

RESEARCH ARTICLE

3D Graph Convolutional Feature Selection and Dense Pre-Estimation for Skeleton Action Recognition

JUNXIAN ZHANG¹, AIPING YANG¹, CHANGWU MIAO¹, XIANG LI²,
RUI ZHANG¹, AND DANG N. H. THANH³

¹Jiangsu Vocational Institute of Commerce, Nanjing 211168, China

²School of Information and Business Management, Dalian Neusoft University of Information, Dalian 116023, China

³College of Technology and Design, University of Economics Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

Corresponding authors: Aiping Yang (yangaiping2025@163.com) and Dang N. H. Thanh (thanhdnh@ueh.edu.vn)


The work of Junxian Zhang and Aiping Yang was supported by the Special Project of the Scientific Research and Development Center of China for Higher Education Institutions of the Ministry of Education under Grant ZJXF2022212. The work of Xiang Li was supported in part by the General Project of Liaoning Provincial Department of Education, China, under Grant LJKMZ20222010; in part by the Joint Fund under Grant LH-JSRZ-202205; and in part by the Liaoning Provincial Education Science Planning Project under Grant JG21EB029. The work of Dang N. H. Thanh was supported by the University of Economics Ho Chi Minh City, Vietnam.

ABSTRACT Action recognition plays an important role in promoting various applications in healthcare and smart education. However, unclear target actions, similar actions, and occluded characters may be encountered in some special scenarios. To solve the issues, a 3D Graph Convolutional Feature Selection and Dense Pre-estimation for Skeleton Action Recognition (3D-GSD) method is proposed to analyze and recognize the motion trajectory of the human skeleton. First, 3DSKNet is designed to adaptively learn and select important features in the skeleton sequence to identify skeleton parts of different importance more accurately according to the size of the input image resolution. It will help to better focus on key skeletal parts, improving the accuracy and robustness of bone recognition. Then, the DensePose algorithm is used to detect the complex key points of the human body posture and optimize the accuracy and interpretability of action recognition for different key points, key channels, and key-frames of the action. The proposed method achieves the best performance on the NTU RGB+D 60, NTU RGB+D 120 datasets, and Kinetics SKelation 400 datasets, with an improvement of 0.02%, 0.06%, and 0.1% in accuracy compared to the state-of-the-art methods.

INDEX TERMS Skeleton action recognition, feature selection, dense pre-estimation, attention mechanism, smart healthcare.

I. INTRODUCTION

Skeleton features are widely used in human action recognition and human-computer interaction. It refers to detecting and tracking key points of a human skeleton from a given image or video. This technology requires depth cameras, sensors, and other equipment to capture the movement trajectory of human bones and analyze and identify them through computer vision and machine learning techniques. Skeleton

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

behavior recognition technology can be applied to games, virtual reality, healthcare, and security. Different types of skeleton data in research and application scenarios increase the difficulty of skeleton recognition tasks. The main difficulties are: 1. The videos in the data set contain multiple characters, and the postures of each character may interfere with each other, making the extraction of skeletons difficult. 2. The videos in the data set come from different perspectives and different cameras, so the expression of the same action may be different, and actions under different perspectives need to be expressed uniformly. 3. The actions in the data set

involve interactions between characters, including occlusion, changes in spatial position, etc. These factors may affect skeleton extraction and action recognition.

Early skeleton-based action recognition methods were based on hand-designed feature extraction and spatiotemporal modeling. The former uses specially designed feature extraction algorithms to extract features representing actions from skeletal joint data [1], [2]. Common features include joint angles, joint distances, joint speeds, etc. The latter treats skeletal joint data as sequences that vary in time and space [3]. Action recognition is achieved by establishing action models, such as dynamic time warping (DTW), hidden Markov models (HMM), or conditional random fields (CRF), to perform spatiotemporal modeling and matching of skeletal sequences. However, these methods ignore the intrinsic relationships between human joints. The iDT algorithm [4] is known as the best performance method without the support of the deep learning technique. There are several methods developed based on iDT [5], [6]. In recent years, the skeleton feature action recognition technology based on deep learning has been roughly classified into two categories: one is based on skeleton key points [7], and the other is based on spatio-temporal feature analysis [8], [9], [10], [11]. The former mainly refers to using key points to describe the movement of the whole human body, and the output is an action category label, which has the advantages of not being disturbed by the environment, and a small amount of data. However, due to the limitations of the information contained in the skeleton points, it is difficult for the algorithms based on the skeleton points to effectively recognize some actions that are closely related to objects or scenes. Methods based on spatio-temporal feature analysis mainly include Two-Stream [12], C3D [7], and convolutional neural network-long short-term memory network (CNN-LSTM) methods [13]. The Two-Stream algorithm is to input the RGB image and the optical flow image into two CNN networks respectively, and then fuse the results of the two networks to obtain the final classification result. The Two-Stream algorithm can use the optical flow information to better capture the motion information of the action and improve the accuracy of action recognition. However, it requires additional GPU computing time and storage space, which has become the bottleneck of the Two-Stream algorithm. C3D extends the mature network structure in 2D CNN to the time domain and then adopts a decomposition strategy of the 3D convolution kernel, which is decomposed into 2D convolution and 1D convolution and adopts different serial and parallel methods combined to obtain the final classification result. C3D can accept the frame of the whole video, and it does not need to process the video into segments, with fast speed and good effect. However, the algorithm is not sensitive to camera viewpoint, noise, and local occlusion, which will affect the acquisition of interest points. For example, Qiu et al. [14] propose a deep neural network architecture called P3D, which aims to better learn the spatiotemporal features in videos, using

