# Action Recognition Method of Basketball Training Based on Big Data Technology

Dongsheng CHEN[1], Zhen Ni[2]*
College of Sports Science, Guangxi College for Preschool Education, Guangxi 530022, China[1]
School of Physical Education & Health, Nanning Normal University, Nanning, Guangxi 530001, China[2]

***Abstract*—Aiming at the problem that improper posture of basketball players leads to not obvious sports effects, the present paper proposes an action recognition method combining computer vision and big data technology and applies it to athletes' daily training and competition. Firstly, based on the current mainstream motion recognition models, 3D graph convolution are used to improve the original 3D convolution to promote the expression ability of spatial structure features and temporal features in skeleton sequences. Secondly, channel and spatial attention mechanisms are introduced to focus on the weight distribution of key points and strong features in different posture recognition processes. Finally, the proposed model is tested in real data, and the test results show that the model runs smoothly while maintaining high recognition performance. It can more effectively direct basketball players to implement comprehensive, systematic, and scientific teaching and training standards that directly support raising the game's general level of performance.**

***Keywords*—Action recognition; computer vision; big data technology; three-dimensional convolution; channel and spatial attention mechanisms***

## I. INTRODUCTION

Accurately identifying and evaluating athlete movements is crucial in basketball training. However, traditional action recognition methods often rely on manual observation or simple video analysis tools, which not only have low efficiency but also cannot guarantee accuracy. Basketball is a highly complex competitive sport that requires athletes to possess superb skills and tactical understanding. In order to achieve the best training results, coaches and athletes need a method that can accurately and efficiently identify and evaluate athlete movements. Traditional action recognition methods mainly rely on manual observation, which is not only time-consuming but also susceptible to subjective factors. With the development of big data technology, the accuracy and efficiency of machine learning and computer vision algorithms have been significantly improved, providing new possibilities for solving this problem. Due to the continuous development of the social economy and science and technology, intelligence has been more and more widely concerned about and studied and is gradually becoming a global trend. With the arrival of the information age, many intelligent devices have become accessible to people, which greatly facilitates people's lives and further promotes the development of relevant research fields. In recent years, research on artificial intelligence has made remarkable progress. Machine vision is the main branch of artificial intelligence, and human behavior recognition is one of the important research directions [1]. Visual human behavior recognition includes the following parts: Acquisition of human behavior video sequence, feature extraction of moving human body, learning and recognition of classifier, among which feature extraction and classification recognition is the main research content of human behavior recognition [2].

Visual human behavior recognition and analysis is a comprehensive research direction; from the theoretical level, human behavior recognition research involves a variety of theoretical disciplines, including pattern recognition, statistics, information processing, computer vision, and so on. The research significance of human motion behavior recognition is mainly reflected in its practical value. With the continuous progress of video acquisition sensors and information science and technology, the research of motion behavior recognition has gradually become a subject with wide application prospects in many fields. It has been successfully used in basketball training, football training, alpine skiing, running, and other sports movement recognition because of its portability, wireless, and easy-to-operate characteristics [3]. However, in competitive competition, the limitation of experimental conditions leads to great challenges in capturing the above movements. Because of the quick development of computer vision and big data technology, it is important to use the new technology to effectively capture the training movements of basketball and football players, help them improve their disharmonious movements quickly, and improve their comprehensive quality.

Recent advancements in computer vision-based motion capture technology have made it possible to recognize human activity in challenging environments. Unmarked motion capture was carried out by camera equipment to obtain kinematics information remotely in the competition. Based on computer vision machine learning algorithms, human actions are displayed as waveforms corresponding to specific actions that are downloaded to a computer terminal. Subsequently, the synchronous video analysis, information extraction, and quick feedback are finished. Motion analysis based on a computer vision image must first predict or estimate the target's position and direction within the image sequence. Real-time tracking and displacement parameter acquisition are then accomplished by locating the target in the continuous image that has the same or comparable features. The human body is typically represented in practical applications as a collection of rigid bodies joined by frictionless revolute joints because this makes machine recognition and tracking easier. However, because human movements involve soft tissues like tendons and

ligaments, they are too complex to be fully understood by a rigid, simple body model. Thus, one of the challenges facing specialists in computer vision, machine learning, and sports science is precise tracking and measurement of dynamic human posture.

A deeper neural network model structure characterizes deep learning. The majority of the algorithms use manually labeled image data to train the neural network and then input the image or video into the trained network to estimate and recognize human posture, joint center, and bone position. In particular, skeleton behavior recognition technology is a method to understand and describe human behavior by extracting the action features in a skeleton sequence. Skeleton behavior recognition is one of the hot research directions in the field of machine vision. It can realize the accurate recognition of the motion of the target object by the computer, then analyze the motion of a human body in the video, and improve the dynamic perception ability of the computer. To evaluate the advantages of our method over existing methods, we conducted experimental comparisons on publicly available datasets. The results show that basketball training action recognition methods based on big data channels and spatial attention mechanism technology have significant advantages in accuracy, efficiency, and applicability to real-world scenarios. Compared with traditional methods based on manual observation or simple video analysis, our method greatly improves the accuracy and efficiency of recognition, reducing the need for manual intervention. Meanwhile, when compared with other deep learning models, our method better processes spatial information in videos by introducing spatial attention mechanisms, further improving the performance of action recognition. The basketball training action recognition method based on big data channels and spatial attention mechanism technology provides an efficient and accurate means of action evaluation for basketball training. This method can not only be applied in the field of basketball training, but can also be extended to other sports or general video action recognition tasks. Future work will further optimize the model structure, explore more effective attention mechanism methods, and apply this technology to practical training scenarios to verify its performance and effectiveness in the real world.

Section I of this article first analyzes the application background of machine learning and computer vision algorithms, which are of great significance for capturing the training movements of basketball players. Section II analyzed the comprehensive temporal dynamic information of bone sequences under different attention mechanisms of network applications. Section III proposes a method for action recognition using computer vision technology and big data technology, and applies it to action recognition in basketball player training. Section IV uses a 3D graph convolution module to extract spatiotemporal information from the skeleton sequence. We have established an attention enhancement structure to help nodes focus on key action information and pay more attention to certain areas. Finally, a behavior recognition model was constructed by combining 3D convolution with attention enhancement structures. Section V summarizes the entire text. The P-R curves of the model in this article can all surround the P-R curves of the comparison model, indicating that the overall action recognition performance of the model in the current study has been improved to varying degrees after using 3D graph convolution and attention mechanism to improve the existing model.

