Neurocomputing 563 (2024) 126903

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

STDM-transformer: Space-time dual multi-scale transformer network for skeleton-based action recognition

Zhifu Zhao [a], Ziwei Chen [b], Jianan Li [c,*], Xuemei Xie [a], Kai Chen [b], Xiaotian Wang [a], Guangming Shi [d]

[a] School of Artificial Intelligence Engineering, Xidian University, Xi'an, Shaanxi, China
[b] Guangzhou Institute of Technology, Xidian University, Guangzhou, Guangdong, China
[c] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China
[d] Pengcheng Laboratory, Shenzhen, China

ARTICLE INFO

Keywords:
Skeleton-based action recognition
Transformer
Dual multi-scale network

ABSTRACT

Transformer-based methods have currently demonstrated impressive results in the field of skeleton-based action recognition. Nevertheless, how to effectively model multi-scale features with transformers remains a challenging problem, which is crucial to distinguish various actions. In this paper, we propose a Space–time Dual Multi-scale transformer (STDM-transformer) to explore the multi-scale collaborative representation employing both fine and coarse scale motion information. In contrast to existing approaches which typically propagate information between scales in a single fusion manner, our Space–time Dual Multi-scale method stratifies the space–time multi-scale into dual levels. One level is to construct fine-grained local motion interactions. In detail, the space–time multi-scale partition strategy and the novel intra-inter space–time transformer module are proposed to extract and aggregate the feature in part scale and body scale, respectively. The other is aimed at modeling coarse-grained global motion contexts, in which the layer-wise multi-scale progressive fusion strategy is designed. Extensive experimental results demonstrate that the proposed STDM-transformer achieves the SOTA performance on large-scale datasets.

1. Introduction

Human action recognition is a crucial and popular research topic in the field of computer vision and has been successfully applied to many applications, such as video surveillance [1,2], human–computer interaction [3–5], and sports analysis [6].

According to the input modality derived from different devices, the study topics are mainly divided into RGB-based [7–10], depth-based [11–14] and skeleton-based action recognition [15,16]. In the early studies of action recognition, RGB-based methods were the mainstream approaches including designing handcrafted features [7,8] and designing neural networks [9,10]. For the depth-based methods, the existing methods focused on exploring the 3D structural information and geometric shape information.

Different from the RGB-based and depth-based methods, skeleton-based action recognition methods have the advantages of being less computationally heavy [17] and being less influenced by environmental changes. Currently, with the maturity of cost-effective sensors [18] and pose estimation algorithms [19,20], skeleton-based methods have been studied extensively and drawn considerable attention. There are mainly two streams in traditional deep learning methods, the CNN-based methods [21–23] and the RNN-based methods [24–26]. For CNN-based methods, the skeleton data is always converted into a pseudo-image based on the designed transformation rules. And then it is input into a convolutional neural network for feature extraction. RNN-based methods usually model the skeleton data as a sequence of the coordinate vectors along both the spatial and temporal dimensions, where each of the vectors represents a human body joint. These CNN-based and RNN-based methods suffer from the drawback of failing to explore the kinematic dependencies and graph topologies of the human body. Recently, Graph Convolutional Networks (GCNs) [27–30] have been adopted to explore the structural connection of skeleton joints by manually designing traversal rules or graph topologies. However, GCN-based methods mainly depend on hand-crafted graph topologies such that there is a lack of flexibility in extracting information.

To this end, the transformer-based methods [31–33] employ the multi-head self-attention mechanism to adaptively explore the potential dependencies between skeleton joints. Zhang et al. [32] designed a spatial transformer block and a directional temporal transformer block

* Corresponding author.
E-mail address: lijianan@xidian.edu.cn (J. Li).

https://doi.org/10.1016/j.neucom.2023.126903
Received 13 December 2022; Received in revised form 14 August 2023; Accepted 3 October 2023
Available online 6 October 2023

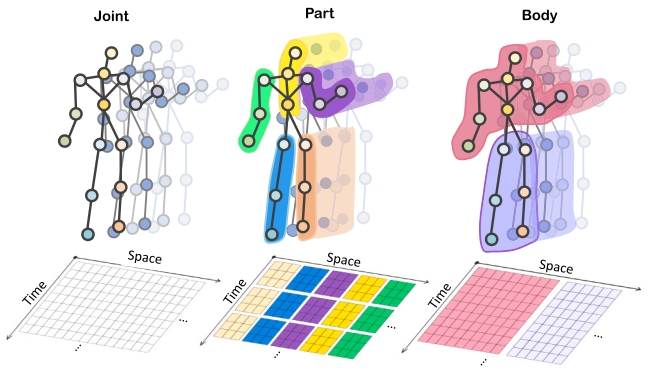


Fig. 1. Illustration of the proposed space–time multi-scale partition strategy. The strategy is employed to split the human action into parts or bodies in spatial–temporal dimension, which is conducive to extracting the multi-scale features.

for modeling skeleton sequences in spatial and temporal dimensions respectively. Shi et al. [31] presented a novel decoupled spatial–temporal attention network to calculate spatial attention and temporal attention for skeleton-based action recognition.

Nevertheless, there are two limitations to existing transformer-based methods. **(1)** Human actions in skeleton sequences are three dimensional (3D) spatio-temporal signals. These methods [34,35] treat spatial and temporal dimensions separately, which fails to explore their concurrence. **(2)** Existing methods [36,37] model the correlations between joint pairs in global topology. This operation only focuses on global information, resulting in the network being influenced by noisy information and ignoring locally important information.

We argue that, firstly, jointly modeling spatiotemporal information via a 3D representation in an end-to-end deep network provides a natural and efficient approach for action recognition. Secondly, besides the movement of individual skeleton joints in global topology, human actions are composed of the movements of different-level body parts, such as low-level parts (e.g., two arms, two legs, and one trunk) and high-level bodies (e.g., upper body and lower body). Modeling human movements at different spatial–temporal levels (as shown in Fig. 1) provides a more robust representation of different human actions. Therefore, it is necessary to model multi-scale features for action recognition.