pseudo-3D convolution operations and residual connections to capture the spatiotemporal information of videos, in multiple videos. Extensive experiments on classification datasets demonstrate its superior performance over state-of-the-art techniques. Yan et al. [15] propose a three-dimensional gesture and action recognition framework called PA3D, which is mainly aimed at video recognition tasks. This method is represented by converting human postures and actions in the video into key point sequences in a three-dimensional coordinate system, and then inputting them into the neural network for recognition. The CNN-LSTM algorithm inputs the video sequence into a convolutional neural network, then inputs the result of the network into an LSTM, and finally classifies the result of the LSTM to obtain the final classification result [16], [17]. Liu et al. [49] propose an end-to-end multi-level long short-term memory (LSTM) network with spatial and temporal attention mechanisms. Its network can automatically select key information from each frame to determine actions, and the network uses spatial and temporal attention modules to assign different importance levels to each frame. Ke et al. [18] propose Global Contextual Attention LSTM (GCA-LSTM), which can selectively focus on discriminating joints. Ke et al. [18] divide the action sequence into several short video clips, then use a 2D convolutional neural network to extract features from each clip, and then input these feature sequences into an LSTM network for sequence modeling to ultimately achieve the classification of action prediction categories. The CNN-LSTM algorithm works well for long-term series and can capture long-term dependencies in time series. However, it is slower and requires more computing resources. In addition, it is not sensitive to factors such as camera viewpoint, noise, and partial occlusion.

Traditional skeletal action recognition methods generally require manual design and feature extraction, and often require the participation of multiple steps and domain experts, making it difficult to adapt to different scenarios and tasks. In addition, most deep learning-based methods perform poorly for pose changes and high motion complexity in skeletal sequences. The MS-G3D network [19] does not need to manually design and extract features, but automatically learns the features of the bone sequence through convolution and pooling operations, which improves the generalization ability and adaptability of the model. At the same time, the network adopts 3D convolution and attention mechanism to process the spatio-temporal information in the skeleton sequence, effectively capturing the key features of the action, while reducing the model parameters and calculation amount, and improving the efficiency and accuracy of the model. However, one of its main drawbacks is the influence of motion being occluded, which may cause the model to fail to capture the key information of the motion correctly and lead to a decline in the performance of the model. Skeleton joints are the key information in the skeleton sequence, but in the MS-G3D network, each skeleton joint is only represented as a coordinate point. This representation cannot fully express the

morphological and dynamic information of the skeletal joints. Therefore, it is necessary to find better ways to strengthen the expressive ability of skeleton joints to further improve performance.

Action recognition methods based on deep learning play a significant role in promoting various applications in health-care and smart education. However, some special scenarios may encounter unclear target actions, similar actions, and occluded characters. To solve the problems of pose change, scaling, and sequence loss in skeleton sequences, we propose a 3D graph convolutional feature selection and dense pre-estimation (3D-GSD) method for action recognition of skeletons. Introducing spatial and temporal attention mechanisms and human prediction models can make the model adaptively focus on key poses and skeletal joints and consider their changes in time. Therefore, our method does not only better capture the local and global information of actions but also analyzes human poses more comprehensively. The main contributions are as follows:

- We design a 3DSKNet to adaptively adjust the model. It can more accurately identify key points of different importance according to the size of the input image resolution. Moreover, it greatly improves the estimation accuracy of skeleton missing key points, reduces the difficulty of skeletal action recognition, and increases the accuracy of skeletal action recognition occluded by objects.
- We introduce a DensePose algorithm to detect the complex key points of human poses and integrate them into the 3D-GSD network model. The 3DSKNet attention mechanism focuses on key skeletal parts, while DensePose can provide more detailed pose and shape information. By combining them, more accurate and complete human motion analysis results can be obtained.
- Extensive quantitative and qualitative experiments are implemented to verify the accuracy of the 3D-GSD. The experiments were evaluated on two different datasets of human recognition.

The rest of the paper is structured as follows: Section II provides a brief review of related work, including the skeletal action recognition, attention mechanism, and human pose estimation algorithm based on CNN. Section III presents the details of the proposed method. Section IV shows the experimental results. Section V is the conclusions.

II. RELATED WORKS

A. SKELETAL ACTION RECOGNITION

Traditional algorithms for skeleton-based action recognition are implemented using hand-designed feature extractors, which can include joint angles, accelerations, velocities, energies, etc. These features are then fed into machine learning models for classification or regression, such as support vector machines (SVM) and hidden Markov models (HMM). With advanced deep learning techniques, models for skeleton-based action recognition are developed and can

be divided into two categories: sequence-based models and graph-based models.

Sequence-based models typically use recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) to model sequence data. These models can handle variable-length bone sequence data and consider the temporal relationship between joints. Liu et al. [20] propose a new gating mechanism to deal with noise in skeleton data by learning the reliability of sequential data and accordingly adjusting the input data's contribution to the long-term contextual representation stored in the cell's memory unit. Wang et al. [21] propose a novel hierarchical attention network with pseudo-meta-paths for skeletal action recognition, which learns discriminative features for action recognition by capturing the long-range dependencies of skeletal joints. Zhang et al. [9] propose a Two-Stream Transform Encoder (TSTE) network utilizing motion spatiotemporal feature embedding and shape transformation. San et al. [22] provide a comprehensive review of deep learning techniques for human activity recognition (HAR) and provide a resource guide for researchers and practitioners. Zhang et al. [23] introduce deep learning-based methods for human action recognition, including methods based on color videos, skeleton sequences, and depth maps. Li et al. [24] propose a new CNN-based action classification and detection framework. Although sequence-based models perform well in skeletal action recognition tasks, there are still some problems and challenges that need to be resolved: when collecting skeletal sequence data, there may be certain noise or missing data, for example, due to sensor failure or human body occlusion, etc.

