Machine Vision and Applications (2023) 34:40
https://doi.org/10.1007/s00138-023-01386-2

ORIGINAL PAPER

ICE-GCN: An interactional channel excitation-enhanced graph convolutional network for skeleton-based action recognition

Shuxi Wang[1] · Jiahui Pan[1,3] · Binyuan Huang[1] · Pingzhi Liu[1] · Zina Li[2] · Chengju Zhou[1]

Received: 3 October 2022 / Revised: 11 February 2023 / Accepted: 20 February 2023

Abstract
Thanks to the development of depth sensors and pose estimation algorithms, skeleton-based action recognition has become prevalent in the computer vision community. Most of the existing works are based on spatio-temporal graph convolutional network frameworks, which learn and treat all spatial or temporal features equally, ignoring the interaction with channel dimension to explore different contributions of different spatio-temporal patterns along the channel direction and thus losing the ability to distinguish confusing actions with subtle differences. In this paper, an interactional channel excitation (ICE) module is proposed to explore discriminative spatio-temporal features of actions by adaptively recalibrating channel-wise pattern maps. More specifically, a channel-wise spatial excitation (CSE) is incorporated to capture the crucial body global structure patterns to excite the spatial-sensitive channels. A channel-wise temporal excitation (CTE) is designed to learn temporal inter-frame dynamics information to excite the temporal-sensitive channels. ICE enhances different backbones as a plug-and-play module. Furthermore, we systematically investigate the strategies of graph topology and argue that complementary information is necessary for sophisticated action description. Finally, together equipped with ICE, an interactional channel excited graph convolutional network with complementary topology (ICE-GCN) is proposed and evaluated on three large-scale datasets, NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton. Extensive experimental results and ablation studies demonstrate that our method outperforms other SOTAs and proves the effectiveness of individual sub-modules. The code will be published at https://github.com/shuxiwang/ICE-GCN.

Keywords Skeleton-based action recognition · Graph convolutional network · Channel-wise attention · Cross-dimensional interaction

1 Introduction

Human action recognition has attracted more and more attention in the area of computer vision and finds its various applications in human–machine interaction, video surveillance, virtual reality, and so on [1–4]. Recently, with the emergence of high-precision depth sensors such as Microsoft Kinect [5] and advanced human pose estimation algorithms [6–8], the skeleton coordinates can be obtained accurately and economically. With its robustness to variations in body size, viewpoints, and complicated backgrounds, as well as efficiency in storage and computational cost, skeleton data have become the mainstream input compared with other modalities, such as traditional RGB videos.

The early-stage deep learning-based approaches directly treated human joint coordinates as sequences of coordinate vectors [9–12] or pseudo-images [13–15] and fed them into convolutional neural networks (CNNs) or recurrent neural

✉ Jiahui Pan
panjiahui@m.scnu.edu.cn

✉ Chengju Zhou
cjzhou@scnu.edu.cn

Shuxi Wang
wangshuxi@m.scnu.edu.cn

Binyuan Huang
binyuanhuang@163.com

Pingzhi Liu
liupingzhi@m.scnu.edu.cn

Zina Li
zina@m.scnu.edu.cn

1 School of Software, South China Normal University, Foshan 528225, China

2 School of Psychology, South China Normal University, Guangzhou 510631, China

3 Pazhou Lab, Guangzhou 510330, China

Published online: 05 April 2023

networks (RNNs). Such representations overlook the intrinsic graph structure relationship among joints, which is crucial for recognizing human action. To solve this issue, recently, Yan et al. [16] proposed a spatio-temporal graph convolutional network (ST-GCN) to model the skeleton data as the graph structure, which represents the joints as graph nodes and the joint connectives as graph edges. In the spatial dimension, joint topology is defined by a sequence of adjacency matrices, and then, a graph convolutional network (GCN) is utilized to capture the joint spatial relationship for each frame. In the temporal dimension, temporal convolution (TCN) is applied to capture the inter-frame relationship for each node. ST-GCN is the first and classical network that introduced GCN to the task of skeleton-based action recognition, which was followed by many improvements and variants [17–24].

To enable the networks to capture various ranges of dependencies and enhance the most discriminative joints in intra-frame space, the spatial attention mechanisms [18,21,22,25] are applied to generate spatial attention maps for each joint. Based on similar considerations, the temporal attention mechanisms [18,22–24] are applied to generate temporal attention maps for each frame. However, these previous joint and frame attention methods have treated feature patterns in different channels equally without considering how to select the more informative and channel-wise features. This limited the representation capability and was not optimal for obtaining the discriminative features.

Since different channels indicate different motion features [20], the importance among joints varies with different motion features. Therefore, exploring various importance of the motion features in different channels can emphasize the informative spatio-temporal feature patterns, which can help the network distinguish confusing actions. Inspired by SENet [26], which is the first to introduce a simple but effective channel attention module for image classification tasks, works [27,28] applied the channel attention to calculate channel-wise modulation weights. But these attention schemes only consider inter-channel information without introducing information from other dimensions. To address this problem, CBAM [29] was proposed to combine channel attention and spatial attention sequentially. And based on this idea, works [18,22,25,27] took spatial and temporal information into account, but these methods treated each single dimension independently and then combined them in a sequential manner; the other dimensions would be globally averaged into a single scalar. Intuitively, the channel and spatio-temporal information is highly related to each other, i.e., feature patterns in each channel are explored from spatio-temporal space. Thus, separately considering channel and spatio-temporal aspects is sub-optimal for exploring finer levels of discriminative joints among intra- and inter-frame.

To address this issue, inspired by works [29–34], an interactional channel excitation (ICE) module is proposed to incorporate both spatial and temporal information into the channel attention with cross-dimensional interactions. ICE is composed of channel-wise spatial excitation (CSE) and channel-wise temporal excitation (CTE) sub-modules. CSE is applied to capture the crucial body global structure patterns to excite the spatial-sensitive channels. CTE is applied to vital temporal dynamics information to excite the temporal-sensitive channels.

Moreover, we also systematically investigate the strategies of graph topology, which is also essential in determining the representation ability of joint relationships in GCN. The topology is represented by the adjacent matrix. Various adjacency matrix schemes are employed to construct graph topology by previous works. They can be mainly summarized into three categories: A_p (physical) means the fixed predefined matrix, which reflects the body natural physical structure [16,35,36]. A_l (learnable) is the learnable matrix, which is parameterized and optimized throughout training [17,37,38]. A_s (similarity) represents the Gaussian similarity matrix, which is used to measure the similarity of pairs of vertexes [17,20]. Based on the experimental observation, we argue that complementary topology is necessary, which can achieve a good balance between adaptation and too large of a searching space.