In this paper, we propose a Space–time Dual Multi-scale transformer (STDM-transformer) to explore the multi-scale collaborative representation employing both fine and coarse scale motion information. The proposed STDM-transformer stratifies the space–time multi-scale into dual levels. One of the two levels is proposed with the aim of building local motion interactions. The first step of this level is to divide the features into different scales using a space–time multi-scale partition strategy. Then, a novel intra-inter space–time transformer module is proposed to extract and aggregate the part-scale and body-scale information from the skeleton data. Moreover, the other level is designed for modeling coarse-grained global motion contexts. It extracts joint-scale, part-scale, and body-scale features at different stages of the network, and finally aggregates multi-scale features to classify actions.

To verify the superiority of the proposed model, extensive experiments are performed on large-scale datasets. Experimental results demonstrate that the proposed STDM-transformer achieves the SOTA performance among existing methods. The main contributions of this work are as follows:

1. We propose a Space–time Dual Multi-scale transformer (STDM-transformer) to explore the multi-scale collaborative representation, which stratifies the space–time multi-scale into dual levels: fine-grained level and coarse-grained level.
2. At the fine-grained level, the Space-Time partition strategy and the novel intra-inter transformer block are proposed to extract and aggregate local interaction. At the coarse-grained level, the layer-wise multi-scale progressive fusion strategy is designed to aggregate joint-scale, part-scale, and body-scale features.
3. Extensive experiments show that the proposed STDM-transformer achieves the state-of-the-art performance on large scale datasets.

2. Related work

2.1. Skeleton-based action recognition

According to the different dimensions of skeleton data, skeleton-based methods are mainly divided into two categories, 2D skeleton-based methods [23,25,38] and 3D skeleton-based methods [16,39,40].

2D skeleton data with 2D coordinates (X, Y) in the pixel coordinate system can be acquired by applying pose estimation algorithms on RGB videos. It can be easily generated even by using a single RGB camera. However, it is insufficient to utilize the skeleton space structure [41]. Conversely, 3D skeleton data with 3D joint locations (X, Y, Z) are mostly obtained based on consumer RGB-D cameras [18]. It is robust when facing complicated backgrounds and changing conditions involving body scales, viewpoints, and motion speeds [42]. However, it also presents several drawbacks (e.g., low spatial resolutions, saturation problems, or optical interferences) due to the inability of depth sensors (e.g., structured light, time-of-flight or stereo-vision) to work properly in uncontrolled outdoor environments [43,44]. Recently, with the maturity of cost-effective sensors, it is easier to gain accurate 3D skeleton data. As a result, 3D skeleton-based methods have attracted many researchers' efforts on human action recognition, and an ever-increasing use of skeleton data can be expected.

For general skeleton-based action recognition, the neural network methods maintain the state-of-the-art and have been proven to be more general and robust than those early handcraft-based methods [45,46]. These methods mainly contain three categories as follows. Recurrent Neural Networks (RNNs) [24,26] have the strong ability to capture temporal dynamic features, which can effectively classify actions. Convolutional Neural Networks (CNNs) [21,23] enabled translation and scale invariance on pseudo images which are mapped in different ways. However, RNNs and CNNs underutilized the graph topology information of the skeleton. To tackle this issue, Graph Convolutional Neural Networks (GCNs) have been successfully applied to this task. The ST-GCN [40] first constructed the space–time graph through the adjacency matrices on the non-euclidean skeleton-based data. Subsequently, many researchers have expanded and improved it. DGNN [27] and DDGCN [28] represented skeleton sequences by a directed graph, which can capture effectively order information of actions. Besides, many GCN-based methods [29,30,47,48] were proposed to enlarge the receptive field, which establishes connections between non-adjacent joints. In general, most GCN-based methods have limitations in adaptively modeling the global dependencies of joints that have potential implications for action classification.

2.2. Transformer-based action recognition

In recent years, most works of transformer-based action recognition focus on RGB or skeleton sequences as input. In the RGB-based action recognition task, many methods [49–51] have attempted to explore various decoupling ways of transformer. And MViT [52] studied a multi-scale transformer for video classification. As for the skeleton-based action recognition task, DSTA [31] and STST [32] designed a decoupled space–time transformer to classify actions. STTFormer [33] utilized the joint space–time attention restricted in a local temporal window to capture space–time dependencies on skeleton data. Besides, [53] proposed an MTT module to extract multi-scale temporal

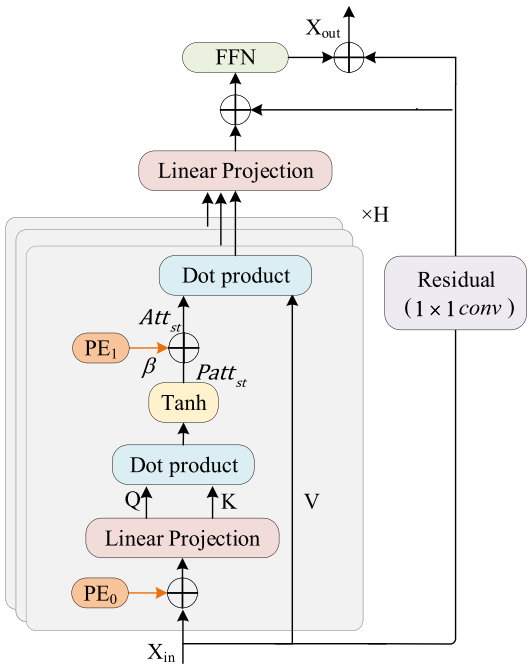


Fig. 2. Structure of the Basic Transformer Module (BTM), which mainly contains a multi-head self-attention module and a feedforward network (FFN). A detailed single-head self-attention is illustrated in the gray rectangle as an example. Q, K, and V are the query, key, and value matrix, respectively. PE_0 and PE_1 are two types of position encoding. H denotes the number of multi-head self-attention modules. Linear projection, dot product, and tanh represent different operators.

features rather than multi-scale space–time features which are significant cues for action recognition. However, these methods ignore the difference between the local and global semantics of joints in skeleton data.