Graph-based models aim to address the limitations of sequence-based models, mainly utilizing graph convolutional neural networks (GCNs) to model the relationship of skeletal sequences. Yan et al. [25] first use ST-GCN to model the problem of skeleton-based action recognition. The AS-GCN network proposed by Li et al. [26] can effectively share information between different actions, and the graph structure can be adaptively optimized through network learning to obtain more behavior details to improve the recognition effect. Shi et al. [27] propose a two-stream network (2s-AGCN) structure using node information and bone information and then construct a two-stream network structure by simultaneously utilizing key points and bone information to obtain more skeleton features for action recognition, significantly improving recognition performance. Shi et al. [28] further propose directed graph networks (DGNN), which can dynamically construct graph connections end-to-end, surpassing other methods in all indicators, and it can effectively identify complex motions in skeletal motions. Graph-based models can naturally capture complex relationships among skeletal sequences and can better handle multi-person actions and object interactions. However, graph models have higher time complexity than sequence-based models since computations need to be performed on all nodes and edges. This can lead to increased training and

inference time, limiting the usefulness of these models in real applications.

B. ATTENTION MECHANISM

In recent years, some new attention mechanisms [29], [30] have been proposed. For example, interactive attention can learn to find key features and salient parts from the input data to achieve better task effects [31], and multi-head attention [32] runs multiple attention mechanisms on the same input data and merges the results. The channel attention [33], [34], [35], [36] that is mainly studied in this paper, can find the specific data in complex data, and improve the accuracy and efficiency of the model by learning how to adjust the weight of each channel in the input data.

C. HUMAN POSE ESTIMATION BASED ON CNN

Human Pose Estimation (HPE) [37] is to obtain human motion information from visual data, including the position of key points, attitude angle, and other information. With the powerful development of CNN, more CNN models are used for human pose estimation, such as the Hourglass [38] model, the Integral Human Pose Regression model [39], the Simple Baseline model [40], etc. Mask R-CNN [41] is a CNN-based target detection and semantic segmentation algorithm, but it does not directly output the position of the key points of the human body but outputs the rectangular frame where the human body is located and the position of each key point in the rectangular frame. Subsequently, the DensePose algorithm [42] appears, the key to which is to train a large number of datasets with pose annotations, so that the model can learn the mapping relationship between the human body surface and pixels and can predict the position of each pixel on the human body surface. However, in practical applications, different scenarios and tasks require different loss functions, which also need to be designed and optimized according to specific problems.

III. PROPOSED METHOD

MS-G3D [19] is a bone recognition method based on a 3D CNN. It can analyze and predict the input 3D skeleton sequence, but it cannot fully express the shape and dynamic information of skeleton joints due to motion occlusion. The proposed 3D-GSD has been modified on this basis, retaining the ms-g3d module to extract the space-time feature representation of the skeleton sequence, and designing a new feature selection module 3DSKNet and dense pre-estimation module DensePose, as shown in Figure 1. 3DSKNet is an attention mechanism for 3D convolutional neural networks, which can adaptively learn important features in skeleton sequences, better focus on key skeleton parts and action sequences, and ignore unimportant parts such as some noise or interference and some irrelevant joints, which helps to improve the accuracy and robustness of skeleton recognition. After skeleton recognition, it is necessary to estimate the pose and shape of the human body in three-dimensional

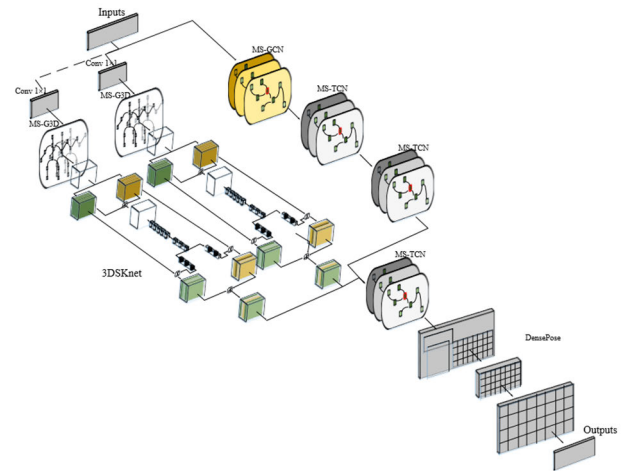


FIGURE 1. Overall architecture diagram of 3D-GSD network.

space. Using DensePose to estimate the human body pose on the input image can analyze the human body pose more comprehensively. Specifically, 3DSKNet can provide position and motion information of key points of interest, while DensePose can estimate more detailed pose and shape information. The combination of them can obtain more accurate and complete human motion analysis results, which is of great significance to many application fields, such as motion analysis and medical diagnosis.

A. 3DSKNET MODULE

SKNet [52] is a lightweight attention mechanism that can enhance the representation ability of the network, but the reason why the SKNet mechanism cannot be directly applied to the 3D network structure is that it is carried out in space, and in the 3D structure, in addition to the spatial dimension (x, y, z), and the time dimension (t), so it is necessary to design the attention mechanism in the time dimension. In addition, the attention mechanism of SKNet needs to operate on feature maps of different scales, and the 3D network structure requires a more complex design to deal with feature maps of different scales due to the larger range of scale changes. Therefore, a 3DSKNet mechanism that can handle joint information in both spatial and temporal dimensions is proposed. The 3DSKNet mechanism adopts a 3D convolution operation and attention mechanism, which can adaptively learn the spatiotemporal features of each joint point and perform a weighted fusion of the features of different time steps to capture the spatiotemporal relationship. In 3DSKNet, the feature learning and feature selection of each joint point are carried out in three-dimensional space, the formula is as follows:

$$y_i = W_2 \cdot \text{relu}(W_1 \cdot X_i) \quad (1)$$

$$s_i = \frac{1}{T} \sum_{t=1}^T y_{i,t} \quad (2)$$

where y_i represents the feature vector of the i -th joint point, X_i is the feature input of the i -th joint point, W_1 and W_2 are

