Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048813

ARTICLE

BCCLR: A Skeleton-Based Action Recognition with Graph Convolutional Network Combining Behavior Dependence and Context Clues

Yunhe Wang[1], Yuxin Xia[2] and Shuai Liu[2,*]

[1]College of Information Science and Engineering, Institute of Interdisciplinary Studies, Hunan Normal University, Changsha, 410081, China

[2]School of Educational Sciences, Institute of Interdisciplinary Studies, Hunan Normal University, Changsha, 410081, China

*Corresponding Author: Shuai Liu. Email: liushuai@hunnu.edu.cn

Received: 19 December 2023 Accepted: 04 March 2024 Published: 26 March 2024

ABSTRACT

In recent years, skeleton-based action recognition has made great achievements in Computer Vision. A graph convolutional network (GCN) is effective for action recognition, modelling the human skeleton as a spatio-temporal graph. Most GCNs define the graph topology by physical relations of the human joints. However, this predefined graph ignores the spatial relationship between non-adjacent joint pairs in special actions and the behavior dependence between joint pairs, resulting in a low recognition rate for specific actions with implicit correlation between joint pairs. In addition, existing methods ignore the trend correlation between adjacent frames within an action and context clues, leading to erroneous action recognition with similar poses. Therefore, this study proposes a learnable GCN based on behavior dependence, which considers implicit joint correlation by constructing a dynamic learnable graph with extraction of specific behavior dependence of joint pairs. By using the weight relationship between the joint pairs, an adaptive model is constructed. It also designs a self-attention module to obtain their inter-frame topological relationship for exploring the context of actions. Combining the shared topology and the multi-head self-attention map, the module obtains the context-based clue topology to update the dynamic graph convolution, achieving accurate recognition of different actions with similar poses. Detailed experiments on public datasets demonstrate that the proposed method achieves better results and realizes higher quality representation of actions under various evaluation protocols compared to state-of-the-art methods.

KEYWORDS

Action recognition; deep learning; GCN; behavior dependence; context clue; self-attention

1 Introduction

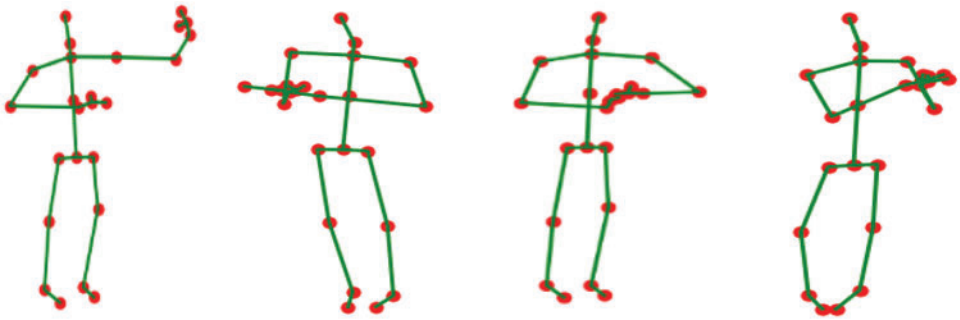
Action recognition is a popular topic and has rich applications in emergency detection [1], educational scenes, and intelligent monitoring [2]. For example, school bullying can be avoided by recognizing students' abnormal actions, and dangerous movements can be predicted by analyzing the athlete's posture, performance, and completion of the movements. Compared to RGB image-based action recognition, the skeleton-based method has strong adaptability under dynamic environments

and complex backgrounds, and avoids the challenges of background occlusion, illumination change, and multiple viewing angles. Therefore, it has attracted considerable attention.

The deep neural network method has been widely used in action recognition field and space-time modeling of skeletal sequence. Conventional action recognition methods are based on convolutional neural network (CNN) and recurrent neural network (RNN) to learn action features from video, with high accuracy and robustness. However, CNN and RNN do not explicitly use skeleton topology information to completely extract features from the irregular natural structure of the human body. Some recent work has captured structural features of actions in spatiotemporal maps through GCN, which have been shown to have superior effects in action recognition.

Because of the strong modeling ability of non-Euclidian structural data, GCN-based methods have attracted significant attention in skeleton-based action recognition. Most existing methods generate graphs with heuristic predefinitions in which joints are defined as vertices and edges are defined according to the physical relations of joint pairs in the human body [3,4]. However, these methods ignore the features of the spatial relationship between non-adjacent joint pairs in special actions and result in the false recognition of these actions with a large number of non-physically dependent joint pairs. The physical relationship between non-adjacent joint pairs affects the judgment of action recognition. For example, in the actions "taking a selfie" and "reading," the non-adjacent joints such as the neck and the hand are important for accurate action recognition, which is called behavior dependence in this paper.

As shown in Fig. 1, comparing the two actions "taking a selfie" in Fig. 1a and "reading" in Fig. 1b, behavior dependence, which is the key feature of an action, cannot be extracted by current methods. In other words, joint pairs that are not adjacent to the physical structure but are adjacent to the spatial structure are not extracted. For example, the dependencies between the two hands, as well as the hand and chest, are extracted to recognize the action. Therefore, these two types of action are incorrectly identified.



(a) Taking a selfie (b) Reading (c) Put palms together (d) Clapping

Figure 1: Visualization of confused actions in recognition

Recognition errors caused by actions with similar poses are common problems in GCNs. Figs. 1c and 1d show the similar frames between the two actions "put palms together" and "clapping." Some poses in these two actions are highly similar, leading to confusion in action recognition. This is because a large number of similar poses mislead the recognition without context clues between adjacent frames.

Therefore, this study proposes a GCN that combines behavioral dependence and contextual clues for skeleton-based action recognition. First, a dynamic learnable graph is constructed by extracting

joint pairs with behavioral dependence for specific actions. With the extracted behavior dependence, specific actions with implicit correlations between nonadjacent joint pairs are satisfactorily recognized. Second, a self-attention module is designed to recognize actions with similar poses by extracting contextual clues between adjacent frames. Finally, several comparative experiments are conducted on a public skeleton-based action dataset to validate the superiority of the proposed model. The contributions of this study are summarized as follows:

1. In this paper, we propose a learnable behavior dependence-based GCN (BDGCN) that models a spatial GCN by extracting joint relationships with strong behavior dependence in actions. The BDGCN recognizes specific actions with many spatio-dependent joint pairs.
2. We design a context clue-based self-attention module (CSM) that extracts pose connections between adjacent frames during actions. In the CSM, self-attention with embedded joint positions is used to extract the spatial joint topology between adjacent frames that accurately recognizes actions with a large number of similar poses.
3. By combining the BDGCN and CSM, we provide a method with GCN combining behavior dependence and context clues (BCCLR) for skeleton-based action recognition based on contrastive learning. Compared to current mainstream algorithms, the BCCLR achieves superior results on public datasets. Ablation experiments reveal that both the BDGCN and CSM improve the effectiveness of action recognition.

