nevertheless, these approaches lack semantic guidance, are challenging to exploit nuanced dependencies fully, and cannot be well interpreted.

Body part-based methods. Li et al. [15] focused on the inter-body semantics of the skeletons in two-person interactive actions and constructed a unified graph through edge labeling strategies. However, this coarse-grained relationship cannot fully mine the joint relationship in the single skeleton. Thakkar et al. [16] divided the human skeleton into four subgraphs and proposed PB-GCN to embed part semantics. However, they impaired the information aggregated across body parts. Song et al. [17] focused on discovering the significant parts and investigated the ResGCN with attention mechanisms, yet the global information is diminished according to the hierarchical structure. Huang et al. [8] learned high-level relations between body parts and highlighted the vital parts using graph pooling and unpooling operations. However, the pooling function neglected the correlation diversity of joints in the same part. Generally, these approaches are disabled to capture elaborate topology, and the performance on challenging actions is unideal consequently. To solve these problems, we propose the part-level refined spatial graph convolution to mine intrinsic joint relations meticulously with the aid of body part semantics. Note that all these part-based models aim to extract features from body parts individually. In contrast, our work focuses on learning and transferring part-level potential relations into joints to refine the topology.

Multi-scale analysis methods. To extract more discriminative features, SEFN [18] proposed the Multi-perspective Attention Fusion Module (MPAFM), which fuses the information from the spatial, channel, and temporal branches by attention mechanism. However, only extracting the single-scale feature fails to capture discriminative action patterns because human movements involve multi-range concurrency. AS-GCN [14] and NAS-GCN [9] tried to capture multi-scale features from non-local neighbors via higher-order polynomials of the adjacency matrix. Generally, these formulations suffered from the biased weighting problem due to self-connection. Liu et al. [11] proposed MS-G3D to disentangle redundant dependencies and first applied differentiable dilated convolutions for multi-scale temporal dynamics. Chen et al. [10] investigated the MS-GC and MT-GC modules for multi-scale spatial-temporal modeling, but the temporal scale is mainly implemented by the scale factor in MS-GC. CTR-GCN [19] adopted the multi-branch structure with different dilations like MS-G3D and achieved better performance. However, the above methods extracted multi-scale features through predefined topological structures or fixed dilations. Due to the inflexibility of scale, they failed to extract discriminative features for analogical and deficient actions. In light of these limitations, we propose the scale-aware mechanism to generate multi-scale features adaptively.

III. Methods

The pipeline of 4-stream SaPR-GCN is depicted in Fig. 2. In this section, we first introduce the GCNs (Sec. III.A) and derive the generalized part-based graph convolution operators (Sec. III.B). Then, we present the basic two modules PR-SGC (Sec. III.C) and Sa-TGC (Sec. III.D), which are responsible for capturing the motion patterns in spatial configuration and temporal dynamics, respectively. Finally, we describe the overall architecture of the network and its implementation (Sec. III.E).

A. Preliminaries

Spatial graph convolutional networks. A human skeleton can be considered as a graph with joints as vertices and bones as edges. The graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \dots, v_V\}$ is the set of V vertices, and $\mathcal{E}$ is the set of edges defined by an adjacency matrix $A \in \mathbb{R}^{V \times V}$. $A_{i,j}$ reflects the connection strength between v_i and v_j. The neighborhood of v_i is represented as $\mathcal{N}_{v_i} = \{v_j \mid A_{i,j} \neq 0\}$. After that, an action with T frames can be described by a tensor $X \in \mathbb{R}^{T \times V \times C}$ or a set of node features $\mathcal{X} = \left\{x_{t,i} \in \mathbb{R}^{C} \cdot \mid t, i \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq i \leq V\right\}$, where $x_{t,i} = X_{t,i,:}$ is the C dimensional feature of v_i at time t. Then, the layer-wise update rule of GCNs can be formulated as

$$X_t^{(l+1)} = \sigma\left(\bar{A} \cdot X_t^{(l)} W^{(l)}\right) \tag{1}$$

where $\sigma(\cdot)$ is the activation function. We define $\tilde{A} = A + I$ as the adjacency matrix with self-loops. Then, $\bar{A} = D^{-\frac{1}{2}} \tilde{A} D^{\frac{1}{2}}$, and D is the diagonal degree matrix of $\tilde{A}$. $X_t^{(l)}$ is the input of the l_{th} layer with the weights $W^{(l)}$.

Temporal graph convolutional networks. As defined in [5], temporal edges exist between the identical vertex in consecutive frames. The neighborhood of v_{ti} represented as $\mathcal{N}_{\mathcal{T}}\left(v_{ti}\right)=\{v_{qi}\|q-t\mid\leq\lfloor\varGamma/2\rfloor\}$, where $\varGamma$ is the temporal window size. The temporal edges can be interpreted as the trajectories of the joints during time T intuitively[30]. The classical implementation of temporal graph convolution can be written as

$$X_{\mathcal{T}}^{(l+1)}=\mathrm{Conv}\,2D[\varGamma\times 1]\left(X^{(l)}\right) \tag{2}$$

where $\varGamma\times 1$ is the kernel size of 2D convolution. Typically, the spatial-temporal convolutional network is built by stacking these layers, i.e., alternately performing Eq. (1) and Eq. (2).

B. Part-based graph convolutional networks

Human actions are the co-movement of various body parts. Intuitively, the skeleton graph can be constructed as a combination of subgraphs with certain properties. Inspired by existing part-based methods [8], [16], [20], [21], we derive a general part-based spatio-temporal graph representation. Let us consider that a graph $\mathcal{G}$ has been divided into P partitions. It can be described as $\mathcal{G}=\bigcup_p^P\mathcal{P}_p\mid\mathcal{P}_p=(\mathcal{V}_p,\mathcal{E}_p)$, where $\mathcal{P}_p$ is the p_{th} part with vertices set $\mathcal{V}_p$ and edges set $\mathcal{E}_p$. Based on the above definitions, a generalized part-based graph convolution is as follows:

$$X^{(l+1)}=\mathcal{F}_{\mathcal{T}}\left(\mathcal{F}_{part}\left(X_{t,p}^{part}\right)\right),1\leq t\leq T,1\leq p\leq P \tag{3}$$

with

$$X_{t,p}^{part}=\mathcal{F}_{joint}\left(x_{t,i}\right),1\leq t\leq T,1\leq p\leq P,1\leq i\leq|v_p| \tag{4}$$

where $\mathcal{F}_{\mathcal{T}}$ is denoted as the temporal graph convolution. $X_{t,p}^{part}$ is the feature vector of part p at time t. $\mathcal{F}_{part}$ and $\mathcal{F}_{joint}$ is the aggregation function for part-level and joint-level information, respectively. We note that the existing methods all extract action features hierarchically in a bottom-to-up manner, i.e., first aggregating low-level joint features and then aggregating high-level part features to generate action representations. These approaches hinder bidirectional information propagation across levels, making it challenging to capture finer-grained features and thus failing to discriminate analogical actions involving the co-movement of identical joints or similar body parts.

C. Part-level Refined Spatial Graph Convolution

Because of this observation, we focus on iteratively capturing refined dependencies between joints guided by high-level body part semantics. Body parts can be considered the smallest action execution unit because they lead to natural constraints between joints, and the joints bound in the same part usually have strong linkage and concurrency, e.g., the leg binds the knee and ankle. In addition, the human body is symmetrical exquisitely, and there are generally consistent or opposite motion patterns between parts. Furthermore, the patterns should be action-specific, e.g., our arms move oppositely when walking, while when jumping, they always show a consistent tendency. Thereupon, high-level body part semantics are essential for low-level joint properties and fine-grained action representations.