2.3. Multi-scale representation on skeleton data

In order to further get discriminative features to improve the performance of skeleton-based action recognition, many methods are proposed to aggregate multi-scale features. These methods can be divided into two categories. The methods [29,30,54–56] in the first category are proposed to capture features from non-local neighbors through multi-scale receptive fields across both spatial and temporal dimensions. For example, [54,56] utilized GCN with the high-order adjacency matrix to extract relations of distant joints. The MST-GCN [30] split the channel of the features into several fragments, and then integrated MS-GC and MS-TC modules by residual connections to capture multi-scale features. In the second category, some approaches [57–60] observed the natural multi-scale representation of skeleton, namely joint, part, and body. DMGNN [57] exploited an MGCU to fuse features across scales for motion feature learning. MSR-GCN [58] proposed a U-Net-like GCN to yield a motion pattern. In contrast, our method is based on the hierarchical transformer structure to explore space–time information from fine to coarse and utilize multi-scale features for action recognition.

3. Preliminary

In this section, we introduce the **Basic Transformer Module (BTM)** in the proposed STDM-transformer, which is the important component of each space–time transformer (ST-transformer). The BTM contains a multi-head self-attention module followed by a feedforward network (FFN), whose structure is shown in Fig. 2. For ease of illustration, we take single-head self-attention as an example in this figure. Specifically, we assume that $X_{in} \in R^{N\times C}$ is the input of the BTM, where N represents the number of spatial–temporal tokens and C represents the number of channels. To retain spatial–temporal positional information, the input X_{in} is added with the position encoding PE_0, whose result is denoted as X'_{in}. Then, we generate the query $Q \in R^{N\times M_q}$, key $K \in R^{N\times M_k}$ and value $V \in R^{N\times M_v}$ by:

$$Q = f_q(X'_{in}), K = f_k(X'_{in}), V = X'_{in}, \tag{1}$$

where $f_q(\cdot)$ and $f_k(\cdot)$ represent two linear projections. And the M_q, M_k and M_v denote the dimensions of queries, keys, and values, respectively. Subsequently, the preliminary spatial–temporal attention map $Patt_{st}$ is obtained by a dot product operator and tanh function. In formulation:

$$Patt_{st} = Tanh(\frac{QK^T}{\sqrt{C}}), \tag{2}$$

To still retain spatial–temporal positional information, $Patt_{st}$ and PE_1 add together to get the spatial–temporal attention map Att_{st}. The PE_1 is the spatial–temporal position encoding of the feature, which is shared for all data samples. It represents a united intrinsic relationship pattern of the human joints. A weight β is multiplied to balance the strength of the spatial–temporal position encoding. Like a typical transformer, a dot product (followed by a linear projection) between Att_{st} and value V is utilized to obtain the enhanced hidden feature. And then a point-wise feedforward neural network is added to get the output X_{out}. Besides, two residual connections with 1×1 kernel are employed to aggregate different features and stabilize the network training.

4. Space-time dual multi-scale transformer

In this section, we first introduce the overall architecture of the STDM-transformer and then describe the space–time multi-scale partition strategy in detail. In addition, we present the technical details of the intra-inter space–time transformer module and multi-scale time aggregation module.

4.1. Overall architecture

The main idea of the proposed STDM-transformer is to construct multi-scale collaborative representations of skeleton data. Through stratifying the space–time multi-scale of skeleton data into dual levels, we implement the fine-grained local interaction and coarse-grained global interaction, respectively. For the fine level, the action features are first transformed to another scale (part or body) by the merging strategy. Then, the intra-inter space–time transformer module extracts the features corresponding to this scale. As illustrated in Fig. 3, the above operations are sequentially performed in the part scale and body scale. For the coarse level, the action features derived from different scales (joint, part, and body) are aggregated for predicting the action class.

As shown in Fig. 3, STDM-transformer contains three stages corresponding to joint, part, and body scale. Given a skeleton sequence, token embedding is utilized to map the data to a high-dimensionality space, which enriches the features of skeleton data. In the first stage, each token represents the feature of one joint. Three Space-Time Transformer Modules (ST-TMs) are performed to utilize BTM and Multi-Scale Temporal Aggregation (MSTA) to model the spatial–temporal relationships between the joints in consecutive frames and generate the joint-scale feature. In the second stage, two intra-inter ST-TMs are employed, which contain partition and merging strategies to divide the joint-scale feature into spatial–temporal parts and further boost the action feature to part scale. In the third stage, the structure takes part-scale features as input and utilizes the body merging strategy to generate spatial–temporal bodies. The same idea as the second stage is adopted to output the body-scale feature. Finally, global average pooling and fully connected layers are designed to fuse the joint-scale, part-scale, and body-scale features for final classification. In the following, we describe the space–time partition strategy and intra-inter multi-scale transformer module in detail.

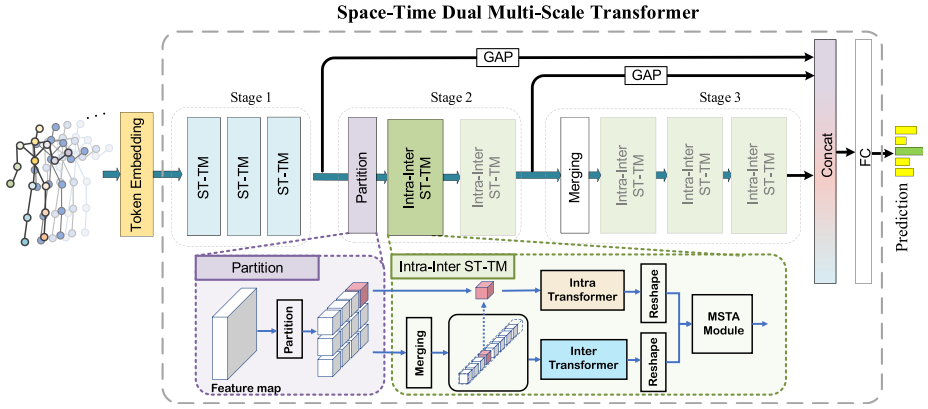


Fig. 3. The overall architecture of the proposed Space–time Dual Multi-scale transformer (STDM-transformer). There are three stages that are denoted as stage 1, 2, and 3 to extract joint, part, and body space–time features, respectively. The multi-scale features that are confused by the concatenate operator are utilized for classifying action categories. The stage 1 contains three Space-Time Transformer Modules (ST-TMs). Stage 2 or 3 is made up of intra-inter Space-Time Transformer Modules (intra-inter ST-TMs). The detailed illustration of intra-inter ST-TM is shown in the green dotted rectangle.