Finally, together equipped with ICE, an interactional channel excited graph convolutional network with complementary topology (ICE-GCN) is proposed. Extensive experiments and ablation studies demonstrate the necessity of the ICE module and the complementary topology scheme. Compared with previous works, the main contributions of our work can be summarized as follows:

- Compared with the existing attention mechanisms, which ignore the cross-dimensional interaction, our interactional channel excitation (ICE) module embeds spatio-temporal information into channel attention, which allows to explore discriminative spatio-temporal features of actions in a finer channel level, adaptively recalibrating spatial-temporal-aware attention maps along channel dimension. ICE, composed of a channel-wise temporal excitation (CTE) and a channel-wise spatial excitation (CSE), can be inserted into any existing graph convolutional networks as a plug-and-play module to enhance the performances notably without light computational cost.
- We systematically investigate the strategies of graphs and argue that complementary topology is necessary. Three adjacency sub-matrices A_p, A_l and A_s are combined to construct the graph topology. This simple but efficient scheme notably improves the performance, which solves the dilemma between adaptation and too large of a searching space.

- Finally, together equipped with ICE, an interactional channel excited graph convolutional network with complementary topology (ICE-GCN) is proposed. Extensive experiments conducted on three large-scale datasets, NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton, demonstrate our ICE-GCN outperforms the state-of-the-art performance. The follow-up ablation experiments and visualization also show the effectiveness of the individual modules in graph convolutional networks.

2 Related works

2.1 Traditional attention mechanisms for skeleton-based action recognition

To model various scales of dependencies and help the network focus on the most informative information, attention modules are integrated into the graph convolutional networks. Work [18] designed a channel attention module based on SENet [26] and generated attentive maps to reweight the channel dimension, which is averaged over all features of the spatial joints and temporal frames. Similarly, work [21] proposed spatial joint attention to measure the importance of each joint. And works [23,24] designed temporal frame attention to enhance the modeling capability of temporal dependencies. These attention mechanisms consider each dimension independently, with all other dimensions being globally averaged. As spatial, temporal and channel dimensions contain the complementary and correlated information for action recognition. More previous works [18,22,25,27,39–42] inspired by the scheme based on CBAM [29] fused single-dimensional attention modules sequentially, such as work [18] fused spatial, temporal, and channel attention modules to construct STC-attention module in a sequential manner.

Both spatial and temporal single-dimensional attention methods ignore the difference in the contribution of different spatio-temporal patterns along different channels. And the channel attention methods based on SENet [26] squeezed global spatio-temporal information into a unit without considering spatial or temporal joint correlations. The methods based on CBAM [29] simply fused channel, spatial, and temporal attention in a sequential manner without cross-dimensional interaction, which is essential to generate channel-wise spatio-temporal selective attention maps.

Thus, there existed some works in the computer vision field adopted cross-dimensional schemes. Coordinate attention [34] embedded positional as well as spatial information into channel attention along the horizontal and vertical directions, which is critical to detecting object structures. [31–33, 43] introduced temporal information into channel attention for video-based action recognition tasks with spatio-temporal data. In more detail, in TEA [32], a motion excitation was proposed to embed temporal dynamic motion patterns that describe the temporal difference between the two adjacent frames into channel attention and then excite these motion-sensitive channels. ACTION-Net [33] inserted one more convolutional layer between two fully connected layers for channel-wise features within temporal information.

Inspired by the considerations mentioned above, we propose our innovative method interactional channel excitation (ICE) module. The difference of our interactional channel excitation (ICE) module is that it is channel-wise and introduces both body global structure patterns and temporal inter-frame dynamics information to channel attention by cross-dimensional interaction. Our ICE is applied for the task of skeleton-based action recognition and focuses on capturing the features of the joint correlation of graphs, which is different from the image and video-based tasks.

2.2 Strategies of graph topology for GCN

Graph topology construction plays a key role in determining the representation ability of joint relationships in GCN. ST-GCN [16] proposed the predefined adjacency matrix A_p, which is based on the body physical structure and manually builds three topologies using three partitioning strategies. Shi et al. [17] proposed a data-driven model called AGCN. This work introduced a learnable adjacency matrix A_l, which is capable of adaptively learning the topologies of the graph, and introduced another Gaussian similarity matrix A_s to measure the similarity of the pairs of vertexes in an embedding space by dot product. Based on AGCN, Chen et al. [20] considered that different channels reflect different types of features, and it is not desirable to only use a single shared topology for all channels. Therefore, they proposed channel-specific A_s (denoted as CA_s) for each channel to calculate the pairs of vertex distances in an embedding space using pairwise subtraction.

Although A_l, A_s and CA_s are more adaptive to capture global graph information, they face too large of a searching space and learn too many "noisy" edges [36]. In this work, to address the dilemma between adaptation and a too large searching space, we systematically investigate the strategies of graphs and argue that complementary topology is necessary.

3 The proposed method

3.1 Interactional channel dimension excitation (ICE)

To solve the problem that features of joint correlation modeling ignores the interaction between spatial-temporal

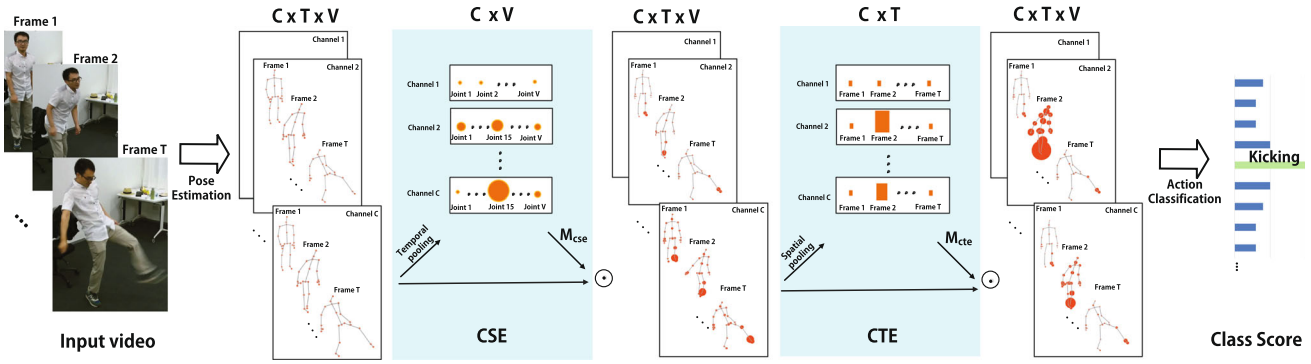


Fig. 1 Schematic diagram of the ICE on the skeleton sequence of action "kicking something," with C, T, and V representing the number of channels, frames, and joints, respectively. The solid orange circles represent the importance of the corresponding joints. The left ankle joint (i.e., joint 15) is emphasized in the "kicking" action. The orange rectangles represent the importance of the corresponding frames. M_{cse} and M_{cte} are attention masks. $\odot$ denotes the element-wise multiplication

dimensions and channel dimension, inspired by previous excitation works [29,31–34], an interactional channel excitation (ICE) module is proposed to capture channel-wise patterns and embed spatio-temporal information into channel attention by cross-dimensional interactions. A schematic diagram of the processing of the ICE on skeleton sequence with the action "kicking something" is shown in Fig. 1. The ICE module consists of two sub-modules, spatial channel excitation (CSE) and temporal channel excitation (CTE), which are described in detail in Sects. 3.1.1 and 3.1.2, respectively. In addition, to more clearly illustrate the innovations of ICE, four schematic diagrams shown in Fig. 2 are to compare our proposed CSE, CTE with classical inter-channel attention mechanism SENet [26] and sequential multi-dimensional attention CBAM [29].

