IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 March 2024, accepted 4 April 2024, date of publication 15 April 2024, date of current version 14 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3389499

RESEARCH ARTICLE

Temporal-Channel Attention and Convolution Fusion for Skeleton-Based Human Action Recognition

CHENGWU LIANG[1,2], JIE YANG[1,2], RUOLIN DU[3], WEI HU[1,2], AND NING HOU[1]

[1]School of Electrical and Control Engineering, Henan University of Urban Construction, Pingdingshan, Henan 467036, China
[2]College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, Hubei 443002, China
[3]School of Transportation and Civil Engineering, Nantong University, Nantong, Jiangsu 226019, China

Corresponding author: Ning Hou (30090807@huuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62176086, and in part by the Science and Technology Development of Henan Province of China under Grant 242102211055 and Grant 232102210163.

ABSTRACT Human Action Recognition (HAR) based on skeleton sequences has attracted much attention due to the robustness and background insensitivity of skeletal data. The convolutional neural network (CNN) for spatio-temporal representation learning has been widely utilized for skeleton-based HAR. However, the long-term spatio-temporal modeling and action category-specific feature attention have not been fully exploited. In order to explore the current potential of CNNs for skeleton-based HAR, a novel CNN architecture with temporal-channel attention and convolution fusion is proposed. Specially, the network architecture is composed of two novel modules, the Temporal-Channels Attention Module (TCA) and Multiscale Temporal Convolution Fusion module (MTCF). TCA module is designed to generate a temporal-channel attention matrix for different visual channels and temporal features, motivating the CNN to focus on the critical category-associated feature representation learning. Along the channels, MTCF module adapts the grouped residual connections to flexibly extend the convolutional temporal receptive field, without introducing additional parameters. By reverse stacking, MTCF module creates a bidirectional information interaction among inter-channels, compensating for the receptive field and information imbalance between subgroups from different branches. The proposed method was evaluated on three benchmark datasets, including NTU RGB-D, NTU RGB-D120 and FineGYM. The results show that the proposed TCA-MTCF method improves the CNNs' ability to model long-term temporal features of skeleton sequences, achieving the state-of-the-art performance for HAR.

INDEX TERMS Skeleton-based, action recognition, attention mechanism, convolutional neural network, multi-scale convolution.

I. INTRODUCTION

Human action recognition, i.e., recognizing and classify human action categories, is one of the most fundamental and challenging tasks in computer vision, and has a wide range of applications in areas such as intelligent surveillance, human-computer interaction, game control and robotics [10], [23], [24], [41]. Compared to the action recognition methods based on the popular modality RGB or gray-scale videos, other data modalities such as depth or skeleton-based methods have received increasing attention in recent years [14], [31], [37]. As an abstract representation of motion, the human skeleton sequences is robust, informative, and has the characteristics of being light and background immunity, which makes it possible to design lightweight and hardware-friendly network models.

Graph Convolutional Networks (GCNs) [18], [37], [47] have become one of the most popular methods for skeleton-based action recognition due to their ability to construct irregular topological information of the

The associate editor coordinating the review of this manuscript and approving it for publication was Ajit Khosla.

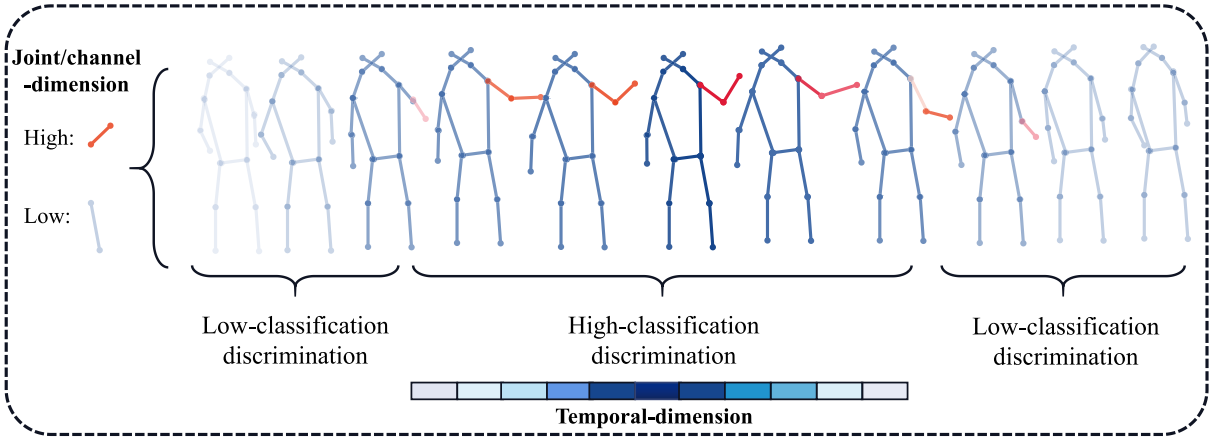


FIGURE 1. Skeleton nodes of different body parts contribute differently to the classification of action. Take skeleton action–"punch" for example, marker red is used to emphasize key skeletal joints, and color shade changes in the temporal dimension indicate the contribution to action recognition.

skeleton. Specifically, the GCN approach has the ability to topologically model non-Euclidean human skeleton data, aggregating spatio-temporal topological dynamic information of the neighbouring skeleton nodes through the design of adjacency matrices. Among the GCN methods and their evolving algorithms, spatial temporal GCNs (ST-GCN) [47] is the classical algorithm for skeleton action recognition, mainly using spatial graph convolution and temporal convolution to model spatial and temporal information. However, on the one hand, GCNs are not easy to model correlation and dynamic changing information between distant and unnaturally connected nodes. Moreover, GCN-based methods are not flexible enough and require complex neural network structure design to achieve fusion with other modal information [7], [49].

Compared to GCN-based methods, CNNs have powerful long-term modelling capabilities and flexible cross-modal fusion, allowing more effective extraction of spatio-temporal feature information and easier fusion with other modal data [7], [25]. Hence, several researchers have started to focus on using CNN methods to process skeleton data [1], [2], [7], [20]. Caetano et al. [2] convert skeleton coordinates into a three-channel pseudo-image and then classify the features extracted from the image. Li et al. [20] convert the skeleton sequence directly into a skeleton matrix and extracted features using a hierarchical aggregated 2D convolutional network. However, these well-designed methods still do not preserve the spatio-temporal information of the skeleton data well and perform slightly worse than GCN methods on mainstream benchmark datasets.

In order to preserve the spatio-temporal structure of the skeleton data, a new approach for skeleton-based action recognition called PoseC3D [7] converts the skeleton sequence into Euclidean data by 3D CNNs, exploiting the spatial information of the skeleton data without fully discovering the temporal dynamic of human motions. However, most 3D CNNs take the RGB video as input and utilize sparse sampling strategy by 8 to 16 spaced frames [3], [9], [42], [43].

In contrast, most publicly available skeleton-based action recognition datasets have a large number of input frames and temporal spans. It is believed that inter-frame interaction and spatio-temporal feature extraction of long time is crucial for HAR. How to effectively improve the long-term modeling capability of CNN models is one motivation for this paper. Hereby we adopt a multi-scale convolutional fusion strategy, inspired by [11], extending the spatial sense field of the 2D convolution to the temporal dimension and transforming it into a $3 \times 1 \times 1$ temporal convolution. Specifically, we divide the input feature maps into several groups. A group of temporal convolution kernels first extracts features from one group of input feature maps. Then output features of the previous group are sent to the next group of temporal convolution kernels along with another group of input feature maps. This process is repeated several times until all input feature maps are processed.