4.2. Space-time multi-scale partition strategy

In this subsection, we present a space–time multi-scale partition strategy, which plays an important role in intra-inter ST-TM. The space–time multi-scale partition strategy (as shown in Fig. 1) is described in detail as follows, so as to introduce the intra-inter ST-TM in the next subsection.

Biological observations from the early works [61–63] suggest that humans are capable of recognizing actions from the motion of skeleton joints of the human body. Besides, the human body can be regarded as an articulated system including rigid bones and hinged joints which are further combined into four limbs and a trunk [64]. Human actions are composed of the short motions of these limbs and trunk (body parts).

Motivated by the above observations, we argue that the different scale features like "joint vs part" and the "part vs body", are complementary to each other. Only single-scale feature for classification is not sufficient [63]. For instance, in the case of distinguishing between the actions of "hopping" and "jumping up", relying solely on body scale information is insufficient to distinguish between these two types of actions. We need to pay attention to the legs and use part-scale information as a supplement to correctly classify them. Furthermore, when we need to distinguish between the actions of "handshake" or "pass", relying solely on part-scale information is also insufficient for correct classification. Therefore, we need to pay attention to the joints of the hand and use more detailed joint-scale information as a supplement to the part-scale information. In our setting, the space–time partition strategy is employed in the part-motion partitions in stage 2 and body-motion partitions in stage 3. Since the joint-motion is already the basic partition in the skeleton sequence, we ignore its partition. For part-motion partitions, the whole human action is split into six parts (two arms, two legs, one head, and one trunk) in the spatial dimension and segmented with a three-frame interval in the temporal dimension. For body-motion partitions, the spatial splits contain the upper body and lower body and the temporal interval is 3.

Mathematically, assuming the input of kth stage is $X'_{k-1} \in R^{C'_{k-1}\times V'_{k-1}\times T'_{k-1}}$, which is the output feature of previous stage. Where V and T denote the partition number of the output feature in the spatial and temporal dimensions, respectively. C denotes the channel number of the output feature. For kth scale, we assume that the partition numbers in the spatial and temporal dimensions are V_k and T_k. Therefore, the partition data $p_k^{i,j}$ with spatial index i and temporal index j is with the shape of $C'_{k-1} \times L^i_{k,s} \times L_{k,t}$, where $L_{k,t}$ is the temporal interval. $L^i_{k,s}$ denotes that the $p_k^{i,j}$ contains $L^i_{k,s}$ spatial partitions of previous stages. Considering that $L_{k,s}{}^i$ be changing in different spatial indexes, we choose $L_{k,s} = \max(L^i_{k,s} | 1 \le i \le V_k)$ and pad $p_k^{i,j}$ into $\hat{p}_k^{i,j}$ with the shape of $C'_{k-1} \times L_{k,s} \times L_{k,t}$ for facilitating transformer processing. The space–time multi-scale partitions corresponding to kth scale are formulated as $X_k = \left\{ \hat{p}_k^{i,j} | 1 \le i \le V_k, 1 \le j \le T_k \right\}$. The space–time multi-scale partitions described above are conducive to extracting the multi-scale features by the designed intra-inter ST-TM in the following.

With the space–time partitions, intra-transformer and inter-transformer are designed to capture Fine-grained multi-scale representations. As shown in the green dotted rectangle of Fig. 3, the intra-transformer utilizes the self-attention mechanism to learn the motion feature within each space–time partition. Given the partition $\hat{p}_k^{i,j}$ with the shape of $C'_{k-1} \times L_{k,s} \times L_{k,t}$, the relationship with the other feature vectors is constructed for each C'_{k-1}-dimension vector in the partition through BTM.

The token number of the BTM is $(L_{k,s} \times L_{k,t})$. The whole process of intra-transformer can be formulated as:

$$Intra\left(X_k\right) = \left[BTM\left(\hat{p}_k^{1,1}\right), \ldots, BTM\left(\hat{p}_k^{i,j}\right), \ldots, BTM\left(\hat{p}_k^{V_k,T_k}\right) \right], \tag{3}$$

where the operator "[ ]" denotes the concatenation operation. For the inter-transformer, merging operation is designed to re-transform the input X_k into $\widetilde{X}_k = \widetilde{p}_k^{i,j} | 1 \le i \le V_k, 1 \le j \le T_k$, where the shape of $\widetilde{p}_k^{i,j}$ is $C_k \times 1$, the $C_k = C'_{k-1} \times L_{k,s} \times L_{k,t}$. Then, BTM explores the space–time dependencies between different partitions for the higher-level action representations, whose token number is $(V_k \times T_k)$. The formulation of the inter transformer is

$$Inter\left(X_k\right) = BTM\left(\widetilde{X}_k\right). \tag{4}$$

Then the outputs of intra-transformer and inter-transformer are reshaped as $C'_{k-1} \times V'_{k-1} \times T'_{k-1}$ and concatenated together.

4.2.1. Intra-inter space-time transformer module

Finally, their concatenation result is fed into the multi-scale time aggregation (MSTA) module. The MSTA module (as shown in Fig. 4) utilizes three different convolution kernels with different kernel sizes to integrate the temporal multi-scale information. The features of the three scales are concatenated with a residual connection.

5. Experiments

In this section, we conducted extensive experiments to evaluate the performance of our proposed STDM-transformer. Firstly, we provided an overview of the large-scale datasets and the implementation details. And then exhaustive ablation studies are conducted to verify the effectiveness of the intra-inter space–time transformer module (intra-inter ST-TM) and dual multi-scale method. Finally, we compare and discuss our proposed model with other state-of-the-art methods.