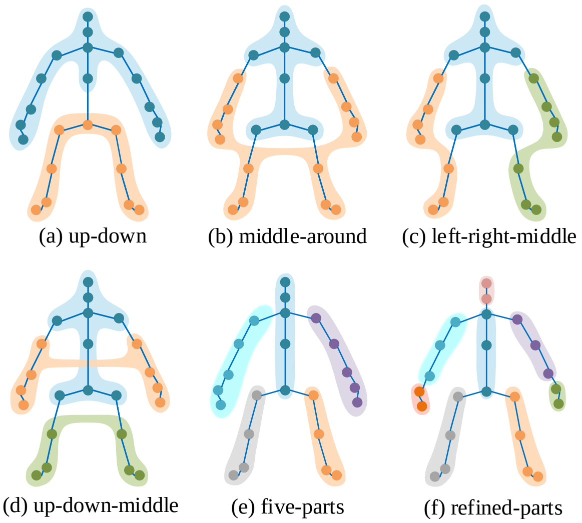


Fig. 3. Part division diagram. According to the composition and structure of the human skeleton, we explore six different strategies for dividing parts. Different colors highlight different parts. (a) The Up-down strategy divides the skeleton into upper and lower parts, while (b) the middle-around divides it into appendicular and axial skeletons. Further, (c) left-right-middle segments the appendicular into left and right parts, and (d) up-down-middle splits the appendicular into upper and lower parts. (e) Five-parts strategy consists of the torso and limbs. (f) Refined-parts strategy separates the skeleton into eight parts: the palms, arms, legs, head, and torso. Experiments in Sec. IV.C show that the refined-parts strategy is better.

Part partition strategies. Based on the physical structure of the human skeleton, we explore six different part partition strategies as depicted in Fig. 3. In this work, we select the refined-parts partitioning through extensive experiments and analysis (Sec. IV.C). Specifically, the human skeleton is categorized into eight parts: head, torso, left and right arms, left and right hands, and left and right legs. Then, we utilize joint-based representations to extend the feature of each body part as follows.

$$X_i^{\text{part}}=\text{Concat}\left(\left\{x_j\mid j\in\mathcal{V}_i{}^{\text{part}}\right\}\right) \tag{5}$$

where $\mathcal{V}_i{}^{part}$ denotes the set of joints contained in part i. $x_j{}^{part}\in\mathbb{R}^{C\times T}$ is the feature of joint j. $Concat$ represents the cascade function. $x_i{}^{part}\in\mathbb{R}^{C\times T\times Q}$, where Q is the joints number of part i. On this basis, we can design the relation modeling and mapping function to refine the interdependence between joints, which allows for a more detailed understanding of how the joints move together.

Relation modeling functions. We investigate two correlation modeling functions $\mathcal{M}(\cdot)$ to measure the dependencies between parts. To reduce computation, we utilize linear transformations $\varphi(\cdot)$ and $\phi(\cdot)$ for compact feature representations. $\mathcal{M}_1$ obtains the instantaneous uniform distribution of each channel through the average pooling in the vertex dimension and performs batch matrix multiplication to compute the distance. Then, the nonlinear transformations $\sigma(\cdot)$ are conducted to obtain part-level topology. Given $part_i$ and $part_j$ with

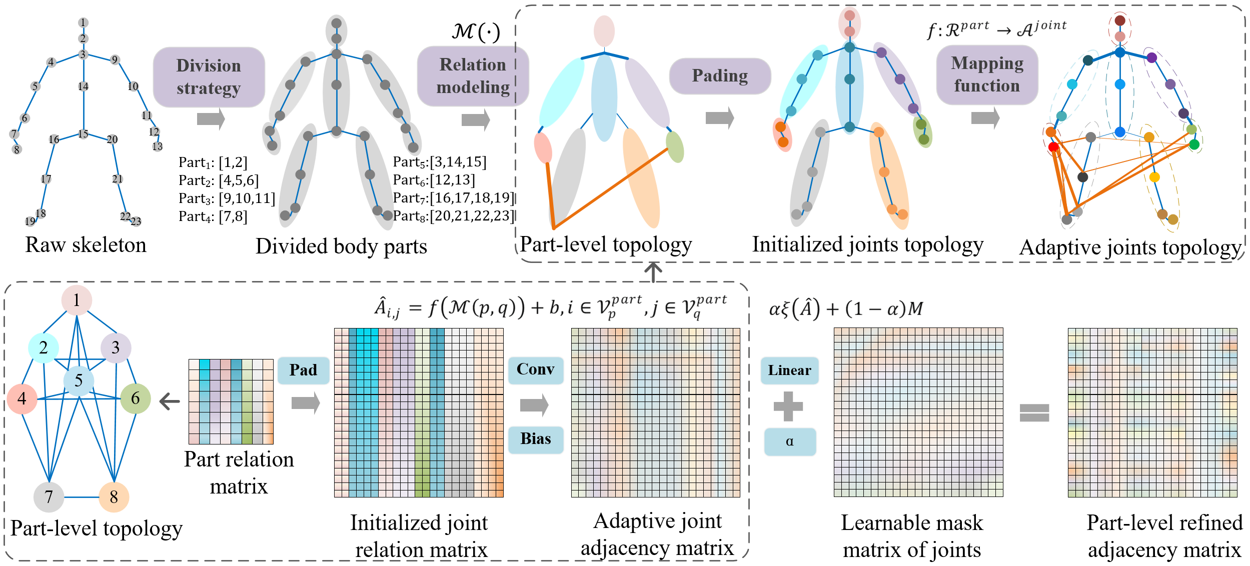


Fig. 4. The topology optimization process. In SaPR-GCN, the raw skeleton is divided into eight parts and constructed as a part-level topology defined by the part relation matrix according to relation modeling functions. After initialization and mapping, the adaptive joint adjacency matrix is derived and visualized by the adaptive joints topology. The thicker the line, the higher the correlation. The color of the joint denotes its weight, and the color of the matrix represents its element. Finally, the part-level refined adjacency matrix is obtained, which embeds body parts semantics gracefully.

corresponding features X_i^{part} and X_j^{part}, the formula of their relationship is as follows:

$$\mathcal{M}_1(i,j) = \sigma\left(\mathrm{Pool}_{1d}\left(\varphi\left(X_i^{part}\right)\right) \otimes \mathrm{Pool}_{1d}\left(\phi\left(X_j^{part}\right)\right)\right) \tag{6}$$

where Pool_{1d} is an adaptive average pooling in the spatial dimension. We use $\otimes$ to denote matrix multiplication in the temporal dimension. Before sending to $\mathcal{M}_2$, we employ two linear transformations for computational efficiency. Different from $\mathcal{M}_1$, $\mathcal{M}_2$ obtains a compact spatio-temporal representation by global average pooling, and the correlation is measured by the nonlinear transformations of distance. Formally:

$$\mathcal{M}_2(i,j) = \sigma\left(\mathrm{Pool}_{2d}\left(\varphi\left(X_i^{part}\right)\right) - \mathrm{Pool}_{2d}\left(\phi\left(X_j{}^{part}\right)\right)\right) \tag{7}$$

where Pool_{2d} is the global average pooling. The proposed relation modeling functions can dynamically reason the context-dependent higher-level relations between body parts in a bottom-to-up manner. Further analysis is presented in Sec. IV.C.

Part-level refined topology. Taking the relation between parts as the correlation between joints is not advisable since the contribution of joints even within the same part toward the action can differ. To solve this issue, we apply body parts semantics to guide the low-level topology of joints and design an up-to-bottom mapping function to substitute intra-parts weight sharing. The optimization process is depicted in Fig. 4. Particularly, the mapping function is defined as $f : \mathcal{R}^{part} \to \mathcal{A}^{joint}, \mathcal{R}^{part} \subseteq \mathbb{R}^{P\times P}, \mathcal{A}^{joint} \subseteq \mathbb{R}^{V\times V}$ which guarantees intra-parts joint communication as well. The refined topology $\hat{A}$ can be calculated as

$$\hat{A}_{i,j} = f(\mathcal{M}(p,q)) + b, i \in V_p^{part}, j \in V_q^{part} \tag{8}$$

where b is the learnable position bias. Particularly, we first initialize joint correlation, i.e., the refined topology, as part relation, and then use 1×1 convolution to realize the learnable dynamic mapping. According to Eq. (8), the physical connections between joints reflected by adjacency matrix A are refined by the embedded part semantics allowing for the weight that indicates connection strength radiating from parts to joints.

Part-semantic guided graph convolution. Finally, the graph convolution is advanced by leveraging the context-dependent intrinsic topology under the guidance of part semantics, and Eq. (1) is changed to

$$X_t^{(l+1)} = \sigma\left(\left[\alpha\xi(\hat{A}) + (1-\alpha)M\right]X_t^{(l)}W^{(l)}\right) \tag{9}$$

where $M \in \mathbb{R}^{C\times V\times V}$ is a learnable mask playing the role of global shared topology. α is a learnable weight that controls the contribution of the part refined topology $\hat{A}$ and M. We use linear transformation $\xi(\cdot)$ for channel consistency and to obtain the channel-wise dynamic topology. Through stacking PR-SGC layers, we iteratively refine the topology of the human skeleton and extract more discriminative spatial features, thus polishing the performance on analogical action.