The remainder of this paper is organized as follows. Section 2 introduces the related research. In Section 3, the proposed BCCLR is explained, including the BDGCN and CSM. Section 4 presents the experimental results to validate the effectiveness of BCCLR. Finally, Section 5 concludes the study.

2 Related Work

2.1 GCNs and Their Use in Different Domains

GCNs are widely used to capture spatial correlations from data in Euclidean space [5,6]. Yan et al. [7] used the same perspective to jointly explain the problem of excessive smoothness and heterogeneity at the node level in a GCN for the first time. Liang et al. [8] proposed a GCN based on a dependence tree and sentiment common-sense knowledge that used sentiment to enhance sentence dependence. Singh et al. [9] proposed convolutional neural network using fast forward quantum optimization algorithm, which can minimize the classification error. Yin et al. [10] proposed a fusion model of a GCN and Long Short-Term Memory (LSTM) for emotion recognition. Zhao et al. [11] proposed the SemGCN model, which combines the channel weights of the implicit prior edges in the learning graph with kernel attributes and significantly improves the convolution ability of the graph. Zhou et al. [12] used a deep mesh relation to generate a GCN. An adaptive adjacency matrix was applied to extract the positive and negative relationships between the joints, and significant results were obtained. Yu et al. [13] proposed a GCN model-based framework to recognize daily lost data and automatically determine the best recovery method. Yue et al. [14] proposed a new attribute fusion model that utilizes attributes through a graph structure to better represent users and projects. Based on TinyJAMBU-128, Rasheed et al. [15] proposed a system which can used for autonomous mobile platforms with CAN bus capability.

2.2 Skeleton-Based Action Recognition

Deep-learning-based methods often use convolutional neural networks (CNN), recurrent neural networks (RNN), and GCNs to learn skeleton sequence representations. CNN-based methods

typically embed skeletons into two-dimensional (2D) pseudo-images for input requirements. It first transforms the skeleton sequence into a skeleton graph with the same target size and then uses a CNN to learn the spatial and temporal features. For example, Rong et al. [16] coupled the correlation between spatial features and motions, and constructed a shape-motion representation using algebraic geometry. RNN-based methods typically extract frame-level features and model sequence correlation [17]. For example, Kumar et al. [18] proposed an ensemble-based deep learning framework comprising a CNN and a dense neural network. ShiftGCN [19] is a shifted convolution method for graph-structured data. Multi-range temporal features were extracted by sequentially stacking the temporal convolutions [20]. The human skeleton was studied according to the skeleton data, and convolution was performed on the graph edge corresponding to the human skeleton [21]. GCN-based methods effectively deal with irregularly structured graphs such as skeleton data. Li et al. [22] introduced an encoder-decoder structure for extracting strong joint correlations and mining potential action correlations. Chen et al. [23] proposed a Channel-wise topology refinement GCN that modeled the joint topology with different embedding channels.

2.3 Self-Attention Mechanism

Inspired by human attention, a self-attention mechanism is used to improve the recognition performance. Several studies have used self-attention as the basic building block for model construction [24]. For example, Pan et al. [25] designed a hybrid model with the elegant integration of self-attention and convolution modules. Li et al. [26] seamlessly integrated convolution and self-attention to stack onto a powerful backbone. Zhang et al. [27] used self-attention in global spatial and temporal dependencies, which are used in the spatial and temporal dimensions, respectively. Ren et al. [28] proposed a shunt self-attention model to distinguish between multiscale features. Du et al. [29] provided a new method for multivariate time-series missing values by learning the lost values from a weighted combination of self-attention blocks. Cerquitelli et al. [30] discussed the topic of machine learning and artificial intelligence algorithms, which providing pointers to the non-expert readers in the field of machine learning to some resources. Shan et al. [31] proposed the NRTSI, a time-series interpolation method that treats a time series as a set of (time, data) tuples. Wang et al. [32] introduced a cross-self-attention model to classify and segment tasks of the origin point cloud by learning features and coordinates.

3 Proposed Method

A detailed description of the proposed model is provided, including a learnable GCN based on behavior dependence and a self-attention module, to explore the intrinsic features of actions through context clues.

3.1 Overall Architecture of the Model

Fig. 2 illustrates the overall framework of the proposed model (BCCLR). In BCCLR, an adaptive BDGCN is first used to construct a learnable graph convolution based on behavioral dependence. The BDGCN extracts graph features based on the input skeleton vertices and constructs a dynamic adjacency matrix using the distance matrix from the training set. During the training process, an adjacency matrix with learnable weights is obtained to express the spatial joint connections through the Bernoulli distribution of every element. Finally, the BDGCN is constructed using the weight relationship between the joint pairs.

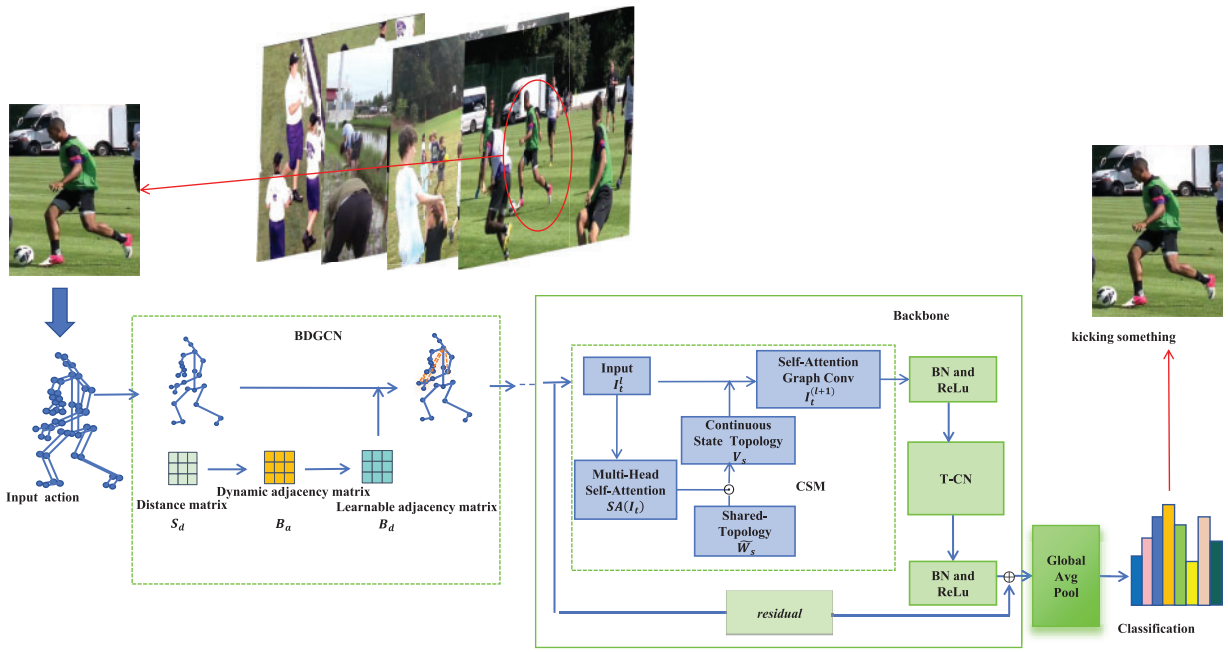


Figure 2: Framework of the proposed BCCLR for action recognition

By using the graph convolution model, a context-clues based self-attention module CSM is proposed to model the spatial convolution network and is added to the backbone multi-layer Spatio-temporal Block [4]. The CSM transforms joint features into vectors with learnable parameters through linear variations and adds position embedding. By combining the shared topology for time and instances, and the multi-head self-attention map, the CSM obtains the context-based clue topology to update the dynamic graph convolution.