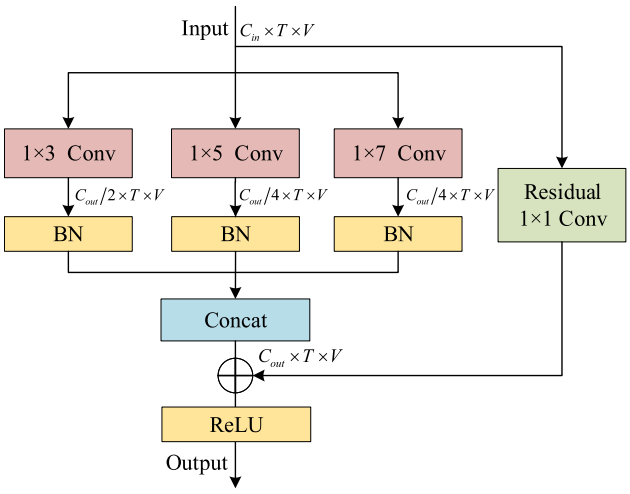


Fig. 4. Illustration of multi-scale time aggregation module (MSTA module). The module contains three scales of 1D temporal convolution kernels (1 × 3, 1 × 5, and 1 × 7) to enhance temporal features. C, T, and V represent the number of channels, number of joints, and time sequence length, respectively. BN (yellow rectangle) denotes the Batch Normalization layer. Residual (green rectangle) represents residual connection.

5.1. Dataset

NTU-RGB+D 60. NTU-RGB+D 60 [65] is a widely used large-scale skeleton-based human action recognition dataset. This dataset contains 56 880 skeleton sequences of 60 action classes performed by 40 distinct subjects. The overall action classes are divided into 40 daily actions, 9 health-related actions, and 11 mutual actions. Each action is captured by cameras at the same height but from three different horizontal angles. The human skeleton is represented as a 3D-coordinate of 25 joints. The dataset is divided into the training set and test set by two evaluation protocols: cross-subject (X-Sub) and cross-view (X-View). In the X-Sub evaluation, the training set contains 40,320 videos that come from one subset of actors, and the rest 16,560 videos that come from remaining actors are utilized for testing. In the X-View evaluation, the training set and test sets are divided into 37,920 videos and 18,960 videos. Training videos in this evaluation come from camera views 2 and 3 while the test videos are from camera view 1. The shortest sequence consists of 32 frames, while the longest sequence contains 300 frames.

NTU-RGB+D 120. NTU-RGB+D 120 is a large-scale skeleton-based human action recognition dataset [66] that extends the previous NTU-RGB+D 60 dataset, which includes 57,367 additional skeleton sequences over 60 additional action classes, resulting in a total of 120 action classes and 114,480 samples. The dataset includes videos performed by 106 subjects from 155 viewpoints, and it is currently the largest available dataset with 3D joint annotations for human action recognition. The dataset has 32 setups, each representing a different location and background. Two evaluation protocols, namely cross-subject (X-Sub) and cross-setup (X-Set), are recommended for this dataset. In the X-Sub setting, 63,026 clips from 53 subjects are used for training, and 50,922 clips from the remaining subjects are used for testing. In the X-Set setting, 54,471 clips with even collection setup IDs are used for training, and the rest of the clips with odd setup IDs are used for testing. It should be noted that 532 bad samples in this dataset are suggested to be ignored in all experiments.

Anubis. Anubis [35] is a large-scale dataset collected by Microsoft Azure Kinect sensors, consisting of 102 action classes and 80 participants. Each collection session involves two participants performing the same action four times, and the actions are divided into interactive and independent categories. The data is collected using five sensors placed in front of the room facing toward the subjects, with one camera in the center and four cameras symmetrically placed on either side with ±30 and ±45-degree angles. The skeleton of Anubis contains 32 joint points and the author padded the sequence to be of the same length of 300 frames by repeating the actions.

Table 1
Accuracy comparisons of different modeling methods with Intra-Inter ST-TM on the X-Sub setting of NTU-RGB+D 60 skeleton dataset.

Method		NTU-60 (%)
Inter only	–	83.81
Inter-intra	At 2nd stage	86.24
	At 3rd stage	87.11
	At 2nd & 3rd stages	**87.37**

5.2. Implementation details

In NTU-RGB+D 60, NTU-RGB+D 120, and Anubis datasets, all skeleton sequences are padded to 128 frames. If the original skeleton sequence contains more frames than 128, we will randomly sample 128 different frames from it as input. If the original skeleton sequence contains fewer frames than 128, we also randomly select some frames from it to be repeated and pad the skeleton sequence to 128 frames.

Our STDM-Transformer is stacked using 3 joint-based ST-TM blocks, 2 part-based Intra-inter ST-TM blocks, and 3 body-based Intra-inter ST-TM blocks with 3 multi-heads. The output channels of blocks are 64, 64, 128, 128, 256, 256, 256, and 256, respectively. We set the number of the body part as 6 (two arms, two legs, one head, and one trunk) and the number of half-bodies as 2 (upper body and lower body). The proposed model is trained for a total of 90 epochs with batch size 32 and SGD as optimizer on NTU-RGB+D 60 with the PyTorch framework (version 1.7). The learning rate is set to 0.1 initially, multiplied by 0.1 after 60 epochs and 80 epochs, and the weight decay is set to 0.00025. We performed a linear warmup of the learning rate during the first 5 epochs. Our experiments are conducted with the PyTorch deep learning framework, and we use 4 Nvidia GeForce Titan V GPUs.

It is worth noting that the skeleton of the Anubis dataset contains 32 joints. Thus, we need to remove 7 joints and keep the topology of the reduced skeleton equivalent to the original one.

5.3. Ablation study

In this subsection, we perform a series of ablation studies to prove the effectiveness of the proposed intra-inter ST-TM and dual multi-scale method. The experiments are conducted on the X-Sub setting of NTU-RGB+D 60.

5.3.1. Effectiveness of intra-inter ST-TM

In this subsection, we evaluate the validity of Intra-Inter ST-TM which is designed to enhance the model's ability to extract fine-grained local motion interactions. The clean model with three stages (joint, part, body) without the multi-scale features fusion is considered as the backbone. We first evaluate the accuracy of the backbone model only using inter-branch. And then we evaluate the backbone model with both inter and intra branches embedded at 2nd, 3rd, 2nd&3rd stages, respectively. The results are collected in Table 1.