In addition, it is true that the channel dimension of the heatmap generated by the PoseC3D [7] method corresponds to the joints dimension of the skeleton data. Studies have shown [12], [13] that skeleton nodes of different body parts contribute differently to the action classification. For example, as shown in Figure 2, in "punching" action, the arm skeleton node is much more important to the classifications result than the other skeleton nodes of the body. And the contributions of key skeleton frames to action recognition varies from different moments. However, the existing CNN-based methods are limited in focusing on important skeleton joints. Therefore, how to effectively improve the CNNs model's ability to mine and model the features of key frames and important nodes is another research motivation of this paper.

With the aforementioned two motivations, we propose a novel CNN architecture, called TCA-MTCF, which consists of Temporal-Channels Attention module (TCA) and Multiscale Temporal convolution Fusion (MTCF) module. TCA is a attention enhancement module, which equivalently pays differential attention to each node at different

temporal stages. Actually, TCA constructs the temporal and channel attention weight coefficient matrices by end-to-end learning, and guides the network to focus on the category-specific feature representation learning. On the other hand, the proposed MTCF module adopts the temporal convolution layer approach of stacking two layered residuals to flexibly extract multi-grain temporal feature information. The MTCF module utilizes temporal convolution with diverse receptive field sizes in different group equivalents through a grouped convolution structure, and then fuses the information from each group to obtain multi-scale temporal information, including long time span information important for skeleton-based action recognition.

The proposed method is evaluated and validated on three benchmark datasets for skeleton-based HAR, NTU RGB-D [32], NTU RGB-D 120 [26] and FineGYM [33]. The results show that the method achieves the state-of-the-art performance for skeleton-based HAR.

The remainder of this paper is organized as follows. Section II gives a review of related work, including skeleton modality, GCN-based HAR and CNN-based HAR methods. Section III provides the details of our proposed TCA-MTCF method. In Section IV, we present the experimental setup, ablation study, attention map visualization, results and comparisons. In Section V, we conclude our paper.

II. RELATED WORK

A. SKELETON MODALITY

Due to the development of affordable depth sensors, pose estimation algorithms on RGB videos, vision motion capture systems and wearable suits with makers, skeleton data or skeleton sequence could acquired and utilized for HAR [38], [41]. The human skeleton data encodes the trajectories of human body joints, dynamic action pose and motion structure evolution, characterizing informative human motions. Therefore, based on skeleton data modality, HAR research community has witnessed anemergence of methods [44].

Generally, the affordable skeleton data estimated from depth maps or RGB videos has noise and varies with vision views. In contrast, the costly accurate skeleton data sensed by motion capture systems or wearable makers is robust for illumination and viewpoint variations. However, skeleton data has less information of human appearance and local detailed texture. In any case, the simple and easily accessible skeleton data has attracted much attention and been popular for computer vision community, especially for HAR researchers.

B. GCN METHOD FOR SKELETON-BASED ACTION RECOGNITION

Skeleton sequences involve with body structure and is naturally represented by graph models. Therefore, GCN takes skeleton sequences as input by joint dependency and structure dynamic learning for skeleton-based HAR [21], [30], [47]. Generally, GCN accomplishes the extraction of spatio-temporal information by modeling the topological relationships of the human skeleton with end-to-end fashion.

ST-GCN [47] is a well-known baseline work, which combines spatial graph convolution and interleaved temporal convolution for action spatio-temporal modeling. The specific GCN is as follows, $\mathbf{X}_{in} \in \mathbb{R}^{n \times d_{in}}$ denotes the input features of all joints (number is assumed as n) in a frame, d_{in} is the input feature dimension; $\mathbf{X}_{out} \in \mathbb{R}^{n \times d_{out}}$ denotes the GCN output features, d_{out} is the output feature dimension.

$$\mathbf{X}_{out} = \sum_{k}^{\mathbf{K}_a} (\mathbf{X}_{in}\mathbf{A}_k)\mathbf{W}_k \tag{1}$$

$$\mathbf{A}_k = \mathbf{D}_k{}^{-\frac{1}{2}}(\widetilde{\mathbf{A}}_\mathbf{k} + \mathbf{I})\mathbf{D}_k{}^{-\frac{1}{2}} \tag{2}$$

where $\mathbf{K}_a$ is the kernel size in spatial dimension, $\mathbf{A}_k$ is the adjacency matrix representing the human joint connections, $\mathbf{W}_k$ is the trainable weight matrix, and $\mathbf{I}$ is the unit matrix. $\mathbf{D}_k$ represents the degree matrix, normalized to the weights of each skeleton point.

Inspired by ST-GCN, AS-GCN [21] proposed a joint encoder-decoder structure to capture the potential dependencies contained in action sequences. Shi et al. [37] leveraged a multibranch architecture to build a two-stream adaptive graph convolutional network 2s-AGCN, which considers both joint information and bone information, and represents the bone information between joint points by calculating the vector difference of coordinates of adjacent joint points. Despite the great success of GCN for skeleton-based action recognition, its robustness and scalability are also limited since the disadvantages of noisy skeleton information and sparse modeling. Moreover, for GCN-based methods, fusing features from the skeleton and other modalities requires complex designs.

C. CNN METHOD FOR SKELETON-BASED ACTION RECOGNITION

In addition to the GCN methods, researchers have made efforts to leverage CNNs as main models for skeleton-based HAR work [1], [2], [7], [16], [20], [25], [50]. CNNs are proficient in processing data with spatial regularity such as images and RGB videos. However, CNNs are helpless to irregular skeleton sequences so that the skeleton data need further processing before taking as inputs.

In the 2D CNN-based approaches, the skeleton sequence is firstly converted into a pseudo image on the basis of a manually designed transformation. Banerjee et al. [1] encoded spatio-temporal features of skeleton sequences using four single-channel greyscale images, including skeleton distance and angle vectors, in order to classify them using four 2D CNNs. Zhang et al. [50] proposed VA-RNN and VA-CNN in order to improve the robustness of skeleton data in viewpoint change, which can self-adjust the observation viewpoint to improve the recognition accuracy. While other work [16], [20] directly converts the skeleton coordinates into a 2D matrix as a pseudo-image, usually generating a 2D