By combining the information on behavior dependence and context clues from the BDGCN and CSM, a graph convolution is constructed. The temporal features are then aggregated using temporal graph convolution. Finally, after global average pooling, the classification results of the actions are obtained through the activation function, and the entire BCCLR model is constructed.

3.2 Learnable Graph Convolution Network with Behavior Dependence (BDGCN)

In 3D coordinates, the GCN constructs a spatiotemporal graph according to the joints as vertices and the physical connectivity in the human body as edges. The pre-defined graph-based spatio-temporal GCN ignores the behavior dependence of the joint pairs, and incorrectly recognizes specific actions with different behavior dependence between joint pairs, such as the actions "salute" and "reading." To address this issue, a learnable graph convolution (BDGCN) is constructed by introducing behavior dependence into the GCN.

3.2.1 Basic GCN Structure

The human skeleton structure is represented by the graph $G = (\mathrm{V}, W)$. In G, the parameters $\mathrm{V} = \{\mathrm{V_i}\}_{i=1\ldots k}$, $W = \left\{W_{i,j}\right\}_{i,j=1\ldots k}$ where k represents the number of joints. The feature graph of a GCN is a tensor of size $c \times t \times k$, where c represents the number of feature channels for the joints, and t represents the total number of frames in the video. A Boolean matrix $W \in [0, 1]_{k\times k}$ is defined as the

adjacency matrix of G as shown in Eq. (1):

$$W_{ij} = \begin{cases} 1 & When\ i\ is\ adjacent\ to\ j\ or\ i = j \\ 0 & When\ i\ is\ not\ adjacent\ to\ j \end{cases} \tag{1}$$

The definition of the graph convolution is expressed as Eq. (2):

$$F^{l+1} = \sigma(MF^l\widetilde{W}) \tag{2}$$

where $\widetilde{W}$ is the symmetric normalization of W, $\sigma(\cdot)$ represents the nonlinear activation function, M is the convolution kernel with size 1×1, and F^l is the representation of the joint in the$l$layer. In a basic GCN, W cannot extract nonadjacent joint pairs with behavior dependence because all nonzero elements exist only in the positions of physically adjacent joints.

3.2.2 GCN Model with Behavior Dependence

Some key joints that influence action recognition are not physically adjacent to each other; therefore, a basic GCN cannot extract the behavioral dependencies between them. To learn the behavioral dependencies between nonadjacent joint pairs in specific actions, this study constructs an adjacency matrix B that dynamically generates edges based on the spatial relation of joints.

First, $\mathrm{S_d} = \{S_d^{ij}\}$ is used to represent the distance matrix of graph G (where $S_d^{ij} = 0$), which is obtained by processing the training set data using Eq. (3):

$$S_d^{ij} = \| V_i - V_j \|_2 . \tag{3}$$

A dynamic adjacency matrix B_a is constructed based on the soft-max function β with scale parameter μ on $\mathrm{S_d}$ in Eq. (4):

$$B_a^{ij} = \begin{cases} \beta\left(\mu \dfrac{1}{S_d^{ij}}\right) & i \neq j \\ 1 & i = j \end{cases}. \tag{4}$$

Based on B_a, a learnable adjacency matrix B_d is constructed according to the Bernoulli distribution Z in Eq. (5):

$$B_d^{ij} \sim Z(x, B_a^{ij}). \tag{5}$$

Each element B_d^{ij} follows a Bernoulli distribution with probability B_a^{ij}. All elements B_d^{ij} change constantly by learning in every iteration, and finally, B_d is obtained. For inference, $B_d^{ij} = 1$ for all elements. Then, through a matrix dot product and operator β, the final graph eigenmatrix B is calculated by Eq. (6):

$$B = \beta(W_s \odot B_\mathrm{d}), \tag{6}$$

where $\odot$ represents the matrix dot product operation; $W_s \in R_{k\times k}$ is a learnable weight matrix of B_d.

W_s is updated using the loss function during training, and the weight of each spatial connection is learned. The basic GCN in Eq. (2) is improved to a learnable GCN using Eq. (7):

$$F^{l+1} = \sigma(MF^l\widetilde{B}) \tag{7}$$

where $\widetilde{B}$ is the symmetric normalization of B.

This module updates W to B based on the behavior dependence to extract the features of nonadjacent joint pairs in the graph and realizes adaptive modeling of a learnable BDGCN. As shown in Fig. 3, the basic GCN extracts only physically adjacent joint pairs, whereas the proposed BDGCN extracts both physically and spatially adjacent joint pairs. Because of the relationship between nonadjacent joint pairs that exhibit behavioral dependence, the BDGCN improves the ability to recognize specific actions with implicit joint correlations. The BDGCN captures deeper spatial dependency features by capturing the rich dependencies between joints, rather than just the physical connectivity of the joints.

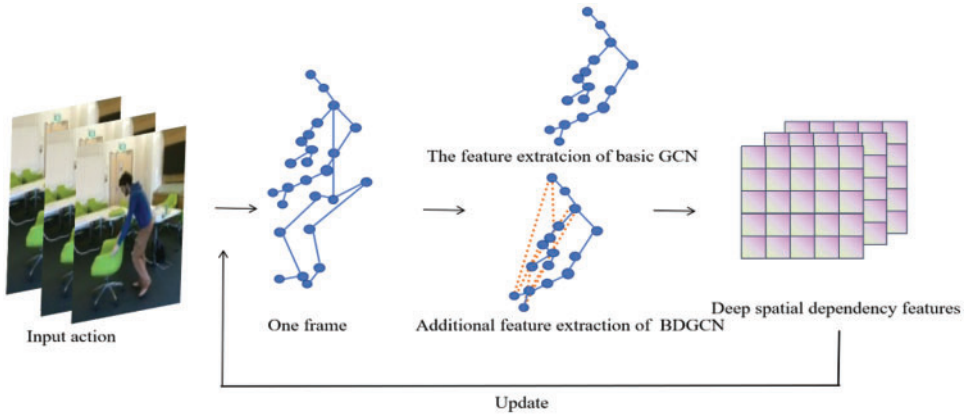


Figure 3: Comparison between the basic GCN and BDGCN. The features between non-adjacent joints of specific actions are strengthened in the BDGCN. The features of both the hand and neck joints have been enhanced in this figure

3.3 Self-Attention Module with Context Clues (CSM)

In action recognition, context clues inside the action play an important role in the recognition of different actions with many similar poses. Here, a self-attention module CSM is provided to model the context clues for adjacent frames in each action. It focuses on the context between adjacent frames and dynamically models the spatial topology between frames. The structure of CSM is proposed in Fig. 4.