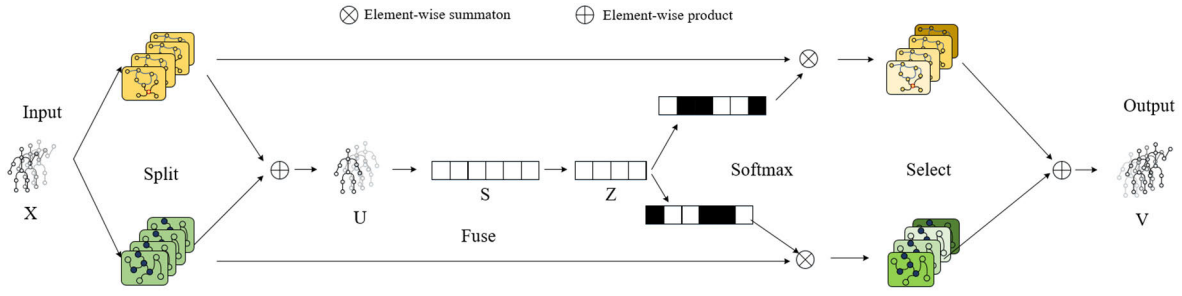


FIGURE 2. 3DSKNet mechanism on 3D skeleton data.

learning parameters, and s_i is the average feature of the i -th joint point in the time dimension. In 3DSKNet, the attention coefficient can be expressed as:

$$z_{i,t} = \text{sigmoid}(W_3 \cdot s_i + b_3) \quad (3)$$

where W_3 is the learning parameter and b_3 is the bias parameter; $z_{i,t}$ represents the attention coefficient of the i -th joint point at time t , and the attention-weighted feature of each joint point at time t can be calculated:

$$a_{i,t} = z_{i,t} \cdot y_{i,t} \quad (4)$$

where $a_{i,t}$ is the attention-weighted feature of the i -th joint point at time t , and $y_{i,t}$ is the feature vector of the i -th joint point at time t . Finally, the attention weight of each joint point at different moments can be weighted for data fusion to obtain the final feature representation:

$$F = \sum_{i=1}^V \sum_{t=1}^T a_{i,t} \quad (5)$$

where V is the number of joint points, T is the number of time steps, and F is the final feature representation, which can be used for subsequent tasks, such as skeleton recognition and human pose estimation. We introduce in detail the mechanism of extending 3DSKNet on 3D skeleton data in Figure 2, which mainly includes three stages, namely the split stage, fuse stage, and select stage.

The split stage mainly performs scale-invariant processing on the input feature map and adds convolution operations of different kernels according to the number of branches. First, the input feature map X is divided to obtain multiple sub-feature maps, and each sub-feature map corresponds to a convolution kernel. According to the number of branches M , the input feature map is divided into M parts as input. Specifically, for the i -th branch, a convolution kernel of $(2i + 1) \times (2i + 1) \times (2i + 1) \times 3$ size is used for the convolution operation, the step size is 1, and the padding is 1. After the convolution operation, the feature maps of all branches are stitched together to obtain a feature map of size $M \times T \times V \times H \times H \times \text{out_channels}$, which T represents the number of time steps, V represents the number of joint points, H represents the size of the spatial dimension (height and width), and out_channels represents output channel dimensions.

In the fusion stage, the features of different scales obtained by all branches are first added element-by-element to generate a mixed feature U with a dimension of $[N, C', T, V, M]$; then, three-dimensional adaptive pooling is performed on U to compress the features to the specified dimension 1 and get S with dimensions $[N, C', 1, 1, M]$. Next, squeeze the results obtained in the previous step into $[N, C']$, and then use the fully connected layer to reduce the dimension to L to get a scalar d . The formula is:

$$d = \text{Max}\{L, W_4(\text{Squeeze}(U))\} \quad (6)$$

where W_4 represents the fully connected layer, and Squeeze represents compressing the dimension of the feature to 2. After performing dimension reduction, dimension increase, and softmax operations on the feature, the correlation between the features is learned, and the weights of different positions are obtained for selecting the appropriate subset of features.

In the selection stage, the feature U output by the fusion stage is first divided into two features a and b through the split operation. Next, compress the second dimension (the number of channels) to obtain two vectors whose length is half the number of channels, denoted as a' and b' , respectively. Then, a' and b' are respectively multiplied elementwise by the weight vector, and the weighted feature V will represent the entire feature U more accurately. This weight vector maps the value to the $[0, 1]$ range according to the softmax function, ensuring that each element is in the $[0, 1]$ range and the sum is 1.

In general, the design idea of the 3DSKNet mechanism is to fuse the output features of the skeleton recognition module with the global features with spatial relationships, to improve the performance of skeleton recognition.

B. DENSEPOSE MODULE

To better capture the characteristics of human motion, we use the DensePose module at the end of 3D-GSD to improve the accuracy of skeletal behavior recognition. By predicting the position of the key points of the human skeleton, more abundant posture information can be provided, and more detailed posture estimation can also be realized, such as the specific position of the hand, the degree

of flexion and extension of the fingers, the posture of the body, etc., thereby improving the accuracy. Specifically, the DensePose human poses prediction module is mainly divided into three stages: feature extraction, feature, and skeleton feature fusion, and pose estimation. As shown in Figure 3.