As shown in the above Table 1, it can be seen that embedding the intra-branch will seriously enhance the performance. The main reason is that the intra-branch can effectively extract the local information within the partitions, which is conducive to classifying the action of slight movement. Moreover, the intra-branch embedded in 2nd&3rd stages obtains higher accuracy than the intra-branch embedded in a single stage. The experimental results prove the effectiveness of the Intra-Inter ST-TM.

Moreover, in order to confirm the effectiveness of this structure more extensively, we also employ the t-distributed Stochastic Neighbor

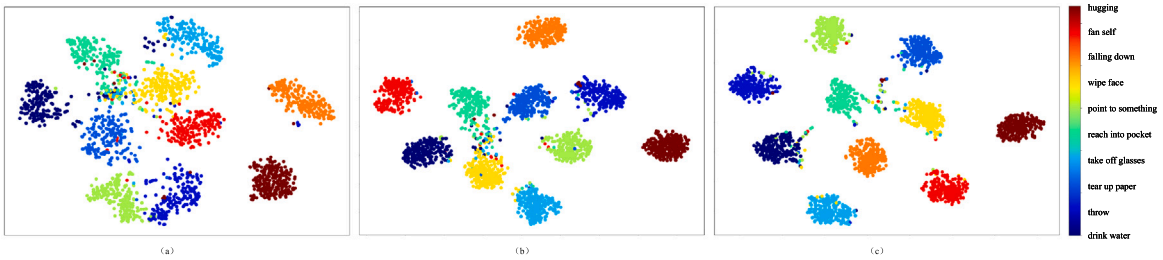


Fig. 5. The t-SNE [67] visualization of sampling classes on NTU-RGB+D 60 dataset with X-Sub evaluations. (a) The backbone model only uses inter-branch. (b) The backbone model with both inter and intra branches embedded at the 2nd stage. (c) The backbone model with both inter and intra branches embedded at the 2nd&3rd stages.

Table 2
The effectiveness of dual multi-scale method the X-Sub setting of NTU-RGB+D 60 skeleton dataset.

w/Joint	w/Part	w/Body	NTU-60 (%)
		✓	87.37
	✓	✓	87.85
✓		✓	88.36
✓	✓	✓	**88.82**

Table 3
Comparison of top-1 accuracy (%) over the state-of-the-art methods on the NTU-RGB+D 60, NTU-RGB+D 120 and Anubis dataset.

Methods	NTU60		NTU120		Anubis
	X-Sub	X-View	X-Sub	X-Set	
Ta-CNN [68]	90.7	95.1	85.7	87.3	51.38
Dynamic GCN [69]	91.5	96.0	87.3	88.6	50.99
AAGCN [70]	90.0	96.2	–	–	45.46
MS-G3D [29]	91.5	96.2	86.9	88.4	54.17
Efficient-GCN [71]	92.1	96.1	88.7	88.9	56.50
DSTA-Net [31]	91.5	96.4	86.6	89.0	55.35
IIP-Transformer [72]	92.3	96.4	88.4	89.7	–
ST-TR-agcn [73]	90.3	96.3	85.1	87.1	44.72
STDM-transformer	**92.6**	**96.4**	**88.9**	**90.8**	**60.89**

Embedding (t-SNE) method to visualize the cluster distribution. In detail, 10 classes of the NTU-RGB+D 60 dataset are randomly selected as samples to analyze, and the class numbers of 10 classes obey the arithmetic progression. From Fig. 5, we can observe that adopting the intra-inter ST-TM in 2nd&3rd can group the intra-class margin more closely and separate the inter-class margin more largely. The qualitative results show that the proposed Intra-Inter ST-TM indeed obtains the discriminative feature to distinguish different actions and then get a better distribution of features for action classification.

5.3.2. *Effectiveness of dual multi-scale method*

Through the above ablation study, we conclude that the multi-scale of fine-grained level is conducive to classifying the action. In our method, we also propose a scheme to implement the multi-scale of coarse-grained level, which fuses the feature derived from different stages (joint-scale, part-scale, and body-scale) for the robust action representation. To demonstrate the effectiveness of the dual multi-scale method, we illustrate the performance comparisons of different fusion methods in Table 2. In detail, all comparison methods embed the intra-inter transformer module in the 2nd&3rd stages. The method in the 2nd lines utilizes the body-scale feature to classify the action, which essentially only applies the multi-scale of fine-grained level. The methods from the 3rd to the 4th lines fuse two or three scales, which reflects the principle of dual multi-scale. From Table 2, we observe that fusing multiple scales improves the model performances compared to using a single body scale. In addition, fusing three scales has the best performance. This indicates that the features of different scales are complementary and the dual multi-scale is essential to classification.

5.3.3. *Comparisons with state-of-the-art methods*

To further illustrate the superiority of our proposed model, we compare the proposed method with the CNN-based, GCN-based, and transformer-based SOTA methods on NTU-RGB+D 60, NTU-RGB+D 120, and Anubis datasets. We use the top-1 accuracy of these methods reported in their original papers. Table 3 shows the comparison of accuracy on X-Sub and X-View splits of NTU-RGB+D 60 and NTU-RGB+D 120 datasets. The comparison of accuracy on the Anubis dataset is also provided in the last column of Table 3. It is worth noting that the notation "-" in the table means the paper did not provide results or code, so we cannot compare our method with it.

As shown in Table 3, our STDM-transformer outperforms the existing methods on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets with X-Sub evaluation and achieves comparable performance on X-View evaluation. On NTU-RGB+D 60, our model with four-stream fusion achieves state-of-the-art (SOTA) performance (92.6%) in the X-Sub setting. However, since the accuracy of the X-View setting in the NTU-RGB+D 60 dataset has already approached saturation, we achieved a SOTA-level of accuracy (96.4%). On NTU-RGB+D 120, our model achieves state-of-the-art performance (top-1 accuracy of 88.9% and 90.8% in the X-Sub and X-Set settings respectively). Our method also achieves SOTA performance (60.89%) on the Anubis dataset. The main benefit is that the STDM-transformer makes effective use of the intra-inter space–time transformer module (intra-inter ST-TM) and dual multi-scale method.