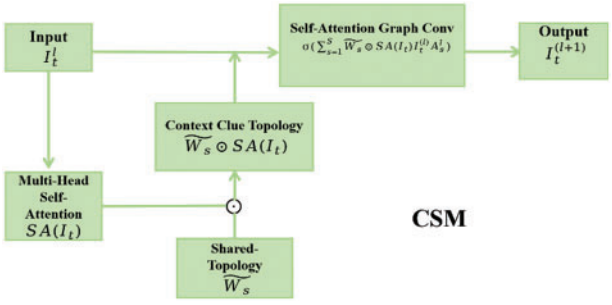


Figure 4: Structure of the proposed CSM

The joint feature $F \in R^{c \times t \times k}$ calculated by Eq. (7) is linearly transformed into a $D^{(0)}$ dimensional vector with learnable parameters. By adding joint position information from position embedding (PE)

to the vector, a joint feature vector $I_t^{(0)}$ is obtained, as shown in Eq. (8):

$$I_t^{(0)} = Linear\,(F_t) + PE \tag{8}$$

where $I_t^{(0)}$, $PE \in R^{V\times D^{(0)}}$, and t is the time index.

I_t is then updated by a spatial convolution with two processes: The average neighborhood vertex feature and the linear transformation aggregation feature. The update of the hidden layer is expressed by Eq. (9):

$$I_t^{(l+1)} = \sigma(\widehat{W} I_t^{(l)} A^{(l)}) \tag{9}$$

where $\widehat{W} = E^{-\frac{1}{2}}(W+I)E^{-\frac{1}{2}}$, E is the degree matrix of $W+I$, $A^{(l)} \in R^{D^{(l)}\times D^{(l+1)}}$ is a learnable parameter matrix in the l-th layer. The joint of learned matrices A_Q and $A_K \in R^{D\times D'}$ represent queries and keys that linearly map from I_t to dimension D'. Self-attentional mapping is expressed in Eq. (10):

$$\mathrm{SA}\,(I_t) = \beta\left(\frac{I_t A_k (I_t A_Q)^T}{\sqrt{D'}}\right) \tag{10}$$

Meantime, SA$(\cdot)$ learns a shared topology with time and instances. Shared topology and self-attention mapping employ S multi-heads in which the models participate jointly from different subspaces. For a head s $\in [1, S]$, the shared topology $\widetilde{W}_s \in R^{k\times k}$ is combined with the self-attentional mapping $SA_s\,(I_t) \in R^{t\times k\times k}$ to obtain the context clue topology $V_s \in R^{t\times k\times k}$:

$$V_s = \widetilde{W}_s \odot SA_s\,(I_t) \tag{11}$$

Using V_s as the neighborhood information of I_t, the CSM module is constructed by updating the joint representation in Eqs. (8) to (12):

$$I_t^{(l+1)} = \sigma\left(\sum\nolimits_{s=1}^{S} \widetilde{W}_s \odot SA_s\,(I_t)\, I_t^{(l)} A_s^l\right) \tag{12}$$

where the modules are residually connected using 1×1 convolution layers. In this paper, $D' = \dfrac{D}{8}$, S $= 3$, $l = 8$.

To date, two modules, the BDGCN and CSM, have been constructed. Subsequently, similar to the ST-GCN [4], the backbone of the modules is updated through the convolution of the time graph T-CN. Finally, the entire BCCLR algorithm is used for action classification with global average pooling and a soft-max function. The algorithm for the proposed network is shown in Algorithm 1.

Algorithm 1: Action Recognition Using Behavior Dependence and Context Clues

Input: The training skeleton feature $F \in R^{c\times t\times k,}$.
Output: The prediction results of the response of the validation set.

1: For training period do:
2: Step 1. Learnable Graph Convolution Network with Behavior Dependence
3: $S_d^{ij} \leftarrow$ Compute distance matrix of graph
4: $B_a^{ij} \leftarrow$ Compute dynamic adjacency matrix
5: $B_d^{ij} \leftarrow$ Compute learnable adjacency matrix by Bernoulli distribution
6: $B \leftarrow$ Compute characteristic matrix
7: $F^{l+1} \leftarrow$ Complete the update of learnable GCN with Eq. (7)

(Continued)

Algorithm 1 (continued)

Step 2. Self-attention Module with Context Clues
  $I_t^{(0)} \leftarrow$ Compute joint feature by adding position embedding from F
  SA $(I_t) \leftarrow$ Compute self-attentional mapping from learned matrices A_Q and A_K
  $V_s \leftarrow$ Compute context clue topology from shared topology $\widetilde{W}_s$ by Eq. (11)
  $I_t^{(l+1)} \leftarrow$Complete the update of joint representation with Eq. (12)
Step 3. features are aggregated by temporal graph convolution.
End for
Step 4. Model prediction
  Global average pooling
  Predict the response of validation set through the updated model

4 Experimental Results and Analysis

To evaluate the proposed method, numerous experiments were conducted using the NTU RGB+D 60 [33]. Relevant experimental results are reported to validate the effectiveness of the proposed method.

4.1 Datasets

NTU RGB+D 60 is a representative skeleton-based action recognition dataset that is also known as NTU-60 and is available at this link[1]. It contains 56,880 clips, consisting of 60 action classes. Each 3D skeleton contains the 3D positions of 25 joints per frame. The dataset is constructed from the views of 40 people captured by three cameras. The dataset is tested using two protocols: 1) x-sub: 40,320 training and 16,560 validation datasets from different persons; 2) x-view: Training and validation data are divided according to different camera views, where the information captured by cameras 2 and 3 is processed as a training set, and the information from camera 1 is the validation set. It consists of 37,920 training and 18,960 validation clips. The classification accuracy of NTU-60 is reported under cross-subject and cross-view settings.

4.2 Implementation Details

4.2.1 Experimental Setting

All experiments are performed using PyTorch 1.7. The invalid frame data for each clip are first removed, and the clips are then adjusted to a length of 50 frames using linear interpolation. The minibatch size is set to 128. The experimental environment is based on Ubuntu 16.04 with a GPU (3090-24G) and CPU (AMD EPYC 7601).

