Algorithm for Skeleton Action Recognition by Integrating Attention Mechanism and Convolutional Neural Networks

Jianhua Liu*
College of Physical Education, Weifang University, Weifang, 261061, China

***Abstract*—An action recognition model based on 3D skeleton data may experience a decrease in recognition accuracy when facing complex backgrounds, and it is easy to overlook the local connection between dynamic gradient information and dynamic actions, resulting in a decrease in the fault tolerance of the constructed model. To achieve accurate and fast capture of human skeletal movements, a directed graph convolutional network recognition model that integrates attention mechanism and convolutional neural network is proposed. By combining spacetime converter and central differential graph convolution, a corresponding central differential converter graph convolutional network model is constructed to obtain dynamic gradient information in actions and calculate local connections between dynamic actions. The research outcomes express that the cross-target benchmark recognition rate of the directed graph convolutional network recognition model is 92.3%, and the cross-view benchmark recognition rate is 97.3%. The accuracy of Top-1 is 37.6%, and the accuracy of Top-5 is 60.5%. The cross-target recognition rate of the central differential converter graph convolutional network model is 92.9%, and the cross-view benchmark recognition rate is 97.5%. Undercross-target and cross-view benchmarks, the average recognition accuracy for similar actions is 81.3% and 88.9%, respectively. The accuracy of the entire action recognition model in single-person multi-person action recognition experiments is 95.0%. The outcomes denote that the model constructed by the research institute has higher recognition rate and more stable performance compared to existing neural network recognition models, and has certain research value.**

Keywords*—Attention mechanism; convolutional neural network; action recognition; central differential network; spacetime converter; directed graph convolution

I. Introduction

Human Action Recognition (HAR) is a research field that has received much attention from scholars both domestically and internationally. HAR plays an important role in public places, healthcare, and safety [1-2]. Early research on HAR was generally based on RGB videos. With the development of technology, algorithms and sensors for estimating human skeletal posture have emerged [3]. Researchers use depth cameras to extract depth information from human skeletal information at different scales, lighting, and backgrounds to construct HAR data in a 3D coordinate system, and construct corresponding algorithm models based on deep learning (DL) [4]. In the research of HAR using 3D skeleton data as the experimental object, HAR and feature capture often use 3D technology to obtain relevant action data information [5]. However, there are still many problems in the rapid acquisition of human actions in related operations. The changes and complexity of the background and environment in the video can result in different skeletal joint information for the same action, and the mutual occlusion between people and background can also reduce the extraction of feature actions. The already built HAR model may also have problems, such as identifying only simple action types, making it difficult to identify more complex behaviors. The above issues will result in the already constructed HAR model being only usable under specific conditions, and the experimentally constructed model cannot be generalized to practical applications. To improve the recognition accuracy of action recognition (AR) models and eliminate interference caused by factors such as environment, background, and occlusion, a directed graph convolutional network (DGCN) recognition model for enhancing attention was proposed. By recognizing the spacetime information extracted from feature extraction, a central differential converter graph convolutional network was constructed to achieve real-time recognition and capture of human actions. The Section II of the study mainly discusses the relevant research on human action models in recent years. The Section III mainly constructs an enhanced information acquisition and enhanced spatiotemporal information conversion graph convolution model. The Section IV tests the performance of the constructed model using different datasets and compares the effectiveness of different HAR models on the same dataset. Section V discusses the results. The Section VI summarizes the research results and draws research conclusions.

II. Related Works

Many scholars have made achievements in the construction of HAR models. Gu et al. constructed a new HAR model using the improved sparse classification model and deep convolutional neural network (CNN), and applied the model to the benchmark dataset. The performance of the model was verified through setting experiments [6]. Huang et al. utilized neural networks for pseudo-image processing of skeletal data. A novel CNN with an adaptive inference framework was constructed by utilizing the dependent joints between skeletal joints [7]. To improve the detection accuracy of human bone models, Li proposed a multi-branch multi-level cascaded CNN structure model to predict the information of occluded parts [8]. Peng et al. proposed a three stream model using two different types of deep CNNs to improve the generalization of HAR models, and verified the effectiveness of the model by

experiments [9]. Yang put forward an algorithm model based on random projection combined with multi-channel 3D CNN to improve the accuracy and recognition speed of HAR models. The effectiveness of the algorithm model was demonstrated through experiments [10]. To raise the accuracy of human action model recognition in video segmentation and its computational efficiency in large-scale datasets, Zhao et al. proposed a multi-dimensional data model for video image AR and non-action based on DL framework, and verified the effectiveness of the model through experiments [11]. Liu and Che used attention spacetime convolutional graph networks to learn the importance of different actions in sports videos and analyze different actions in sports videos [12]. To solve the problems of background clutter, scene diversity, viewpoint change, occlusion, etc. in the HAR, He et al. put forward a closely connected bidirectional long and short-term memory (LSTM) network DL model to capture the temporal and spatial patterns of human action in videos. The research findings also indicated that the effectiveness of the proposed model was good [13]. Wenbo et al. put forward a protocol recognition method based on CNNs to improve the accuracy of feature acquisition, and it was proved that the proposed algorithm had high accuracy and fast convergence speed [14]. Ma et al. proposed a novel deep convolutional generative adversarial network to recognize human action posture, and verified through experiments that its model had significant advantages over existing models [15]. Chen et al. put forward a multi-radar collaborative HAR model based on transfer and integrated learning to solve the view limitation in AR. Through experiments, the proposed model had higher recognition accuracy compared to the single-view radar fusion model [16]. To promote the accuracy of action combination training AR model, Jiang and Tsai proposed a model based on sequential minimal optimization model and artificial intelligence. And in subsequent experiments, it has been proven that this model could improve the recognition rate of actions and meet the recognition needs of online actions [17]. To establish a better HAR model, Chen et al. proposed a model that integrated LSTM weight convolution neural networks. It was validated that the model had an authentication accuracy of 98.0% [18].

In summary, research on constructing HAR models using CNNs has become mature, but there is still room for improvement in accuracy. Based on this, a CNN human recognition model integrating attention mechanism (AM) is proposed, which combines central difference graph convolution network (CDGN) and spacetime converter (SC) to achieve accurate and fast human actions recognition.