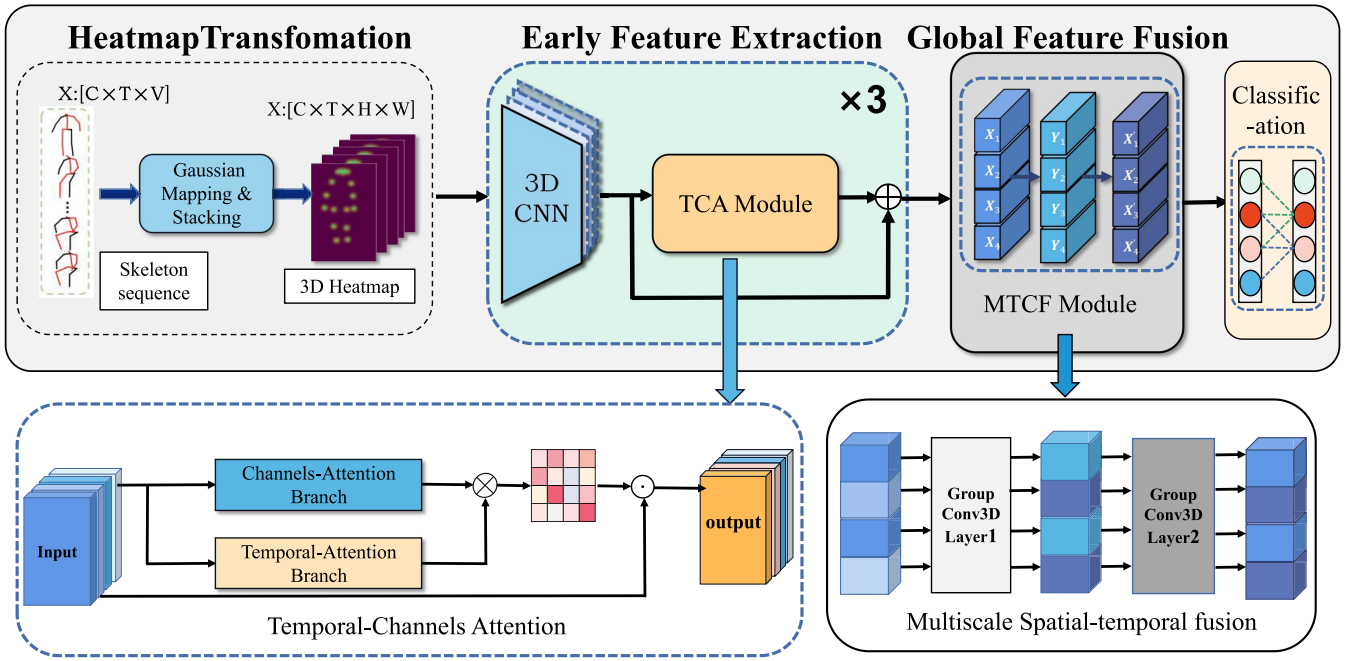


FIGURE 2. The proposed network architecture. This CNN architecture has two novel modules: temporal-channel attention (TCA) and multiscale temporal convolution fusion modules (MTCF).

input of shape K × T, where K is the number of joints and T is the length of time. Such inputs do not take advantage of the localization of convolutional networks, resulting in the inability to preserve the complete spatial structure when convolution aggregates the information.

Despite careful design, the aforementioned methods struggle to address the shortcoming of information loss in processing skeleton data, which results in lower performance than GCNs on mainstream datasets. In this paper, skeleton sequences are leveraged to generate 3D heatmap groups containing spatio-temporal information, which can solve the aforementioned deficiencies.

However, the current CNN methods for skeleton action recognition do not take into account the difference in the spatio-temporal dimension between the heatmap (generated by the skeleton) and the RGB video. Most of them directly use existing 3D-CNNs as backbone to process heat maps, without adapting the model to be more suitable for skeleton action recognition tasks. Therefore we incorporate Temporal-Channel attention mechanisms and multi-scale convolution to improve the ability of CNN models to extract features over long time spans. We therefore incorporate Temporal-Channel attention mechanisms and multi-scale convolution to improve the CNN model's ability to extract long time-span features. This can make CNNs more adaptable to the challenge of modelling global features over long temporal distances brought about by skeleton sequences.

III. PROPOSED METHOD

In this paper, a novel CNN-based network model with temporal-channels attention and multiscale temporal convolution fusion modules, is proposed for the task of skeleton-based action recognition. Generally, due to its spatial irregularity, transformation of the input skeleton sequence is required since the skeleton modal data is struggling to be effectively modeled by general CNN-based methods.

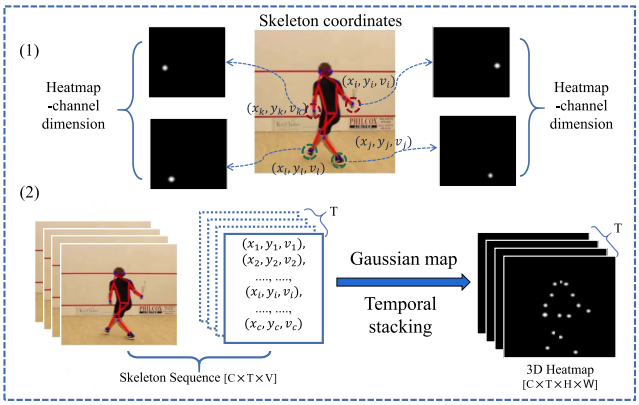


FIGURE 3. 3D heatmap generation. (1)Generation of single-channel heatmaps from individual skeleton coordinates. (2)Generation of a ($C \times T \times H \times W$) 3D heatmap group from the entire skeleton sequence.

As shown in Figure 2, the novel network architecture proposed in this paper consists of three key components: heatmap transformation by Gaussian mapping and stacking, early feature extraction by 3D CNN and the proposed TCA module, and global feature fusion by the proposed MTCF module. Specially, the method of generating heatmaps is to transform the skeleton sequences into regular Euclidean data for CNN models. The TCA module is composed of channel and temporal attention branches. The MTCF module is composed of two group 3D convolutional layers for

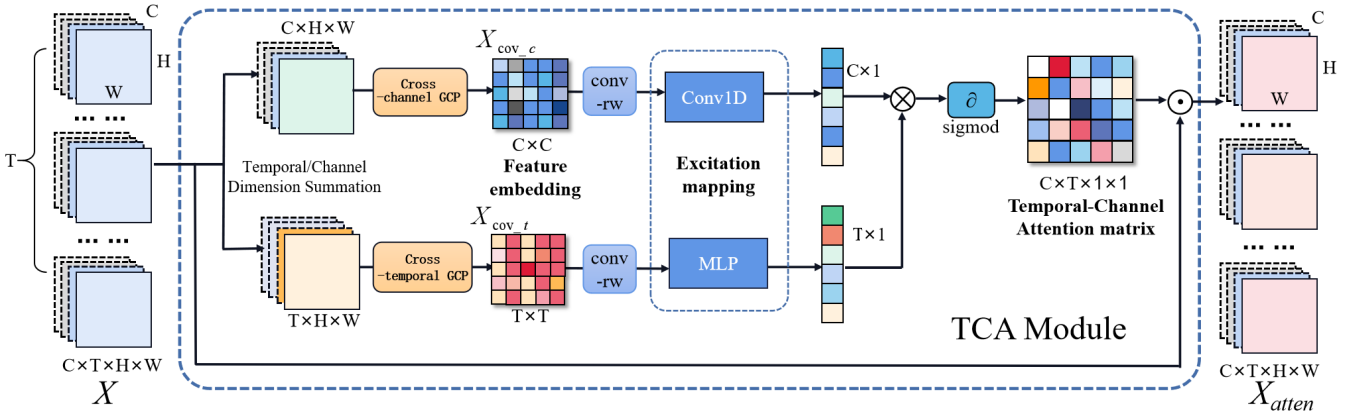


FIGURE 4. The proposed Temporal-channels Attention module (TCA). The TCA module consists of two branches, temporal-wise attention and channel-wise attention branches, each of which includes feature embedding (global covariance matrix and row convolution) and excitation mapping.

multiscale spatial-temporal fusion. Below is the detailed introduction to the methods proposed in this article.