## II. Related Works

Human motion recognition research is getting more and more advanced as a result of the steady advancement of deep learning, machine learning, and other related technologies. The research and application of human motion recognition based on attitude sequence are different from tasks such as image recognition and target detection. The research on human motion recognition is related to time sequence. Input data includes spatial dimension and time dimension, so compared with other fields of computer vision, the difficulty and challenge of action recognition are greater.

### *A. Motion Recognition Based on Human Bone Sequence*

Dynamic human skeletons often contain a wealth of information and best represent human movement and behavior. The motion recognition algorithms based on the human bone sequence are usually divided into four categories: manual feature-based method, RNN/LSTM-based method, CNN-based method, and graph convolutional Neural network (GCN) based method.

Using the geometric relationships found in the space structure of the human skeleton for motion recognition is the aim of the traditional manual feature design method. Literature [4] listed nine geometric features, including eight static features and one time feature. The static feature encodes the form of motion and posture and uses the time feature to represent the change in time. The study in [5] proposed the use of the rotation and displacement of human bones to represent the three-dimensional transformation relationship between various body parts. The research in [6] proposed an integral invariant used to represent the motion trajectory of bone points and matched the motion trajectory. A collection of geometric characteristics, such as the separation between joints and the distance between joints to the plane formed by joints, were taken from study [7] and used to characterize posture and movement. While designing features, it is impossible to account for every factor so that most experimental results could be improved. Deep learning and other data-driven methods have gained popularity recently. Among them, the most popular models are CNN, GCN, and RNN/LSTM.

The main advantage of RNN/LSTM is that context dependencies can be modelled in the time domain. In addition, in the RNN/LSTM-based approach, the bone sequence is modeled as a coordinate vector of a series of joints, each coordinate representing a human joint. The study in [8] proposed the STA-LSTM network, which applies an attention mechanism to choose discriminating patio-temporal features, key joints, and keyframe information, respectively. The research in [9] proposed a VA-LSTM network. In VA-LSTM, two sub-networks were used to return parameters of rotation and translational matrix for rotating and translational bone coordinates to appropriate observation directions. Then, the new observed bone was input into the three-layer LSTM main network for motion recognition. The study in [10] proposed a

GCA-LSTM network and introduced global context memory to generate attentional representations for optimizing global context information. The SR-TSL approach was first presented in study [11] in 2018. It uses a time stack learning network (TSLN) to gather comprehensive temporal dynamic information on the bone sequence and a spatial inference network to gather high-level spatial structure information. The study [12] converted the input skeleton into several possible visual observation values, which were respectively processed by the attention LSTM network and finally fused with the output to generate recognition results.

In contrast to RNN/LSTM, the CNN-based method can learn spatial and temporal features simultaneously. The method based on CNN is used to encode bone sequences as pseudo-images (RGB or grayscale images) and time series as rows of bone joints in the image as columns. The study in [13] proposed to encode five spatial skeletal features as pseudo images and further explore space-time information by using CNN. The study in [14] proposed a new bone sequence representation method, which transformed a bone sequence into three fragments corresponding to the joint coordinate channel, and each fragment was composed of several grayscale images. The generated fragments are then fed into a deep CNN model for motion recognition. The research in [15] converted the transformed bone sequence into an RGB image, regarding the coordinates of the bone sequence (X, Y, Z) as the coordinates of the color image (R, G, B), and designed an arrangement network for data rearrangement. The study in [16] proposed an HCN model that can learn global co-occurrence features from skeletal sequences. This network combines graph convolution with LSTM, replaces the internal LSTM operations with graph convolution operations, and uses an attention mechanism to strengthen key node information while weakening non-key node information, highlighting more discriminative spatial features.

Recently, the GCN-based graph convolution network has attracted extensive attention because of its more natural representation of bone structure than based on RNN and CNN. In 2018, the research in [17] first developed a new deep learning model, namely space-time graph convolutional network (ST-GCN), which directly modeled bone data as graph structure, in which natural connections of human bones constitute spatial edges and corresponding joints in adjacent frames constitute temporal edges. Based on ST-GCN, the study in [18] used a frame distillation network to select keyframes and then sent the selected keyframes into a graph convolution network for action recognition. However, the spatial diagram in ST-GCN is fixed, and only human joints with natural connections are considered at the spatial edges. Moreover, it also proposed a multi-flow adaptive graph convolutional network (MS-AAGCN), which introduced an attention module and multi-flow network into 2S-AGCN. The study in [19] used the residual attention module to identify key joints. In contrast, the attention module used the original RGB image as input to generate attention masks to emphasize the areas that are important for emotion recognition in a frame. The research in [20] proposed that context information in original RGB videos should be used to extract joints with not only richer information but also highly relevant context information. At the same time, it used a neural structure search (NAS) algorithm to construct a graph convolutional network based on bone action recognition.

### B. Behavior Recognition in Sports

The integration of human behavior recognition and attitude estimation technologies into display application scenes led to the gradual introduction of these technologies into intelligent sports analysis systems. Hoop Tracker is a smart basketball analytics system that works with smart wear. The hoop tracker has a speed sensor that detects every shot. The shot detector is made into a patch, fixed inside the basket, through which the ball passes after each goal. Each time the shooter takes a shot, the watch and the shot detector communicate in real-time to see how far the shooter is from the basket and whether the ball hits. The system displays data on shots, three-pointers, and free throws, as well as field goal percentage and points. Shot Tracker's smart system, called Shot Tracker Team, has smart sensors attached to basketballs and players' shoes. In addition, the basketball court and the top of the court are surrounded by sensors, so the players are in a space without dead space [21]. These sensors give real-time feedback on the position of the player and the ball. Through the intelligent equipment, real-time analysis is made according to the data of players and ball movement on the court. The comparative information on the advantages and disadvantages of each player in the game is displayed in the form of data, including the analysis of players' shooting times, mistakes, assists, steals, dunks, and other actions. This data not only shows the players on the court to the audience but also provides the on-court coach with a data-backed tactical plan. In addition, in daily training, players can also use this system to let them understand their strengths and weaknesses to help adjust their targeted training. Most of the above products rely on smart wearable devices to capture and analyze athletes' movements. However, in general competitions, smart wearable devices will have a certain influence on athletes' competitions, and the sensors of smart wearers will have various uncertainties of system accuracy. NBA, as the highest professional basketball game in the world, mainly uses computer vision as a single information capture method of the Sport-VU system. The system has six 3D high-definition cameras fixed to the ceiling of each arena. Each camera takes pictures at high speed and sends them to a computer for data analysis. These cameras work at 25FPS. The software records and analyzes the player's movement trajectory and other information to obtain all kinds of technical and tactical data such as scoring, steals, rebounds, assists, instant speed, and so on.