6. Conclusion

In this work, we propose a space–time dual multi-scale transformer to explore the multi-scale collaborative representation employing both fine and coarse-scale motion information. A novel space–time intra-inter transformer module is presented to extract the fine-grained local motion information, and a dual multi-scale method is proposed to construct the coarse-grained global motion contexts. Through extensive experiments on NTU-RGB+D 60, we demonstrated the effectiveness of the proposed intra-inter ST-TM and dual multi-scale method. Compared with the other SOTA methods, the proposed STDM-transformer achieves better performance on large-scale datasets.

Declaration of competing interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2020AAA0109301), Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62202356, 62 302373), National Natural Science Foundation of China under grant 61836008, Science and Technology Program of Guangzhou, China (202201011287), Guangdong Provincial Key Field Research and Development Plan Project (2021B0101400002).

References

[1] X. Wang, Intelligent multi-camera video surveillance: A review, Pattern Recognit. Lett. 34 (1) (2013) 3–19.
[2] R. Sharma, A. Sungheetha, et al., An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance, J. Soft Comput. Paradigm 3 (02) (2021) 55–69.
[3] F. Karray, M. Alemzadeh, J. Abou Saleh, M.N. Arab, Human-computer interaction: Overview on state of the art, Int. J. Smart Sens. Intell. Syst. 1 (1) (2017).
[4] S. Ahmed, K.D. Kallu, S. Ahmed, S.H. Cho, Hand gestures recognition using radar sensors for human-computer-interaction: A review, Remote Sens. 13 (3) (2021) 527.
[5] H. Liu, H. Nie, Z. Zhang, Y.-F. Li, Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction, Neurocomputing 433 (2021) 310–322.
[6] L. Ke, K.-C. Peng, S. Lyu, Towards to-at spatio-temporal focus for skeleton-based action recognition, 2022, arXiv preprint arXiv:2202.02314.
[7] D.K. Vishwakarma, C. Dhiman, A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel, Vis. Comput. 35 (11) (2019) 1595–1613.
[8] D. Vishwakarma, A. Dhiman, R. Maheshwari, R. Kapoor, Human motion analysis by fusion of silhouette orientation and shape features, Procedia Comput. Sci. 57 (2015) 438–447.
[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L.V. Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.
[10] Y. Meng, C.-C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, K. Saenko, R. Feris, Ar-net: Adaptive frame resolution for efficient action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 86–104.
[11] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P.O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Hum.-Mach. Syst. 46 (4) (2015) 498–509.
[12] A. Sanchez-Caballero, S. de López-Diz, D. Fuentes-Jimenez, C. Losada-Gutiérrez, M. Marrón-Romera, D. Casillas-Perez, M.I. Sarker, 3Dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information, Multimedia Tools Appl. (2022) 1–25.
[13] Y. Song, J. Tang, F. Liu, S. Yan, Body surface context: A new robust feature for action recognition from depth videos, IEEE Trans. Circuits Syst. Video Technol. 24 (6) (2014) 952–964.
[14] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1809–1816.
[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, Learning clip representations for skeleton-based 3d action recognition, IEEE Trans. Image Process. 27 (6) (2018) 2842–2855.
[16] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.
[17] B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3d skeleton-based action recognition using learning method, 2020, arXiv preprint arXiv:2002.05907.
[18] Z. Zhang, Microsoft kinect sensor and its effect, IEEE Multimedia 19 (2) (2012) 4–10.
[19] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[20] Y. Wang, M. Li, H. Cai, W.-M. Chen, S. Han, Lite pose: Efficient architecture design for 2d human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 13126–13136.
[21] P. Wang, Z. Li, Y. Hou, W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 102–106.
[22] W. Nie, W. Wang, X. Huang, SRNet: Structured relevance feature learning network from skeleton data for human action recognition, IEEE Access 7 (2019) 132161–132172.
[23] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.
[24] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.
[25] D. Avola, M. Cascio, L. Cinque, G.L. Foresti, C. Massaroni, E. Rodola, 2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs, IEEE Trans. Multimed. 22 (10) (2019) 2481–2496.
[26] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.
[27] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with directed graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7912–7921.
[28] M. Korban, X. Li, Ddgcn: A dynamic directed graph convolutional network for action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 761–776.
[29] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 143–152.
[30] Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1113–1122.
[31] L. Shi, Y. Zhang, J. Cheng, H. Lu, Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[32] Y. Zhang, B. Wu, W. Li, L. Duan, C. Gan, STST: Spatial-temporal specialized transformer for skeleton-based action recognition, in: Proceedings of the ACM International Conference on Multimedia, 2021, pp. 3229–3237.
[33] H. Qiu, B. Hou, B. Ren, X. Zhang, Spatio-temporal tuples transformer for skeleton-based action recognition, 2022, arXiv preprint arXiv:2201.02849.
[34] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[35] Z. Qin, Y. Liu, M. Perera, S. Anwar, T. Gedeon, P. Ji, D. Kim, ANUBIS: Review and benchmark skeleton-based action recognition methods with a new dataset, 2022, arXiv preprint arXiv:2205.02071.
[36] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, Int. J. Comput. Vis. 130 (5) (2022) 1366–1401.
[37] R. Yue, Z. Tian, S. Du, Action Recognition based on RGB and skeleton data sets: A survey, Neurocomputing (2022).
[38] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action Transformer: A self-attention model for short-time pose-based human action recognition, Pattern Recognit. 124 (2022) 108487.
[39] C. Caetano, J. Sena, F. Brémond, J.A. Dos Santos, W.R. Schwartz, Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition, in: International Conference on Advanced Video and Signal Based Surveillance, IEEE, 2019, pp. 1–8.
[40] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[41] L. Weng, W. Lou, X. Shen, F. Gao, A 3D graph convolutional networks model for 2D skeleton-based human action recognition, IET Image Process. 17 (3) (2023) 773–783.
[42] G. Johansson, Visual perception of biological motion and a model for its analysis, Percept. Psychophys. 14 (1973) 201–211.
[43] C. Jing, J. Potgieter, F. Noble, R. Wang, A comparison and analysis of RGB-D cameras' depth performance for robotics application, in: International Conference on Mechatronics and Machine Vision in Practice, IEEE, 2017, pp. 1–6.
[44] H. Yao, C. Ge, J. Xue, N. Zheng, A high spatial resolution depth sensing method based on binocular structured light, Sensors 17 (4) (2017) 805.
[45] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1290–1297.
[46] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in: International Conference on Pattern Recognition, IEEE, 2014, pp. 4513–4518.
[47] J. Li, X. Xie, Y. Cao, Q. Pan, Z. Zhao, G. Shi, Knowledge embedded GCN for skeleton-based two-person interaction recognition, Neurocomputing 444 (2021) 338–348.
[48] H.-g. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, InfoGCN: Representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.