A. HEATMAP TRANSFORMATION

CNN models is stumblingly directly extract the topological relationship of the skeleton sequence.To overcome the shortcoming of CNNs, the method of heatmap mapping are adopted to transform the skeleton node coordinates of each frame to a heatmap X of size 56×56 through Gaussian mapping. Through the heatmap generated by the skeleton node coordinates, the spatial relationship and natural structure between the skeleton nodes can be clearly constructed, and the general CNN network can also extract spatial features from it. The specific formula is as follows:

$$\mathbf{X}_{c_{i,j}} = e^{-\frac{(i-x_c)^2+(j-y_c)^2}{2*\sigma^2}} * V_c \tag{3}$$

where σ is used to control the variance of Gaussian mapping, with i and j representing the pixel positions in the heat map space, and x_c,y_c,V_c are the positions and confidence scores of the c-th joint, respectively. Actually, C is the number of channels in a single frame representing the total number of nodes. In different channels of the heatmap coming from different node mappings, V_c represents the node confidence score. Eq.3 indicates that the c th node is used as the centre of the mapping to generate the heatmap of the corresponding channel. Then the generated heatmaps of each frame along the time dimension T are stacked to obtain a 3D heatmap group of shape $[C \times T \times H \times W]$.

As shown in Figure 3, takes the 2D pose obtained by the modern pose estimator as input. The 2D pose is represented by the heatmap stack of bone joints, rather than the coordinates operated on the human skeleton map. The 3D heatmap group is completely preserved as a pseudo video. The spatio-temporal information of the skeleton sequence converts the irregular skeleton sequence into Euclidean data that can be processed by the CNN network, which can be further sent to the backbone networks for feature extraction and then actions identification.

B. TEMPORAL-CHANNELS ATTENTION MODULE (TCA)

As shown in Figure 4, the TCA module is proposed to improve the representation learning of important skeleton joints. 3D heatmap sets are first fed into 3D convolutional layers to obtain middle-level features $\mathbf{X} \in \mathbb{R}^{C\times T\times H\times W}$, and then fed into the TCA module for attention enhancement. The TCF module proposed comprises of three stages, including feature embedding, excitation mapping, and attention matrix generation. The two stages of feature embedding and excitation mapping employ a two-branch structure of the temporal and channel dimensions, and finally aggregate to generate the attention weight matrix. In the feature embedding stage, we derive feature vectors through the summation of features along the temporal and channel dimensions, covariance pooling, and row-wise convolution, respectively. Then, we produce the attention weight vector through 1D convolution and MLP linear mapping using temporal or channel branching. Subsequently, the weight vectors of the two branches are multiplied through matrix broadcast and activation function to obtain the attention matrix.The specific steps are as follows:

- Step 1: The input feature $\mathbf{X} \in \mathbb{R}^{C\times T\times H\times W}$ is accumulated and summed along the temporal ($\mathbf{T}$ dimension) and the channel ($\mathbf{C}$ dimension) respectively. Thus, the input features $\mathbf{X}$ are aggregated into $\mathbf{X}_c \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{X}_t \in \mathbb{R}^{T\times H\times W}$, then sent to different branches for processing.
- Step 2: Calculate the cross-channel and cross-temporal covariance matrices $\mathbf{X}_{cov\_c} \in \mathbb{R}^{C\times C}$ and $\mathbf{X}_{cov\_c} \in \mathbb{R}^{T\times T}$ for $\mathbf{X}_c$ and $\mathbf{X}_t$. Specifically, max pooling or average pooling only utilizes the first-order information of the features, whereas global covariance pooling is done by calculating the covariance matrix (second-order information) of the feature map to select this value that is representative of the distribution of the data in the feature map. Inspired by [25], we use cross-channel and cross-time covariance pooling. For the channel attention branch, first transform the features $\mathbf{X}_c \in \mathbb{R}^{C\times H\times W}$ into $\mathbf{X}'_c \in \mathbb{R}^{C\times N}$. Then the features $\mathbf{X}'_c$ are grouped along the

channel dimension C to get $\mathbf{f}_i \in \mathbb{R}^{1\times N}, (i = 1, 2, .., c)$. $\mathbf{X}_{cov\_c}$ is defined as follows:

$$\mathbf{X}_{cov\_c} = \begin{bmatrix} \text{cov}(\mathbf{f}_1, \mathbf{f}_1) & \text{cov}(\mathbf{f}_1, \mathbf{f}_2) & \cdots & \text{cov}(\mathbf{f}_1, \mathbf{f}_c) \\ \text{cov}(\mathbf{f}_2, \mathbf{f}_1) & \cdots & \cdots & \text{cov}(\mathbf{f}_2, \mathbf{f}_c) \\ \vdots & \ddots & \ddots & \vdots \\ \text{cov}(\mathbf{f}_c, \mathbf{f}_1) & \cdots & \cdots & \text{cov}(\mathbf{f}_c, \mathbf{f}_c) \end{bmatrix} \tag{4}$$

$$\text{cov}\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{1}{N-1}\sum_{k=1}^{N}\left[\mathbf{f}_i(k) - E\left(\mathbf{f}_i\right)\right]\left[\mathbf{f}_j(k) - E\left(\mathbf{f}_j\right)\right] \tag{5}$$

where $E(\mathbf{f}_i)$ represents the feature average of the i-th channel. $\text{cov}(\mathbf{f}_i, \mathbf{f}_j)$ calculates the covariance between the features of channels i and j to represent the inter-channel correlation. The same is true for the cross-time covariance matrix $\mathbf{X}_{cov\_c}$ of the temporal branch. The process of obtaining $\mathbf{X}_{cov\_t}$ from $\mathbf{X}_t$ is consistent with the aforementioned process of $\mathbf{X}_c$ to X_{cov_c}. Therefore, $\mathbf{X}_{cov\_t}$ is defined as follows:

$$\mathbf{X}_{cov\_t} = \text{Cov}_{temporal}(\text{reshape}(\mathbf{X}_t)) \tag{6}$$

where $\text{Cov}_{temporal}$ represents the function of generating covariance matrix in the temporal dimension, and reshape(.) operation transforms $\mathbf{X}_t \in \mathbb{R}^{T\times H\times W}$ to $\mathbf{X}'_t \in \mathbb{R}^{T\times N}$.

- Step 3: The next step is the attention excitation process. In the channel-wise attention branch, this paper adopts row-wise convolution, cross-channel 1D convolution and Relu activation function to map the channel embedding covariance matrix into a low-dimensional weight vector. In the temporal attention branch, the work of this paper adopts the method of row-wise convolution and 3-layers MLP for embedding mapping, in order to obtain more accurate prediction of weight coefficients. Then, we perform matrix multiplication and sigma activation functions on the two sets of weight coefficient vectors for the temporal and channel dimensions to obtain the temporal channel attention matrix. Finally, the temporal channel attention matrix was used to multiply the elements with the original features for attentional activation.

$$\mathbf{X}_{Atten} = \sigma\left(\text{Conv 1D}_1\left(\text{Conv\_rw}\left(\mathbf{X}_{\text{cov}\_c}\right)\right) * \text{MLP}\left(\text{Conv\_rw}\left(\mathbf{X}_{\text{cov}\_t}\right)\right)\right) \otimes \mathbf{X} \tag{7}$$

where Conv_rw(.) represents row-wise convolution, $\mathbf{X}, \mathbf{X}_{Atten} \in \mathbb{R}^{C\times T\times H\times W}$ are the input and output of the module respectively, Conv1D(.) is a 1D convolutional layer, and $\sigma(.)$ is an activation function employed to generate attention weights.