## III. Behavior Recognition Model based on 3D Graph Convolution and Attentional Enhancement

### A. Three-Dimensional Convolution with Graph Convolution

*1) 3D-convolution:* The basketball training action recognition method based on big data technology mainly relies on data collection, feature extraction, and action classification. Firstly, this article collects a large amount of basketball training video data, including various movements of athletes. Then, computer vision algorithms are used to preprocess these videos and extract key motion features. Finally, machine

learning algorithms are used to classify and recognize these features.

Specifically, this method combines deep learning and computer vision techniques, utilizing convolutional neural networks (CNNs) to identify and extract key information from videos. In order to better understand and recognize various actions in basketball training, 3D Convolutional Neural Network (3D CNN) was adopted, which can better process spatial and temporal information in video sequences. In the action classification stage, attention mechanism is introduced. The attention mechanism allows the model to focus on key information regions when processing complex basketball training videos, thereby improving the accuracy of action recognition. By integrating 3D graph convolution and attention mechanism, the method proposed in this paper has higher efficiency and accuracy in handling basketball training action recognition tasks. Sample areas at the same location in several consecutive frames make up the 3D sampling space of 3D convolution [22], which contains two dimensions: time and space. Through the three-dimensional convolution kernel, stack and sum the data of the sampling area in multiple consecutive frames to generate multidimensional data, thus realizing the convolution operation of the three-dimensional sampling space, as shown in Fig. 1. Given that the convolution kernel size of the three-dimensional convolution kernel is $[P_i, Q_i, R_i]$; thus, the position response of the JTH feature graph in the layer I network can be expressed as Formula (1).

$$u_{ij}^{xyz} = \sigma\left(b_{ij} + \sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ij}^{pqr} u_{(i-1)j}^{(x+p)(y+q)(z+r)}\right) \quad (1)$$

where, $P_i$ and $Q_i$ are the two spatial dimensions of the three-dimensional convolution kernel, $R_i$ is the time dimension of the three-dimensional convolution kernel, $u_{ij}^{xyz}$ represents the sampling weight in the three-dimensional convolution kernel, and $b_{ij}$ represents the bias value; The A function contains operations such as batch standardization and activation functions.

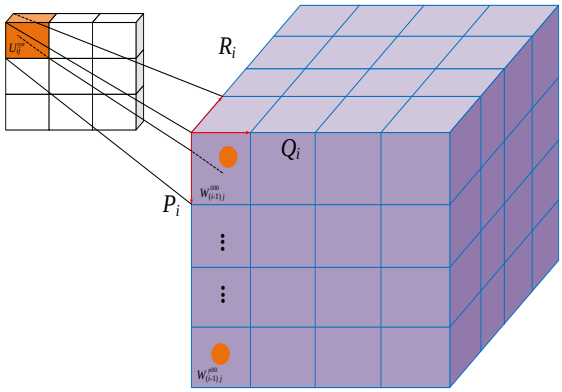



Fig. 1. Schematic diagram of the three-dimensional convolution operation.

3D sampling can not only collect spatial information but also build the connection between the current feature map and multiple consecutive frames in the output of the previous layer by weighted superposition of multiple consecutive frames in the output of the previous layer, realizing the capture of time information in the range of multiple frames. Therefore, 3D convolution can not only realize the collection of spatial and temporal information at the same time but also retain the correlation between the two. Therefore, the 3D convolution can be applied to the collection of spatial-temporal features of 3D sequential data in European space, such as continuous motion video frame sequence.

*2) Graph convolution:* Graph convolution is a general and effective way to learn graph structure data. Graph convolution aggregates information of neighbor nodes by weighted summation of hidden states of neighbor nodes through graph convolution kernel, which can process variable length neighbor nodes, realizes the convolution operation of graph structure data, and extracts information on a graph. Therefore, graph convolution can process graph structure data with generalized topological structure, so it is widely used in skeleton behavior recognition and attitude estimation.

Suppose there are m nodes in the output graph of layer *L* network, and the n-dimensional hidden state from the first node to the *m*-th node is represented $h_1^l, h_2^l, ..., h_m^l$, as shown in Fig. 2. The node states in the figure are denoted as $H^l[h_1^l, h_2^l, ..., h_m^l] \in R^{m\times n}$, and an adjacency matrix can represent the connection relationship $A \in R^{m\times n}$, so the first node in the output of the layer $l+1$ responds with 1, which is expressed as Formula (2).

$$h_1^{l+1} = \sigma(b + D^{-1/2} \otimes A \otimes D^{-1/2} \otimes H^l \otimes W) \quad (2)$$

where, D represents the degree matrix of A, a is the element of A to judge whether the node is A neighbor node with connection, W refers to the weight matrix of graph convolution, B is the bias value, $\sigma(\cdot)$ represents the activation function of nonlinear variation.

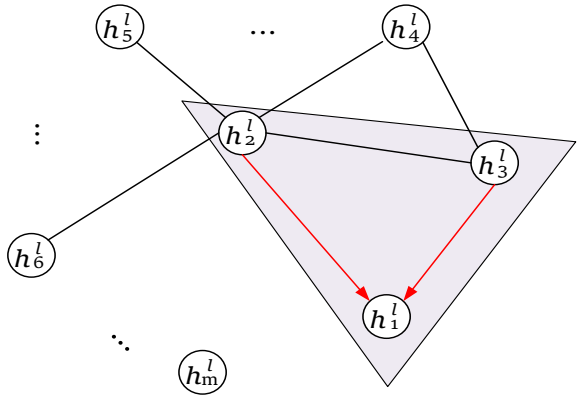



Fig. 2. Schematic diagram of the graph convolution operation.