III. The Construction of Skeleton AR Model Based on AM and CNN

This study mainly uses graph convolutional neural networks (GCNN) as the skeleton and integrates AMs to design a DGCN model to enhance the recognition and capture of action information. To prevent the omission of dynamic gradient information in actions and calculate local connections between dynamic actions, CDGN and SC are introduced to construct an enhanced graph convolution model for spacetime information conversion.

A. The Construction of ADGCN Skeleton AR Model Integrating AM and Neural Network

HAR, as an emerging research project in computer vision, has great impact on many fields. Medical assistance, human-computer interaction, intelligent driving, and intelligent security can be achieved through the recognition of HAR. The method of skeleton AR is generally completed using 3D skeleton data, using skeleton joint points to form corresponding groups, and then connecting the corresponding groups to complete the representation of the entire body structure. At present, the commonly used methods for HAR include two types: extraction based on traditional manual features and extraction based on DL methods. DL methods mainly include methods based on recurrent neural network (RNN), CNN, fusion based on RNN and CNN, and GCNN. GCNN takes skeleton data modeling as a blueprint, joints as fixed points, and draws skeleton edge maps. Compared to traditional manual feature extraction and CNN and RNN methods, it has a more intuitive description of skeleton data. Therefore, the study selects the GCN method as the basis to construct a model that is more convenient for skeleton AR. The flow chart of GCN object diagram convolution is shown in Fig. 1.

From Fig. 1, the GCN method involves the temporal structure and continuity of human skeleton actions recognition. By decomposing a certain action information into corresponding data, and then placing the decomposed data into different channels according to dimensional relationships for processing and operation, the final action analysis result is obtained. Due to the existing skeleton models constructed based on GCN, their skeleton AR accuracy may be reduced in complex backgrounds and dynamic environments. And the skeleton model constructed by GCN is often an undirected graph, which can only determine whether there is a connection relationship between adjacent vertices and edges in the graph, and cannot capture more feature information.

To address the above issues, an AM is introduced to improve the GCN skeleton model. A multi-stream framework is used to capture the positions of joints and bones, identify their action trajectories and features, establish the dependency relationship between vertices and edges in the skeleton graph, introduce an AM to focus on feature channels, and construct an attention enhanced direct graph convolutional network (ADGCN) skeleton AR model. The workflow diagram of the ADGCN model is displayed in Fig. 2.

From Fig. 2, the ADGCN skeleton AR model is divided into five parts: multi-stream framework input, DGCN processing, attention enhancement network, data fusion, and output data. The data input by the multi-stream framework corresponds to the action information of joints and bones. The related motion information is judged by the adjacency of DGCN and captured by the key frame action of the attention enhancement network. The data processed by the first four information flow are weighted and fused to complete the prediction of related actions. The update and aggregate functions are used to determine the connection of vertices and edges in the ADGCN model to represent the connected vertices between two vertices and the relationship between vertices and

edges after edge update. The update function of vertices is shown in equation (1).

$$v_i' = u^v([v_i, \bar{e}_j, \bar{e}_j]) \quad (1)$$

In equation (1), v_i' is the updated vertex, v_i is the vertex; [.] means the connection operation; u indicates the update function; $\bar{e}_j, \bar{e}_j$ denote the position of the updated vertex v_i' and its adjacent edge e. Similarly, it obtains the function equation (2) between the updated edge e' and adjacent vertices.

$$e_j' = u^e([e_j, v_j^{s'}, v_j^{t'}]) \quad (2)$$

In equation (2), $v_j^{s'}, v_j^{t'}$ stand for the updated source vertex and target vertex, respectively. Let the joint coordinate at time t be $v_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$, and the calculation function for the same node in two consecutive frames is shown in equation (3).

$$m_{i,t+1} = v_{i,t+1} - v_{i,t} \quad (3)$$

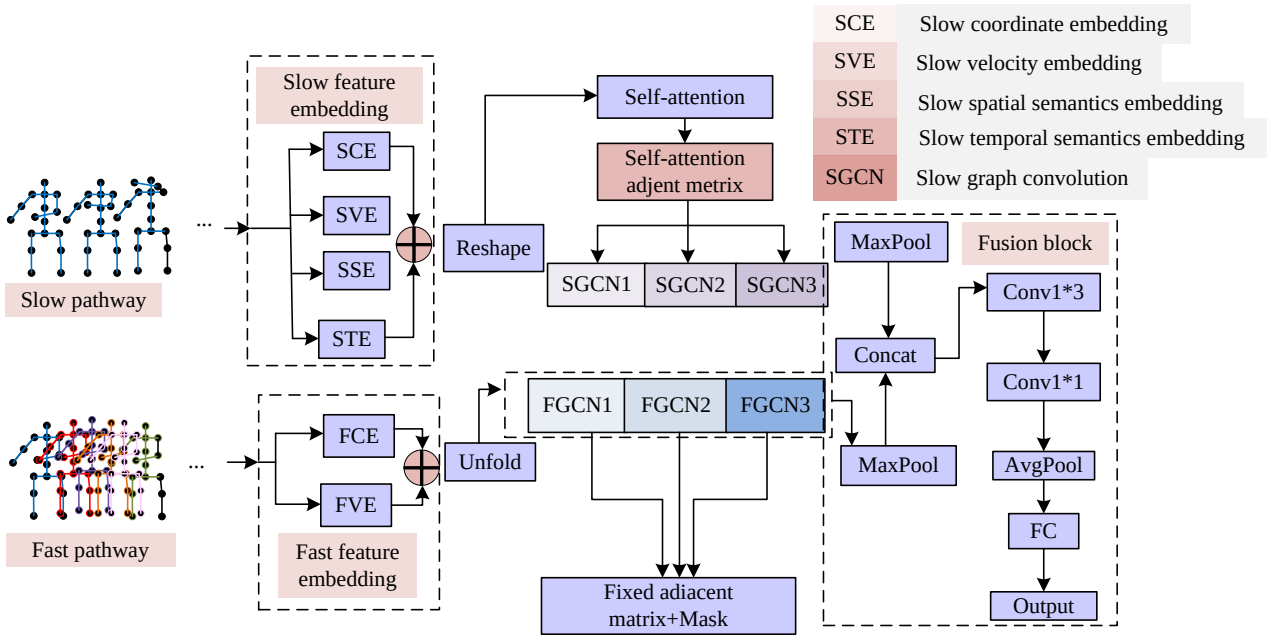


Fig. 1. Flow chart of GCN object diagram convolution.