4.2.2 Self-Supervised Pretext Training

The AimCLR settings are kept in the proposed experiments, where the momentum coefficient m is set to 0.999, and the temperature hyperparameter τ is set to 0.07. SGD with momentum (0.9) and weight decay (0.0001) is used for training optimization. The model is trained for 300 epochs, where the learning rate is 0.1 during the first 250 epochs, and decreased to 0.01 during the last 50 epochs. The nonlinear activation function ReLU is used for model training. A weight fusion vector [0.6,0.6,0.4] is defined for the three skeletal sequence streams: Joint, bone, and motion.

[1] https://rose1.ntu.edu.sg/dataset/actionRecognition/

4.2.3 Linear Evaluation

The proposed BCCLR is validated using a linear evaluation of action tasks. Specifically, a linear classifier (a fully connected layer combined with a soft-max layer) is trained and supervised by a fixed encoder.

4.3 Ablation Experiment

Generalized ablation experiments are conducted using both evaluation protocols for the NTU-60 dataset to validate the effectiveness of the BDGCN and CSM models. The results are presented in Tables 1 and 2.

Table 1: Ablation experiments by linear evaluation of a single-stream on NTU-60

Methods	x-sub	x-view
Baseline	74.3	79.7
Baseline + BDGCN	74.6	80.1
Baseline + CSM	74.5	80.0
Baseline + BDGCN + CSM	**74.8**	**80.2**

Table 2: Ablation experiments by linear evaluation of a three-stream fusion on NTU-60

Methods	x-sub	x-view
Baseline	78.9	83.8
Baseline + BDGCN	80.4	84.6
Baseline + CSM	80.3	84.4
Baseline + BDGCN + CSM	**80.8**	**84.9**

Table 1 shows that the recognition accuracy of the single-stream with baseline in the x-sub protocol was 74.3%, and when BDGCN is used, the accuracy improves by 0.3%, reaching 74.6%. After continuously adding the CSM, the accuracy of the proposed model improved to 74.8%, which is an improvement of 0.5%. Under the x-view protocol, when BDGCN was used, the accuracy improved to 80.1%, with an increase of 0.4%. With the addition of CSM, the accuracy of the proposed model improved to 80.2%, with an improvement of 0.5%.

Then, in the ablation experiment with three-stream fusion in Table 2, when using the BDGCN, the accuracy under the x-sub and x-view protocols increased by 1.5% and 0.8%, respectively. When the CSM is further introduced, the third-stream model achieves the highest accuracy under the two evaluation protocols, increasing by 1.9% and 1.1% to 80.8% and 84.9%, respectively.

Tables 1 and 2 show that the proposed model outperforms baseline method on the NTU-60, whether under three single-stream or multi-stream fusion. The BDGCN extracts the behavior dependence of joint pairs to model convolution graph, and better combines the important feature of non-adjacent joint pairs for action recognition. Besides, CSM extracts the topology structure between frames by context clues to avoid interference from many similar poses within different actions. Both two modules provide better recognition accuracy than baseline.

4.4 Quantitative Analysis

To validate the effectiveness of the proposed method, evaluations and comparisons with the existing methods are conducted for NTU-60. The experimental results are listed in Table 3.

Table 3: Comparisons of skeleton-based action recognitions with linear evaluation on NTU-60

Method	Encoder	NTU-60	
		x-sub	x-view
PCRP (TMM'21) [34]	GRU	54.9	63.4
AS-CAL (InS'21) [35]	LSTM	58.5	–
EnGAN-PRNN (WACV'19) [36]	LSTM	68.6	77.8
H-Transformer (ICME'21) [37]	Transformer	69.3	72.8
SeBiReNet (ECCV'20) [38]	GRU	–	79.7
ST-Graph CMRL(CVIU'23) [39]	GCN	74.7	82.6
3s-SkeletonCLR (CVPR'21) [40]	GCN	75.0	79.8
3s-Colorization (ICCV'21) [41]	GCN	75.2	83.1
3s-CrossSCLR (CVPR'21) [40]	GCN	77.8	83.4
GL-Transformer (ECCV'22) [42]	Transformer	76.3	83.8
3s-AimCLR (AAAI'22) [43]	GCN	78.9	83.8
HiCo-GRU(AAAI'23) [44]	GRU	80.6	–
FoCoViL (Neurocomputing'23) [45]	RNN	–	83.2
3s-BCCLR (Ours)	GCN	**80.8**	**84.9**

Table 3 shows that the proposed method provides the best recognition accuracy for both two protocols x-sub and x-view. Comparing the SOTA AimCLR, the recognition accuracy of BCCLR under x-sub protocol and x-view protocol increased by 1.9% and 1.1%, respectively.

Table 4 presents the linear evaluation results for different periods of the x-sub protocol for the NTU-60 dataset.

Table 4: Linear evaluation for x-sub on NTU-60 with different epochs

Method	100ep	150ep	200ep	300ep
3s-SkeletonCLR [40]	71.3	73.8	74.1	74.1
3s-CrossSCLR [40]	70.0	72.8	76.0	77.2
3s-AimCLR [43]	76.5	77.4	78.3	78.9
3s-BCCLR (Ours)	**76.9**	**78.4**	**79.7**	**80.8**

As shown in Table 4, the proposed 3s-BCCLR achieved the best recognition accuracy during each period. After the 100-th epoch, 3s-BCCLR had an accuracy of 76.9%, which was 0.4% higher than that of SOTA 3s-AimCLR. After the 150-th epoch, 3s-BCCLR had an accuracy of 78.4%, which was 1.0% higher than that of 3s-AimCLR. At the 200-th epoch, the difference between the accuracies of the models was 79.7%–78.3% = 1.4%; finally, with the final accuracy of 3s-BCCLR reaching 80.8%,

the difference also reached its highest value of 1.9%. As the number of epochs increased, the accuracy increment of 3s-BCCLR continuously improved relative to the existing algorithms.

The linear evaluation results with the three single streams and the fusion of the three streams on both the x-sub and x-view for AimCLR and BCCLR are presented in Table 5.

Table 5: Linear evaluation results with different streams on NTU-60

Method	Stream	NTU-60	
		x-sub	x-view
AimCLR	Joint	74.3	79.7
BCCLR	Joint	**74.8**	**80.2**
AimCLR	Motion	66.8	70.6
BCCLR	Motion	**70.6**	**73.3**
AimCLR	Bone	73.2	77.0
BCCLR	Bone	**74.1**	**77.4**
3s-AimCLR	Joint + Motion + Bone	78.9	83.8
3s-BCCLR	Joint + Motion + Bone	**80.8**	**84.9**

Table 5 shows that the recognition accuracy of BCCLR is higher than that of AimCLR with both single and multi-stream fusion. In detail, compared to those of AimCLR, the results of BCCLR with every single stream improved from 0.4%–3.8%. Furthermore, the result of three-stream fusion improved by 1.9% under the x-sub protocol and 1.1% under x-view protocol. This method focuses on the behavior dependence between joints and the context clues in actions, which makes the recognition of single flow more accurate, improving the accuracy of joint flow. As the features extracted from the joint flow play a leading role in multi-flow fusion, the proposed method achieves better multi-flow fusion results.