Firstly, the output feature map of the previous stage is used as input, and a series of convolutional layers (including 3 Conv2d and 3 ConvTranspose2d) are used for feature extraction and dimensionality reduction. In the recognition of skeletal sequences, 384 feature points are extracted from the output of the skeleton network and used as the representation of skeletal sequences. Specifically, perform a Conv2d operation to reduce the number of the feature channels from 384 to 256, and then perform two downsampling operations (Conv2d with stride=2) to reduce the feature size to 1/4 of the original. Subsequently, perform two more Conv2d operations to reduce the number of channels of the feature map to 128 and 64 respectively. Finally, perform a Conv2d operation to reduce the number of channels of the feature map to 32 again, and then use ConvTranspose2d three times to increase the dimension to obtain the final feature map. Which purpose is to reduce the number of feature channels while keeping the size of the feature map constant, thereby improving the abstraction ability of features. By fusing the DensePose feature with the skeleton feature, more comprehensive human pose information can be obtained. The fused formula is as follows:

$$F_{fuse} = \frac{1}{T} \sum_{t=1}^T [F_{DP,t}; F_{ske,t}] \quad (7)$$

where $F_{DP,t}$ is the DensePose feature of the t -th frame, $F_{ske,t}$ is the skeleton feature of the t -th frame, $[\cdot; \cdot]$ represents the splicing operation in the feature channel dimension, T is the total number of frames in the video, and F_{fuse} is the fused feature vector. Then, F_{fuse} performs global average pooling to obtain the final feature vector f :

$$f = \frac{1}{T} \sum_{t=1}^T F_{fuse,t} \quad (8)$$

where $F_{fuse,t}$ is the fused feature vector of the t -th frame, and the fused features are input into two fully connected layers for classifying actions. The final output is the probability value for each category, which is obtained by the softmax:

$$y_k = \frac{\exp(h, k)}{\sum_{i=1}^K \exp(h, i)}, k = 1, \dots, K \quad (9)$$

where y_k represents the probability of belonging to the k -th category, and h is the output of two fully connected layers.

IV. EXPERIMENTS

Experiments are implemented on a Windows system equipped with an Intel Xeon(R) 4210R CPU and an NVIDIA RTX 3090 GPU. The network framework is also based on the PyTorch platform. The full source code is available at the address <https://github.com/wizardbo/3D-GSD>.

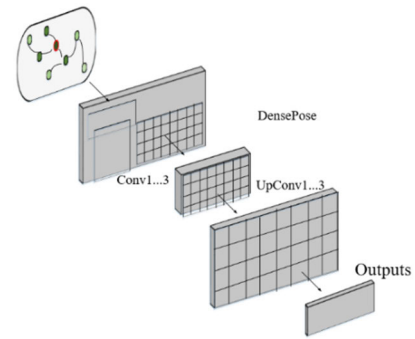


FIGURE 3. Human body posture prediction DensePose module.

A. DATASETS

The NTU RGB+D 60 Skeleton dataset is a skeleton dataset for human action recognition, including 60 different action categories. The data set uses the Microsoft Kinect v2 depth camera and inertial measurement unit (IMU) to collect bone data. The bone sequence of each action includes the data of 3 people, and each person's action execution has different angles and distances.

The NTU RGB+D 120 Skeleton dataset is a commonly used human action recognition dataset. It includes a total of 120 human action categories, which are divided into 60 single-person actions and 60 double-person actions. Each action category consists of multiple different instances, each instance includes 300 frames of skeleton data and depth image data. The NTU RGB+D 120 dataset is widely used in many studies because it contains a large number of action sequences and many different types of actions, such as jumping, stretching, waving, walking, crossing arms, making fists, etc. In addition, it also provides a variety of difficulty levels of data set division to suit different types of research and application scenarios.

B. COMPARISON TO STATE-OF-THE-ART METHODS

To confirm the effectiveness of the proposed method for the skeleton action recognition task, we conduct a comparative study with several state-of-the-art techniques, including IndRNN [43], HCN [44], ST-GR [45], 2s-AGCN [27], AGC-LSTM [46], DGNN [28], MST (joint) [47], QCYHZ (Qin-Cai-Yu-He-Zhang) [48], MS-G3D [19], ST-LSTM [49], GCA-LSTM [18], RotClips+MTCNN [50], Body Pose Evaluation Map [51]. Meanwhile, we use two common performance metrics: X-Sub(%) and X-Set(%), where the higher metric value indicates better performance.

We quantitatively compare our method with the other competing deep learning-based methods on NTU RGB+D 60 Skeleton, NTU RGB+D 120 Skeleton datasets, and Kinetics SKELETON 400 datasets. Table 1 and Table 2 display the statistical outcomes of X-Sub and X-Set for all the competing methods. It can be seen that the proposed method achieves the best X-Sub and X-Set results on all datasets. Moreover, The X-Sub and X-Set values of our method are 0.02%~0.06% higher than those of the baseline MS-G3D. Table 3 shows the

TABLE 1. Quantitative comparison (X-Sub and X-Set) of the NTU RGB+D 60 SKELETON dataset. The top-performing result is highlighted in bold, while the second-best result is underlined.

Methods	NTU RGB+D 60	
	X-Sub(%)	X-Set(%)
IndRNN ^[43]	81.8	88.0
HCN ^[44]	86.5	91.1
ST-GR ^[45]	86.9	92.3
2s-AGCN ^[27]	88.5	95.1
AGC-LSTM ^[46]	89.2	95.0
DGNN ^[28]	89.9	96.1
MST (joint) ^[47]	89.0	95.1
QCYHZ ^[48]	90.5	96.1
MS-G3D ^[19]	<u>91.50</u>	<u>96.20</u>
3D-GSD (Ours)	91.52	96.23

TABLE 2. Quantitative comparison (X-Sub and X-Set) of the NTU RGB+D 120 SKELETON dataset. The top-performing result is highlighted in bold, while the second-best result is underlined.