D. *Scale-aware Temporal Graph Convolution*

As spatial graph convolution is a local operation, it can only utilize the spatial configurations between joints but fails to model the temporal dynamics vital to action recognition. For example, when recognizing actions putting on and taking off shoes that involve similar interactions of parts, temporal dynamic information is particularly crucial. Furthermore, focusing on multi-scale information for skeleton sequences can effectively extract sufficient discriminative features, especially for deficient actions with subtle movements or short durations. Different from previous methods [10], [11] that extract multi-scale features through predefined topological structures or

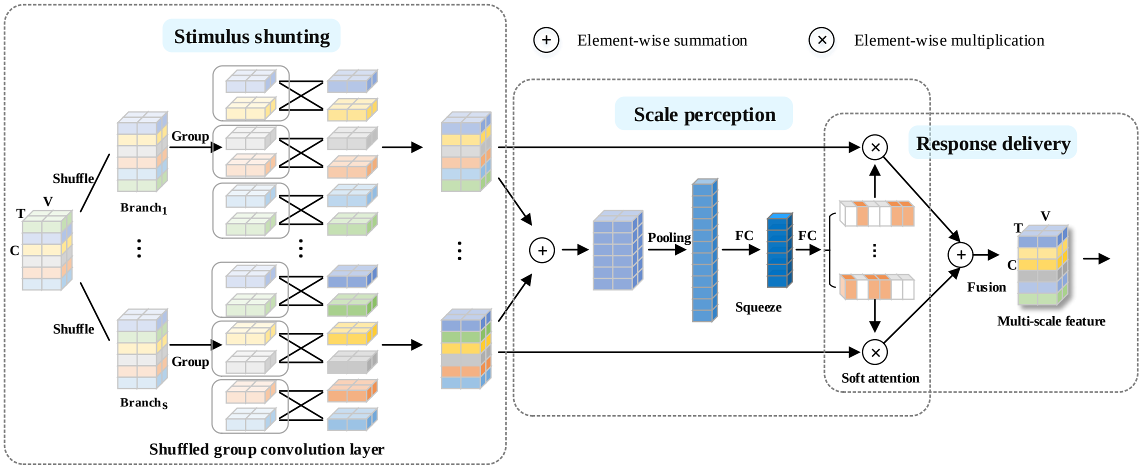


Fig. 5. The architecture of proposed Sa-TGC that mimics the mechanism of the visual cortex. It deploys a multi-pathway design to capture multi-scale temporal dynamics simultaneously. Stimulus shunting is responsible for sending input to branches for feature extraction. Each branch contains a shuffled group convolution layer for the sake of channel communication and efficiency. Scale perception helps the model to reason the critical features according to action context. Response delivery plays the role of feature selection, fusion, and transmission.

fixed-size convolution kernels, we propose a Sa-TGC for adaptive multi-scale feature extraction and fusion. We employ the bottleneck [22] structure with multiple branches to build the scale-aware temporal modeling module. Specifically, it can be divided into three stages, i.e., stimulus shunting, scale perception, and response delivery, as illustrated in Fig. 5.

Stimulus shunting. To reduce computation cost, we utilize the 1×1 convolution layer at both ends of Sa-TGC, which is responsible for reducing and then restoring dimensions. Afterward, the batch-normalized feature called stimuli is shunted through a multi-branch structure. Unlike [11], [19], we use convolution kernels with varying sizes in each branch to enrich the receptive field instead of adjusting the dilation. We build each branch using point-wise convolutions [23] and group convolutions for low complexity and introduce the channel shuffling operator to enable cross-channel communication and capture contextual information outside the group region [24]. The shuffled multi-scale temporal convolution can be illustrated in Eq. (10).

$$\mathcal{F}_{\mathcal{T}}^{(k)}\left(X^{l}\right)=\operatorname{Conv} 2D[k \times 1]\left(\mathcal{S}\left(X^{(l)}\right)\right) \tag{10}$$

where $\mathcal{S}(\cdot)$ is the channel shuffle function and k is a factor that controls the receptive fields of temporal convolution. Note that the value of convolution parameter padding is constrained by the shape invariance rule with variable k. The output of each branch can be obtained by the following equation

$$U_s=\mathcal{F}_{\mathcal{T}}^{(k_s)}(X), 1 \leq s \leq S \tag{11}$$

where $U_s \in \mathbb{R}^{C \times T \times V}$ is the result of s_{th} branch. S is the number of branches and controls the diversity of receptive fields. k_s is the kernel size of branch s.

Scale perception. Multi-scale analysis should be action-specific since actions have diverse durations, such as “writing” and “sneeze”. To this end, the scale-aware mechanism is designed to compute the importance of adaptive multi-scale feature fusion, which simulates neurons’ excitation and inhibition when perceiving various signals. We employ the addition operation to perform the prophase fusion of features shunted from branches denoted as $\widehat{U} \in \mathbb{R}^{C \times T \times V}$. Then, the global average pooling is adopted to squeeze spatial-temporal information into a channel descriptor. The scale perception process can be described as follows.

$$Q=\sigma\left(\mathcal{F}_{fc}\left(\frac{1}{T \times V} \sum_{i=1}^{T} \sum_{j=1}^{V} \widehat{U}_{i,j}\right)\right) \tag{12}$$

where $\mathcal{F}_{fc}$ is the sequence of fully connected layers with reduction ratio r for dimension transformation and better efficiency. We employ the softmax function to obtain a soft attention vector $Q \in \mathbb{R}^{S \times C}$ that describes the multi-scale channel-wise importance, endowing the model with scale-aware capabilities.

Response delivery. The response delivery phrase mainly uses the above clues to adaptively fuse the information and pass outputs considered as responses to the next layer. Instead of concatenating the original results of branches, we perform weighted aggregation of them guided by scale-aware matrix Q to get final results $U^{+} \in \mathbb{R}^{C \times T \times V}$. Mathematically,

$$U^{+}=\sum_{s}^{S} U_s \otimes Q_s \tag{13}$$

where $Q_s \in \mathbb{R}^{S \times C}$ is the channel-wise importance for feature fusion. Due to the above efforts, Sa-TGC can adaptively recalibrate multi-scale feature responses by explicitly modeling inter-dependencies of branches and channels. With the tricks of group convolution, channel shuffle, and scale-aware mechanism, the fused features are more flexible and not confined to the fixed receptive field of convolutional filters. That is why Sa-TGC is competent for extracting discriminative features of continuous and terminating actions.

Note that the proposed PR-SGC and Sa-TGC are compatible with other graph convolutional networks, e.g., ST-GCN [5], 2S-AGCN [3], MS-G3D [11]. By replacing the spatial or temporal convolution block with PR-SGC and Sa-TGC, we can achieve better performance, as depicted in Table IV.