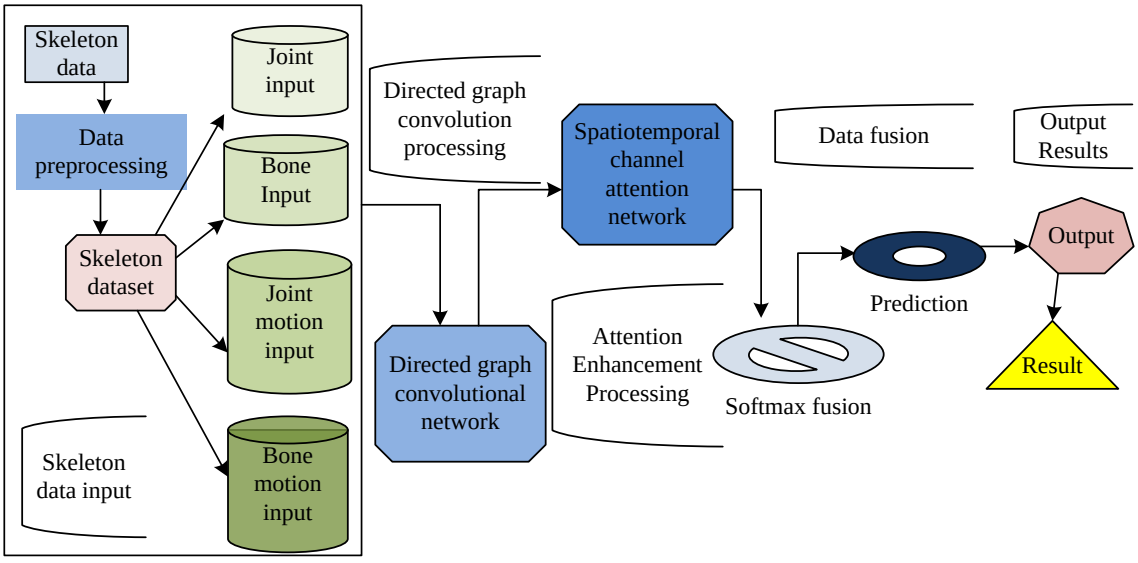


Fig. 2. ADGCN model workflow diagram.

Designing a spacetime channel attention network (CAN) is to strengthen the information connection between key joints and bones, while taking care not to weaken the information intensity of other joints. The spacetime CAN constructed by the research institute includes three types: spatial attention network (SAN), temporal attention network (TAN), and CAN. Among them, SAN mainly allocates different degrees of attention to joints and the bones connected to them. The attention of vertices and edges in the graph is M_s, and the set of M_s is expressed as $\{M_{s,v}, M_{s,e}\}$. The corresponding calculation method is shown in equation (4).

$$M_s = \delta(g_s(AvgPool(fin))) \tag{4}$$

In equation (4), δ means the sigmoid activation function; g_s indicates a one-dimensional convolution operation; fin refers to the input data information of vertices and edges in the directed graph. As a dynamic time selection mechanism, TAN mainly determines when attention begins and sets the attention time of vertices and edges in a directed graph to M_t. The relevant function expressions are shown in equation (5).

$$M_t = \delta(g_t(AvgPool(fin))) \tag{5}$$

In equation (5), $M_t \in R^{1*T*1}$, the explanation of parameter meanings refers to equation (4). After allocating time attention and attention to joints and edges, it considers enhancing the model's feature description of input samples, and thus introduces CAN. The relevant calculation is shown in equation (6).

$$M_c = \delta(W_2(\tanh(W_1(AvgPool(fin))))) \tag{6}$$

In equation (6), W_1 and W_2 are the weights of two full connection layers, and the functions used by TAN network are tanh and Sigmaid activation functions. SAN, TAN, CAN three kinds of attention enhancing networks are serially arranged to enhance the recognition ability of spacetime CAN and highlight its attention enhancement effect.

B. Constructing a Graph Convolutional Model for Enhanced Spatiotemporal Information Transformation Based on CDTN

After the construction of the enhanced information acquisition model, its ability to obtain spacetime channel information has been raised to a certain degree, but it ignores the dynamic gradient information in the actions and the local connections between the corresponding dynamic actions. Introducing CDGN can further improve the model constructed by the research institute, by obtaining corresponding dynamic gradient information and using a converter to obtain fixed point connections between nodes. CDGN, as an image processing algorithm, is mainly used for image edge detection and feature extraction. It uses the difference between the central and adjacent pixels to calculate the value of new pixels, reducing noise interference in the image and enhancing its features. The CDGN algorithm is shown in equation (7).

$$\begin{cases} g\_x(x,y) = f(x+1,y) - f(x-1,y) \\ g\_y(x,y) = f(x,y+1) - f(x,y-1) \end{cases} \tag{7}$$

In equation (7), $g\_x(x,y)$ and $g\_y(x,y)$ express different gradient values in the horizontal and vertical directions, respectively, and $f(x,y)$ means the pixel values in the original image. The edge values and features of the image are calculated by calculating the gradient value of each pixel point. CDGN includes two parts: sampling and aggregation, and its feature aggregation is shown in Fig. 3.

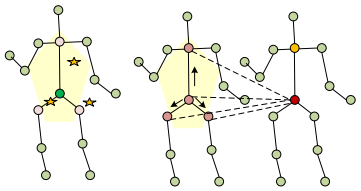


Fig. 3. CDGN feature aggregation principle.

From Fig. 3, the sampling of CDGN takes into account both feature differences and gradient features. After taking into account the differences between certain node and central node and the gradient features of the corresponding central node, the connection is made according to the corresponding gradient direction. From this, the center gradient of the sampling vertices is aggregated and represented by equation (8).

$$O(v_i) = \sum_{v_j \in Ri} \frac{1}{Z_{ij}} w(l_i(v_j)) * (I(v_j) - I(v_i)) \tag{8}$$

In equation (8), I indicates the input feature; O means the output feature; w indicates the weight function; Ri expresses the first order adjacency joint distance of vertex v_i; l_i is the partition function. The converter network can coordinate the global self-attention, self-attention operation and convolution operation. Combining the CNN with the converter can allow the model to mine more relevant information from the captured features. The SC used in the research institute consists of three parts: joint embedding, SC attention module, and time converter attention module. The network structure of the SC is shown in Fig. 4.