Methods	NTU RGB+D 120	
	X-Sub(%)	X-Set(%)
ST-LSTM ^[49]	55.7	57.9
GCA-LSTM ^[18]	61.2	63.6
RotClips + MTCNN ^[50]	62.2	61.8
Body Pose Evolution Map ^[51]	64.6	66.9
SGN ^[37]	79.2	81.5
2s-AGCN ^[27]	82.9	84.9
MST (joint) ^[47]	82.8	84.5
QCYHZ ^[47]	85.7	86.8
MS-G3D ^[19]	<u>86.90</u>	<u>88.40</u>
3D-GSD (Ours)	86.96	88.44

TABLE 3. Quantitative comparison (Top 1 and Top 5) of the kinetics SKELETON 400 dataset.

Methods	Kinetics SKELETON 400	
	Top 1(%)	Top 5(%)
ST-GCN ^[25]	30.7	52.8
AS-GCN ^[26]	34.8	56.5
ST-GR ^[45]	33.6	56.1
2s-AGCN ^[27]	36.1	58.7
DGNN ^[28]	36.9	59.6
MS-G3D ^[19]	38.0	60.9
3D-GSD (Ours)	38.1	61.0

statistical results of the Top 1% and Top 5% on the Kinetics SKELETON 400 dataset, and our methods own the best results.

These results indicate that our proposed method achieves better performance for various datasets and improves the action recognition performance of the model by focusing on key parts and action details.

For the complexity, the proposed 3D-GSD contains 5,012,643 parameters and MS-G3D – 3,194,595 parameters. For the training time cost, both MS-G3D and 3D-GSD took around 1 week, and the difference is only a few hours. This is understandable because the number of parameters of the proposed model is larger. However, the difference in time for

TABLE 4. Ablation study of 3D-GSD for different modules on NTU RGB+D 60 SKELETON dataset.

Methods	NTU RGB+D 60	
	X-Sub(%)	X-Set(%)
3D-GSD (Joint Only) [w/ 3DSKNet]	89.44	95.03
3D-GSD (Joint Only)	89.51	95.06
3D-GSD (Bone Only) [w/ 3DSKNet]	90.12	95.35
3D-GSD (Bone Only)	90.15	95.38
3D-GSD (Ours)	91.52	96.23

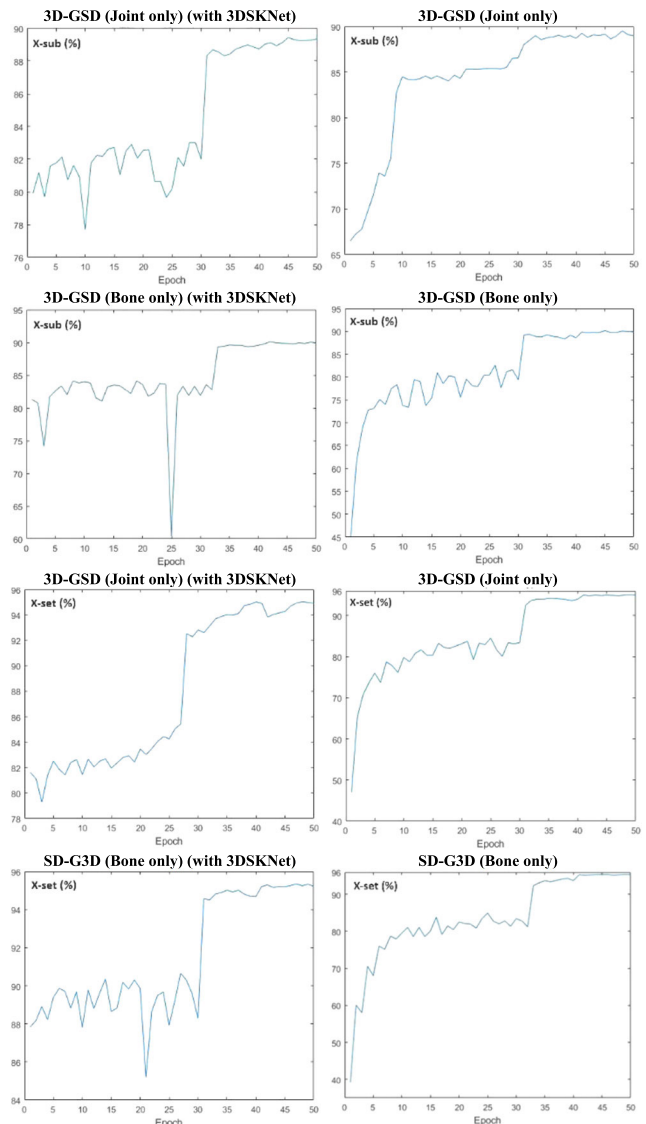


FIGURE 4. X-Sub (%) and X-Set (%) rise curve of NTU 60 dataset.

the training procedure is acceptable. Moreover, for the testing time, the difference is just very minor: the proposed 3D-GSD took 13.668 seconds and MS-G3D took 12.918 seconds for the data of NTU RGB+D 120 Skeleton dataset.

C. ABLATION STUDY

To further validate the proposed 3D-GSD, we analyze the contribution of each module to the 3D-GSD method

by removing different network modules, including removing 3DSKNet and removing Densepose, respectively. Here, we also perform tests on joint only and bone only on NTU RGB+D 60 Skeleton dataset, as shown in Table 4.