C. MULTI-SCALE TEMPORAL CONVOLUTION FUSION MODULE (MTCF)

The architecture of proposed MTCF module is shown in Figure 5. The input feature, obtained by previously mentioned TCA module, is divided into four subsets along the channel dimension $[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4]$, where $\mathbf{X}_i \in \mathbb{R}^{C/4\times T\times H\times W}$. This module designs a four-grouped 3D convolutional structure with two segments. In the first hierarchical residual convolutional layer, the branch uses a 3D convolution with a convolution kernel size of 1 to output $\mathbf{Y}_1$, and the temporal receptive field is 1. The grouped features in the $\mathbf{X}_2$ branch are convolved with a kernel size of $3 \times 1 \times 1$ resulted in $\mathbf{Y}_2$, and then fused with the 3th group of $\mathbf{X}_3$ branch to continue the convolution operation with kernel size $3 \times 1 \times 1$. This process is equivalent to deepening the number of convolutional layers and expanding the size of the receptive field. The 4th group of $\mathbf{X}_4$ branches is the same to $\mathbf{X}_3$. The receptive field of the $\mathbf{X}_4$ branch is 7, which means that features modelled over the time span of 7 frames are obtained.

The detailed multi-scale convolution equivalent process of $\mathbf{Y}_4$ branch features is demonstrated in Figure 6. The $\mathbf{X}_4$ is fused with the previous two groups of features. This process further deepens the convolution to obtain the informative features of three distinctive receptive fields with different temporal scales. This processing enhances the multi-granularity representation ability and long-term construction of the convolution model. $\mathbf{Y}_i$ represents output subset feature of the i-th branch. Specifically we can obtain $\mathbf{Y}_i$ from $\mathbf{X}$ according to the following equation:

$$\mathbf{Y}_i = \begin{cases} \text{Conv 3D}\left(\mathbf{X}_i\right) & i = 1, 2 \\ \text{Conv 3D}\left(\mathbf{X}_i + \alpha_{1,i-2} \cdot \mathbf{Y}_{i-1}\right) & i > 2 \end{cases} \tag{8}$$

where the trainable parameter $\alpha_{1,i-2}$ is to control the fusion scale coefficient of other branch.

The receptive fields of the output features of branches 1 to 4 are progressively larger, which leads to an imbalance in the temporal modelling of the individual grouped channels (the first branch only interacts with itself). In addition, the convolutional fusion strategy we used only the latter branch can fuse the feature information from the former branch (the last $\mathbf{Y}_4$ branch fuses the feature information from $\mathbf{X}_1$ to $\mathbf{X}_4$), which leads to the lack of bidirectional transfer of information resulting in an imbalance in the amount of feature information of each subgroup. Therefore the second segment of hierarchical residual convolution is added, as shown in Figure 5. Taking the output feature $\mathbf{Y}$ in the first section as input, we perform hierarchical residual convolution from $\mathbf{Y}_4$ to $\mathbf{Y}_1$ feature subsets in the reverse order in the first section.

Subsequently, the feature subsets $\mathbf{X}'_i \in \mathbb{R}^{C/4\times T\times H\times W}$ are concatenated by channel-wise and then residually concatenated with the original feature $\mathbf{X}$ to obtain the final output $\mathbf{X}_{out} \in \mathbb{R}^{C\times T\times H\times W}$.

$$\mathbf{X}'_i = \begin{cases} \text{Conv 3D}\left(\mathbf{Y}_i\right) & i = 3, 4 \\ \text{Conv 3D}\left(\mathbf{Y}_i + \alpha_{2,i+2} \cdot \mathbf{X}'_{i+1}\right) & i < 2 \end{cases} \tag{9}$$

$$\mathbf{X}_{out} = \text{concat}(\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3, \mathbf{X}'_4) + \mathbf{X} \tag{10}$$

It should be noted that compared with the general 3D convolutional layer, our proposed module not only flexibly improves the temporal receptive field and enhances

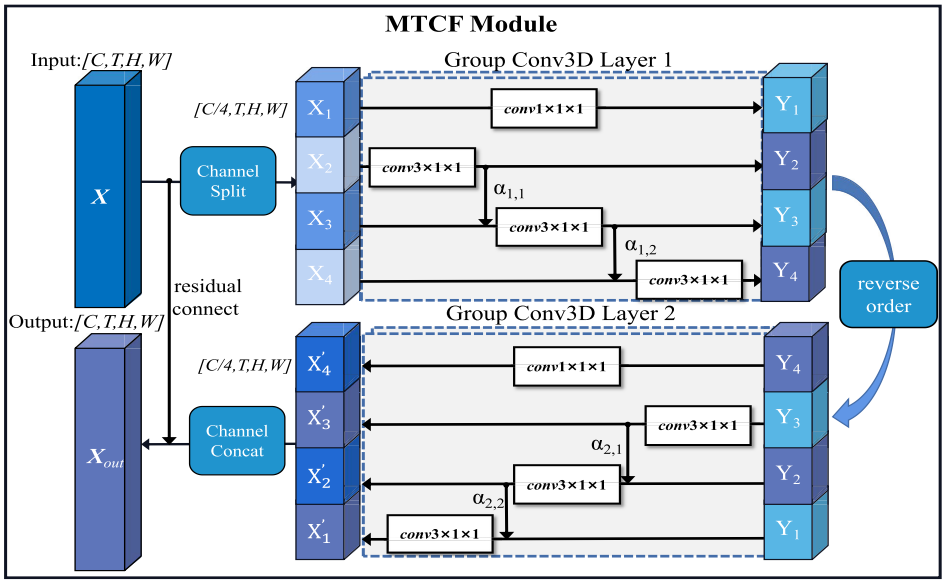


FIGURE 5. Illustration of the porposed Multiscale Temporal convolution Fusion module (MTCF).

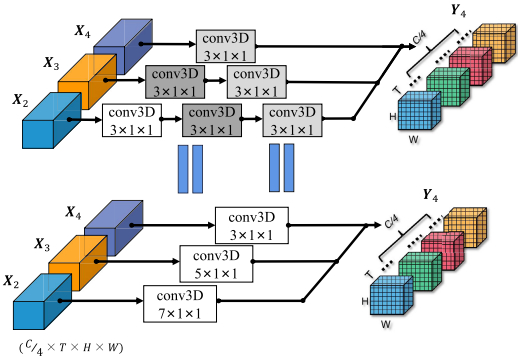


FIGURE 6. Illustration of the multiscale convolution equivalent process of Y_4 branch features generation.

cross-channel feature interaction, but also reduces the amount of parameters to a certain extent. For a convolution with channel number C, after dividing N groups along the channel, the number of channels in each grouped convolution layer is C/N, and the parameter amount of each grouped convolution becomes $1/N^2$ of the original number. Consequently, the overall parameter amount of the proposed convolution layer consisting of n grouped convolutions is reduced to $1/N$. As a result, the MTCF module saves more parametric counts than the fashion manner of adding normal 3D convolutional layers.

IV. EXPERIMENTS

A. DATASETS AND IMPLEMENTATION DETAILS

1) DATASETS

The experiments and performance evaluation of the proposed method is conducted on public datasets including NTU RGB-D [32], NTU RGB-D 120 [26] and FineGYM [33].

NTU RGB-D [32] is a large-scale available human action recognition datasets. It contains over 56K video samples of 60 human action classes performed by 40 different human subjects. Following the authors of this dataset recommendation, we process this dataset into two benchmarks: cross-subject(X-sub) and cross-view(X-view). In the cross-subject setting, sequences of 20 subjects are for training, and the sequences of the rest 20 subjects are for validation. In the cross-view setting,skeleton sequences are split by camera views. Samples from two camera views are used for training, and the rest are used for evaluation.