From Fig. 4, different modules implement different functions. Among them, the joint marker embedding module includes a multi-head self-attention (MHSA) layer and a feedback neural network (FNN) layer to extract spacetime features. The SC attention module is applied to the acquisition of labeled space, while the time converter attention module corresponds to the calculation of spatial and temporal dependencies. Combining CDGN and SC network models, a central difference transformer graph network (CDTG) is developed to capture action gradient information and calculate local dependencies between nodes. The CDTG network structure is shown in Fig. 5.

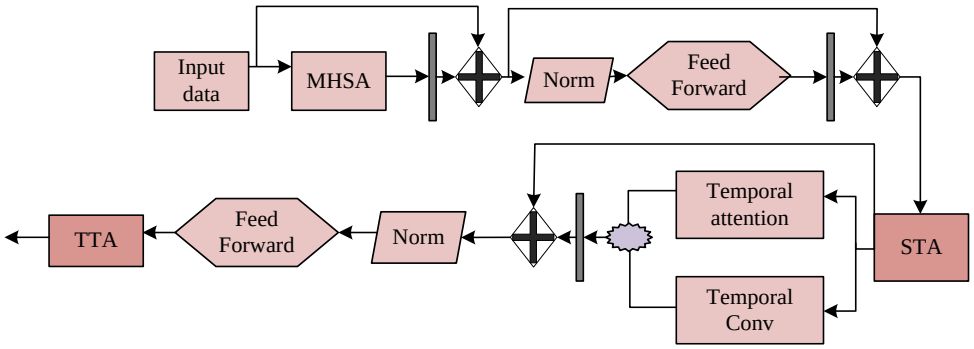


Fig. 4. Diagram of SCNetwork.

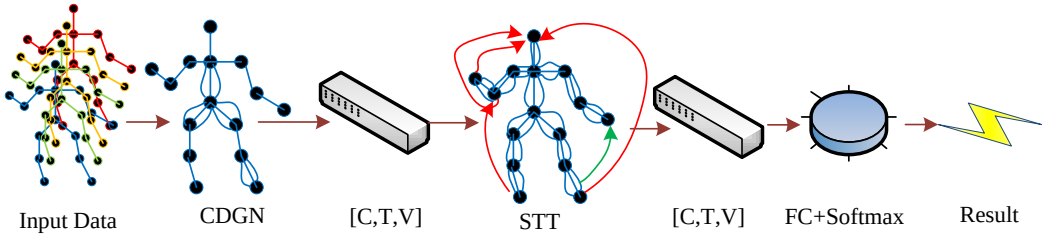


Fig. 5. CDTG Network architecture.

From Fig. 5, CDTG is divided into four parts: skeleton data input, CDGN, SC attention network, and fully connected layer recognition information. The last one will use the Softmax function, and the relevant expressions are shown in equation (9).

$$P(y=j)=\frac{e^{x^T W_j}}{\sum_{k=1}^{k} e^{x^T W_k}} \quad (9)$$

In equation (9), $P(y=j)$ expresses the probability that the sample x belongs to the j classification; k denotes the result of linear function, and is the weight value. The expression of cross entropy function is shown in equation (10).

$$H(p,q)=\sum_{x} p(x)\log q(x) \quad (10)$$

In equation (10), p means the true labeled distribution, and q refers to the pre labeled distribution of the trained model. It inputs the skeleton action information into CDGN, updates the spatial gradient information of the next layer's vertices based on the position of the central node and its adjacent nodes. The dependency relationship of the global spacetime nodes is calculated based on the attention network of the SC, and finally the data are processed through the fully connected layer recognition information processing and weighted average to complete the prediction of the entire action behavior.

IV. Analysis of AR Results for ADGN and CDTG Models

The experiment used NTU-RGB+D, MSR-Action 3D, and SBU datasets collected by the Kinect v2 3D tactile camera as the research dataset. Experiments were organized to evidence the effectiveness of the proposed ADGCN algorithm model. The NTU-RGB+D dataset contained 56880 skeleton action video sequences, with each skeleton action video covering 25 nodes and each node providing corresponding 3D coordinate positions. MSR-Action 3D contained 20 types of actions and 567 data sequences. The SBU dataset contained 8 types of actions, 284 videos, and 15 skeleton joint points. The environmental parameters for the skeleton AR experiment are shown in Table I.

TABLE I. Experimental Parameter Settings for Skeleton AR

Experimental setup	Experimental parameters
Experimental system	Linux Ubuntu18.04
GPU	GeForce RTX2080Ti
Computing Platform	CUDA10.0

The batch size of the model is set to 64 and the initial weight attenuation value to 0.0005. The model was trained in the NTU-RGB+D dataset. The accuracy of the trained model was tested on the NTU-RGB+D dataset for crosstarget and view benchmarks. And it compared the accuracy of manual feature extraction algorithms Lie Group, RNN-based feature extraction methods (GCA-LSTM and STA-LSTM), CNN-based methods (3SCNN and TCN), and GCN-based methods (ST-GCN, 2sAGCN, DGCN) on CV and CS. The research outcomes are expressed in Fig. 6.