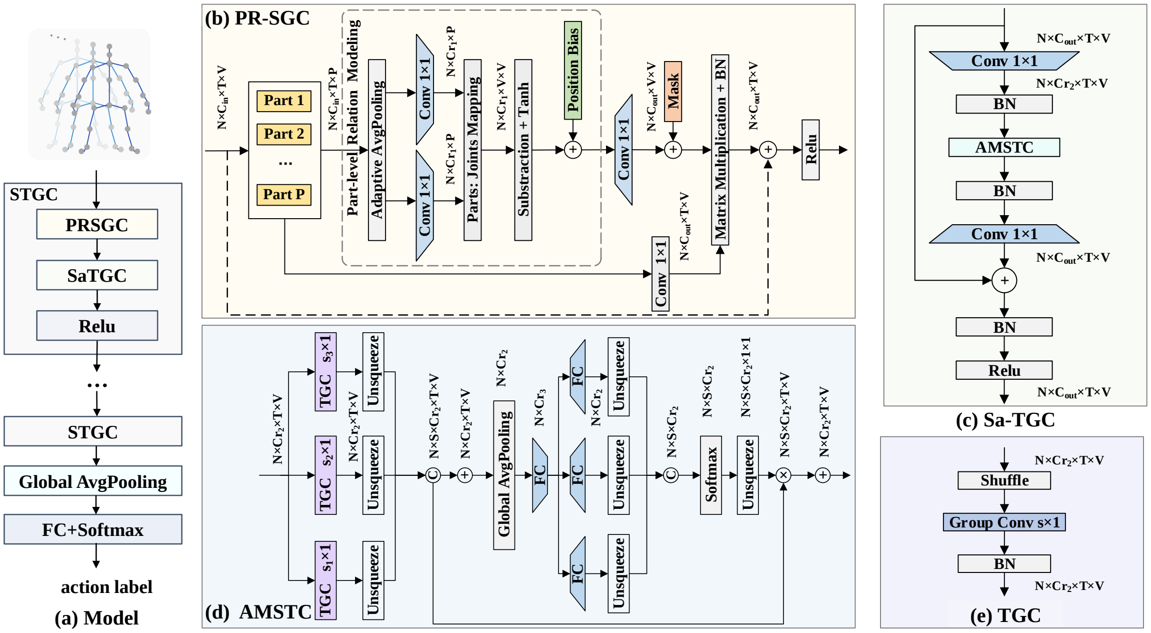


Fig. 6. The architecture of the proposed SaPR-GCN. The dotted arrow represents the residual connection. ⊗ is batch matrix multiplication. ⊕ is element-wise addition. ⓒ is concatenation operation. (a) The SaPR-GCN model that stacks ten basic STGC blocks for spatial-temporal modeling. The STGC block combines PR-SGC (b) and Sa-TGC (c). The Sa-TGC is symmetric and consists of AMSTC (d) and TGC (e). After global average pooling, the features extracted by the last STGC layer will be passed to the fully connected layer and softmax classifier. The dimension transformation of the input feature X has been marked. C_{in} is the input channel and C_{out} is the output channel.

E. Network Architecture

The network structure of the proposed SaPR-GCN is presented in Fig. 6. It stacks ten basic STGC blocks for spatial-temporal modeling as depicted in Fig. 6(a). Each block has two modules that complement and cooperate in endowing the network with both refined topology and scale awareness, followed by a global average pooling and a softmax classifier for action recognition.

Spatial modeling. We devise the PR-SGC module with a shortcut connection for spatial modeling as illustrated in Fig. 6(b). We divide each skeleton into eight body parts and apply relational modeling functions to obtain the dependencies between them. After the adaptive average pooling layer, the cascaded parts are put into two 1×1 convolutional layers with the same reduction rate r_1. r_1 is equal to 16. The part-level dependencies are permeated to the joint-level through part-joint mapping. Then, we acquire the part-wise refined adjacency matrix through a sequence of subtraction operation, tanh activation function, position bias addition, and 1×1 convolution. We introduce a learnable mask as the global shared topology and perform weight fusion with the normalized adjacency matrix. Afterward, the graph convolution is conducted according to Eq. (9). After the ReLU activation, the spatial representation is obtained.

Temporal modeling. As shown in Fig. 6(c), the Sa-TGC module employs a bottleneck structure followed by batch normalization and activation layer. Unlike the original bottleneck [22], we embed the AMSTC module between linear transformations, enabling the model to select the optimal receptive fields adaptively. As illustrated in Fig. 6(d), AMSTC contains three branches, each consisting of a temporal group convolution (TGC) block with different kernel sizes, 3, 7, and 11, respectively. At the beginning of TGC, we add a channel shuffle operation in Fig. 6(e). The results of each branch are added in the branch dimension by the unsqueeze operator and then perform cascade and sum operations. The global information embedding is obtained through an average pooling layer, and then a fully connected layer with the reduction of r_2 is built for compact representation where $r_2 = 2$. Then the results pass through different linear transformations, respectively. Afterword, a concatenation, softmax activation function, and unsqueeze operator are conducted sequentially for channel importance. Finally, the recalibrated multi-scale feature is obtained by weighted aggregating according to learned attention vectors.

Multi-stream fusion. Multi-stream fusion framework is widely utilized in [19], [25], [26]. Inspired by this, we adopt a four-stream framework where a separate model with identical architecture is trained using four input streams. They are the raw coordinates of joints, called "joint stream"; the difference between first-order adjacent joints, called "bone stream"; the difference of the above two streams between adjacent frames, called "joint motion stream" and "bone motion stream", respectively. Each model captures stream-specific motion patterns through stacked spatio-temporal graph convolutional layers. Finally, the softmax scores of multiple streams are ensembled as the final decision.

IV. Experiments

To demonstrate the advantages of SaPR-GCN, we compare our model with mainstream baselines on three public datasets and conduct ablation studies to analyze the components and parameter selections.

TABLE I
THE RESULTS OF PARTITION STRATEGIES

ID	Strategy	Part number	Para.(M)	Acc.(%)
(a)	up-down	2	1.80	90.30
(b)	middle-around	2	1.80	90.17
(c)	left-right-middle	3	1.87	90.25
(d)	up-down-middle	3	1.87	90.27
(e)	five-part	5	2.02	90.20
(f)	refined-part	8	2.07	90.43

[1] Each srategy is illustrated in Fig. 3 corresponding to its ID.

TABLE II
THE RESULTS OF TOPOLOGY REFINING MODULE

Model	$\mathcal{M}$	$Mask$	r_1	Acc.(%)	Para.	FLOPs
ST-GCN*		×		81.07	3.08M	3.49G
ST-GCN*		✓		82.32	3.08M	3.49G
2s-AGCN*		✓		87.21	3.45M	3.99G
SaPR-GCN	$\mathcal{M}_1$	×	16	84.27	2.07M	1.68G
SaPR-GCN	$\mathcal{M}_2$	×	16	89.57	2.07M	1.65G
SaPR-GCN	$\mathcal{M}_1$	✓	16	88.52	2.07M	1.68G
SaPR-GCN	$\mathcal{M}_2$	✓	16	90.43	2.07M	1.65G
SaPR-GCN	$\mathcal{M}_2$	✓	2	89.59	6.51M	2.02G
SaPR-GCN	$\mathcal{M}_2$	✓	8	89.79	2.71M	1.07G

[1] Those marked with * are methods we reproduced.

A. Datasets

NTU RGB+D 60. NTU RGB+D 60 [27] contains 56,880 action samples categorized into 60 classes. The actions are conducted by 40 volunteers. The skeleton sequences are captured by three Microsoft Kinect v2 cameras from three views. There are two popular benchmarks: (1) cross-subject (X-sub): training data comes from half of the subjects and testing data comes from the other. (2) cross-view (X-view): training set comes from camera IDs 2 and 3, and the testing set comes from camera ID 1.

NTU RGB+D 120. NTU RGB+D 120 [28] is an extension of NTU RGB+D 60, which has 120 action classes and 114,480 samples. The samples are collected in various locations and backgrounds denoted as 32 setups. In addition to the original cross-subject (X-sub), the cross-setup (X-set) evaluation is introduced, where the training set comes from samples with odd setup IDs, and the testing set comes from the rest.

NW-UCLA. NW-UCLA [29] is a multi-view dataset captured by three Kinect cameras at the same time. It contains 1494 video clips and covers 10 action labels. Each action is performed by 10 subjects. We follow the same evaluation protocol in [25]: the samples captured by the first two cameras are grouped as a training set, and the residual makes a testing set.

B. Implementation Details

All experiments are conducted on two RTX 3090 GPUs with the PyTorch deep learning framework. Our models are trained for 60 epochs by stochastic gradient descent (SGD) with a momentum of 0.9. We apply a warmup strategy in the first 5 epochs for training stability. The weight decay is 0.0004. For NTU RGB+D and NTU RGB+D 120, the batch size is 64, and we set the learning rate to 0.1 divided by 10 at epoch 40 and 50. For NW-UCLA, we set the batch size to 32, and the learning rate is ten times smaller at epoch 50. We adopt the data pre-processing following [19] on these datasets.

C. Ablation Studies

In this subsection, we examine the effect of the proposed components of SaPR-GCN. For fairness and brevity, all the ablation experiments are conducted on NTU RGB+D 60 dataset with the X-sub setup, and we only use bone stream as the input for all benchmarks. We discuss from two perspectives: one is the parameters selection, and the other is the comparison with alternative components.