3.1.1 Channel-wise spatial excitation (CSE)

CSE is applied to capture the crucial body global structure patterns to excite the spatial-sensitive channels, which adaptively recalibrates the importance of joints along different channels. The architecture of the CSE module is illustrated in Fig. 2 c.

Given an input feature $X \in \mathbb{R}^{C\times T\times V}$, the average pooling is applied to summarize the temporal information for CSE and focuses on the interaction between channel and spatial dimensions. It also helps to reduce the computational cost in this way.

$$X_{\text{t\_pool}} = \frac{1}{T}\sum_{j=1}^{T} X[:, j, :],\ X_{\text{t\_pool}} \in \mathbb{R}^{C\times V} \tag{1}$$

where $X_{\text{t pool}}$ denotes the feature after temporal pooling and T is the number of frames.

Second, a 1D convolution layer conv_{spa} with the kernel size K set to V is applied to enable CSE to have a global receptive field covering all joints in a frame and facilitating the extraction of global structural features, which is ignored in previous spatial attention works for they considered joints independently. conv_{spa} also reduces the number of channels to alleviate computational and model complexity at the same time.

$$X_{\text{spa}} = \text{conv}_{\text{spa}} * X_{\text{t\_pool}},\ X_{\text{spa}} \in \mathbb{R}^{C/r\times V} \tag{2}$$

where X_{spa} denotes the global structure feature in the spatial dimension and its channel-reduced, r is the scale ratio (set to 16 in this work), and $*$ indicates the convolution operation.

Third, after feeding X_{spa} to ReLU for nonlinearity, another 1D convolution layer conv_{exp} with kernel size set to 1 is applied to expand the channel dimension back to the original channel dimension. Then, the tensor X is reshaped as $[C, 1, V]$ and fed into a Sigmoid activation to obtain the spatial-attentive mask M_{cse}.

$$\begin{aligned} M_{\text{cse}} &= \text{Sigmoid}(\text{conv}_{\text{exp}} * \text{ReLU}(X_{\text{spa}})) \\ M_{\text{cse}} &\in \mathbb{R}^{C\times 1\times V} \end{aligned} \tag{3}$$

Finally, the spatial-sensitive channels and crucial joints are excited by multiplication between the input features X and M_{cse} along the channel dimension. Furthermore, a residual connection is applied to preserve the original information and ensure network stability.

$$F_{\text{cse}} = X * M_{\text{cse}} + X,\ F_{\text{cse}} \in \mathbb{R}^{C\times T\times V} \tag{4}$$

Therefore, by interacting the spatial dimension with the channel dimension, CSE excites the beneficial spatial-sensitive channels and adaptively recalibrates the importance of joints simultaneously. Finally, we obtain the excited output

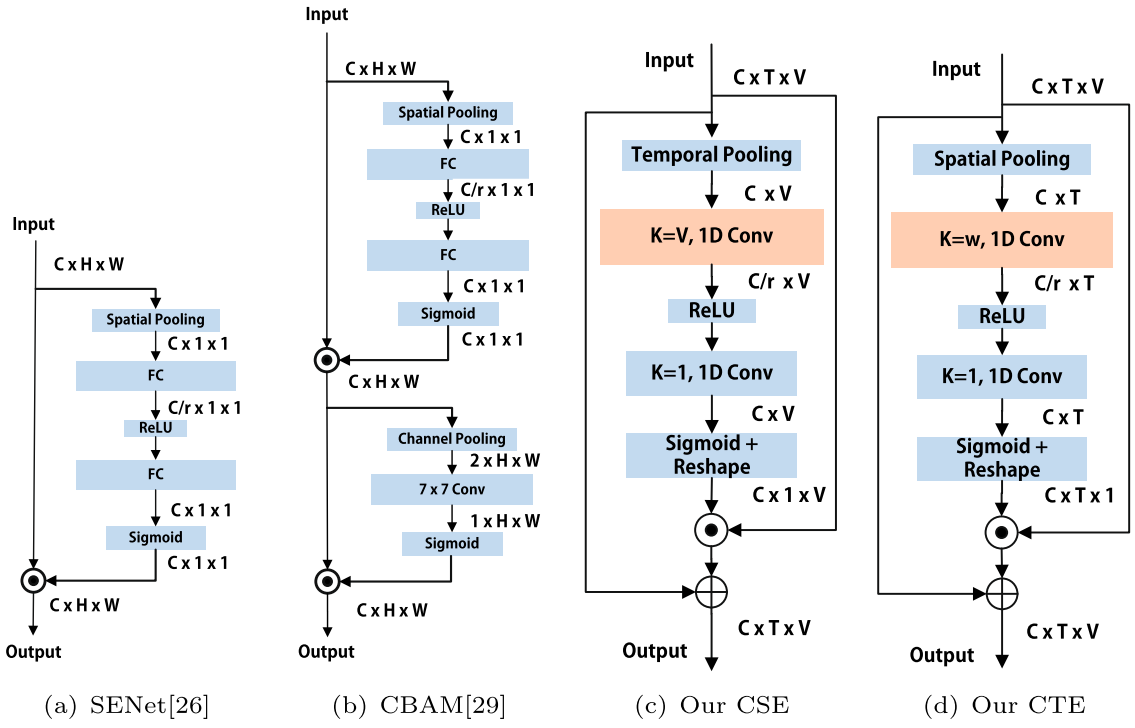


Fig. 2 Schematic comparison between the proposed CSE (**c**), CTE (**d**) and the classical attention module of SENet (**a**), CBAM (**b**). C, T, and V denote the number of channels, frames, and joints, respectively. FC denotes a fully connected layer. r is the reduction ratio, and K means kernel size. w denotes the size of sliding temporal window; $\odot$ indicates the element-wise multiplication. $\oplus$ denotes the element-wise summation

feature F_{cse} and input the following channel-wise temporal excitation (CTE) sub-module.

3.1.2 Channel-wise temporal excitation (CTE)

Like CSE, CTE aims to utilize temporal dynamics information to discriminate and excite the vital temporal-sensitive channels and frames. The architecture of the CTE module is illustrated in Fig. 2d.

Given an input feature as $X \in \mathbb{R}^{C\times T\times V}$, the average pooling is applied to summarize the spatial information.

$$X_{s\_pool} = \frac{1}{V}\sum_{i=1}^{V} X[:,:,i],\ X_{s\_pool} \in \mathbb{R}^{C\times T} \tag{5}$$

where $X_{\text{s\_pool}}$ denotes the feature after spatial pooling.