NTU RGB-D 120 [26] is an extension of NTU RGB-D datasets, and it is currently the largest datasets by adding 57k video samples of 60 action classes, containing 113k samples of 120 human action classes performed by 106 human subjects. The authors offered the cross-subject(X-sub) and cross-setup(X-set) as two benchmark evaluations. In the cross-subject setting, sequences from 53 subjects are for training, and sequences from the other 53 subjects are for testing. In the cross-setup setting, skeleton sequences are split by setup ID. Samples from even set-up IDs are used for training, and the odd setup IDs are used for evaluation.

FineGYM [33] is a fine-grained action recognition dataset. It contains 29K videos of 99 fine motor categories collected from 300 professional gymnastics competitions. As shown in Figure 7, the FineGYM dataset is extracted using the HRNet [40] pose estimation algorithm to obtain 2D skeleton data and perform heatmap transformation. The FineGYM dataset differs from existing action recognition datasets in several ways, including high-quality and action-centric data, semantically and temporally consistent annotations across multiple granularities, and diverse and informative rich action examples.

We choose Mean Top-1 accuracy for FineGYM and Top-1 accuracy for NTU RGB-D and NTU RGB-D 120 datasets. In order to compare the fairness of the experiments, we used only skeleton points as a single modal input in all the experiments and not skeleton bone, velocity as supplements.

Order	body parts
1	nose
2	l-eye
3	r-eye
4	l-ear
5	r-ear
6	l-shoulder
7	r-shoulder
8	l-elbow
9	r-elbow
10	l-wrist
11	r-wrist
12	l-hip
13	r-hip
14	l-knee
15	r-knee
16	l-ankle
17	r-ankle

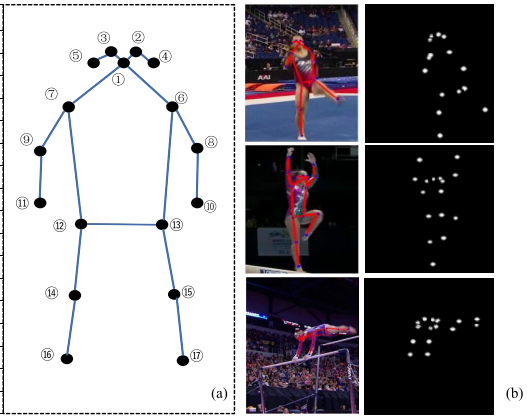


FIGURE 7. Illustration of skeleton joints and corresponding human body parts. (a) The order of the joint points corresponds to the human body parts. (b) FineGYM dataset skeleton extraction and heatmap transformation.

2) IMPLEMENTATION DETAILS

Our proposed method is implemented on Pytorch, using NVIDIA RTX 3090 GPUs for training and testing, with the BatchSize of 32. The model is trained for 30 epochs with the SGD optimizer and the decays using the cosine annealing optimizer. The initial learning rate is set to 0.15 and the momentum of the SGD optimizer is set to 0.9. Weight decay is set to 0.0003. Skeleton joint points are estimated by HRNet [40] to obtain 2D coordinates, in which 17 joint points are selected, and the corresponding human body parts are shown in Figure 7(a).

3) MODEL DESIGN

As shown in Figure 8, we instantiated 3D-CNN using slow-only as the backbone and inserted three TCA and MTCF modules. Where the TCA is inserted into Conv1, ResNet layer2 and ResNet3 of the backbone respectively following setting the size of the Temporal-Channel Attention Maps of the TCA as (C=32, T=48), (C=128, T=48), (C=256, T=24) separately. For the MTCF module, we divide it into a two-stage, four-group structure, with a $3 \times 1 \times 1$ temporal convolution for each sub-convolution kernel.

B. ABLATION EXPERIMENT

Table 1 and Table 2 show the performance improvement brought by the TCA and MTCF modules on FineGYM dataset. The performance evaluations are tested with different backbone networks to verify the generalization of the proposed method. We evaluated the mean Top-1 accuracy index using baseline models with Slow-Only [9] and C3D-light [42] as backbones, to which TCA and MTCF modules are added.

1) TCA MODULE ABLATION EXPERIMENT

As shown in Table 1, compared with the basedline methods, the mean Top1 recognition accuracy of two different backbone networks are improved by 0.5% and 0.9% respectively. The experimental results validate that TCA is effectiveness. During heatmap generation, the point dimension of skeleton

TABLE 1. TCA module ablation experiments on FineGYM datasets.

Method	Backbone:C3D-light		Backbone:slow-only	
	TOP-1(%)	Params	TOP-1(%)	Params
Baseline	90.9	3.40M	93.2	2.04M
+TCA	91.8	3.55M	93.7	2.19M
+TCA (no-covariance)	91.4	3.55M	93.4	2.19M
+TCA+MTCF	**93.3**	**3.93M**	**94.1**	**2.57M**

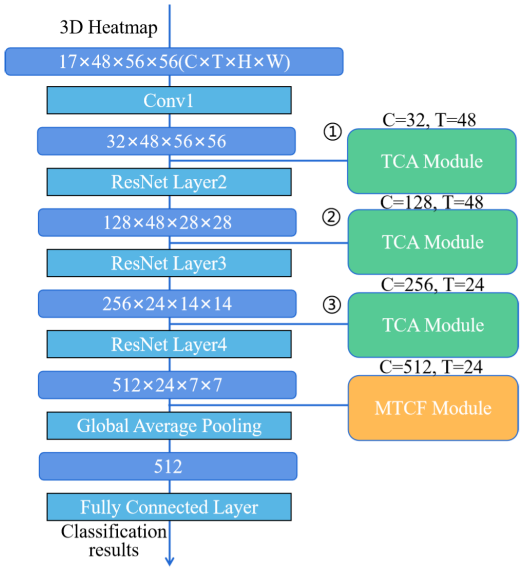


FIGURE 8. TCA-MTCF model design.

sequences corresponds to the heatmap channel-dimension. And it is generally believed that focusing on the features of the key skeleton points is necessary for skeleton-based HAR. Experimental results also demonstrate that temporal-channel attention can effectively improve recognition accuracy, corroborating the aforementioned viewpoints.

In order to verify the performance improvement by the global covariance of TCA module, consequent experiments in which average pooling is utilized directly (instead of global covariance) to compress feature information. As shown in the Table 1, the performance of TCA decreases from 91.8% to 91.4%, and 93.7% to 93.4% respectively. It shows that the global covariance matrix (instead of average pooling) effectively preserves the global statistical information across channels and temporal dimensions, and better reflects the data distribution of the input features. Furthermore, the overall amount of parameters of the proposed method is increased into 3.55M, which is comparable to the C3D-light baseline model (3.40M).

2) MTCF MODULE ABLATION EXPERIMENT

Table 2 records the ablation experiments of the MTCF module. In order to verify that the MTCF performance improvement is not simply by deepening the number of network layers, we set up the "Baseline+1layer" experiment. "Baseline+1layer" indicates the experimental results of adding a layer of 3D convolution(the convolution kernel

TABLE 2. MTCF module ablation experiments on FineGYM datasets.

Method	Backbone:C3D-light TOP-1(%)	Params	Backbone:slow-only TOP-1(%)	Params
Baseline	90.9	3.04M	93.2	2.04M
Baseline(+1layer)	91.1	5.17M	93.0	3.81M
+MTCF ($3 \times 3 \times 3$)	92.5	4.29M	93.5	2.93M
+MTCF ($3 \times 1 \times 1$)	**93.0**	**3.79M**	**93.9**	**2.43M**

is $3 \times 3 \times 3$ and the number of channels is 256) at the end of the backbone, compared with MTCF. The experimental results show that mean top1 only changes from 90.9% to 91.1% and from 93.2% to 93.0% at different backbone(C3D-light and slow-only) respectively. Compared with "Baseline+1layer", the experiment results demonstrate that the MTCF module adopts the convolutional fusion strategy to obtain performance improvement by improving the long time modelling capability and the inter-channel feature information interaction.