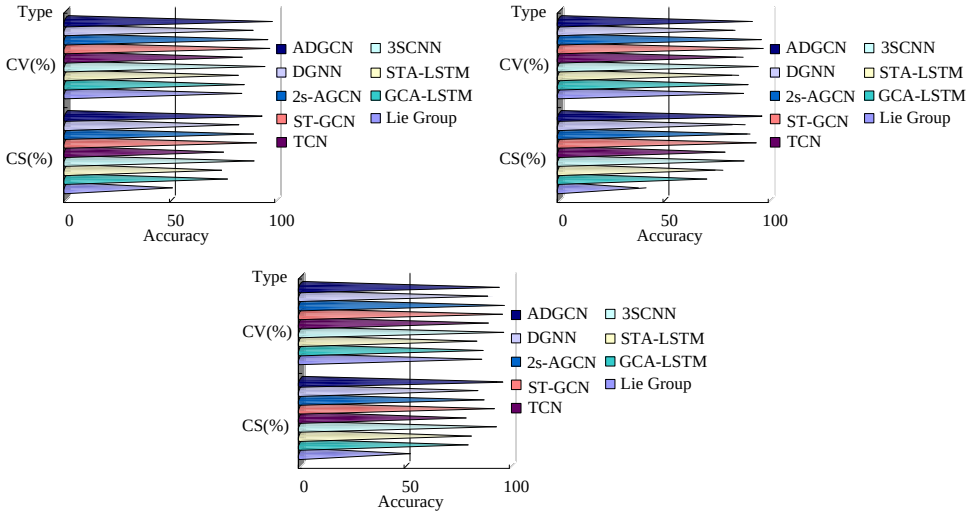
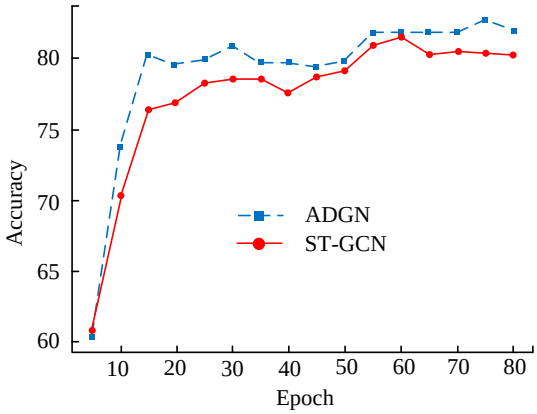
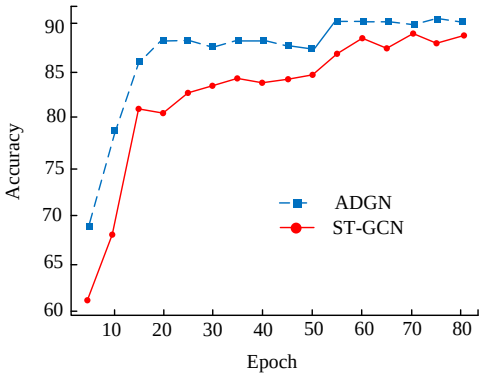


Fig. 6. Performance comparison between algorithms on the NTU RGB+D dataset.

From Fig. 6, in the three datasets of NTU RGB+D, MSR-Action 3D, and SBU, the artificial feature extraction method Lie Group had the lowest CS recognition accuracy, with 50.1%, 47.5%, and 50.5%, respectively, while other DL-based methods had an accuracy of over 73% in CS benchmark recognition. The CS accuracy based on the GSCN network structure was over 80%, with the highest being ST-GCN89.90%. The ADGCN model proposed by the research institute scored the highest among the four network models, with CS benchmark recognition accuracy of 92.33%, 93.21%, and 94.12% in the three datasets, respectively. In cross-view benchmark recognition, the recognition accuracy of the four algorithms was higher than that of cross-target benchmark recognition. The model based on the GSCN network structure still had higher CV recognition accuracy than the other three network models, while the ADGCN model proposed by the research institute had the highest CV recognition accuracy in the past, with 97.3%, 97.8%, and 96.7% in the three datasets, respectively. To further validate the effectiveness of the model, the ST-GCN model with the second highest score was selected to compare the results of the proposed ADGN model CS and CV benchmark recognition accuracy changes with the training. The laboratory findings are shown in Fig. 7.



(a)Comparison of contrast rates between two models under CS



(b)Comparison of contrast rates between two models under CV

Fig. 7. Changes in the accuracy of CS and CV benchmark recognition for two algorithms during training.

From Fig. 7, as the number of iterations continued to increase, the accuracy of the two bone AR models would show an upward trend. Under the CS benchmark, after the number of iterations reached 15, the recognition accuracy of the model tended to stabilize. However, the recognition accuracy of the ADGN model has always been higher than that of the ST-GCN model, with a maximum difference of 0.5% and a maximum difference of 3.42%. Under the CV benchmark, the recognition accuracy of both models has increased compared to the CS benchmark, with an increase of about 5%. After 15 iterations, the recognition accuracy of the model tended to a relatively stable accuracy range, with the maximum recognition accuracy of ADCN being 90%, ST-GCN being 88.5%, and the ADCN recognition accuracy curve consistently above ST-GCN. The experimental results indicated that the ADCN model constructed by the research institute had a certain degree of stability and could ensure good recognition accuracy. Experiments were designed to verify the recognition accuracy of the ADGCN model for different levels of actions on the Kinetics Skelton dataset, and it compared the recognition accuracy of Deep LSTM, TCN, DGCN, and SAN algorithms for the same action. The research outcomes are displayed in Fig. 8.

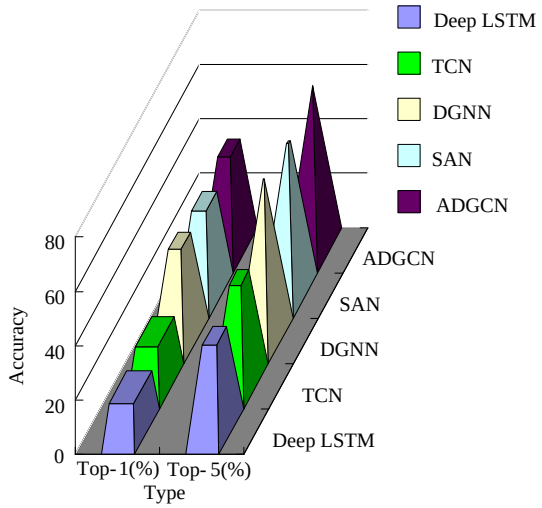


Fig. 8. Comparison of recognition accuracy of the kinetics skeleton dataset.

Due to the fact that the action videos in the Kinetics Skelton dataset are sourced from YouTube, with a larger variety and complexity compared to the previously used NTU RGB+D dataset, the recognition accuracy of the five validated models is lower. From Fig. 8, in the type of Top-1, the recognition accuracy of the SAN, DGNN, and ADGCN models was around 35%, belonging to the three branches with good performance among the five algorithms. The ADGCN model's Top-1 AR accuracy was higher than DGNN with a 0.7% advantage, ranking first. Compared to Top-1, in Top-5 indicator recognition, the recognition accuracy of the five research algorithms has improved, with an increase of 18.6% to 23.0%. Among them, DGNN had the largest increase, the Top-5 accuracy of the model was 56.5%, and ADGCN had an increase of 22.9%. However, the recognition accuracy of ADGCN was the highest, with 60.5%. It designed experiments to evidence the effectiveness of the CDTG model proposed by the research institute. In the NTU-RGB+D dataset, it compared the recognition accuracy obtained by CNN-based methods (HCN, TCN, GCNN, Clips+ CNN+MTLN), RNN-based methods (ST-LSTM, LSTM-CNN), and GCN-based methods (ST-GCN STGR-GCN GR-GCN GCST Dynamic GCN). The research findings are expressed in Fig. 9.