*3) 3D graph convolution:* It is impossible to analyze the correlation between the spatial structure features and temporal features of skeleton sequences separately because they can describe all of the action information in the sequences together. Accordingly, the three-dimensional graph convolution method needs to be investigated to achieve the effective extraction of spatial-temporal information from the skeleton sequence.

In particular, the 3D sampling space in 3D convolution is rasterized sampling, which is only suitable for the feature collection of 3D sequential data in Euclidean space [23]. For 3D data in non-Euclidean space, the number of neighbor nodes in the sampling space is not fixed. Therefore, 3D convolution cannot extract spatial-temporal information from skeleton sequences with non-Euclidean three-dimensional data; Graph convolution can only extract spatial information on a graph through graph convolution kernel, which is capable of handling neighbor nodes with varying lengths. A three-dimensional graph convolution method is proposed in this paper to extract spatial and temporal information of three-dimensional skeleton sequences in non-Euclidean space. The technique is predicated on the graph convolution kernel, which is capable of managing neighbor nodes with varying lengths in graph convolution. Using the 3D sampling space in 3D convolution as an improvement idea, the 2D graph convolution kernel is improved to the graph convolution kernel with three-dimensional sampling space.

In the model process of the three-dimensional graph transformation based on the skeleton order, the adjacent nodes in the three-dimensional model space contain two adjacent nodes connected to the node in the skeleton stream and nodes close to the same point in several consecutive frames. Based on the three-dimensional graph convolution kernel, the three-dimensional graph convolution of the skeleton sequence is realized by weighted stack summation of neighbor node data in three-dimensional sampling space to generate multidimensional data, and the spatial-temporal information of the skeleton sequence is extracted effectively. As shown in Fig. 3, suppose L continuous skeleton frames in the three-dimensional sampling space. From frame 1 to frame L is denoted as $G^0, G^1, ..., G^{L-1}$ then the output result of three-dimensional graph convolution can be expressed as Formula (3).

$$x' = \sigma\left(b + \sum_{t=0}^{L-1}\sum_{c=0}^{C-1}\sum_{k=0}^{K-1} D^{-1/2} \otimes A \otimes D^{-1/2} \otimes G_{c,k}^{t} \otimes W_{c,k}^{t}\right) \tag{3}$$

where, A indicates the adjacency matrix of the connection relation, D denotes the degree matrix of A, $G_{c,k}^{t}$ is the characteristic value of channel C of the KTH neighbor node of frame T in the three-dimensional sampling space, $W_{c,k}^{t}$ refers to the weight matrix of three-dimensional graph convolution, b is the bias value; The $\sigma(\cdot)$ function contains operations such as batch standardization and activation functions.

## B. Attentional Mechanism

*1) Channel grouped attention:* Channel grouping attention groups channel features and highlight salient features in each group by using the similarity between both local and global features. Fig. 4(a) presents the network layer structure of channel-grouped attention. Features are divided into G groups according to channels, and the input features of each group are fused with the feature map G containing a semantic vector after global average pooling to form a new feature map. After normalization and sigmoid activation function operation, the input features of each group are integrated with the original feature by the site. According to the definition of the dot product, $g \cdot x_i$ can be written as $\| g \| \| x_i \| cos\theta_i$, where $\theta_i$ is the Angle between $g$ and $x_i$. Therefore, the feature of the modulus length and the feature that is close to the direction of the global feature vector will get a larger initial attention coefficient. At the same time, attention values of different samples vary greatly, so they need to be normalized to the same range to give accurate attention weight.

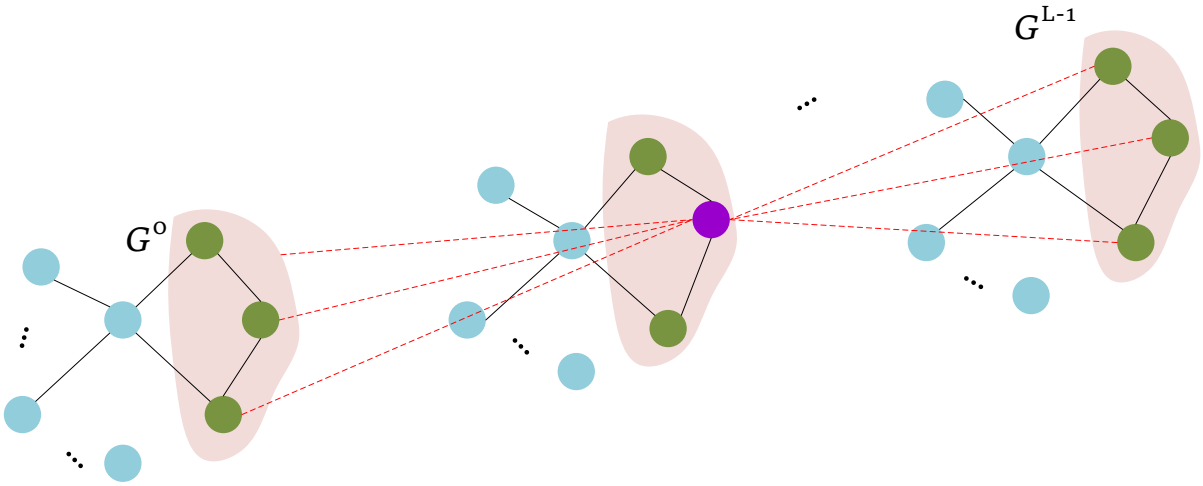



Fig. 3. Schematic diagram of 3D graph convolution operation in skeleton sequence.