Part partition strategies. Firstly, we discuss six partition strategies that refine the joint dependence and affect the feature aggregating process in SGC. The results of such methods are shown in Table I. The refined-part strategy, which divides the human body into eight parts, achieves optimal results. We find that more body parts do not always indicate better performance. As depicted in Table I, the Up-Down approach surpasses the five-part strategy and outperforms the middle-around method, which also separates the body into two parts. Moreover, the more parts reflect, the more parameters. Overall, it is essential for GCNs to employ appropriate partition strategies and embed part semantics for the refined skeleton structure.

Part-level refined topology module. Analogical actions tend to have similar parts interactions. Therefore, we refine the topology by introducing body parts and action semantics to capture higher-level motion patterns. The topology refining module is mainly affected by three factors: (1) The correlation modeling function $\mathcal{M}$ which determines the evaluation criteria for part-level dependencies. (2) The global shared topology ($Mask$) is complementary to the correlation modeling function to generate potential edges. (3) The feature reduction rate r_1 that controls the complexity of the model. As shown in Table II, the accuracy of $\mathcal{M}_2$ is higher than $\mathcal{M}_1$ regardless of whether the mask is added or not, and it has lower computational complexity when r_1 is the same. Besides, the mask brings a gain of 4.25% and 0.86% to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. It can be concluded that the global

TABLE III
THE RESULTS OF SCALE-AWARE MECHANISM

Model	$S(K)$	G	$\mathcal{S}$	R_2	Acc.(%)
ST-GCN*					82.32
2s-AGCN*					87.21
2s-AGCN	3(1, 3, 5)	8	✓	2	89.68
SaPR-GCN	3(1, 3, 5)	8	✓	16	89.05
SaPR-GCN	3(7, 9, 11)	8	✓	2	89.80
SaPR-GCN	3(1, 5, 9)	8	✓	2	90.02
SaPR-GCN	3(3, 7, 11)	8	✓	2	90.43
SaPR-GCN	3(3, 7, 11)	8	×	2	90.17
SaPR-GCN	4(3, 5, 7, 9)	8	✓	2	89.84
SaPR-GCN	5(3, 5, 7, 9, 11)	8	✓	2	90.28
SaPR-GCN	3(3, 7, 11)	1	✓	2	90.12
SaPR-GCN	3(3, 7, 11)	16	✓	2	89.96
SaPR-GCN	3(3, 7, 11)	32	✓	2	90.05

[1] Those marked with * are methods we reproduced.

TABLE IV
THE COMPARATIONS OF VARIOUS COMPONENTS

Model	Components	Acc.(%)
A	AGCN* (ASGC w/ TC)	87.21
B	AGCN* (ASGC w/ Sa-TGC)	90.13
C	2s-AAGCN*(AAGC w/ TC)	86.02
D	2s-AAGCN* (AAGC w/ Sa-TGC)	88.15
E	MS-G3D*(MS-GC w/ MS-TC)	88.64
F	MS-G3D* (MS-GC w/ Sa-TGC)	90.06
G	2s-AGCN* (PR-SGC w/ TC)	89.67
H	MS-G3D * (PR-SGC w/ MS-TC)	90.02
I	SaPR-GCN1(PR-SGC w/ Sa-TGC w/o S)	90.17
J	SaPR-GCN2(PR-SGC w/ Sa-TGC)	90.43

[1] Those marked with * are methods we reproduced.

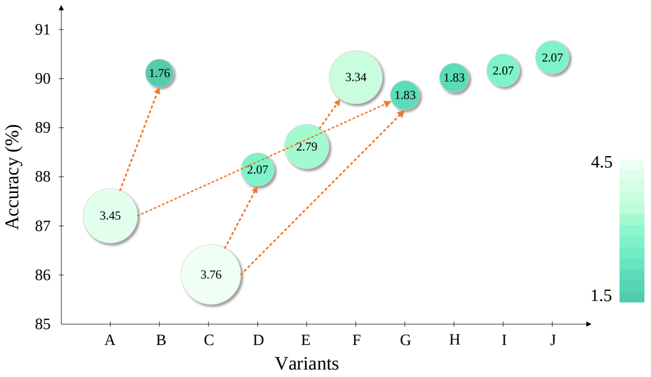


Fig. 7. The analysis of components. A-J corresponds to the models in Table III. We use circles of different sizes to indicate the amount of parameter (M), and the specific values are marked on them. The darker the circle, the smaller the value of FLOPs (G). The arrows indicate the gains achieved by replacing the corresponding components.

TABLE V
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE NW-UCLA DATASET. WE REPORT THE ENSEMBLED STREAMS, TOP-1 ACCURACY(%), FLOPS(G) AND PARAMETERS(M).

Methods	Publication	S	Acc.	FLOPs	Para.
HBRNN-L [21]	CVPR'15	1	78.5	-	-
TS-LSTM [30]	ICCV'17	4	89.2	-	-
AGC-LSTM [31]	CVPR'19	2	93.3	-	-
Shift-GCN [32]	CVPR'20	4	94.6	0.70*	1.28*
DC-GCN+ADG [33]	ECCV'20	4	95.3	3.56*	9.84*
ShiftGCN++ [25]	TIP'21	4	95.0	0.11*	0.44*
CTR-GCN [19]	ICCV'21	4	96.5	2.32*	5.68*
RGCA [34]	ICME'21	1	85.3	-	-
Graph2Net [6]	TCSVT'22	2	95.3	0.64*	1.62*
FGCN [35]	TIP'22	2	95.3	-	-
Ta-CNN [4]	AAAI'22	2	96.1	0.16	1.06
Ta-CNN+ [4]	AAAI'22	2	97.2	0.32	2.12
GAP [36]	ICCV'23	4	97.2	-	-
SaPR-GCN (ours)	-	4	96.6	1.31	2.06

[1] S is the number of ensembled streams.
[2] Those marked with * are the results we reproduced.

shared topology is essential for capturing the dependencies between joints, especially for weaker models. Experiments on ST-GCN [5] also verify this claim. In addition, we observe that r_1 greatly influences the model complexity and causes minor accuracy fluctuations. The performance is the best when $r_1 = 16$. Compared to ST-GCN [5] and 2s-AGCN [3] methods without utilizing parts semantics, our model achieves 8.11% and 1.97% improvement with less than 50% complexity, respectively. The experimental results demonstrate that the part-level refined topology module can effectively learn the intrinsic dependencies between joints, thus improving the overall accuracy.

Scale-aware mechanism. Scale-aware mechanism endows the model with dynamic multi-scale analysis for multi-range actions, which are mainly constrained by the following factors: (1) The number of branches and the convolution kernel size of each branch $S(K)$. S determines the scale diversity of the model, and K determines the granularity of multi-scale fused representations. (2) Temporal convolution. The parameters of temporal convolution include the number of groups (G) and the channel shuffling operation. G is a vital factor affecting the complexity of the model, and S is introduced for the communication between groups. (3) Reduction rate r_2 controls the squeezed rate of feature channels. We first briefly evaluate the parameter settings of r_2. The results are reported in Table III. We find that a lower compression rate can improve performance at an acceptable parameter cost. Therefore, we set r_2 equal to 2. Next, we explore the impact of kernel sizes. The results show that the larger the difference between convolution kernel sizes of multiple branches, the better the model's performance. The accuracy is optimal when $S(K)$ = 3(3, 7, 11). It is worth noting that increasing the number of branches can improve the accuracy but expand the complexity of the model. The five-branch model is only 0.2% better than the four-branch model. Thus, S and K are significant parameters worth tuning. Experiments show that using group convolutions can reduce the complexity but note that not the more groups, the better. After weighing, we set G to 8. We also study the effect of channel shuffling operations, and experiments show that it can raise the accuracy by 2.6% without additional parameters and FLOPs. The models combined with the scale-aware mechanism all outperform baselines, and the accuracy can be improved by at least 6.73% compared with ST-GCN [5]. Furthermore, the complexity of SaPR-GCN is also competitive. The above experiments demonstrate the effectiveness of the scale-aware mechanism, indicating that adaptive multi-scale feature fusion can significantly improve model performance.

PR-SGC & Sa-TGC vs. other components. We select the components AGC, AAGC, and TC modules in 2s-AGCN [3], and multi-scale MS-GC, MS-TC modules in MS-G3D [11] and then replace them with our proposed PR-SGC and Sa-TGC to generate variants, as shown in Table IV. We default to apply the optimal experimental settings described in Sec. IV.C. It can be seen that the PR-SGC embedded in the part-level refined module outperforms the SGC, AGC, and AAGC that only consider physical connections and importance masks. Moreover, the Sa-TGC with the scale-aware mechanism is more optimal than TC and MS-TC. Fig. 7 intuitively illustrates our findings. According to the results above, SaPR-GCN outperforms other derived models, demonstrating the superiority of PR-SGC and Sa-TGC.