Visual comparisons of the three single streams and stream fusion of different methods on NTU-60 are shown in Figs. 5 and 6. In joint flow, the proposed method achieved the best recognition accuracy. As the features extracted from the joint flow play a leading role in multi-flow fusion, the proposed method achieved the best multi-flow fusion results.

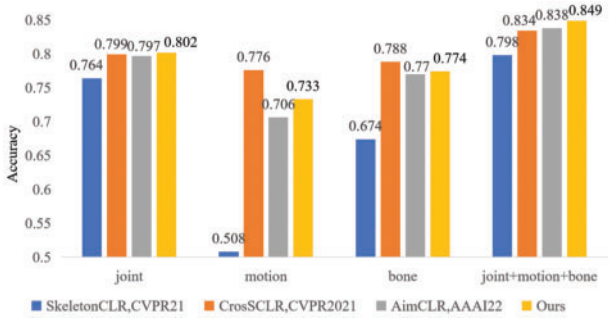


Figure 5: Visual comparisons of x-view with different streams on NTU-60

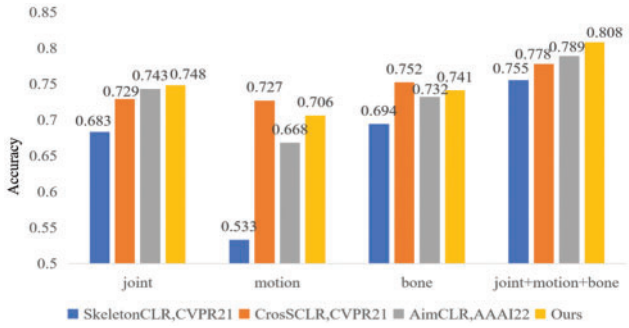


Figure 6: Visual comparisons of x-sub with different streams on NTU-60

Accuracy comparisons between 3s-BCCLR and SOTA 3s-AimCLR for actions with behavior dependence and context clues are shown in Fig. 7.

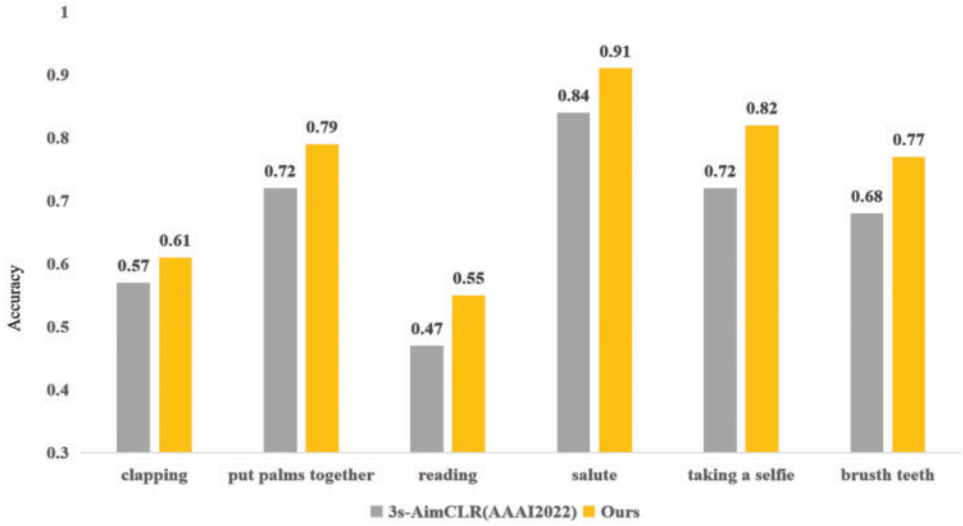
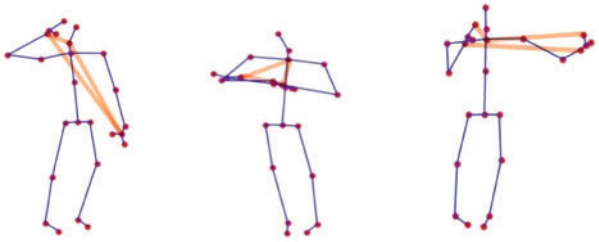


Figure 7: Comparison of x-view accuracy on actions with behavior dependence and context clues

In Fig. 7, the first two actions "reading" and "taking a selfie" have behavior dependence between non-adjacent joint pairs, the 3rd and 4th actions "clapping" and "put palms together" have context clues, and the last two actions "salute" and "brush teeth" have both behavior dependence and context clues. For actions with behavioral dependence between the hands, head, and neck, the recognition accuracy improved by 8.3% and 10.5%, respectively. This is because the proposed method extracts and models the behavior dependence between joint pairs in actions and then improves the accuracy of such specific actions with implicit correlation between non-adjacent joints. For actions with context clues between consecutive frame sequences, the recognition accuracy improved by 4.5% and 7.6%, respectively. Actions with many similar poses are accurately recognized by the proposed method because it focuses on the context connection between consecutive frame sequences in the action. Similarly, for the actions "salute" and "brush teeth," with both behavior dependence and context clues, the accuracy of the proposed method improved from 84.5% and 68.0% to 91.5% and 76.9%, respectively.

4.5 Qualitative Analysis

To validate the relationships between joint pairs with strong behavior dependences, particular actions are extracted with the BDGCN. Fig. 8 shows the topological visualization of the joint pairs in actions with behavior dependence. Figs. 8a–8c correspond to topology visualizations of the actions "salute", "reading", and "taking a selfie", respectively. In the BDGCN, the original physical connections between the joints are represented by blue lines, and the joint pairs with strong behavioral dependence are extracted as additional orange lines. Fig. 8 shows that the BDGCN extracts behavior dependence in specific actions, obtains connections between nonadjacent joint pairs, and connects the potential joint pairs. Based on the relationship between these nonadjacent joint pairs, the BDGCN accurately recognizes the actions in which existing methods encounter errors.



(a) Salute (b) Reading (c) Taking selfies

Figure 8: Topology visualization of joint pairs in actions with behavior dependence inferred by our BDGCN method

Fig. 8 shows that the three actions are mainly accomplished in the upper extremity, and the entire joint participates in them. During different actions, many different connections are established between the hand, neck, and other body parts, which exhibit different action dependencies between joints. These results confirm that the proposed method captures abundant information on action dependence, shows that behavior dependence better distinguishes different actions, and improves recognition accuracy.