Different from CSE, a 1D convolution layer $conv_{\text{tmp}}$ with the kernel size K set to w is applied to interact with the temporal dimension based on a sliding temporal window to capture the inter-frame temporal relationship of w frames. We set w as a hyperparameter (3, 5, etc.) according to the frames of different datasets to obtain the most appropriate temporal receptive field.

$$X_{\text{tmp}} = \text{conv}_{\text{tmp}} * X_{\text{s\_pool}},\ X_{\text{tmp}} \in \mathbb{R}^{C/r\times T\times 1} \tag{6}$$

where X_{tmp} denotes the contextual feature among w frames and its channel-reduced, r is the scale ratio (set to 16 in this work), and $*$ indicates the convolution operation.

$$\begin{aligned} &M_{\text{cte}} = \text{Sigmoid}(conv_{\text{exp}} * ReLU(X_{\text{tmp}})) \\ &M_{\text{cte}} \in \mathbb{R}^{T\times C\times 1} \end{aligned} \tag{7}$$

Like CSE, CTE adaptively recalibrates the importance of frames and excites the temporal-sensitive channels simultaneously by the interaction between the temporal dimension and channel dimension. Finally, the excited output feature F_{cte} is obtained.

$$F_{\text{cte}} = X * M_{\text{cte}} + X,\ F_{\text{cte}} \in \mathbb{R}^{C\times T\times V} \tag{8}$$

3.2 Complementary topology scheme

By rethinking various adjacency matrix schemes from previous works, primarily focusing on ST-GCN [16] and its two variants 2 s-AGCN [17] and CTR-GCN [20], we can summarize the various adjacency matrices as three different types: A_{p}(physical), A_{l} (learnable), and A_{s}(similarity).

A_p denotes the predefined matrix reflecting the physical structure of the human body, which is fixed during the training process. ST-GCN [16] applies predefined A_p of spatial configuration partitioning, dividing the neighbor set into three subsets according to their distances to the skeleton gravity center.

A_l denotes the learnable matrix covering the global graph. This indicates whether the connections exist between each pair of two joints and how strong they are. The ST-GCN utilizes the attention matrix M_k to learn edge importance weighting and dot multiplies to A_p. The 2 s-AGCN builds an adjacency matrix with the same shape of A_p and makes it parameterized without any constraints, which can be optimized during the training process.

A_s denotes the Gaussian similarity matrix between two vertexes, and A_s is data dependent, which is different from A_l. The 2 s-AGCN measures the similarity of pairs of vertexes in an embedding space by the dot product. CTR-GCN uses pairwise subtraction to calculate the distances along the channel dimension. Most importantly, CTR-GCN makes A_s channel-wise and learns channel-specific A_s for each channel, leading to stronger representation capability than channel-shared A_s. We mark the channel-specific A_s as CA_s.

In this study, we found that, however, A_l, A_s, and CA_s are more adaptive to capture global graph information for different input samples compared with A_p. However, it is not appropriate to neglect the necessity of A_p, especially on large-scale datasets. Although A_l and A_s can automatically capture global graph information, they face a too large search space to find the most appropriate topology. The optimization process will be confused if each topology has too many "noisy" edges [36]. To address this issue, we find it necessary to take them all into account. We combine three sub-matrices by simple summation as $A_p + A_l + CA_s$. This simple but efficient scheme achieves a better performance.

3.3 ICE-GC block and ICE-GCN

An efficient interactional channel excited graph convolution (ICE-GC) block, which is equipped with ICE and a complementary topology scheme elaborated above, is proposed. The structure of our ICE-GC is depicted in Fig. 3a. The operation of ICE-GC is formulated as follows:

$$F_{out} = M_{cte} M_{cse} \sum_{k}^{K_v} (A_p + A_l + CA_s) W_k F_{in} \tag{9}$$

where the input feature map F_{in} is a 3D tensor as $C \times T \times V$. A 1×1 2D convolutional layer is utilized to transform input features into high-level representations, where W_k is the $C \times C' \times 1$ weight vector. K_v denotes three subsets according to three partition strategies proposed by ST-GCN [16]. A_p and A_l are both $V \times V$ adjacency matrices, which are the same for each channel. CA_s is the $C' \times V \times V$ adjacency matrix which contains C' channel-specific $V \times V$ adjacency matrices for C' channels, and the final refined A is obtained by element-wise summation as $C' \times V \times V$. After matrix multiplication with high-level representations, the graph convolution is accomplished. Three graph convolution blocks are utilized in parallel to extract the latent representations. The output excited feature map F_{out} as $C' \times T \times V$ is obtained by CTE and CSE, which is a significant complementary approach after GCN, since the adjacency matrices can only define the existence of connections between joints, which cannot adaptively reflect the importance between joints along channel dimension.

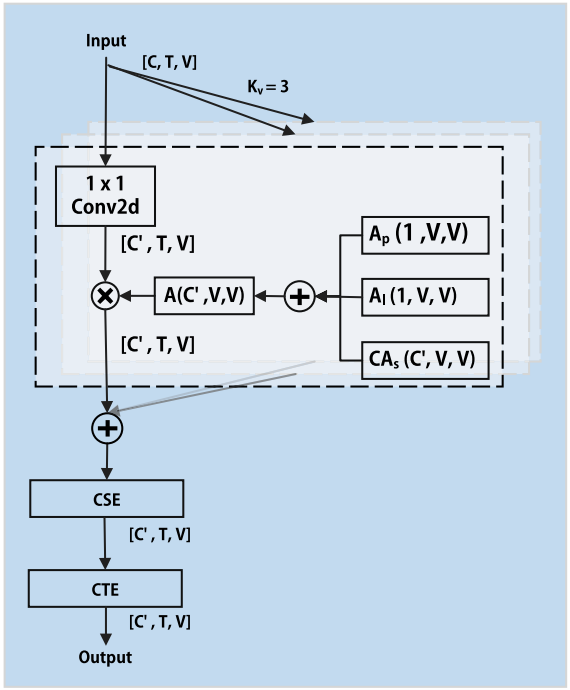
Based on the ICE-GC block, an interactional channel excitation enhanced graph convolutional network (ICE-GCN) is constructed. The basic block of our ICE-GCN is shown in Fig. 3b. A multi-scale temporal convolutional module (MS-TCN) is applied following the design of [20,35] for multiple receptive fields and temporal pooling, which is different from CTE. For the residual connection, a 1×1 convolution is worked when C is not equal to C'. Therefore, our proposed ICE-GCN has powerful characterization capabilities in spatial, temporal, and channel dimensions.

As shown in Fig. 3c, the architecture of our ICE-GCN is similar to most of the improved ST-GCN frameworks. First, a batch normalization layer (BN) is added to normalize the input data. Then, ten basic ICE-GCN blocks mainly constitute the entire network. From block 1 to block 10, the input channel and output channel are (3,64), (64,64), (64,64), (64, 64), (64, 128), (128, 128), (128, 128), (128, 256), (256, 256), and (256, 256). The frames T will be halved after block 5 and block 8. After ten main basic ICE-GCN blocks, a global average pooling (GAP) layer is added to pool output feature maps. Finally, a fully connected layer (FC) receives the pooled output and generates predictions for the action class through scores.