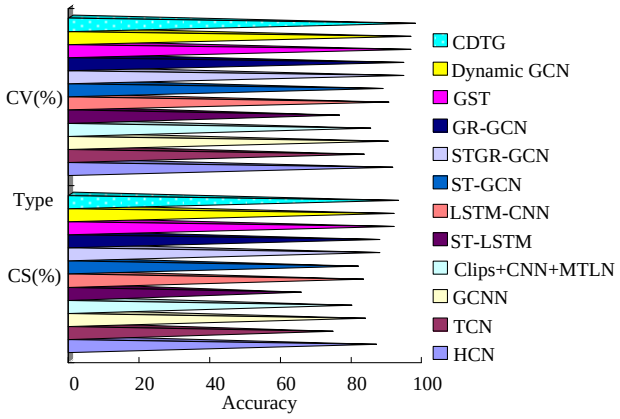
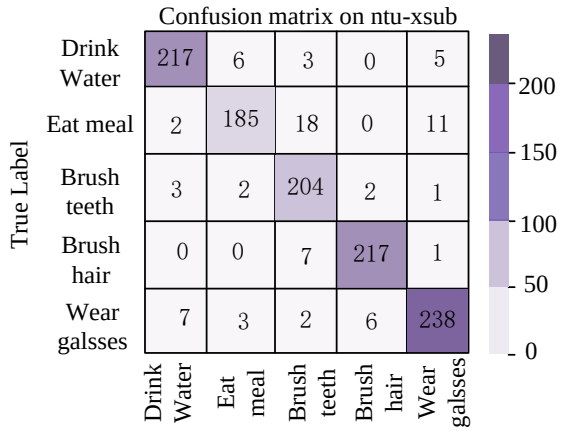
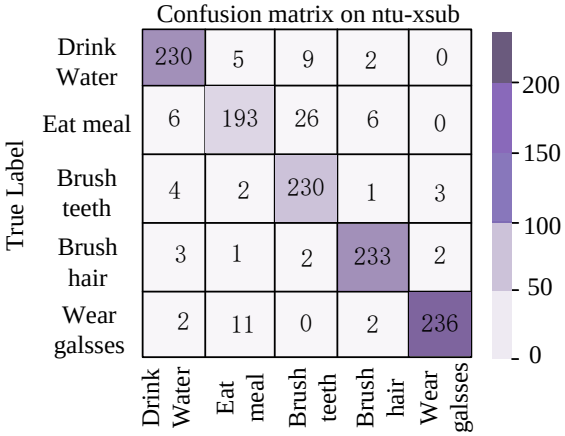


Fig. 9. Comparison of recognition accuracy between different algorithms.

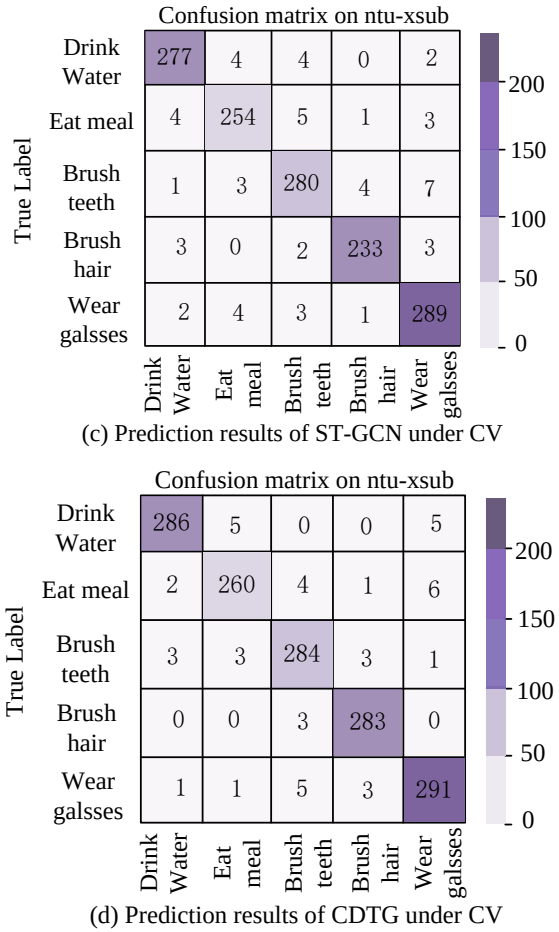
From Fig. 9, the recognition accuracy of the model based on GCN network structure on CS and CV benchmarks was higher than that of the model built on CNN and RNN. Among the models constructed using GCN networks, GCST, Dynamic GCN, and the CDTG model constructed by the research institute had relatively good recognition accuracy. Among them, the CDTG model had the highest recognition accuracy, with recognition accuracy rates of 92.87% and 97.52% in CS and CV, respectively. Compared with other types of AR models, its accuracy improvement could reach a maximum of 27.67% (CS benchmark) and 21.42% (CV benchmark). Compared to the GCST and Dynamic GCN models with higher AR accuracy in recent years, the recognition accuracy of CDTG in CS benchmark had increased by 1.36% and 1.37%, respectively. The recognition accuracy of CV benchmark has been improved by 1.32%. This indicated that the CDTG model proposed by the research institute had good recognition performance in skeleton AR. To further evidence the recognition performance of the CDTG model, five categories with similar action content were selected for comparison in the NTU-RGB+D dataset. The selection of similar action content was divided into five categories: "drinking water", "eating", "brushing teeth", "washing hair", and "wearing glasses". The experimental results were set up to verify the recognition results of CDTG and ST-GCN for the above five action categories. The experimental results are shown in Fig. 10.