Also the large number of convolutional kernels in the added 3D convolutional layer leads to a significant increase in the overall parameter count, boosting more than 1M parameters. As shown in Table 2, in the case of "Baseline(+1layer)" experiment, the number of parameters with C3D-light and slow-only as the backbone increases from 3.04M to 5.17M, and 2.04M to 3.81M, respectively. In contrast MTCF as a novel 3D convolutional layer, just increases 0.75M and 0.39M additional parameters, respect to two different backbones.

In addition, the influence of using convolution kernels of different sizes ($3 \times 3 \times 3$ and $3 \times 1 \times 1$) in the MTCF module is also evaluated. It is worth noting that MTCF with $3 \times 1 \times 1$ convolutional kernel achieves 0.5% and 0.4% higher performance than MTCF with $3 \times 3 \times 3$ convolutional kernel on different backbone, respectively. Moreover MTCF with $3 \times 1 \times 1$ convolutional kernels has higher fewer parameters than $3 \times 3 \times 3$. The reason is that MTCF module is at the end of the backbone, and the input data are high-level features with very small spatial dimensions, so there is no need to expand the spatial receptive field.

C. ATTENTION MAP VISUALIZATION

As shown in Figures 9 and 10, we selected a sample of the FineGYM dataset with the category "Gymnastics" for action recognition, and output a visualization of the attention matrix of the TCA module. Specifically, as shown in Figure 10, three TCA modules are inserted in our proposed network. TCA attention map visualizations ($T = 48, C = 32$) in the first feature extraction phase are displayed in Figure 10, where the color depth indicates the magnitude of the attention weight coefficients. The darker the color, the more important this channel or frame is in action. The x-axis represents the temporal dimension of the features feeding into the TCA module and the y-axis represents the channel dimension of the features.

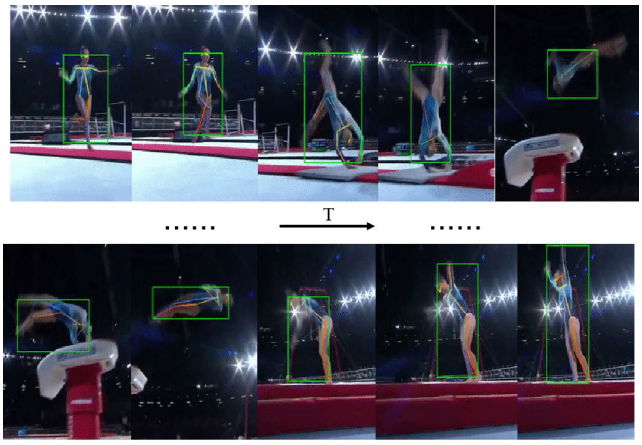


FIGURE 9. FineGYM dataset "Gymnastics" category sample.

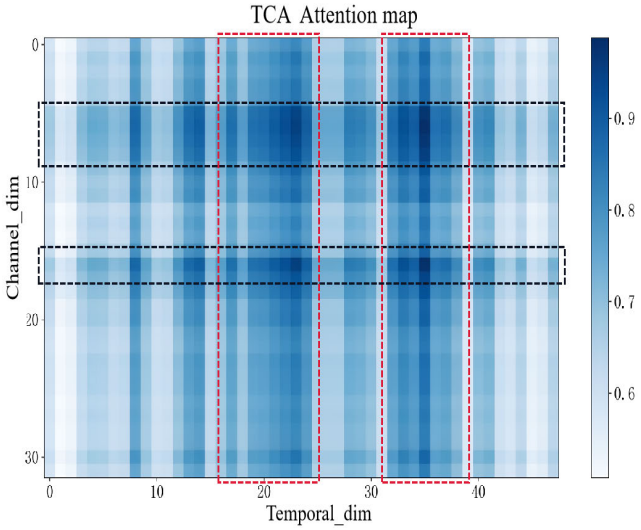


FIGURE 10. TCA attention map visualization ($T = 48, C = 32$) in the FIRST feature extraction phase.

It can be seen in Figure 10 that the attention matrix given by TCA gives higher response at $T = 16 - 25$ and $T = 34 - 38$ (marked by the red box in Figure 10). Also at $Channel = 5 - 8$ and $Channel = 17 - 19$ (marked by the black box in Figure 10), the TCA module gives higher attention weight coefficients. It can be observed that for "gymnastic" action, the mid-time "jumps" and "falls" are critical phases and are of interest to the TCA. The "walking" process in the beginning of the action is almost ignored.

In conclusion, the TCA module outputs the attention weight matrix with discretization and focusing, which highlights the important parts of the features extracted by the enhanced CNN in the temporal and channel dimensions, reducing the noise interference of the non-correlated features for HAR.

D. COMPARATIVE ANALYSIS OF EXPERIMENTAL PERFORMANCE

In this section, we evaluate the performance of the proposed methods on three benchmark datasets: NTU RGB-D, NTU

TABLE 3. Performance comparison on NTU RGB-D datasets.

Type	Method	X-sub(%)	X-View(%)
RNN – Based	GCA-LSTM [28]	74.4	82.8
	TS-LSTM [17]	74.6	81.3
	VA-LSTM [51]	79.4	87.6
	VA-RNN [50]	79.8	88.9
	dense-IndRNN [22]	86.7	93.7
	MANs [19]	82.7	93.2
CNN – Based	DSTA-Net [36]	91.5	96.4
	Ta-CNN+ [46]	90.7	95.1
	RotClips+MTCNN [15]	81.1	87.4
	SkeleMotion [2]	76.5	84.7
	Skepxel [25]	81.3	89.2
	liu *et al.* [29]	80.0	87.2
	VA-CNN [50]	88.7	94.3
	Banerjee *et al.* [1]	84.2	89.7
	HCN [20]	86.5	91.1
GCN – Based	ST-GCN [47]	81.5	88.3
	AS-GCN [21]	86.8	94.2
	RA-GCN [39]	87.3	93.6
	Shift-GCN [5]	90.7	96.5
	SGN [52]	89.0	94.5
	MS-G3D [37]	91.5	96.2
	FGCN [48]	90.2	96.3
Ours	**CNN with TCA-MTCF**	**93.8**	**96.9**

TABLE 4. Performance comparison on NTU RGB-D 120 datasets.

Type	Method	X-sub(%)	X-set(%)
RNN – based	Trust Gate ST-LSTM [27]	25.5	26.3
	ST-LSTM [27]	58.2	60.9
	GCA-LSTM [28]	58.3	59.2
CNN – Based	Banerjee *et al.* [1]	74.8	76.9
	Clips+CNN+MTLN [16]	58.4	57.9
	Liu *et al.* [29]	60.3	63.2
	SkeleMotion [2]	67.7	66.9
GCN – based	ST-GCN [47]	70.7	73.2
	AS-GCN [21]	78.3	79.8
	3s-AdaSGN [35]	85.9	86.8
	Shift-GCN [5]	85.9	87.6
	MS-G3D [37]	86.9	88.4
Ours	**CNN with TCA-MTCF**	**86.6**	**89.9**

RGB-D 120 and FineGYM. Many state-of-the-art methods employ multi-stream fusion models [7], namely fusing the information with skeleton joints and nodes. In order to make a fair comparison, the evaluation model in this paper is compared with the state-of-the-art methods obtained by single-stream models on each datasets, and the results show that the proposed methods achieve fine performance.