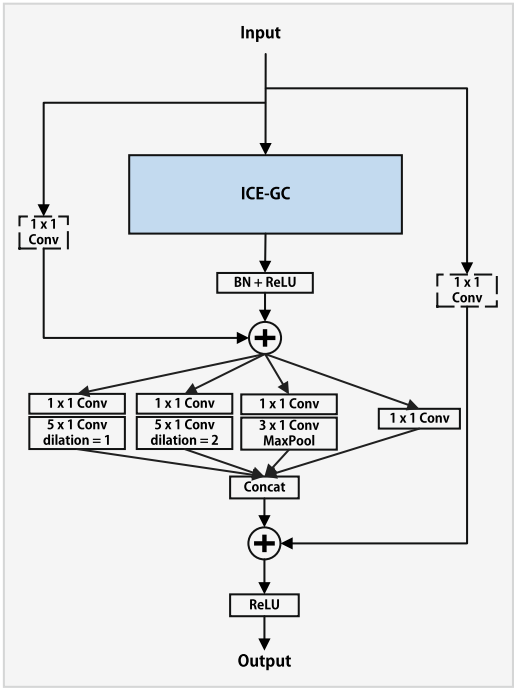
4 Experiments and results

4.1 Datasets

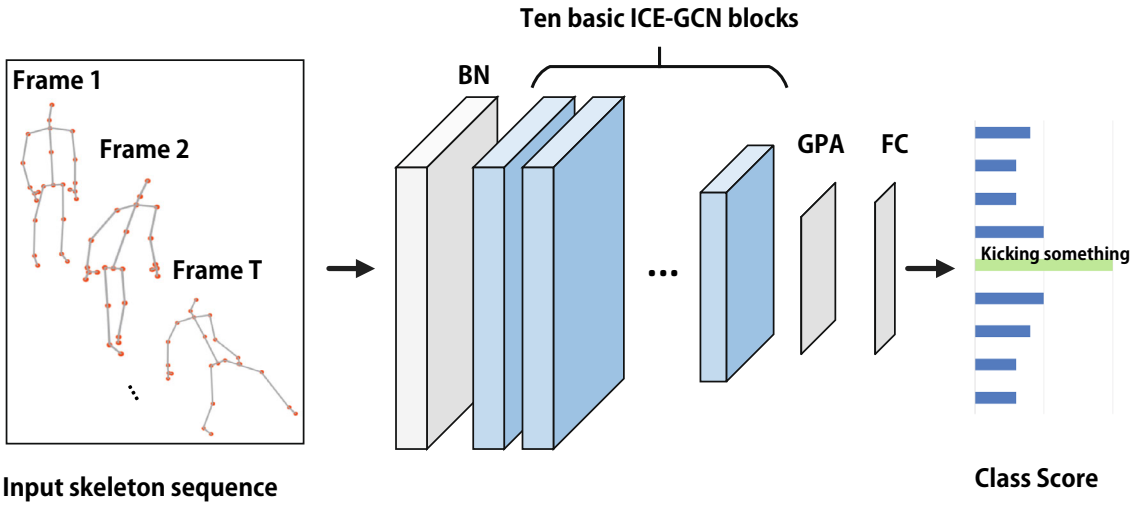
Kinetics-Skeleton The Kinetics-Skeleton [44] dataset includes approximately 300,000 video clips and 400 human action classes, which are collected from the YouTube video website. However, it only offers raw video clips and does not provide skeleton data. Thanks to the work ST-GCN [16] and OpenPose [6] toolbox, which estimated the locations of 18 joints on every frame of the clips. There are 240,000 clips for training and 20,000 clips for testing. According to the conventional evaluation method of the ST-GCN, we report the top-1 and top-5 accuracies to evaluate our model.



(a) ICE-GC block



(b) ICE-GCN block



(c) ICE-GCN network

Fig. 3 **a** Illustration of ICE-GC block. **b** The basic block of our ICE-GCN. **c** The architecture of our ICE-GCN network. C (channels), C' (output channels), T (number of frames), T' (output number of frames), V (joints), r (channel reduce ratio). $\oplus$ denotes the element-wise summation. $\otimes$ denotes the matrix multiplication

NTU RGB+D 60 NTU RGB+D 60 [45] is a large and widely used human action recognition dataset, which has 56880 human 3D skeleton action sequences, 40 volunteers and 60 classes collected by three Kinect v2 [5] cameras with different views. Each frame contains one or two actors, and each skeleton has 25 joints. There are two recommended benchmarks for this dataset: Cross-subject (X-sub) and cross-view (X-view). X-sub: 20 subjects for training and 20 subjects for testing. X-view: Camera views 2 and 3 for training and camera view 1 for testing.
NTU RGB+D 120 With 57,367 additional samples and more than 60 action classes based on NTU RGB+D 60, NTU RGB+D 120 [46] is the largest dataset with 3D skeleton action sequences for human action recognition available right now. It contains 114,480 action samples and 120 action classes in total, which were recorded by 106 volunteers using three different camera views. Cross-subject (X-Sub) and cross-setup (X-Set) are two recommended benchmarks. X-sub: A total of 106 subjects were split into two groups of 53 for training and 53 for testing. X-setup: Dividing the samples into training and testing groups half and half based on the camera setup IDs.

4.2 Implementation details

All our experiments are conducted on the PyTorch deep learning framework, trained on one RTX 3090 GPU. The optimization strategy is SGD with a momentum of 0.9. For Kinetics-Skeleton, we follow the same data processing method as [17], which has 150 frames with two bodies in each frame. We set the batch size to 114 and the temporal receptive field of CTE to 3 frames. The training phase is completed at the 65th epoch. For NTU RGB+D 60 and NTU RGB+D 120, we follow the same data processing method as [20], which resized each sample to 64 frames. We set the batch size to 64 and the temporal receptive field of CTE to 1 frame. The training phase is completed at the 80th epoch.

4.3 Ablation studies

4.3.1 Effectiveness of three excitation modules.

For a fair comparison, we choose the widespread framework AGCN [17] on the Cross-View of the NTU RGB+D 60 using joint coordinates as the only input data stream. We separately test the contributions of two sub-modules CTE and CSE and find that they both can improve the accuracy by 0.4% and 0.6%, respectively. Then, it is observed that ICE module improves AGCN by 1.1% by connecting two sub-modules, more than either of them as shown in Table 1. It illustrates that both spatial and temporal channel-wise features are necessary for distinguishing different action categories, and they are complementary to each other.