(a) Prediction results of ST-GCN under CS



(b) Prediction results of CDTG under CS

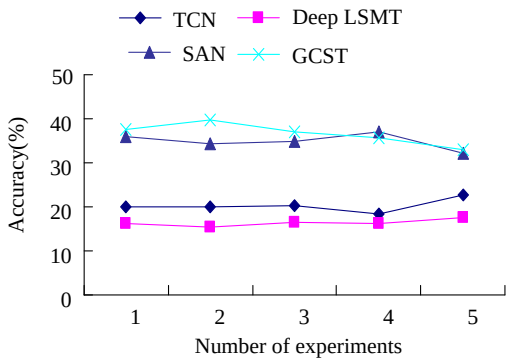


(c) Prediction results of ST-GCN under CV

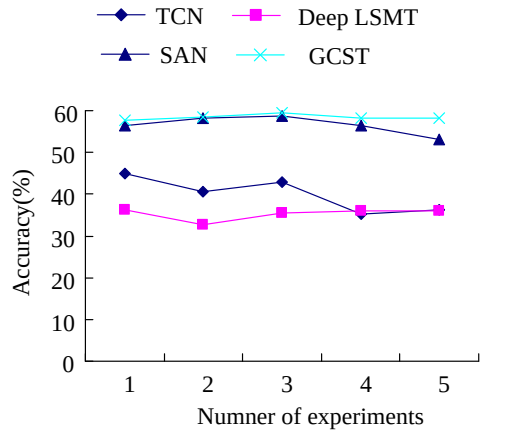
(d) Prediction results of CDTG under CV

Fig. 10. Comparison of recognition results of similar actions between ST-GCN and CDTG models.

From Fig. 10, the recognition performance of the two models under the CV benchmark was better than that under the CS benchmark. There were 276 CS benchmark samples to be tested, among which the recognition numbers for the five actions of "drinking water", "eating", "brushing teeth", "washing hair", and "wearing glasses" in ST-GCN were 217, 185, 204, 217, and 238, respectively, with recognition accuracy rates of 78.6%, 67.0%, 73.9%, 78.6%, and 86.2%. The average recognition rate was 76.9%. The recognition rates of the CDTG model for the above five actions corresponded to 83.3%, 69.9%, 83.3%, 84.4%, and 85.5%, with an average recognition rate of 81.3%. The number of samples to be tested under the CV benchmark was 316, indicating that the recognition accuracy of the CDTG model was better than that of ST-GCN. The recognition rates for "drinking water", "eating", "brushing teeth", "washing hair", and "wearing glasses" actions were 90.5%, 82.3%, 89.9%, 89.6%, and 92.1%, respectively. The average rate of AR was 88.9%. The average recognition rate of ST-GCN was 86.3%. It designed experiments to verify the recognition accuracy of CDTG on Top-1 and Top-5 in the Kinetics Skelton dataset, and compared the recognition accuracy of TCN, Deep LSTM, and SAN GCST models. The experimental results are shown in Fig. 11.



(a) Comparison of Model Performance Regarding Top-1 Indicators



(b) Comparison of Model Performance Regarding Top-5 Indicators

Fig. 11. Comparison of accuracy of five algorithms on the kinetics skeleton dataset.

From Fig. 11, the CDTG model had better recognition accuracy in Top-1 and Top-5 compared to other models, with the highest accuracy among the comparison models, at 37.58% and 59.61%, respectively. Compared to other types of AR models, its Top-1 had a maximum improvement accuracy of 21.18%. The minimum improvement accuracy was 0.88%. The highest improvement in recognition accuracy of Top-5 was 24.31%, and the lowest improvement accuracy was 1.25%. Four experimental testers (two males and two females) were identified and predicted for 3D movements using ADGN and CDTG models. The research outcomes are shown in Table II.

TABLE II. RESULTS OF ADGN AND CDTG MODELS ON PHYSICAL EXPERIMENTS

Action label	Drinking	Smoking	Walking	Jumping
Male tester 1	True	True	True	True
Male tester 2	True	True	True	True
Female tester 3	True	False	True	True
Female tester 4	True	True	True	True
Action label	Shaking hands		Hugging	
Tester 1+2	True		True	
Tester 3+4	True		True	

From Table II, the combination model constructed by the research institute can accurately recognize single and multi-person movements, with a recognition accuracy of 100% for "drinking water", "walking", and "jumping" movements. However, in the recognition of "smoking" movements, female test subject 3 made an error in identifying the movements, which might be due to the insufficient fine-grained model constructed by the research institute. On the other hand, it might also be that female testers were not familiar with the 'smoking' action, which led to model recognition errors. The experimental results indicated that the action model constructed by the research institute had good recognition performance.

TABLE III. RESPONSE TIME AND SATISFACTION BETWEEN DIFFERENT MODELS

Indicator Name	ADGN Model	CDTG Model	ST-GCNModel	GCSTModel
Response time (ms)	91	86	156	137
Satisfaction (%)	90	92	76	78
Feasibility (%)	91	93	84	81

Table III shows the response time and satisfaction between different models. The response times of the ADGN model and CDTG model proposed in the article were 91ms and 86ms, respectively. Compared with the ST-GCN model and GCST model, they had faster reaction speed and better low latency. Meanwhile, the above two models achieved 90% and 92% user satisfaction during the application process, and the feasibility under user evaluation also reached 91% and 93%, respectively. Therefore, the ADGN model and CDTG model proposed in this study had relatively superior practical application value in different regions.

V. DISCUSSION

Skeleton AR can be used in HAR to promote the expression generalization ability of human behavior. At present, research on skeleton AR has received widespread development and attention. Fang Z et al. used GCN to capture the activity status of skeleton joints and identify their movements. The accuracy of the skeleton AR model constructed in this experiment was only 87.79% [19]. Considering the background complexity of skeleton AR, this study optimized the skeleton AR model using AM and CDGN, and the final recognition accuracy of the models was above 90%. To improve the robustness of the skeleton AR model, Liu Y et al. used the KA-AGTN algorithm to construct a skeleton AR model mainly based on time series. Although it improved the recognition accuracy by 1.9% compared to traditional GCN models, it was lower than the model recognition accuracy proposed in this study [20]. The reason why the model proposed in the study was more excellent was that AMcould enhance the model's ability to capture key actions, while CDGN could reduce noise interference in AR. Therefore, the optimized skeleton AR model had higher accuracy.