TABLE VI
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE NTU RGB+D 60 & 120 DATASET. WE REPORT THE ENSEMBLED STREAMS, TOP-1(%) ACCURACY, FLOPS AND PARAMETERS.

Methods	Publication	Stream	NTU RGB+D 60		NTU RGB+D 120		FLOPs	Para.
			X-sub	X-view	X-sub	X-set		
ST-GCN [5]	AAAI'18	1	81.5	88.3	70.7	73.2	16.32G*	3.10M*
PB-GCN [16]	BMVC'18	1	87.5	93.2	-	-	-	-
AS-GCN [14]	CVPR'19	1	86.8	94.2	77.9	78.5	26.76G*	9.50M*
2s-AGCN [3]	CVPR'19	2	88.5	95.1	82.9	84.9	37.32G*	6.94M*
NAS-GCN [9]	AAAI'20	2	89.4	95.7	-	-	72.30G*	13.00M*
PL-GCN [8]	AAAI'20	1	89.2	95.2	-	-	-	20.70M
SGN [37]	CVPR'20	1	89.0	94.5	79.2	81.5	27.46G*	0.69M
PA-ResGCN-B19 [17]	ACM MM'20	1	90.9	96.0	87.3	88.3	18.52G	3.64M
Shift-GCN [32]	CVPR'20	4	89.7	96.0	85.9	87.6	10.00G	2.76M
MS-G3D [11]	CVPR'20	2	91.5	96.2	86.9	88.4	48.88G*	6.44M*
DDGCN [38]	ECCV'20	1	91.1	97.1	-	-	-	-
Dynamic GCN [26]	ACM MM'20	4	91.5	96.0	87.3	88.6	7.96G	14.40M
SEFN [18]	TCSVT'21	2	90.7	96.4	86.2	87.8	152.30G	34.70M
Js-CTR-GCN* [19]	ICCV'21	1	90.1	94.6	84.9	87.0	1.97G*	1.46M*
RA-GCN [13]	TCSVT'21	3	87.3	93.6	81.1	82.7	32.80G	6.21M
MST-GCN [10]	AAAI'21	4	91.5	96.6	87.5	88.8	26.60G*	11.68M*
ShiftGCN++ [25]	TIP'21	4	90.5	96.3	85.6	87.2	1.70G	1.80M
KA-AGTN [39]	KBS'22	2	90.4	96.1	86.1	88.0	-	5.4M
FGCN [35]	TIP'22	2	90.2	96.3	85.4	87.4	-	-
Graph2Net [6]	TCSVT'22	2	90.1	96.0	86.0	87.6	9.90G*	1.64M*
AimCLR [40]	AAAI'22	3	86.9	92.8	80.1	80.9	1.71G*	2.46M*
Ta-CNN [4]	AAAI'22	2	90.4	94.8	85.4	86.8	0.16G	1.06M
Ta-CNN+ [4]	AAAI'22	2	90.7	95.1	85.7	87.3	0.32G	2.12M
SMotif-GCN+TBs [41]	TPAMI'22	1	90.5	96.1	87.1	87.7	-	-
MS&TA-HGCN-FC [7]	TCSVT'23	2	90.8	96.4	87.0	88.4	-	-
EfficientGCN(B4) [42]	TPAMI'23	1	92.1	96.1	88.3	89.1	8.36G	1.10M
ActCLR [43]	CVPR'23	3	88.2	93.9	82.1	84.6	1.71G*	2.52M*
HiCLR [44]	AAAI'23	3	90.4	95.7	85.6	87.5	3.54G*	4.68M*
SkeAttnCLR [45]	IJCAI'23	3	89.4	94.5	83.4	92.7	10.44G*	9.24M*
GAP [36]	ICCV'23	4	92.9	97.0	89.9	91.1	-	-
RVTCLR+ [46]	ICCV'23	3	87.5	93.9	82.0	83.4	3.33G*	2.46M*
Js-SaPR-GCN (ours)	-	1	90.1	94.9	85.4	87.0	1.65G	2.07M
SaPR-GCN (ours)	-	4	92.4	96.4	88.7	90.3	6.60G	8.28M

[1] Those marked with * are the results from the corresponding methods we reproduced.

D. Comparison with the State-of-the-Art methods

We adopt the multi-stream fusion framework as [19], [25], [26], including four modalities, i.e., joint, bone, joint motion, and bone motion. We compare our model with the state-of-the-art methods on the NW-UCLA, NTU RGB+D 60, and NTU RGB+D 120 dataset, and the corresponding results are reported in Table V and Table VI. Js refers to the raw data, and Bs refers to the bone data. 2-stream(2s) denotes the ensemble of joint and bone, and 4-stream(4s) denotes the fusion of all streams in Sec. III.E. SaPR-GCN adopts the 4-stream ensemble schema by default. On the NW-UCLA dataset, SaPR-GCN is superior to most of the state-of-the-art models, especially those without graph convolutions (e.g., HBRNN-L [21] and TS-LSTM [30]). While the result is slightly lower than Ta-CNN+ [4], SaPR-GCN surpasses it in the other two datasets even without the extra data augment trick as shown in Table VI. For the NTU RGB+D 60 & 120 datasets, SaPR-GCN achieves more competitive performance than mainstream approaches, especially compared with part-based approaches [8], [16], [17], [46]. Although our accuracy was slightly lower in the X-view experimental setting than that in DDGCN [38], our accuracy was 1.3% higher in the X-sub setting. Compared to SkeAttnCLR [45], SaPR-GCN outperforms it in all but the X-set setting, with up to 5.3% over the x-sub protocol for the NTU RGB+D 120 dataset. Unfortunately, our approach still has certain limitations and is slightly inferior to GAP [36], which applies the large generative model GPT-3 offline action description generation. We believe it is meaningful to introduce large models to action recognition tasks and investigate foundation models. Last but not least, SaPR-GCN is energy-saving for storage and computation. This advantage is more noticeable compared to other GCN-based methods from the perspective of FLOPs. Overall, the results above demonstrate the superiority of our method SaPR-GCN.

V. DISCUSSION

We discuss the proposed blocks PR-SGC and Sa-TGC to further illustrate the interpretability [47] of SaPR-GCN and analyze its performance on analogical and deficient actions. All analyses are based on the model trained with the joint stream on NTU RGB+D 60 in X-sub setting as default.

A. Part activation maps and refined topology

To be more intuitive, we illustrate the interpretability of SaPR-GCN by showing the attention of SaPR-GCN through part activation maps which is the visualization of part attention. We take drinking water as an example, as depicted in Fig .8. The first row is the corresponding skeleton sequence of

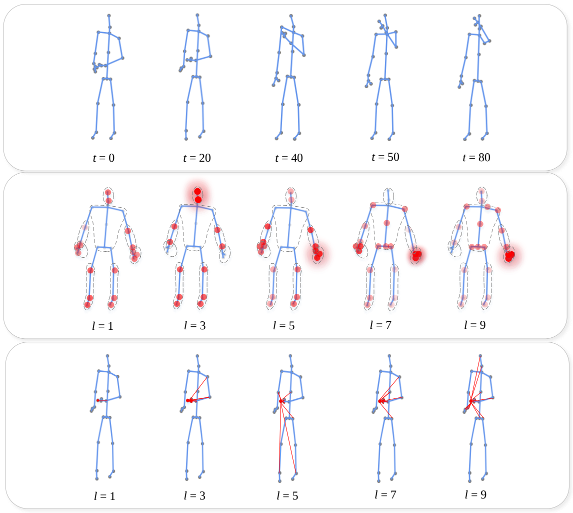


Fig. 8. Part activation maps and refined topology of drinking water. The first line shows the sampled skeleton sequence of "drink water". The second line is the part activation map obtained by accumulating activation joints in each part. The activation joints labeled in solid red circles are derived by CAM. The deeper the color, the higher the activation degree. The third line is the intrinsic connections to other joints of the left-hand tip (>0.04).