Table 1 Effectiveness of three excitation modules

Methods	X-view (%)	Δ Acc. (%)
AGCN	93.7	–
AGCN with CTE	94.1	+ 0.4
AGCN with CSE	94.3	+ 0.6
AGCN with ICE	94.8	+ 1.1

Table 2 Comparison with other excitations

Methods	X-View (%)	Δ Acc. (%)
AGCN [17]	93.7	–
AGCN with SENet [26]	94.2	+ 0.5
AGCN with STC-attention [18]	94.3	+ 0.6
AGCN with ICE (ours)	**94.8**	**+ 1.1**

4.3.2 Comparison with other excitation modules

To validate the superiority of our ICE, we compare the performance of ICE with other channel dimension excitation on the Cross-View of the NTU RGB+D 60 using the joint stream as shown in Table 2. SENet [26] was proposed to be embedded into 2D CNNs for the image classification task; it is a classic and popular channel attention mechanism, and we adapted and applied it to our skeleton-based task. STC-attention [18] applies channel attention in a skeleton-based action recognition task, concatenating spatial attention, and temporal attention in a sequential manner without cross-dimensional interactions, as shown in Table 2. Note that the enhancements brought about by SENet and STC-attention for AGCN (+0.6% and +0.7%, respectively) are both less than our ICE for AGCN (+1.1%). This validated the rationality and superiority of our ICE, which introduces both spatial and temporal information into the channel dimension to capture the cross-dimensional interactions.

4.3.3 Transferring to other backbones

We verify the generality, adaptability, and complexity of our proposed ICE module on both the Cross-View and Cross-Subject of NTU RGB+D 60 using joint stream. We also chose the well-known and widespread backbone AGCN [17], the lightweight model Shift-GCN [19], and the latest proposed optimal model CTR-GCN [20]. Our ICE module is simply equipped with those models in a plug-and-play way. As shown in Table 3, the backbones equipped with our ICE outperform themselves notably, and the computation cost (measured by Floating Point Operation per second (FLOPs) and the number of parameters) does not change too much (only increase about 0.03 GFLOPs and 0.42 M parameters).

Table 3 Study on the impact of transferring to different backbones in accuracy, FLOPs, and the number of parameters

Methods	X-sub (%)	X-view (%)	Param.(M)	FLOPs(G)
AGCN [17]	86.5	93.7	3.47	37.38
AGCN with our ICE	87.4 (+ 0.9)	94.8 (+1.1)	3.89 (+ 0.42)	37.42 (+ 0.04)
Shift-GCN [19]	87.8	95.1	0.69	2.50
Shift-GCN with our ICE	88.4 (+ 0.6)	97.4 (+2.3)	1.11 (+ 0.42)	2.52 (+ 0.02)
CTR-GCN [20]	89.9	94.9	1.45	16.40
CTR-GCN with our ICE	90.2 (+ 0.3)	95.0 (+ 0.1)	1.87 (+ 0.42)	16.43 (+ 0.03)

Table 4 Comparisons of accuracies when removing A_p, A_l, and CA_s from ICE-GCN

Methods	X-view (%)	Δ Acc. (%)
ICE-GCN	**95.3**	–
ICE-GCN w/o A_p	95.0	− 0.3
ICE-GCN w/o A_l	95.1	− 0.2
ICE-GCN w/o CA_s	94.8	− 0.5

4.3.4 Effectiveness of adjacency matrix schemes

Then, we verify the necessity of the three adjacency matrices by removing A_p, A_l, and CA_s from ICE-GCN on the Cross-View of NTU RGB+D 60 using joint steam. As described in Sect. 3.4., A_p denotes the physical adjacency matrix, A_l denotes the learnable adjacency matrix, and CA_s denotes channel-wise similarity adjacency matrix. As shown in Table 4, the performance of our ICE-GCN can reach 95.3%. When removing A_p, A_l and CA_s, the performance drops 0.3%, 0.2%, and 0.5%, respectively. It verified that all three adjacency matrices are efficient and complementary to each other and verified the rationality of our refined an efficient complementary topology scheme $A_p + A_l + CA_s$.

4.4 Comparison with the state-of-the-arts

Finally, we compare our ICE-GCN model with the state-of-the-art methods in skeleton-based action recognition on three large-scale datasets Kinetics-Skeleton [44], NTU RGB+D 120 [46], and NTU RGB+D 60 [45] in Tables 5, 6, and 7, respectively. 2 s-AGCN [17] proposed the bone stream (the lengths and directions of bones) of the skeleton data, which shows notable improvement in the recognition accuracy. Generally, most state-of-the-art methods adopt multi-stream fusion strategies. For a fair and comprehensive comparison, we use both single- and multi-stream fusion strategies for comparison. *Js* denotes only the "joint stream" using the original skeleton coordinates as input. *Bs* denotes only the "bone stream" using the differential spatial coordinates as input. 2*s* denotes using both "joint stream" and "bone stream."

Table 5 Comparisons of the Top-1 and Top-5 accuracy with the state-of-the-art methods on the Kinetics-Skeleton dataset

Model	Year	Kinetics-Skeleton	
		Top-1 (%)	Top-5 (%)
ST-GCN [16]	2018 (AAAI)	31.6	53.7
AS-GCN [37]	2019 (CVPR)	34.8	56.3
Js-AGCN [17]	2019 (CVPR)	35.1	57.1
Js-NAS-GCN [38]	2020 (AAAI)	35.5	57.9
Js-AAGCN [18]	2020 (TIP)	36.0	58.4
Js-MST-GCN [47]	2021 (AAAI)	36.0	58.5
Js-ASE-GCN [48]	2022 (TCSVT)	35.8	58.4
Js-ICE-GCN (ours)	–	**37.2**	**60.2**
Bs-AGCN [17]	2019 (CVPR)	33.3	55.7
Bs-NAS-GCN [38]	2020 (AAAI)	34.9	57.1
Bs-AAGCN [18]	2020 (TIP)	34.7	57.5
Bs-MST-GCN [47]	2021 (AAAI)	32.3	58.2
Bs-ASE-GCN [48]	2022 (TCSVT)	34.1	57.4
Bs-ICE-GCN (ours)	–	**36.9**	**59.6**
2 s-AGCN [17]	2019 (CVPR)	36.1	58.7
2 s-DGNN [35]	2019 (CVPR)	36.9	59.6
2 s-NAS-GCN [38]	2020 (AAAI)	37.1	60.1
2 s-AAGCN [18]	2020 (TIP)	37.4	60.4
2 s-MST-GCN [47]	2021 (AAAI)	37.8	60.3
2 s-ASE-GCN [48]	2022 (TCSVT)	36.9	59.7
2 s-ICE-GCN (ours)	–	**38.8**	**61.8**