Many similar poses belonging to different actions can cause confusion in recognition. The intrinsic topology of an action is inferred by context clues from adjacent frames, which better distinguish actions with similar poses. In Fig. 9, the original physical connections between joints are represented by purple lines, and the intrinsic topology inferred by context clues in action is represented by orange lines, where the strength of the inferred relationship is represented by the thickness of the lines and the size scale of the joints.

Fig. 9 shows that similar poses have different intrinsic topologies for different actions when context clues are considered. In Fig. 9, when $t = 0$, the orange line and joint size show that these two poses have similar intrinsic topology. As the poses changes during the action, when $t = 30$, the intrinsic topology of different actions shows significant differences because of the different previous poses. Finally, when $t = 50$, the attention from the left hand to the head is more concentrated in the action "put palms together" than in the "clapping", while the attention from the left hand to right hand and elbow is stronger in the action "clapping" than in the action "put palms together". This is due to the amplitude of "clapping" is larger and the action frequency is more intensive than that of "put palms together". Moreover, the coupling between the hands is stronger, while "put palms together" establishes more connections with the head and neck. The proposed method accurately extracts the features of these time-based topologies and thus accurately distinguishes actions with similar poses.

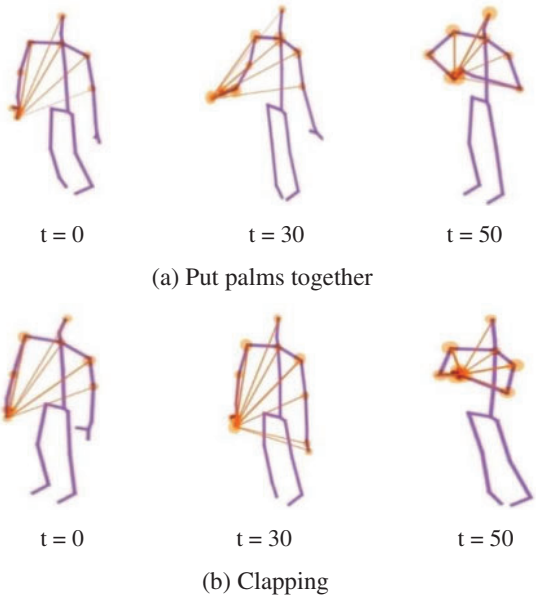


Figure 9: Topology visualization with context clues in actions. The orange lines represent the inferred topology from the joint "left hand" to other joints. The thickness of the orange lines and the size of joints are proportional to the strength of the inferred relationship

Fig. 10 shows the attention map of the joints, which confirms that the joints with more weights in BCCLR are the key joints in action recognition. In Fig. 10, the upper row shows the attention mapping results of AimCLR, and the lower row shows the results of our BCCLR. The first line is the action "put palms together", the middle is the action "clapping", and the last is the action "salute".

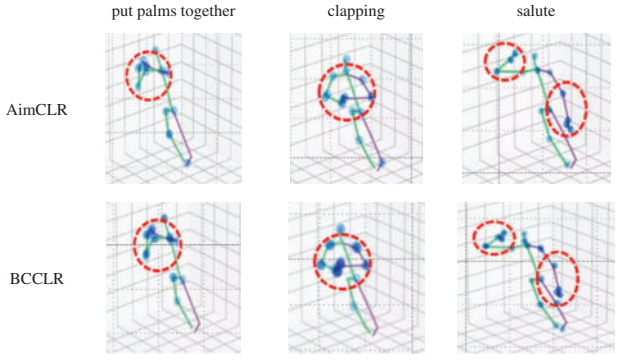


Figure 10: Comparison between AimCLR [43] and BCCLR by attention map of joints

In Fig. 10, the green and purple lines present the left and right sides of the skeleton, respectively. The color depth of the vertices indicates the attentional strength of the joint. In the attention map, the strength of the attention to a joint represents the sum of the weights of each connected joint, which reflects the attention paid to the joints in action recognition.

For the similar actions "put palms together" and "clapping", BCCLR focuses more attention on the joints with behavior dependence than AimCLR, such as the hands and shoulders of the upper body. Then, the action is distinguished by adding the context clues of actions, which AimCLR fails to

recognize. Similarly, for the action "salute", BCCLR assigns higher weights to the key joints of both hands and elbows than AimCLR and achieves a more accurate recognition.

5 Conclusions and Future Work

In this study, we propose BCCLR, which is a skeleton-based action recognition with a GCN combining behavior dependence and context clues. We first propose a learnable Graph Convolution Network (BDGCN) based on behavior dependence that extracts joint relationships with strong dependence on specific actions, thereby increasing the flexibility of the GCN. Experiments revealed that it had a positive effect on the action recognition of the BDGCN, which promotes the recognition application of action. In addition, this study extracted the internal topological features of actions using the self-attention mechanism (CSM), which focuses on the context clues of actions and links the spatial topological structures of human joints. The CSM solves the problem of incorrect identification of actions owing to the interference of similar poses, which provides more accurate recognition for abnormal detection. According to experiments on the widely used NTU RGB+D 60 dataset with many ablation and comparison results, the proposed BCCLR, which combines BDGCN with CSM, showed a good effect on action recognition based on skeleton data and was conducive to the realization of stronger action recognition. For the future work, the use of multi-modal information for action recognition is a future direction. In addition, the hybrid GCN-CNN architecture introduced will also be a good choice for skeleton-based analysis and other aspects of research.

Acknowledgement: The authors would like to thank the editors and reviewers for their valuable work, as well as the supervisor and family for their valuable support during the research process.

Funding Statement: This work was supported in part by the 2023 Key Supported Project of the 14th Five Year Plan for Education and Science in Hunan Province with No. ND230795.

Author Contributions: Study conception and design: Y. Wang, S. Liu; data collection: Y. Wang, Y. Xia; analysis and interpretation of result: Y. Wang, S. Liu; draft manuscript preparation: Y. Wang, S. Liu, Y. Xia. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this paper can be requested from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Zhang and T. Zhao, "P-2.31: An emergency rescue action recognition method based on improved spatiotemporal decomposition network," in *SID Symp. Dig. Tech. Papers*, 2023, vol. 54.

[2] T. M. Alzubi, J. A. Alzubi, A. Singh, O. A. Alzubi, and M. Subramanian, "A multimodal human-computer interaction for smart learning system," *Int. J. Hum.-Comput. Int.*, vol. 2022, no. 4, pp. 1–11, 2023. doi: 10.1080/10447318.2023.2206758.

[3] Y. Liu, H. Zhang, D. Xu, and K. Kang, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," *Knowl-Based Syst.*, vol. 40, no. 1, pp. 108146, 2022. doi: 10.1016/j.knosys.2022.108146.

[4] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, 2018, vol. 32.