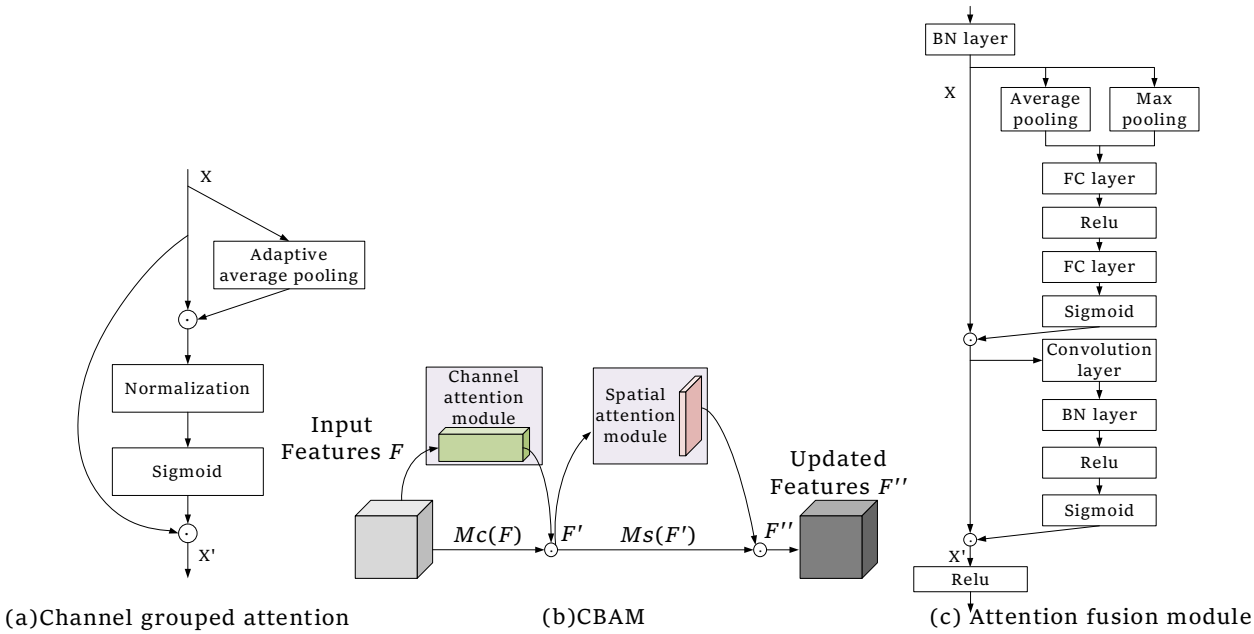



Fig. 4. Each attention mechanism module component.

*2) Channel-spatial attention:* CBAM attention network [24] is analogous to the human visual attention mechanism, which reconstructs the feature matrix through channels and Spaces. A self-learning method is adopted to recalibrate the weights of features. The overall network structure of CBAM is shown in Fig. 4(b). The feature F extracted from the input image was dotted with the feature $M_c(f)$ after the action of the channel attention module to obtain the feature $F'$. Similarly, the improved feature $F''$ was obtained through the action of the spatial attention module. The sequential connection of the channel attention module and spatial attention module is more effective than the parallel connection.

Unlike the channel attention module, which concentrates on the context in which the information is meaningful, the spatial attention module is location-specific. The channel attention module and the spatial attention module complement each other, focusing on location and content respectively and linking sequentially [25]. Similar to the channel attention module, for the H×W×C input feature $F'$, the maximum pooling and average pooling of channel dimensions are firstly carried out to obtain two features of 1×1×C dimensions and then splice them according to channels. Then, the feature is convolved with a 7×7 convolution, and a spatial matrix with the same dimension as the sigmoid activation function obtains the original feature. The new feature after scaling can be obtained by multiplying the spatial attention matrix with the original feature.

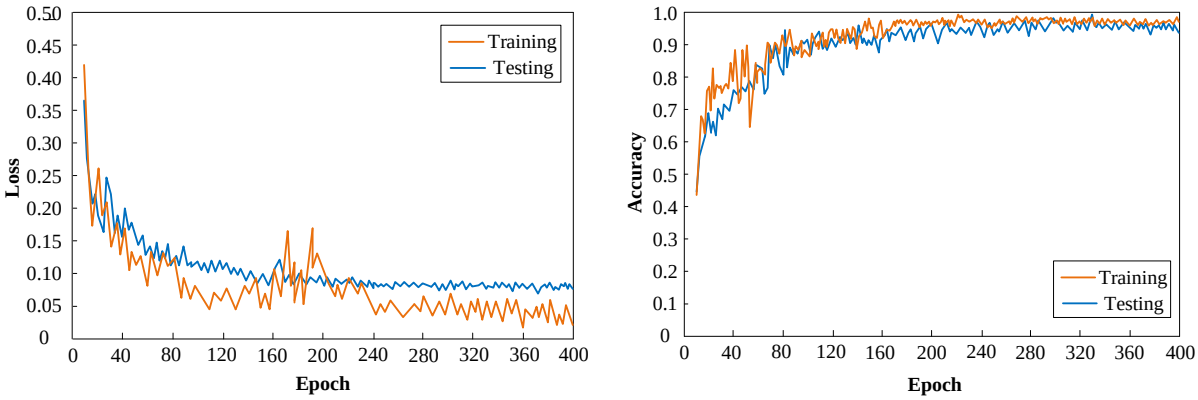
*3) Attention fusion:* Fig. 4(c) sows the attention fusion module, in which the CBAM module is added after the BN layer of the backbone network. In the training process, the fusion attention module divided the features into G groups according to the channels. After obtaining G from the global average pooling, the features of each group were integrated with the original group features by the site. After the normalization and activation function operation, the features were dotted with the original group features to obtain the activation of significant semantic regions based on the above operations [26]. The upper and lower parts in the figure respectively represent channel attention and spatial attention to features, and the feature $W'$ of a channel or space with different weight distribution can be obtained through the action of the attention module.

## IV. Experimental Design and Result Analysis

### A. Experimental Environment and Evaluation Indicators

The experimental environment of this work is the Ubuntu 16.04 operating system. The experimental platform is Intel(R)Core I7-7800X processor, six-core 3.5ghz, and NVIDIA GTX1080Ti graphics card. The development language is Python 3.7. The testing platform is Pycharm, and the PyTorch deep learning framework is adopted. Set the initial learning rate as 0.0001, the weight attenuation term as 0.0005, and the random discard rate as 0.5. The batch size is set to 16.

The curves of training and testing performance and iteration times of the model in this paper are represented in Fig. 5. It is evident that as the number of iterations reaches 240, the training and test accuracy rate and Loss curve region are stable, and the model achieves stable convergence.