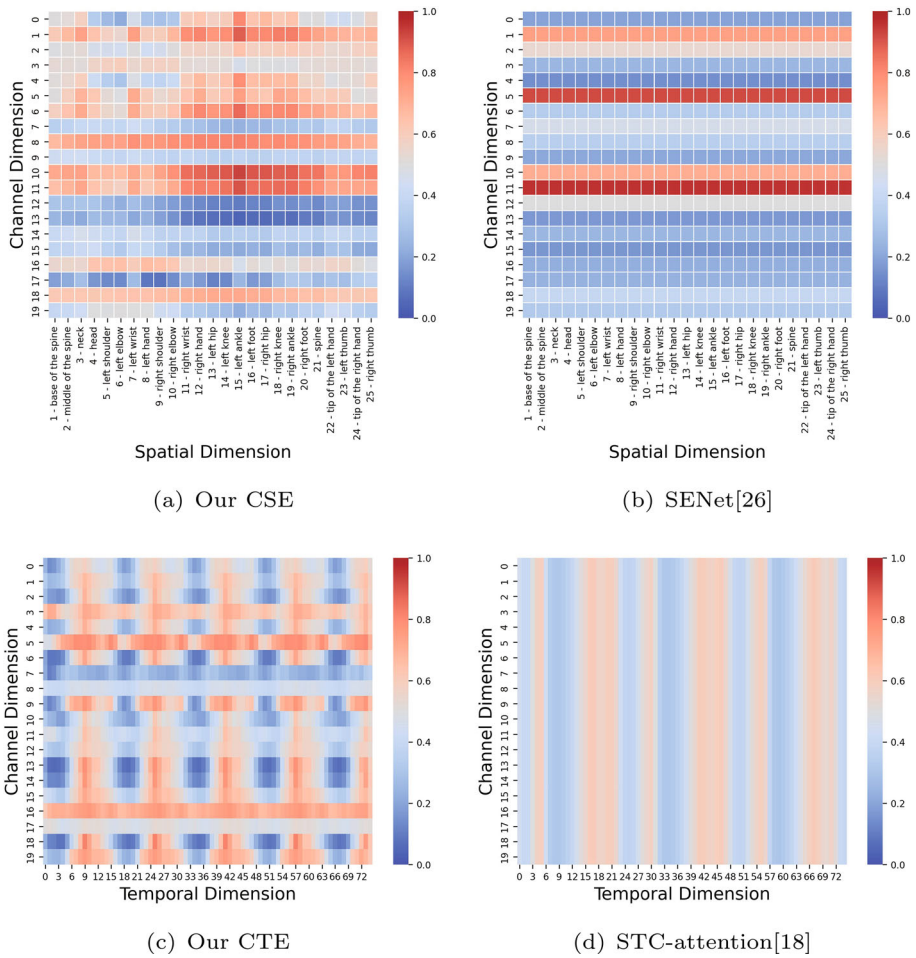
On Kinetics-skeleton, the ICE-GCN notably outperforms the existing methods by about 2% on Top-1 and Top-5 for all the fusion strategies. On NTU RGB+D 60 and NTU RGB+D 120 datasets, ICE-GCN also outperforms the existing methods in most cases on both Cross-Subject and Cross-view. As shown in the results of the comparisons on all three large-scale datasets, the state-of-the-art and competitive results verify the superiority of our ICE-GCN model. It demonstrates that our model has stronger modeling capability and performance on a larger dataset Kinetics-skeleton.

Table 6 Comparisons of the Top-1 accuracy with the state-of-the-art methods on the NTU RGB+D 120 dataset

Model	Year	NTU RGB+D 120	
		X-Sub (%)	X-Set (%)
FSNet [49]	2019 (TPAMI)	59.9	62.4
SGN [50]	2020 (CVPR)	79.2	81.5
Js-Shift-GCN [19]	2020 (CVPR)	80.9	83.2
Js-MST-GCN [47]	2021 (AAAI)	82.8	84.5
Js-CTR-GCN [20]	2021 (ICCV)	84.9	86.6
Js-ML-STGNet [51]	2022(TIP)	84.9	86.5
Js-ICE-GCN (ours)	–	**85.1**	**86.9**
Bs-MST-GCN [47]	2021 (AAAI)	84.8	86.3
Bs-CTR-GCN [20]	2021 (ICCV)	85.7	**87.5**
Bs-ML-STGNet [51]	2022 (TIP)	85.7	87.1
Bs-ICE-GCN (ours)	–	**86.8**	87.1
2 s-MS-G3D [52]	2020 (CVPR)	86.9	88.4
2 s-Shift-GCN [19]	2020 (CVPR)	85.3	86.6
2 s-MST-GCN [47]	2021 (AAAI)	87.0	88.3
2 s-CTR-GCN [20]	2021 (ICCV)	88.7	90.1
2 s-ML-STGNet [51]	2022 (TIP)	88.6	90.0
2 s-ICE-GCN (ours)	–	**89.1**	**90.2**

Table 7 Comparisons of the Top-1 accuracy with the state-of-the-art methods on the NTU RGB+D 60 dataset

Model	Year	NTU RGB+D 60	
		X-Sub (%)	X-View (%)
ST-GCN [16]	2018 (AAAI)	81.5	88.3
AS-GCN [37]	2019 (CVPR)	86.8	94.2
Js-AGCN [17]	2019 (CVPR)	86.5	93.7
Js-AGC-LSTM [21]	2019 (CVPR)	87.5	93.5
Js-NAS-GCN [38]	2020 (AAAI)	–	94.6
Js-AAGCN [18]	2020 (TIP)	88.0	95.1
Js-Shift-GCN [19]	2020 (CVPR)	87.8	95.1
Js-MS-G3D [52]	2020 (CVPR)	89.4	95.0
Js-MST-GCN [47]	2021 (AAAI)	89.0	95.1
Js-CTR-GCN [20]	2021 (ICCV)	89.9	94.9
Js-ML-STGNet [51]	2022 (TIP)	89.8	94.9
Js-ICE-GCN (ours)	–	**90.1**	**95.2**
Bs-AAGCN [18]	2020 (TIP)	88.4	94.7
Bs-NAS-GCN [38]	2020 (AAAI)	–	94.7
Bs-MS-G3D [52]	2020 (CVPR)	90.1	**95.3**
Bs-MST-GCN [47]	2021 (AAAI)	89.5	95.2
Bs-CTR-GCN [20]	2021 (ICCV)	90.6	–
Bs-ML-STGNet [51]	2022 (TIP)	90.2	94.6
Bs-ICE-GCN (ours)	–	**90.6**	94.9
2 s-AAGCN [18]	2020 (TIP)	89.4	96.0
2 s-NAS-GCN [38]	2020 (AAAI)	–	95.7
2 s-MS-G3D [52]	2020 (CVPR)	91.5	96.2
2 s-MST-GCN [47]	2021 (AAAI)	91.1	**96.4**
2 s-ML-STGNet [51]	2022 (TIP)	91.8	96.1
2 s-ICE-GCN (ours)	–	**92.0**	96.2



(a) Our CSE (b) SENet[26]

(c) Our CTE (d) STC-attention[18]

Fig. 4 Visualization comparisons between CSE **a** and SENet **b**, CTE **c** and STC-attention **d**. The vertical axis denotes the randomly selected 20 channels. For **a** and **b**, the horizontal axis denotes the 25 joints. For **c** and **d**, the horizontal axis denotes the final output feature of 75 frames

4.5 Visualization

To illustrate how ICE affects the final performance and highlight the difference of cross-dimension interactions, the attention maps are visualized. One real evaluation sample of "kicking something" is randomly selected from NTU RGB+D 60 and visualized. As shown in Fig. 4, CSE and SENet are compared on channel and spatial dimensions, and CTE and STC-attention are compared on channel and temporal dimensions.