The joint SD-G3D network that only adds 3DSKNet has a lower recognition rate of 0.07% than the joint SD-G3D network that adds both the 3DSKNet attention mechanism and the DensePose pre-estimation module. For the skeletal SD-G3D network with the 3DSKNet attention mechanism and DensePose pre-estimation module, the recognition rate is 0.03% higher than that with only the 3DSKNet skeletal SD-G3D network. The human body pre-estimation DensePose module can estimate the key points and pose information of the human body in the input image, thereby improving the recognition accuracy of the occluded parts. It can also map the two-dimensional points on the image to the surface of the three-dimensional human body and mark them so that the model can understand the posture and shape of the human body more accurately, and effectively solve the problem of being occluded. The rising curves of X-Sub(%) and X-Set(%) for each part of the NTU RGB+D 60 dataset are shown in Figure 4.

In the experiments of NTU RGB+D 60 (Joint Only) and NTU RGB+D 60 (Bone Only), since the network only considers joint points and bone information, adding the 3DSKNet mechanism can improve the connection between joint points and bones so that the model can better understand the skeleton information and better distinguish different actions. At the same time, the 3DSKNet mechanism can effectively reduce noise interference and improve the robustness of the model, thereby improving the accuracy of the model.

V. CONCLUSION

This paper proposes a 3D graph convolutional feature selection with a dense pre-estimation (3D-GSD) method for action recognition of skeletons. This method is mainly to design the 3DSKNet attention mechanism in the MS-G3D network of bone recognition and introduce the human body pose estimation DensePose. Specifically, the designed 3DSKNet attention mechanism can make the network pay more attention to important features, improving the accuracy while keeping the computational cost small. Secondly, the introduction of the DensePose module can provide pose information on the skeletal sequence, further enhancing the performance of skeletal behavior recognition. The 3D-GSD network has advantages in processing spatiotemporal sequence data, so it can better handle bone sequence data. Finally, the paper conducts extensive experimental validation on several commonly used action recognition datasets. The results show that the proposed method achieves the best performance on the NTU RGB+D 60, NTU RGB+D120 datasets, and Kinetics SKeletion 400 datasets, achieving accuracy gains of 0.02%, 0.06%, and 0.1% compared to the best-performing methods. The SD-G3D network model may be more effective for specific datasets and tasks, while the generalization performance on other datasets or tasks may be degraded. This is because

features and fusion strategies for multimodal data are usually designed for specific problems and may not be applicable to other scenarios.

REFERENCES

- [1] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.
- [2] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., Sep. 2009, pp. 121–124.
- [3] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1996–2003, doi: [10.1109/CVPR.2009.5206744](https://doi.org/10.1109/CVPR.2009.5206744).
- [4] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.
- [5] K. Ohnishi, M. Hidaka, and T. Harada, "Improved dense trajectory with cross streams," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 257–261, doi: [10.1145/2964284.2967222](https://doi.org/10.1145/2964284.2967222).
- [6] L. Li and S. Dai, "Action recognition with spatio-temporal augmented descriptor and fusion method," *Multimedia Tools Appl.*, vol. 76, no. 12, pp. 13953–13969, Jun. 2017.
- [7] 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)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.
- [8] J. Lee, M. Lee, S. Cho, S. Woo, S. Jang, and S. Lee, "Leveraging spatio-temporal dependency for skeleton-based action recognition," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Jun. 2023, pp. 10255–10264.
- [9] H. Zhang, H. Geng, and G. Yang, "Two-stream transformer encoders for skeleton-based action recognition," *Proc. 6th Int. Tech. Conf. Adv. Comput., Control Ind. Eng. (CCIE)*, in Lecture Notes in Electrical Engineering, vol. 920, Y. S. Shmaliy and A. A. Zekry, Eds. Singapore: Springer, 2022, pp. 272–281, doi: [10.1007/978-981-19-3927-3_26](https://doi.org/10.1007/978-981-19-3927-3_26).
- [10] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal segments attention for skeleton-based action recognition," *Neurocomputing*, vol. 518, pp. 30–38, Jan. 2023.
- [11] Q. Zhang, T. Wang, M. Zhang, K. Liu, P. Shi, and H. Snoussi, "Spatial-temporal transformer for skeleton-based action recognition," in *Proc. China Autom. Congr. (CAC)*, Beijing, China, Oct. 2021, pp. 7029–7034, doi: [10.1109/CAC53003.2021.9728206](https://doi.org/10.1109/CAC53003.2021.9728206).
- [12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 568–576.
- [13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, SB, Australia, Apr. 2015, pp. 4580–4584.
- [14] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542.
- [15] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "PA3D: Pose-action 3D machine for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7914–7923, doi: [10.1109/CVPR.2019.00811](https://doi.org/10.1109/CVPR.2019.00811).
- [16] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018, doi: [10.1109/TIP.2017.2778563](https://doi.org/10.1109/TIP.2017.2778563).
- [17] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 3671–3680, doi: [10.1109/CVPR.2017.391](https://doi.org/10.1109/CVPR.2017.391).
- [18] 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, doi: [10.1109/TIP.2018.2812099](https://doi.org/10.1109/TIP.2018.2812099).
- [19] 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)*, Seattle, WA, USA, Jun. 2020, pp. 140–149, doi: [10.1109/CVPR42600.2020.00022](https://doi.org/10.1109/CVPR42600.2020.00022).