(a) Loss curve (b) Accuracy curve

Fig. 5. Curves of the training and testing phases.

Fig. 5 shows the curves during the training and testing stages. (a) Indicates the loss curve. (b) Represents an accuracy curve. Time Overhead (TO) of single image motion recognition, Accuracy, Precision, Recall, F1-score, and other mainstream evaluation indices were employed to assess the model performance in order to confirm the efficacy of the suggested algorithm. Formula (4) to Formula (7) displays the computation expressions. In Table I, the confusion matrix is displayed. In particular, the calculation of Precision and Recall is contradictory, so the precision-recall curve is used for comparison in the current paper. The model's classification performance improves with increasing area under the curve.

$$Accuracy = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \tag{4}$$

$$Precision = \frac{Tp}{Tp + Fp} \tag{5}$$

$$Recall = \frac{Tp}{Tp + Fn} \tag{6}$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{7}$$

TABLE I. CONFUSION MATRIX CALCULATION

| Actual | Predicted | |
|---|---|---|
| | ***Positive*** | ***Negative*** |
| Positive | TP | FP |
| Negative | FN | TN |

## B. Result Analysis

Fig. 6 shows the confusion matrix generated by this method in three groups of experiments, where the actual action sequence is represented by the rows of the matrix. In contrast, the columns show the action sequence recognized by the algorithm. The confusion matrix states that 307, 311, 301, 318, and 307 times of the six movements were successfully recognized in the three groups of experiments, and the recognition accuracy was 97.07%, 96.14%, 98.34%, 94.03%, and 97.39% respectively. In addition, Fig. 7 represents the average Accuracy, Precision, Recall, and F1 curves of the proposed model in multiple experiments. Also, the performance of the presented model tends to be stable on several experimental results, indicating the robustness of the model presented in the present study.

## C. Ablation Experiment

To assess the effect of different components in the model on the overall recognition performance, three ablation experiments were designed, respectively. 1) The original recognition model using skeleton analysis only; 2) Replace original graph convolution with 3D graph convolution; 3) Introduce channel-spatial attention mechanism. Fig. 8 displays the experimental results, where Original represents the first group of experiments; 3D-GC represents that only a three-dimensional graph convolution network is used; ATT means only channel-spatial attention mechanism is used; 3DGC-ATT represents the final model of this paper. It can be seen that compared with experiment 1) of the original control group, the overall performance of the model is improved by 3.91% in recognition accuracy by using 3D graph convolution instead of original graph convolution. The main reason is that the spatial structure features and time features of the skeleton sequence are introduced to improve the expression ability of the action sequence in the skeleton sequence further. In addition, after the introduction of the channel-spatial attention mechanism, the recognition accuracy of the model is improved by 5.18%, which is 1.31% higher than that of the recognition model after the introduction of three-dimensional graph convolution. The reason is that the attention mechanism can focus on the weight distribution of strong features, further, increase the contribution of the largest feature to the overall recognition performance, and suppress the weight of edge features.

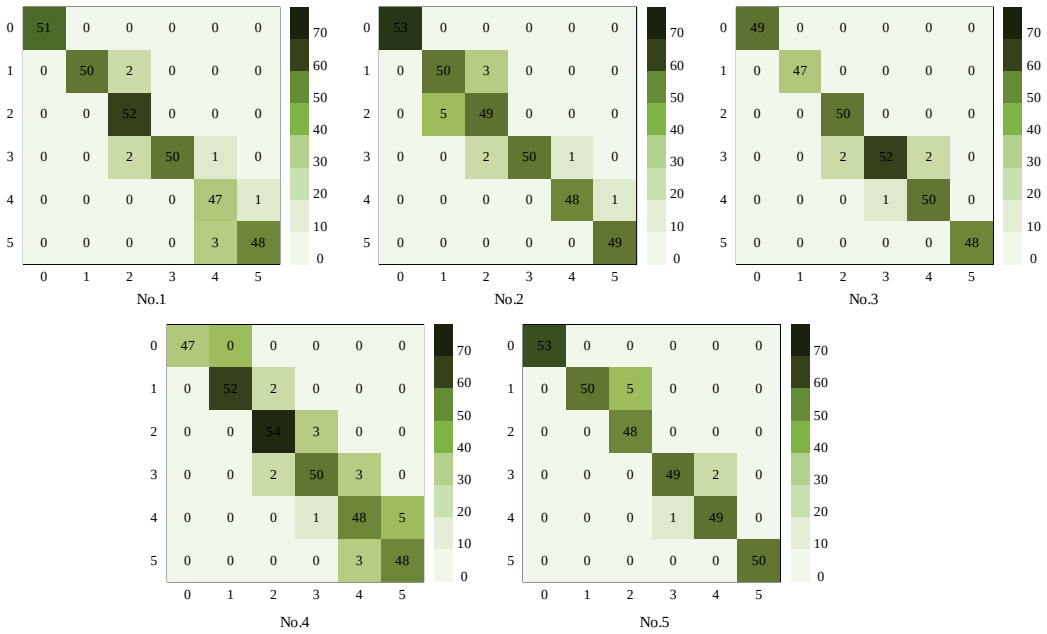



Fig. 6. Confusion matrix.

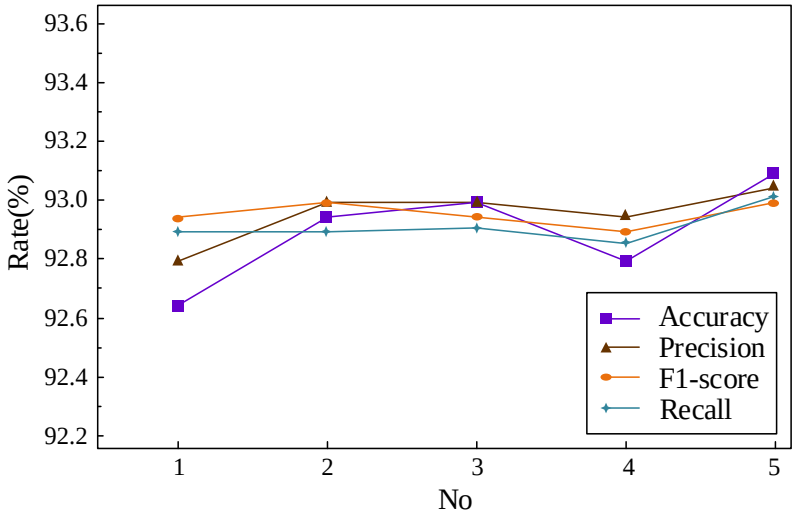



Fig. 7. Curves under different evaluation indexes.