As shown in Fig. 4a, CSE can not only reweight crucial joints along each channel but also excite those spatial-sensitive channels (such as channels 1, 10 and 11). CSE focuses on the joints of the legs like joint 15 "left knee" and joint 16 "left foot" which are relevant to the action "kicking something." The importance of joint 15 "left knee" is consistently strong in channels 0, 1, 6, etc., indicating that the spatial information of these related joints is generally important for the current action in the excited channels. But SENet reweights each joint constantly for each channel without channel-wise difference and interactions with spatial dimension, as shown in Fig. 4b.

As shown in Fig. 4c, CTE can not only reweight vital frames, respectively, but also excite those temporal-sensitive channels (such as channels 3, 5, and 16). CTE focuses on the frames (such as the frames from 21 to 30) which are most informative to the action of "kicking something." It is worth noting that the importance of these frames is consistently strong, indicating that the temporal relationship of these frames is generally important for the current action in the excited channels. As shown in Fig. 4d, STC-attention only reweights frames constantly for each channel without channel-wise discriminative consideration.

5 Conclusion

In this paper, we propose an interactional channel excited graph convolutional network with complementary topology for skeleton-based action recognition. The interactional channel excitation module (ICE) consists of CSE and CTE sub-modules. CSE is applied to capture the crucial body global structure patterns along different channels and then adaptively recalibrate the importance of joints and excite the

spatial-sensitive channels. CTE is applied to capture vital temporal inter-frame dynamics information along different channels and then adaptively recalibrate the importance of frames and excite the temporal-sensitive channels. In addition, to solve the dilemma between the adaptation ability and too large of a searching space, to avoid too many "noisy" graph edges, a complementary topology scheme is refined as $A_p + A_l + C A_s$. By coupling the ICE module and topology strategy, we propose an interactional channel excitation-enhanced graph convolutional network with complementary topology (ICE-GCN), which is a powerful network to help extract optimal features covering three dimensions (spatial, temporal, and channel). Extensive experiments conducted on three large datasets NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton demonstrate that our ICE-GCN outperforms state-of-the-art methods and proves the effectiveness of each sub-modules. In the future, the efficiency of ICE-GCN still needs more consideration.

Acknowledgements This work was partially supported by the STI 2030-Major Projects under grant 2022ZD0208900, the National Natural Science Foundation of China under grant 62076103, and the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) under grant pdjh2022a0125.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Malik, Z., Shapiai, M.I.B.: Human action interpretation using convolutional neural network: a survey. Mach. Vision Appl. **33**(3), 1–23 (2022)
2. Kong, Y., Fu, Y.: Human action recognition and prediction: a survey. Int. J. Comput. Vision **130**(5), 1366–1401 (2022)
3. Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H., Moon, H.: Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognit. **108**, 107561 (2020)
4. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vision Image Underst. **115**(2), 224–241 (2011)
5. Zhang, Z.: Microsoft Kinect sensor and its effect. IEEE Multimed. **19**(2), 4–10 (2012)
6. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
7. Yang, H., Guo, L., Zhang, Y., Wu, X.: U-shaped spatial-temporal transformer network for 3d human pose estimation. Mach. Vision Appl. **33**(6), 1–16 (2022)
8. Ocegueda-Hernández, V., Román-Godínez, I., Mendizabal-Ruiz, G.: A lightweight convolutional neural network for pose estimation of a planar model. Mach. Vision Appl. **33**(3), 1–21 (2022)
9. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457–5466 (2018)
10. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2117–2126 (2017)
11. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
12. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision, pp. 816–833 (2016). Springer
13. Soo Kim, T., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28 (2017)
14. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3288–3297 (2017)
15. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit. **68**, 346–362 (2017)
16. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence (2018)
17. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)
18. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans. Image Process. **29**, 9532–9545 (2020)
19. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 183–192 (2020)
20. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368 (2021)
21. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1227–1236 (2019)

22. Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., Lu, H.: Decoupling gcn with dropgraph module for skeleton-based action recognition. In: European Conference on Computer Vision, pp. 536–553 (2020). Springer
23. Qiu, H., Wu, Y., Duan, M., Jin, C.: GLTA-GCN: Global-local temporal attention graph convolutional network for unsupervised skeleton-based action recognition. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022). IEEE
24. Xie, Y., Zhang, Y., Ren, F.: Temporal-enhanced graph convolution network for skeleton-based action recognition. IET Comput. Vision **16**(3), 266–279 (2022)
25. Gao, B.-K., Dong, L., Bi, H.-B., Bi, Y.-Z.: Focus on temporal graph convolutional networks with unified attention for skeleton-based action recognition. Appl. Intell. **52**(5), 5608–5616 (2022)
26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
27. Yang, H., Gu, Y., Zhu, J., Hu, K., Zhang, X.: PGCN-TCA: pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. IEEE Access **8**, 10040–10047 (2020)
28. Sun, N., Leng, L., Liu, J., Han, G.: Multi-stream slowfast graph convolutional networks for skeleton-based action recognition. Image Vision Comput. **109**, 104141 (2021)
29. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
30. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3139–3148 (2021)
31. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: Stm: Spatiotemporal and motion encoding for action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2000–2009 (2019)
32. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 909–918 (2020)
33. Wang, Z., She, Q., Smolic, A.: Action-net: Multipath excitation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13214–13223 (2021)
34. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)
35. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7912–7921 (2019)
36. Wang, M., Ni, B., Yang, X.: Learning multi-view interactional skeleton graph for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
37. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3595–3603 (2019)
38. Peng, W., Hong, X., Chen, H., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 2669–2676 (2020)
39. Ding, C., Liu, K., Cheng, F., Belyaev, E.: Spatio-temporal attention on manifold space for 3d human action recognition. Appl. Intell. **51**(1), 560–570 (2021)
40. Xing, Y., Zhu, J., Li, Y., Huang, J., Song, J.: An improved spatial temporal graph convolutional network for robust skeleton-based action recognition. Applied Intelligence, 1–17 (2022)
41. Xie, J., Miao, Q., Liu, R., Xin, W., Tang, L., Zhong, S., Gao, X.: Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition. Neurocomputing **440**, 230–239 (2021)
42. Zhu, J., Zou, W., Zhu, Z., Hu, Y.: Convolutional relation network for skeleton-based action recognition. Neurocomputing **370**, 109–117 (2019)
43. Liu, Z., Wang, L., Wu, W., Qian, C., Lu, T.: Tam: Temporal adaptive module for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13708–13718 (2021)
44. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
45. Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
46. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019)
47. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1113–1122 (2021)
48. Xiong, X., Min, W., Wang, Q., Zha, C.: Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition. IEEE Trans. Circuits Syst. Video Technol. **33**(1), 342–353 (2022)
49. Liu, J., Shahroudy, A., Wang, G., Duan, L.-Y., Kot, A.C.: Skeleton-based online action prediction using scale selection network. IEEE Trans. Pattern Anal. Mach. Intell. **42**(6), 1453–1467 (2019)
50. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)
51. Zhu, Y., Shuai, H., Liu, G., Liu, Q.: Multilevel spatial-temporal excited graph network for skeleton-based action recognition. IEEE Transactions on Image Processing (2022)
52. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)