- [20] 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.
- [21] M. Wang, X. Li, X. Zhang, and Y. Zhang, "Hierarchical graph attention network with pseudo-metapath for skeleton-based action recognition," *Neurocomputing*, vol. 501, pp. 822–833, Aug. 2022.
- [22] P. P. San, P. Kakar, X. L. Li, S. Krishnaswamy, J. B. Yang, and M. N. Nguyen, "Deep learning for human activity recognition," in *Big Data Analytics for Sensor-Network Collected Intelligence* (Intelligent Data-Centric Systems). Amsterdam, The Netherlands: Elsevier, 2017, ch. 9, pp. 186–204.
- [23] Z. Zhang, X. Ma, R. Song, X. Rong, X. Tian, G. Tian, and Y. Li, "Deep learning based human action recognition: A survey," in *Proc. Chin. Autom. Congr. (CAC)*, Jinan, China, Oct. 2017, pp. 3780–3785, doi: [10.1109/CAC.2017.8243438](https://doi.org/10.1109/CAC.2017.8243438).
- [24] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Hong Kong, Jul. 2017, pp. 597–600, doi: [10.1109/ICMEW.2017.8026285](https://doi.org/10.1109/ICMEW.2017.8026285).
- [25] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
- [26] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Action-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3590–3598, doi: [10.1109/CVPR.2019.00371](https://doi.org/10.1109/CVPR.2019.00371).
- [27] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.
- [28] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7904–7913.
- [29] D. Wang, J. Liu, R. Liu, and X. Fan, "An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection," *Inf. Fusion*, vol. 98, Oct. 2023, Art. no. 101828.
- [30] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022.
- [31] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 4068–4074.
- [32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.
- [33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11531–11539.
- [34] D. Wang, H. Tang, J. Pan, and J. Tang, "Learning a tree-structured channel-wise refinement network for efficient image deraining," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.
- [35] D. Wang, J.-S. Pan, and J.-H. Tang, "Single image deraining using residual channel attention networks," *J. Comput. Sci. Technol.*, vol. 38, no. 2, pp. 439–454, Apr. 2023.
- [36] D. Wang, L. Ma, R. Liu, and X. Fan, "Semantic-aware texture-structure feature collaboration for underwater image enhancement," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Philadelphia, PA, USA, May 2022, pp. 4592–4598.
- [37] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1653–1660.
- [38] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Amsterdam, The Netherlands, Jun. 2021.
- [39] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11210, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, doi: [10.1007/978-3-030-01231-1_33](https://doi.org/10.1007/978-3-030-01231-1_33).
- [40] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11210, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, doi: [10.1007/978-3-030-01231-1_29](https://doi.org/10.1007/978-3-030-01231-1_29).
- [41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2980–2988, doi: [10.1109/ICCV.2017.322](https://doi.org/10.1109/ICCV.2017.322).
- [42] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7297–7306, doi: [10.1109/CVPR.2018.00762](https://doi.org/10.1109/CVPR.2018.00762).
- [43] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," 2018, *arXiv:1803.04831*.
- [44] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Newport Beach, CA, USA, Jul. 2018.
- [45] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019.
- [46] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1227–1236, doi: [10.1109/CVPR.2019.00132](https://doi.org/10.1109/CVPR.2019.00132).
- [47] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2021, pp. 1113–1122.
- [48] X. Qin, R. Cai, J. Yu, C. He, and X. Zhang, "An efficient self-attention network for skeleton-based action recognition," *Sci. Rep.*, vol. 12, no. 1, p. 4111, Mar. 2022, doi: [10.1038/s41598-022-08157-5](https://doi.org/10.1038/s41598-022-08157-5).
- [49] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Computer Vision—ECCV 2016* (Lecture Notes in Computer Science), vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, doi: [10.1007/978-3-319-46487-9_50](https://doi.org/10.1007/978-3-319-46487-9_50).
- [50] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1159–1168.
- [51] 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)*, Seattle, WA, USA, Jun. 2020, pp. 1109–1118, doi: [10.1109/CVPR42600.2020.00119](https://doi.org/10.1109/CVPR42600.2020.00119).
- [52] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 510–519.



JUNXIAN ZHANG is currently a Lecturer with the School of Health, Jiangsu Vocational Institute of Commerce, China. Her research interests include smart elderly care technology and elderly care technology.



AIPING YANG is currently a Professor with the School of Health, Jiangsu Vocational Institute of Commerce, China. Her research interests include artificial intelligence, smart elderly care technology, and nutrition allocation technology.



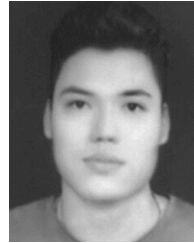
CHANGWU MIAO is currently a Professor with the Jiangsu Vocational Institute of Commerce, China. His research interests include intelligent management and artificial intelligence.



RUI ZHANG is currently an Experimentalist with the School of Health, Jiangsu Vocational Institute of Commerce, China. Her research interests include smart elderly care technology and nutrition allocation technology.



XIANG LI is currently an Associate Professor with the Dalian Neusoft University of Information. Her research interests include artificial intelligence and computer vision.



DANG N. H. THANH received the B.Sc. and M.Sc. degrees in applied mathematics, and the Ph.D. degree in computer science.

He is currently an Assistant Professor with the Department of Information Technology, College of Technology and Design, University of Economics Ho Chi Minh City (UEH), Vietnam. He has about 120 works in international peer-reviewed journals, conference proceedings, book chapters, books, and one European Patent. His research interests include image processing, computer vision, machine learning, data mining, computational mathematics, fuzzy mathematics, and optimization. He is a member of the Scientific Organization INSTICC, Portugal; ACM, USA; and IAENG, Taiwan. He is also a member of international conferences committees, such as IEEE ICCE, Vietnam; IWBBIO, Spain; IEEE ICIEV, USA; IEEE ICEEE, Turkey; ICIEE, Japan; ICoCTA, Australia; and ICMTEL, U.K. He is an Associate Editor of *The Journal of Engineering* (Wiley), a Scientific Editor of *PLOS One*, and an Editorial Member of the *Health Informatics Journal* (SAGE). He served as a Guest Editor for the *Current Medical Imaging* and *Frontiers in Applied Mathematics and Statistics*.

...