VI. CONCLUSION

To accurately capture and recognize spatiotemporal information and human actions in AR models, an ADGN recognition model based on spatiotemporal CNN was proposed. By introducing the AM, the ADGN model enhanced the ability to obtain joint information. By integrating the central differential and spatiotemporal transformation networks simultaneously, a CDTG model could be constructed to compensate for the insufficient local dependency of joint information in the ADGN model. The experimental results showed that the ADGN model constructed in this study had CS benchmark recognition accuracy of 92.33%, 93.21%, and 94.12% in NTU RGB+D, MSR-Action 3D, and SBU datasets, respectively. The accuracy of CV recognition was 97.3%, 97.8%, and 96.7%, respectively. And the recognition accuracy of the ADGN model has reached 90%. The CDTG model had the highest recognition accuracy, with recognition accuracy rates of 92.87% and 97.52% in CS and CV, respectively. The response times of ADGN and CDTG modelsin application were 91ms and 86ms, respectively, with satisfaction rates of 90% and 92%, and feasibility rates of 91% and 93%, respectively. Therefore, ADGN and CDTG models proposed in this study had relatively superior practical application value in different regions. However, due to the small sample size of the entity and the fact that the model has not yet been AR validated in complex backgrounds, there is still room for improvement. At the same time, the combination of RGB or depth information is beneficial for optimizing the performance of the model, so other parameter conditions can be used to optimize the research model in the future.

VII. FUTURE RESEARCH WORK

Due to the fact that this study was not tested in a complex background, the test results of the model's performance were idealized. Considering the practical application of the model, the study will set up different complex environmental scenarios for model performance evaluation in subsequent experiments. The performance results were summarized and analyzed in complex scenarios, and the model is adjusted and optimized by setting different parameter conditions to improve its accuracy and efficiency. In addition, because the algorithm used in this study is an AM to enhance the performance of ADGN and CDTN models, and multimodal fusion forms such as RGB or depth information can also improve model performance. As a result, model performance tests can be conducted under different modalities of fusion to select the optimal multimodal fusion method to construct skeleton AR algorithms.

REFERENCES

[1] H. Zhao, W. Xue, X. Li, Z. Gu, L. Niu, and L. Zhang, "Multi-mode neural network for human action recognition," IET Comput. Vis., vol. 14, no. 8, pp. 587-596, Nov. 2020.

[2] C. Peng, H. Huang, A. C. Tsoi, S. L. Lo, Y. Liu, and Z. Yang, "Motion boundary emphasised optical flow method for human action recognition," IET Comput. Vis., vol. 14, no. 6, pp. 378-390, Jul. 2020.

[3] W. R. Ko, M. Jang, J. Lee, and J. Kim, "AIR-Act2Act: Human‐human interaction dataset for teaching non-verbal social behaviors to robots," Int. J. Robot. Res., vol. 40, no. 4-5, pp. 691-697, Jan. 2021.

[4] Y. Yang and X. Song, "Research on face intelligent perception technology integrating deep learning under different illumination intensities," J. Comput. Cogn. Eng., vol. 1, no. 1, pp. 32-36, May. 2022.

[5] Y. Lei, "Research on microvideo character perception and recognition based on target detection technology," J. Comput. Cogn. Eng., vol. 1, no. 2, pp. 83-87, May. 2022.

[6] B. Gu, W. Xiong, and Z. Bai, "Human action recognition based on supervised class-specific dictionary learning with deep convolutional neural network features," Comput., Mater. Contin., vol. 63, no. 1, pp. 243-262, Mar. 2020.

[7] H. Huang, H. Su, Z. Chang, M. Yu, J. Gao, X. Li, and S. Zheng, "Convolutional neural network with adaptive inferential framework for skeleton-based action recognition," J. Vis. Commun. Image Represent., vol. 73, no. 11, pp. 102925.3-102925.10, Nov. 2020.

[8] Z. Li, "Three-dimensional diffusion model in sports dance video human skeleton detection and extraction," Adv. in Math. Phys., vol. 2021, no. 3, pp. 3772358.5-3772358.15, Sept. 2021.

[9] C. Peng, H. Huang, A. Tsoi, S. Lo, Y. Liu, and Z. Yang, "Motion boundary emphasised optical flow method for human action recognition," IET Comput. Vis., vol. 14, no.6, pp. 378-390, Jul. 2020.

[10] J. Yang, "Study of human motion recognition algorithm based on multichannel 3D convolutional neural network," Complex., vol. 2021, no. 18, pp. 7646813.54-7646813.62, May. 2021.

[11] P. Zhao, D. Zhao, and X. Chen, "Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework," IET Image Process., vol. 14, no. 7, pp. 1257-1264, Apr. 2020.

[12] J. Liu and Y. Che, "Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network," J. Electron. Imaging, vol. 30, no. 3, pp. 33017.3-33017.16, Jun. 2021.

[13] J. He, X. Wu, Z. Cheng, Z. Yuan, and Y. Jiang, "DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition," Neurocomputing, vol. 444, no. 15, pp. 319-331, Jul. 2021.

[14] W. Feng, Z. Hong, L. Wu, M. Fu, Y. Li, and P. Lin, "Network protocol recognition based on convolutional neural network," China Commun., vol. 17, np. 4, pp. 125-139, Apr. 2020.

[15] R. Ma, Z. Zhang, and E. Chen, "Human motion gesture recognition based on computer vision," Complex., vol. 21, no. 5, pp. 6679746.15-6679746.26, Feb. 2021.

[16] P. Chen, S. Guo, H. Li, X. Wang, G. Cui, C. Jiang, and L. Kong, "Through-wall human motion recognition based on transfer learning and ensemble learning," IEEE Geoscie. Remote Sens. Lett., vol. 191, no. 5, pp. 66-74, Apr. 2022.

[17] H. Jiang and S. Tsai, "An empirical study on sports combination training action recognition based on smo algorithm optimization model and artificial intelligence," Math. Probl. Eng.: Theory, Meth. Appl., vol. 2021, no. 31, pp. 7217383.46-7217383.51, Jul. 2021.

[18] Z. Chen, X. Chen, Y. Ma, S. Guo, Y. Qin, and M. Liao, "Human posture tracking with flexible sensors for motion recognition," Comput. Animat. Virtual Worlds, vol. 32, no. 5, pp. 1993.56-1993.63, Apr. 2021.

[19] Z. Fang, X. Zhang, T. Cao, Y. Zheng, and M. Sun, "Spatial-temporal slowfast graph convolutional network for skeleton-based action recognition," IET Comput. Vis., vol. 16, no. 3, pp. 205-217, Nov. 2021.

[20] Y. Liu, H. Zhang, D. Xu, and K. He, "Graph transformer network with temporal kernel attention for skeleton-based action recognition," Knowl.-B. Syst., vol. 340, no. 15, pp. 1-16, Mar. 2022.