[5] Z. Lin, J. Feng, Z. Lu, and Y. Li, "DeepSTN+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, Hawaii, USA, 2019, vol. 33, no. 1, pp. 1020–1027.

[6] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Min.*, Anchorage, AK, USA, 2019, pp. 1720–1730.

[7] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra, "Two sides of the same coin: Heterophily and oversmoothing in graph convolutional eural networks," in *2022 IEEE Int. Conf. Data Min. (ICDM)*, Orlando, FL, USA, IEEE, 2022, pp. 1287–1292.

[8] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowl.-Based Syst.*, vol. 235, pp. 107643, 2022. doi: 10.1016/j.knosys.2021.107643.

[9] P. Singh and M. K. Muchahari, "Solving multi-objective optimization problem of convolutional neural network using fast forward quantum optimization algorithm: Application in digital image classification," *Adv. Eng. Softw.*, vol. 176, pp. 103370, 2023. doi: 10.1016/j.advengsoft.2022.103370.

[10] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, pp. 106954, 2021. doi: 10.1016/j.asoc.2020.106954.

[11] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3D human pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3425–3435.

[12] S. Zhou, M. Jiang, S. Cai, and Y. Lei, "DC-GNet: Deep mesh relation capturing graph convolution network for 3D human shape reconstruction," in *Proc. 29th ACM Int. Conf. Multimed.*, China, 2021, pp. 171–180.

[13] Y. Yu, V. O. K. Li, J. C. K. Lam, and K. Chan, "GCN-ST-MDIR: Graph convolutional network-based spatial-temporal missing air pollution data pattern identification and recovery," *IEEE Trans. Big Data*, vol. 9, no. 5, pp. 1347–1364, 2023. doi: 10.1109/TBDATA.2023.3277710.

[14] G. Yue, R. Xiao, Z. Zhao, and C. Li, "AF-GCN: Attribute-fusing graph convolution network for recommendation," *IEEE Trans. Big Data*, vol. 9, no. 2, pp. 597–607, 2023. doi: 10.1109/TBDATA.2022.3192598.

[15] A. Rasheed, M. Baza, M. Khan, N. Karpoor, C. Varol and G. Srivastava, "Using authenticated encryption for securing controller area networks in autonomous mobile platforms," in *2023 26th Int. Symp. Wirel. Pers. Multimed. Commun. (WPMC)*, Tampa, FL, USA, IEEE, 2023, pp. 76–82.

[16] Y. Li, R. Xia, X. Liu, and Q. Huang, "Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition," in *2019 IEEE Int. Conf. Multimed. Expo (ICME)*, Shanghai, China, IEEE, 2019, pp. 1066–1071.

[17] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 1227–1236.

[18] A. Kumar, S. Saumya, and A. Singh, "Detecting dravidian offensive posts in MIoT: A hybrid deep learning framework," in *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, New York, NY, USA, 2023.

[19] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 183–192.

[20] F. Li, J. Li, A. Zhu, Y. Xu, H. Yin and G. Hua, "Enhanced spatial and extended temporal graph convolutional network for skeleton-based action recognition," *Sens.*, vol. 20, no. 18, pp. 5260, 2022. doi: 10.3390/s20185260.

[21] Y. 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, 2022.

[22] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3595–3603.

[23] 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/CVF Int. Conf. Comput. Vis.*, Montreal, Canada, 2021, pp. 13359–13368.
[24] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[25] X. Pan *et al.*, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 815–825.
[26] K. Li *et al.*, "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600. 2023.
[27] Y. Zhang, X. Wei, X. Zhang, Y. Hu, and B. Yin, "Self-attention graph convolution residual network for traffic data completion," *IEEE Trans. Big Data*, vol. 9, no. 2, pp. 528–541, 2022.
[28] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 10853–10862.
[29] W. Du, D. Côté, and Y. Liu, "Self-attention-based imputation for time series," *Expert. Syst. Appl.*, vol. 219, pp. 119619, 2023.
[30] T. Cerquitelli, M. Meo, M. Curado, L. S. Kapov, and E. E. Tsiropoulou, "Machine learning empowered computer networks," *Comput. Netw.*, vol. 109807, pp. 109807, 2023.
[31] S. Shan, Y. Li, and J. B. Oliva, "NRTSI: Non-recurrent time series imputation," in *ICASSP 2023–2023 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Rhodes Island, Greece, IEEE, 2023, pp. 1–5.
[32] G. Wang, Q. Zhai, and H. Liu, "Cross self-attention network for 3D point cloud," *Knowl.-Based Syst.*, vol. 247, no. 5, pp. 108769, 2022. doi: 10.1016/j.knosys.2022.108769.
[33] A. Shahroudy, J. Liu, 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.*, Las Vegas, NV, USA, 2016, pp. 1010–1019.
[34] S. Xu, H. Rao, X. Hu, J. Cheng, and B. Hu, "Prototypical contrast and reverse prediction: Unsupervised skeleton-based action recognition," *IEEE Trans. Multimed.*, vol. 25, pp. 624–634, 2021.
[35] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, 2021. doi: 10.1016/j.ins.2021.04.023.
[36] J. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan, "Unsupervised feature learning of human actions as trajectories in pose embedding manifold," in *2019 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, IEEE, 2019, pp. 1459–1467.
[37] Y. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang and L. Lin, "Unsupervised representation learning for skeleton-based human action recognition," in *2021 IEEE Int. Conf. Multimed. Expo (ICME)*, Shenzhen, China, IEEE, 2021, pp. 1–6.
[38] Q. Nie, Z. Liu, and Y. Liu, "Unsupervised 3D human pose representation with viewpoint and pose disentanglement," in *Computer Vision–ECCV 2020: 16th European Conf.*, Glasgow, UK, Springer International Publishing, Aug. 23–28, 2020, pp. 102–118.
[39] C. Bian, W. Feng, F. Meng, and S. Wang, "Global-local contrastive multiview representation learning for skeleton-based action recognition," *Comput. Vis. Image Und.*, vol. 229, pp. 103655, 2023. doi: 10.1016/j.cviu.2023.103655.
[40] L. Li, M. Wang, B. Ni, and H. Wang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 4741–4750.
[41] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3D action representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 13423–13433.
[42] B. Kim, H. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *European Conf. Comput. Vis.*, Cham, Springer Nature Switzerland, Tel Aviv, Israel, 2022, pp. 209–225.

[43] 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.*, vol. 36, no. 1, pp. 762–770, 2022. doi: 10.1609/aaai.v36i1.19957.
[44] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu and X. Wang, "Hierarchical contrast for unsupervised skeleton-based action representation learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, pp. 525–533, 2023.
[45] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leung, "Focalized contrastive view-invariant learning for self-supervised skeleton-based action recognition," *Neurocomput.*, vol. 537, pp. 198–209, 2023.