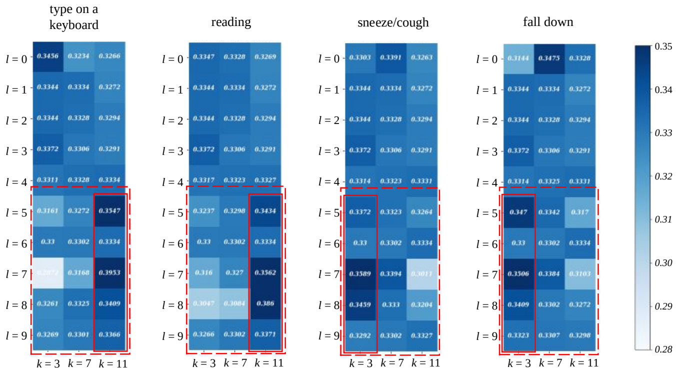


Fig. 9. Scale attention map of different actions. The scale attention map can reflect the concerned scale in specific layers. We show the results of two continuous actions "type on a keyboard" and "reading" and two terminating actions "sneeze/cough" and "fall down". k is the kernel size of the temporal convolution in each branch.

drinking water (mirrored). We can see that the right-hand part is the vital part of "drink water", which is more discriminative for identifying the action. The second line is the part activation map obtained by accumulating activation joints in each part. We leverage class activation mapping [20] to acquire activation joints. In order to picture the activated parts more clearly, we show the frontal human skeleton to avoid overlapping joints, as presented in the second row. We observe that the shallow network tends to capture the features of each part uniformly. In the third layer of the network, the model shifts its attention to the upper body and focuses on head changes. As the number of layers continues to increase, the model pays more attention to the movement of the hand. Moreover, it is worth noting that the last layer can clearly distinguish the importance of the left and right hands, which indicates that our model has good interpretability and can capture the most critical hand according to the action semantics. In addition, we analyze the topology after part refinement. We select the right-hand tip from the activation parts as the analysis object. The third row of Fig. 8 shows its intrinsic connections to other joints. We only show edges with weights above the threshold of 0.04. The thickness of the lines indicates the strength of connectivity. For the action of drinking water, we find that the model can pay more attention to the relationship between the left hand and the upper body, which is in line with human cognition. We analyze the topology of various actions and conclude that the information transfer between joints is asymmetric. The self-connection is weakened by the attention mechanism, which coincides with [11]. Additionally, we observe that shallow layers tend to capture close-range neighbors. In contrast, deep layers can expand the radiation range of joints, which explains the phenomenon that the deeper the network is, the higher the recognition accuracy. It can be seen that our model can focus on the most active parts and extract more discriminative features through the refined topology, thereby improving the performance of the model.

B. Context-dependent multi-scale feature extractor

To illustrate that our proposed multi-scale temporal convolution module Sa-TGC can dynamically adjust the weights of each branch according to the behavior context, we visualize the branch importance vectors inferred by the model as scale attention maps, as shown in Fig. 9. We can see that each action has its unique attention map. Moreover, each layer has unequal scale weights for a specific action. The inferred scale attention map is diverse, even for similar actions such as typing on a keyboard and reading. In addition, we find that Sa-TGC tends to use larger convolution kernels for continuous actions with subtle changes, such as typing on a keyboard and reading. In contrast, for the terminating actions of sneezing/coughing and falling down, Sa-TGC prefers to use a smaller convolution kernel. The above analyses indicate that the model has adaptive scale awareness, and different layers exhibit different scale diversity, especially the last five layers. It is the fundamental reason for the higher accuracy of SaPR-GCN on deficient actions.

C. Performance on analogical and deficient actions

Existing methods still have challenges recognizing analogical and deficient actions. Specifically, analogical actions can be further classified as (1) actions involving interactions between the same parts, e.g., "drink water" and "eat a meal" (2) actions covering interactions between similar parts, e.g., "neck pain" and "headache", "sneeze/cough" and "nausea/vomiting" (3) actions with reverse temporal dynamics, e.g., "put on a shoe" and "take off the shoe". In addition, deficient actions can be summarized as (1) continuous action with subtle movements, e.g., "reading", "writing" and "play with phone". (2) terminating actions with short durations, e.g., "sneeze", "nausea". We further compare the performance of SaPR-GCN and its variants on specific actions. We select twelve categories of actions from the NTU RGB+D 60 dataset and further divide them into six pairs according to the characteristics of actions,

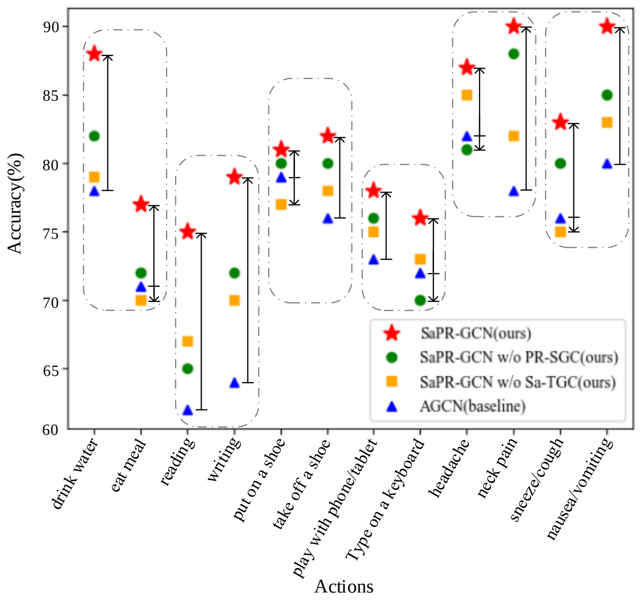


Fig. 10. The effectiveness of PR-SGC and Sa-TGC module in recognizing analogical and deficient actions. We select twelve analogical and deficient actions as test samples and divide them into six pairs marked in dotted rectangles. The figure mirrors the results of 2s-AGCN and our method, i.e., SaPR-GCN and its variants.

as shown in Fig. 10. For clarity, we have added double-headed arrows as auxiliary lines to the boundaries of the baseline method. It can be seen that SaPR-GCN outperforms the other two variants and outperforms the baseline method marked by blue triangles. This further demonstrates the effectiveness of the proposed modules PR-SGC and Sa-TGC.

Moreover, we compare SaPR-GCN with other approaches, including AS-GCN [14], AGCN [3], CTR-GCN [19], MS-G3D [11], and Ta-CNN [4]. We reproduce the above methods on the NTU RGB+D 60 dataset under the X-sub setting and show the average recognition accuracy for specific actions in Figure 11. Since Ta-CNN [4] only uses joint data, we train these models by the joint data for fairness. It can be seen that our method is superior to the existing methods on the whole. It is worth noting that compared with AS-GCN [14], which introduces the global action semantic structure, our part-level refined topology has apparent advantages in recognizing subtle actions such as writing and reading, bringing up to 15% accuracy gain. The possible reason is that these actions are coupled with two hands, which asks the model to capture more fine-grained dynamics of hands. Besides, our method is more robust than MS-G3D [11] and CTR-GCN [19], which employ static multi-scale mechanisms. In summary, our method achieves superior performance by embedding part semantics and exploiting adaptive multi-scale analysis. However, the average recognition accuracy of the above methods for recognizing challenging samples is lower than 90%. In future work, we will strive to improve the accuracy of analogical and deficient actions and other challenging samples.

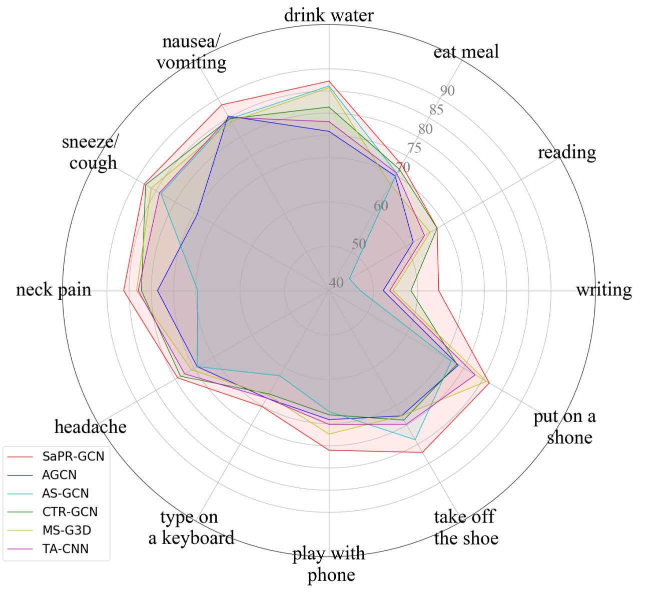


Fig. 11. Comparison with competitive methods on analogical and deficient actions. The above methods are trained by the joint data in NTU-RGB+D 60 dataset on the X-sub setting. Different colored lines indicate the accuracy of different methods. Among them, the red-filled polygon area is the largest, which proves that our method is more advantageous in identifying analogical and deficient actions.