On NTU RGB-D dataset, the results shown in Table 3 demonstrate our model achieves 93.8%, 96.9% on the X-Sub and X-View settings respectively, which is significantly 3.6%, 7.3%, 9.6%,7.0% better than methods [1], [20], [21], [48]. Although the performance of the X-View benchmark is close to saturation, the proposed model still achieves remarkable performance, achieving an accuracy of 93.8% on the X-Sub benchmark. Compared with other CNN-based methods, our proposed method achieves significant enhancement and effectively improves the applicability of CNN in skeleton action recognition.

TABLE 5. Performance comparison on FineGYM datasets.

Method	Fine-GYM:Mean Top-1 Accuracy (%)
ST-GCN [47]	25.2
Jing Shi *et al.* [34]	83.7
InfoGCN [6]	92.0
CTR-GCN [4]	91.9
PoseC3D [7]	93.7
MS-G3D+ [37]	92.6
Ours (CNN with TCA-MTCF)	**94.1**

On the challenging NTU RGB-D 120 dataset, our method also has favorable performance indicators. As shown in Table 4, we obtain 1.5% improvements for X-Set benchmarks compared with the state-of-the-arts. Furthermore, as shown in Table 5, the results evaluated on FineGYM dataset show that our proposed method achieves the state-of-the-art performance for skeleton-based HAR with 94.1% accuracy, which is 1.5% higher than the state-of-the-art GCN-based methods. It means that when addressing with large motion deformation and fast displacement transformation, our model has higher performance than GCN-based methods.

V. CONCLUSION

In skeleton-based HAR, long-term spatiotemporal modeling and action category-specific feature attention mechanisms are not fully exploited in CNN. To address the challenges, a novel CNN network architecture consists attention and multiscale temporal convolution fusion modules, is proposed for skelton-based HAR. This novel CNN network, called TCA-MTCF, composes two new modules, i.e., Temporal-Channels Attention Module (TCA) and Multiscale Temporal Convolution fusion Module (MTCF). The performance evaluations on three benchmark datasets (NTU RGB-D, NTU RGB-D120 and FineGYM), including ablation study, attention map visualization and the comparisons, demonstrate that proposed method is effective for skeleton-based HAR with CNN fashion.

REFERENCES

[1] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2206–2216, Jun. 2021.

[2] C. Caetano, J. Sena, F. Brémond, J. A. D. Santos, and W. R. Schwartz, "SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[3] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[4] 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. (ICCV)*, Oct. 2021, pp. 13339–13348.

[5] 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. (CVPR)*, Jun. 2020, pp. 183–192.

[6] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20154–20164.

[7] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.

[8] N. E. Elmadany, Y. He, and L. Guan, "Improving action recognition via temporal and complementary learning," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 3, pp. 1–24, Jun. 2021.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Mali, Oct. 2019, pp. 6201–6210.

[10] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson, and C. Fookes, "Continuous human action recognition for human-machine interaction: A review," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–38, Dec. 2023.

[11] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[12] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 601–610.

[13] P. Han, Z. Ma, and J. Liu, "Topology-embedded temporal attention for fine-grained skeleton-based action recognition," *Appl. Sci.*, vol. 12, no. 16, p. 8023, Aug. 2022.

[14] K. Hu, J. Jin, F. Zheng, L. Weng, and Y. Ding, "Overview of behavior recognition based on deep learning," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 1833–1865, Mar. 2023.

[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3D action recognition," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2842–2855, Jun. 2018.

[16] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.

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

[18] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8561–8568.

[19] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, and J. Chen, "Memory attention networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4800–4814, Sep. 2022.

[20] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, *arXiv:1804.06055*.

[21] 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. (CVPR)*, Jun. 2019, pp. 3590–3598.

[22] S. Li, W. Li, C. Cook, and Y. Gao, "Deep independently recurrent neural network (IndRNN)," 2019, *arXiv:1910.06251*.

[23] C. Liang, L. Qi, Y. He, and L. Guan, "3D human action recognition using a single depth feature and locality-constrained affine subspace coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2920–2932, Oct. 2018.

[24] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7083–7093.

[25] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," in *Proc. CVPR Workshops*, 2019, pp. 10–19.

[26] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[27] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[28] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1647–1656.

[29] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[30] M. Liu, F. Meng, C. Chen, and S. Wu, "Novel motion patterns matter for practical skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023, pp. 1–11.

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

[32] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.

[33] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2613–2622.

[34] J. Shi, Y. Zhang, W. Wang, B. Xing, D. Hu, and L. Chen, "A novel two-stream transformer-based framework for multi-modality human action recognition," *Appl. Sci.*, vol. 13, no. 4, p. 2058, Feb. 2023.

[35] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "AdaSGN: Adapting joint number and model size for efficient skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13393–13402.

[36] L. Shi, Y. Zhang, and J. Cheng, "Decoupled spatial–temporal attention network for skeleton-based action-gesture recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–12.

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[38] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *J. Vis. Commun. Image Represent.*, vol. 76, Apr. 2021, Art. no. 103055.

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

[40] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[41] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[42] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[43] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5551–5560.

[44] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020.

[45] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.

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

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

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

[49] W. Yang, J. Zhang, J. Cai, and Z. Xu, "HybridNet: Integrating GCN and CNN for skeleton-based action recognition," *Appl. Intell.*, vol. 53, no. 1, pp. 574–585, 2023.

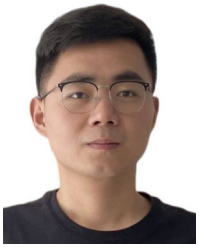
[50] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[51] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2136–2145.
[52] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1109–1118.

CHENGWU LIANG received the M.A.Sc. degree in information and communication engineering from the University of Electronic Science and Technology of China, China, and the Ph.D. degree in information and communication engineering from Zhengzhou University, China. From 2015 to 2017, he was a Visiting Ph.D. Student with Ryerson University, Toronto, ON, Canada. He is currently an Associate Professor with Henan University of Urban Construction, China. His current research interests are interpretable artificial intelligence, including statistical machine learning, computer vision, and statistical pattern recognition, especially video understanding. He was a recipient of the top five papers award from the 2017 IEEE International Conference on Visual Communication and Image Processing.

JIE YANG was born in Hubei, China. He is currently pursuing the M.S. degree in electrical engineering with China Three Gorges University, Yichang, China. He is also a Visiting Student with Henan University of Urban Construction, China. His research interests include computer vision, pattern recognition, and multi-modality information fusion.

RUOLIN DU received the bachelor's degree in traffic equipment and control engineering from Nantong University, China, where she is currently pursuing the master's degree in artificial Intelligence with the School of Transportation and Civil Engineering. She is also a Visiting Student with Henan University of Urban Construction, China. Her research interests include machine learning, computer vision, including object detection, image defogging, de-rain, and transfer learning.



WEI HU was born in Anhui, China. He is currently pursuing the M.S. degree in electrical engineering with China Three Gorges University, Yichang, China. He is also a Visiting Student with Henan University of Urban Construction, China. His research interests include computer vision, deep learning, and human action recognition.

NING HOU received the Ph.D. degree in electrical engineering from Hefei University of Technology, China. He is currently an Associate Professor with Henan University of Urban Construction, China. His current research interests include artificial intelligence, including statistical pattern recognition and deep learning and its hardware acceleration.