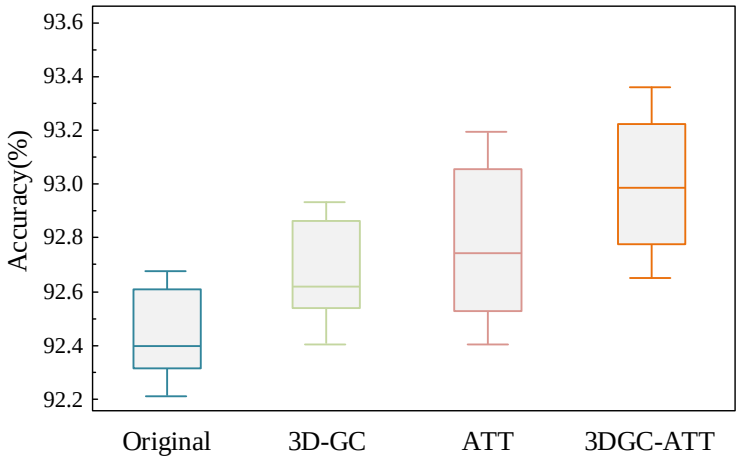



Fig. 8. Ablation experiment.

### D. Comparison of Similar Related Works

To verify the effectiveness of the proposed model, the same data set, environment, and evaluation indicators were compared with the current mainstream model. Fig. 9 shows the curves of various models under Accuracy, Precision, Recall, and F1. Fig. 10 shows the comparison results of different models' running times. It can be seen that in Accuracy, Precision, Recall, F1, and other evaluation indicators, the model presented in the present work has obvious competitive advantages compared with mainstream models, and it is also competitive in identifying time costs. Although the time cost is improved compared with the model in literature [27], the comprehensive performance of this model is: The model in this paper performs well. In addition, to integrate the calculation contradiction between Precision and Recall, the precision-recall curve is adopted here for comparison. Fig. 11 shows the comparison results of P-R curves of different models. The model's classification performance improves with the increasing area under the curve.

The P-R curve of the model in this paper can all surround the P-R curve of the comparison model in the basketball action recognition results, which indicates that the performance of the overall action recognition of the model in the current study has been improved to varying degrees after the existing model is improved by using three-dimensional graph convolution and attention mechanism. The above data further verify the robustness of the proposed model.

The basketball training action recognition method based on big data 3D convolution technology is a very promising research direction. The accuracy and efficiency of this method have been validated in many cases, but there are still some areas that need improvement. Especially when dealing with large-scale data, how to improve computational efficiency and reduce the consumption of computing resources is an urgent problem that needs to be solved. In addition, how to improve the generalization ability of the model is also an important research direction. In practical applications, different basketball training scenarios and individual differences among athletes may lead to a decrease in model performance. Therefore, studying how to better adapt the model to these differences is a challenging task. I believe that with the continuous progress of technology and in-depth research, the basketball training action recognition method based on big data 3D convolution technology will be further optimized and improved, and will play a greater role in practical applications.

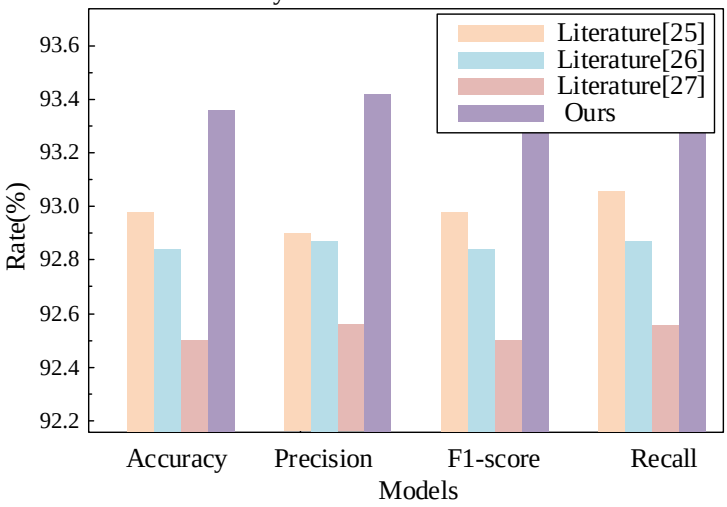



Fig. 9. Performance comparison of different models.

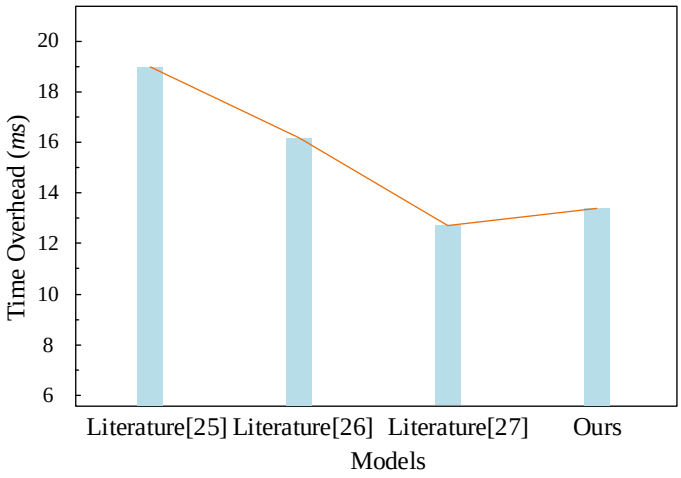



Fig. 10. Comparison of recognition time costs of different models.