VI. Conclusion

In this work, we propose a practical learning framework SaPR-GCN for skeleton-based human action recognition. We introduce two portable modules, PR-SGC and Sa-TGC, to obtain an effective spatial-temporal motion representation, especially for analogical and deficient actions. Specifically, PR-SGC embeds body parts semantics to refine topology, and Sa-TGC leverages scale perception mechanism to acquire context-dependent multi-scale features adaptively. Taking advantage of these modules, SaPR-GCN can dynamically extract more discriminative features and thus achieve superior performance on three public datasets. We will focus on fine-grained actions with subtle hand movements in future work.

References

[1] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, pp. 147–170, 2019.

[2] H. Sang and Q. Tian, "Rapid action recognition system for human-computer interaction," *Comput. Eng. and Appl.*, vol. 55, no. 6, pp. 101–107,167, 2019.

[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12 018–12 027.

[4] K. Xu, F. Ye, Q. Zhong, and D. Xie, "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 3, 2022, pp. 2866–2874.

[5] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 32, no. 1, 2018, pp. 7444–7452.

[6] C. Wu, X. Wu, and J. Kittler, "Graph2net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2120–2132, 2022.

[7] 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 Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1868–1883, 2023.

[8] L. Huang, Y. Huang, W. Ouyang, L. Wang, and I. Assoc Advancement Artificial, "Part-level graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 7, 2020, pp. 11 045–11 052.

[9] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 3, 2020, pp. 2669–2676.

[10] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 2, 2021, pp. 1113–1122.

[11] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 140–149.

[12] L. Shi, Y. Zhang, J. Cheng, H. Lu, and Soc, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 7904–7913.

[13] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1915–1925, 2021.

[14] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3590–3598.

[15] Z. Li, Y. Li, L. Tang, T. Zhang, and J. Su, "Two-person graph convolutional network for skeleton-based human interaction recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3333–3342, 2023.

[16] K. C. T. andP. J. Narayanan, "Part-based graph convolutional network for action recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 270–283. [Online]. Available: http://bmvc2018.org/contents/papers/1003.pdf

[17] Y. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020, pp. 1625–1633.

[18] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4394–4408, 2021.

[19] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 13 339–13 348.

[20] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020, p. 1625–1633.

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

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[23] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[24] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. Springer Eur. Conf. Comput. Vis. (ECCV)*, vol. 11218, 2018, p. 122–138.

[25] K. Cheng, Y. Zhang, X. He, J. Cheng, and H. Lu, "Extremely lightweight skeleton-based action recognition with shiftgcn plus," *IEEE Trans. Image Process.*, vol. 30, pp. 7333–7348, 2021.

[26] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020.

[27] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1010–1019.

[28] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, 2020.

[29] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 2649–2656.

[30] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1012–1020.

[31] C. Si, W. Chen, W. Wang, L. Wang, T. Tan, and Soc, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1227–1236.

[32] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 180–189.

[33] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *Proc. Springer Eur. Conf. Comput. Vis. (ECCV)*, vol. 12369, 2020, p. 536–553.

[34] H. Yao, S.-J. Zhao, C. Xie, K. Ye, and S. Liang, "Recurrent graph convolutional autoencoder for unsupervised skeleton-based action recognition," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[35] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank, "Feedback graph convolutional network for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 164–175, 2022.

[36] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Generative action description prompts for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vision(ICCV)*, October 2023, pp. 10 276–10 285.

[37] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1109–1118.

[38] M. Korban and X. Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in *Proc. Springer Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[39] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowledge-Based Syst.*, vol. 240, p. 108146, 2022.

[40] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 1, 2022, pp. 762–770.

[41] Y. Wen, L. Gao, H. Fu, F. Zhang, S. Xia, and Y.-J. Liu, "Motif-gcns with local and non-local temporal blocks for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2022.

[42] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1474–1488, 2023.

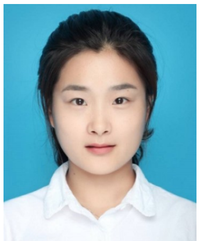
[43] L. Lin, J. Zhang, and J. Liu, "Actionlet-Dependent Contrastive Learning for Unsupervised Skeleton-Based Action Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 2363–2372.

[44] J. Zhang, L. Lin, and J. Liu, "Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 3, Jun. 2023, pp. 3427–3435. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25451

[45] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part Aware Contrastive Learning for Self-Supervised Action Recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*. Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 855–863. [Online]. Available: https://www.ijcai.org/proceedings/2023/95

[46] Y. Zhu, H. Han, Z. Yu, and G. Liu, "Modeling the relative visual tempo for self-supervised skeleton-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vision(ICCV)*, October 2023, pp. 13 913–13 922.

[47] J. Fu, J. Gao, and C. Xu, "Learning semantic-aware spatial-temporal attention for interpretable action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5213–5224, 2022.



Chang Li received the B.Sc. degree from Jiangsu Ocean University, Lianyungang, China, in 2019. She received the M.Sc. degrees in computer science and technology from Hohai University, Nanjing, China, in 2021. She is currently pursuing Ph.D. degree in Hohai University, Nanjing, China. Her research interests include media computing, deep learning, distributed computing, smart sensor device, and computer vision especially action recognition and pose estimation.

Yingchi Mao was born in China. She received the B.Sc. and M.Sc. degrees in computer science and technology from Hohai University, Nanjing, China, in 1999 and 2003, respectively, and the Ph.D. degree in computer science and technology from Nanjing University, China, in 2007. She is currently a Professor with the College of Computer and Information, Hohai University. Her research interests include distributed computing, wireless sensor networks, and distributed data management.

Qian Huang received the B. Sc. degree in computer science from Nanjing University, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2010. From 2010 to 2012, he was a deputy technical manager of Mediatek (Beijing) Incorporation, Beijing, China. Since Dec. 2012, he serves as the dean of Computer Science and Technology Department, Hohai University, Nanjing, China. His research interests include media computing, data mining, and intelligent education.

Xiaowei Zhu received the B.Sc. degree from China Jiliang University, Hangzhou, China, in 2021. He is currently pursuing M.Sc. degree in computer science and technology from Hohai University, Nanjing, China. His research interests include deep learning, action recognition and pose estimation.

Jie Wu is the Director of the Center for Networked Computing and Laura H. Carnell professor at Temple University. He also serves as the Director of International Affairs at College of Science and Technology. He served as Chair of Department of Computer and Information Sciences from the summer of 2009 to the summer of 2016 and Associate Vice Provost for International Affairs from the fall of 2015 to the summer of 2017. Prior to joining Temple University, he was a program director at the National Science Foundation and was a distinguished professor at Florida Atlantic University. His current research interests include mobile computing and wireless networks, routing protocols, network trust and security, distributed algorithms, and cloud computing. Dr. Wu regularly publishes in scholarly journals, conference proceedings, and books. He serves on several editorial boards, including IEEE Transactions on Mobile Computing, IEEE Transactions on Service Computing, Journal of Parallel and Distributed Computing, and Journal of Computer Science and Technology. Dr. Wu is/was general chair/co-chair for IEEE IPDPS'08, IEEE DCOSS'09, IEEE ICDCS'13, ACM MobiHoc'14, ICPP'16, IEEE CNS'16, WiOpt'21, and ICDCN'22 as well as program chair/cochair for IEEE MASS'04, IEEE INFOCOM'11, CCF CNCC'13, and ICCCN'20. He was an IEEE Computer Society Distinguished Visitor, ACM Distinguished Speaker, and chair for the IEEE Technical Committee on Distributed Processing (TCDP). Dr. Wu is a Fellow of the AAAS and IEEE. He is the recipient of the 2011 China Computer Federation (CCF) Overseas Outstanding Achievement Award.