[49] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding, 2021, p. 4, arXiv preprint arXiv:2102.05095. 2 (3).

[50] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 6836–6846.

[51] X. Zha, W. Zhu, L. Xun, S. Yang, J. Liu, Shifted chunk transformer for spatio-temporal representational learning, Adv. Neural Inf. Process. Syst. 34 (2021).

[52] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 6824–6835.

[53] J. Kong, Y. Bian, M. Jiang, MTT: Multi-scale temporal transformer for skeleton-based action recognition, IEEE Signal Process. Lett. 29 (2022) 528–532.

[54] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.

[55] K. Hu, Y. Ding, J. Jin, L. Weng, M. Xia, Skeleton motion recognition based on multi-scale deep spatio-temporal features, Appl. Sci. 12 (3) (2022) 1028.

[56] Z. Zheng, Y. Wang, X. Zhang, J. Wang, Multi-scale adaptive aggregate graph convolutional network for skeleton-based action recognition, Appl. Sci. 12 (3) (2022) 1402.

[57] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 214–223.

[58] L. Dang, Y. Nie, C. Long, Q. Zhang, G. Li, MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 11467–11476.

[59] W. Xu, M. Wu, J. Zhu, M. Zhao, Multi-scale skeleton adaptive weighted GCN for skeleton-based human action recognition in IoT, Appl. Soft Comput. 104 (2021) 107236.

[60] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction, IEEE Trans. Image Process. 30 (2021) 7760–7775.

[61] K. Thakkar, P. Narayanan, Part-based graph convolutional network for action recognition, 2018, arXiv preprint arXiv:1809.04983.

[62] L. Huang, Y. Huang, W. Ouyang, L. Wang, Part-level graph convolutional network for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11045–11052.

[63] W. Li, X. Liu, Z. Liu, F. Du, Q. Zou, Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network, IEEE Access 8 (2020) 144529–144542.

[64] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.

[65] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.

[66] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.

[67] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[68] K. Xu, F. Ye, Q. Zhong, D. Xie, Topology-aware convolutional neural network for efficient skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2866–2874.

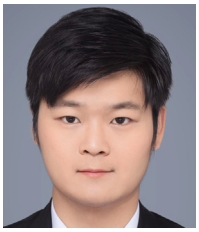
[69] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, H. Tang, Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 55–63.

[70] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Trans. Image Process. 29 (2020) 9532–9545.

[71] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[72] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, R. Weng, Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition, 2021, arXiv preprint arXiv:2110.13385.

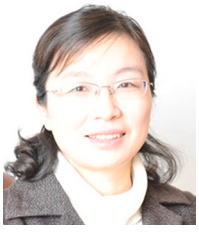
[73] C. Plizzari, M. Cannici, M. Matteucci, T. Liu, Skeleton-based action recognition via spatial and temporal transformer networks, Comput. Vis. Image Underst. 208 (2021) 103219.



Zhifu Zhao is a Lecturer in the School of Artificial Intelligence at Xidian University, Xi'an, China. He received his Ph.D. degree in School of Artificial Intelligence from Xidian University in 2020. His research interests are deep learning, video understanding and compressive sensing.

Ziwei Chen was born in Maoming, Guangdong Province on May 16, 1998. He is currently pursuing the master's degree with the Guangzhou Institute of Technology at Xidian University, China. His current research interests are computer vision and action recognition.

Jianan Li is a Lecturer in the School of Computer Science and Technology at Xidian University, Xi'an, China. She received his Ph.D. degree in School of Artificial Intelligence from Xidian University in 2020. Her research interests are deep learning, video understanding and action recognition.



Xuemei Xie received the M.S. degree in Electronic Engineering from Xidian University, Xi'an, China, in 1994 and the Ph.D. degree in Electrical & Electronic Engineering from the University of Hong Kong in 2004. She is currently a professor with the School of Artificial Intelligence at Xidian University, Xi'an, China. Her research interests include artificial intelligence, compressive sensing, deep learning, image and video processing, and multirate filter banks.

Kai Chen was born in Hengyang, Hunan Province on June 7, 2000. He is currently pursuing the master's degree with the Guangzhou Institute of Technology at Xidian University, China. His current research interests are computer vision and action recognition.

Xiaotian Wang received her Ph.D. degree in the School of Electronic Engineering from Xidian University in 2011. She is currently an associate professor with the School of Artificial Intelligence at Xidian University, China. Her research interests include Intelligent Signal Processing and computer vision.

Guangming Shi received the B.S. degree in Automatic Control in 1985, the M.S. degree in Computer Control and Ph.D. degree in Electronic Information Technology, all from Xidian University in 1988 and 2002, respectively. He joined the School of Electronic Engineering, Xidian University, in 1988. From 1994 to 1996, as a Research Assistant, he cooperated with the Department of Electronic Engineering at the University of Hong Kong. Since 2003, he has been a Professor in the School of Electronic Engineering at Xidian University, and in 2004 the head of National Instruction Base of Electrician & Electronic (NIBEE). From June to December in 2004, he had studied in the Department of Electronic Engineering at University of Illinois at Urbana-Champaign (UIUC). His research interests include compressed sensing, theory and design of multi-rate filter banks, image denoising, low-bit-rate image/video coding and implementation of algorithms for intelligent signal processing.