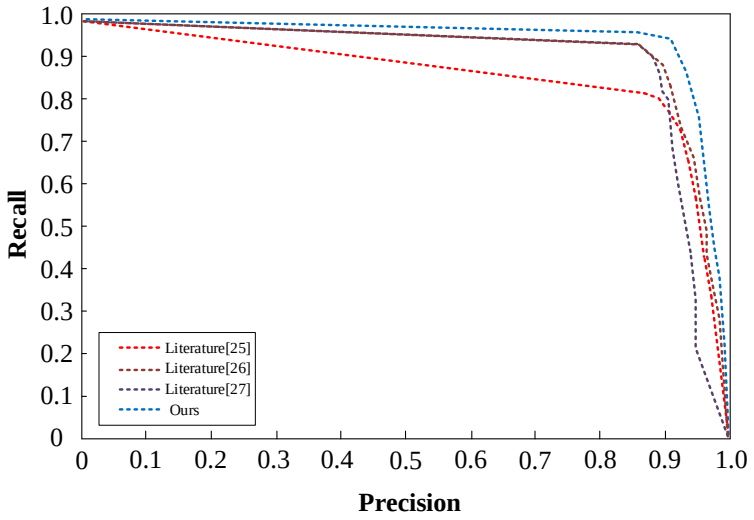



Fig. 11. Comparison of P-R curves of different models.

## V. Conclusion

This paper explores the motion capture technology of computer vision and large data sets in advanced training technology application status in the field of motion gesture recognition. Using three-dimensional figure convolution and attention mechanism to improve the existing model, through the test in the practical data, this model is verified in recognition accuracy and time cost is increased. Thus, this model can be applied to the daily training of basketball players and provide a reference for the overall evaluation and decision-making of athletes and coaches.

However, the research has certain limitations. The method based on 3D CNN requires a large amount of computing resources, such as GPU memory and computing power, when processing large-scale basketball training video data. This may result in very time-consuming training and inference processes, which cannot meet the needs of real-time processing. Therefore, how to improve the efficiency of algorithms and the utilization of computing resources is also another challenge faced by current methods.

To address the issues of data scale and quality, future research can explore data augmentation techniques, such as using Generative Adversarial Networks (GANs) to generate high-quality simulated data or using transfer learning methods to acquire knowledge from other relevant datasets. In addition, adaptive learning algorithms can adaptively adjust model parameters based on individual characteristics of different athletes, improving adaptability to individual differences.

## Competing of Interests

The authors declare no competing of interests.

## Authorship Contribution Statement

Zhen Ni: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Dongsheng Chen: Methodology, Software, Validation.

## References

[1] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," Artif Intell Rev, vol. 54, pp. 2259–2322, 2021.

[2] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," Pattern Recognit Lett, vol. 118, pp. 14–22, 2019.

[3] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," Int J Comput Vis, vol. 130, no. 5, pp. 1366–1401, 2022.

[4] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention models in graphs: A survey," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 13, no. 6, pp. 1–25, 2019.

[5] X. Luo, H. Li, X. Yang, Y. Yu, and D. Cao, "Capturing and understanding workers' activities in far - field surveillance videos with deep action recognition and Bayesian nonparametric learning," Computer - Aided Civil and Infrastructure Engineering, vol. 34, no. 4, pp. 333–351, 2019.

[6] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 4, pp. 447–453, 2020.

[7] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," IEEE Transactions on Image Processing, vol. 29, pp. 9532–9545, 2020.

[8] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," IEEE Trans Pattern Anal Mach Intell, 2022.

[9] L. Zhu, H. Fan, Y. Luo, M. Xu, and Y. Yang, "Temporal cross-layer correlation mining for action recognition," IEEE Trans Multimedia, vol. 24, pp. 668–676, 2021.

[10] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, and K.-K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," IEEE Internet Things J, vol. 6, no. 6, pp. 9280–9293, 2019.

[11] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," Int J Multimed Inf Retr, vol. 7, pp. 87–93, 2018.

[12] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," Artif Intell Rev, vol. 54, pp. 137–178, 2021.

[13] J. Kim et al., "Rotational Variance - Based Data Augmentation in 3D Graph Convolutional Network," Chemistry‐An Asian Journal, vol. 16, no. 18, pp. 2610‐2613, 2021.

[14] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," Neurocomputing, vol. 337, pp. 325–338, 2019.

[15] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439–2450, 2018.

[16] S.-Y. Shih, F.-K. Sun, and H. Lee, "Temporal pattern attention for multivariate time series forecasting," Mach Learn, vol. 108, pp. 1421–1441, 2019.

[17] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," IEEE Trans Neural Netw Learn Syst, vol. 32, no. 10, pp. 4291–4308, 2020.

[18] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, "GAFNet: Group attention fusion network for PAN and MS image high-resolution classification," IEEE Trans Cybern, vol. 52, no. 10, pp. 10556–10569, 2021.

[19] K. Sangeetha and D. Prabha, "Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM," J Ambient Intell Humaniz Comput, vol. 12, pp. 4117–4126, 2021.

[20] Q. Lyu, M. Guo, and M. Ma, "Boosting attention fusion generative adversarial network for image denoising," Neural Comput Appl, vol. 33, pp. 4833–4847, 2021.

[21] G. Huo, Y. Zhang, J. Gao, B. Wang, Y. Hu, and B. Yin, "CaEGCN: Cross-attention fusion based enhanced graph convolutional network for clustering," IEEE Trans Knowl Data Eng, 2021.

[22] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3d point clouds," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 10, pp. 3664–3680, 2020.

[23] W.-Z. Nie, M.-J. Ren, A.-A. Liu, Z. Mao, and J. Nie, "M-GCN: Multi-branch graph convolution network for 2D image-based on 3D model retrieval," IEEE Trans Multimedia, vol. 23, pp. 1962–1976, 2020.

[24] Y. Chen, X. Zhang, W. Chen, Y. Li, and J. Wang, "Research on recognition of fly species based on improved RetinaNet and CBAM," IEEE Access, vol. 8, pp. 102907–102919, 2020.

[25] M. Xia, T. Wang, Y. Zhang, J. Liu, and Y. Xu, "Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery," Int J Remote Sens, vol. 42, no. 6, pp. 2022–2045, 2021.

[26] X. Wang et al., "Self-paced feature attention fusion network for concealed object detection in millimeter-wave image," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 224–239, 2021.

[27] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," IEEE Trans Instrum Meas, vol. 69, no. 12, pp. 9645–9656, 2020.