Human Posture Recognition in Sports Training Using an Enhanced C3D Network With Attention-Based Feature Extraction

Liang Yan
Department of Sports, Guangdong University of Finance and Economics, Guangzhou 510320, Guangdong, China
E-mail: 15920538612@163.com

Keywords: attention mechanism, target recognition, action recognition, C3D network, residual network

Received: April 24, 2025

Traditional sports scene human posture recognition technology faces problems such as low accuracy and complex structure. To solve the above problems, an efficient sports training human posture recognition technology is proposed. Among them, the research is based on target recognition algorithms for modeling, introducing attention mechanisms to enhance feature attention, and improving the training loss function. In addition, a posture action recognition model is constructed based on a three-dimensional convolutional network, in which a residual network is used to improve the model with asymmetric convolutional layers, and an attention mechanism combining convolutional blocks is introduced to extract target input features. In the experimental testing of action target recognition, the UCF101 dataset and a self-made dataset are used to compare the accuracy recall performance of different techniques. The precision value of the research method on the UCF101 dataset was 0.988, indicating the best performance. In the training loss test of the self-made dataset, the research model showed the best performance with a loss of 0.061. In the frame rate detection test, a self-made dataset was used for testing, and the average frame rate of the research model was 28.5FPS, which performed the best. In the posture and action recognition test, the research model achieved the highest recognition accuracy in scenes such as swimming, high jump, and basketball. In table tennis recognition, the action recognition accuracy was 92.54%, which was the best overall performance. In addition, the research technology focused on floating-point operations, and the research model had floating-point values of 38.456MB and 37.458MB on the Kinetic400 dataset and UCF101 dataset, respectively, which were superior to other techniques. The research technology has good application effects in sports training scenes, providing technical support for sports standardization training and action recognition technology.

Povzetek: Tehnologija združuje izboljšano mrežo C3D z večnivojskim pozornostnim mehanizmom za kvalitetno prepoznavo telesne drže in gibanja v športnem treningu z nizko porabo virov.

1 Introduction

Posture and action recognition is crucial in computer vision, which can recognize human behavior from video sequences [1]. Driven by deep learning related technologies, action recognition technology has made significant achievements in fields such as security monitoring, healthcare, and human-computer interaction. Especially Convolutional Neural Network (CNN) and recurrent neural networks have demonstrated excellent capabilities in processing video data and time series information, resulting in meaningful improvements in the accuracy and robustness of action recognition [2]. Related scholars have conducted extensive research on action recognition. Khan M A et al. built a fully automatic human action recognition scheme that integrated neural networks to address the human behavior recognition in fields such as video surveillance. This scheme took pre-trained models to obtain features and selected the optimal feature subset through relative entropy and mutual information. Finally, the training was conducted on public data. The research method performed excellently [3]. Zhou L et al. conducted research on the insufficient semantic action recognition. A method based on joint learning of multiple body parts was proposed. This technology established a probability mapping from hidden visual postures to observed video frames by learning multiple visual posture models and posture dictionary models, analyzed posture actions. In the result analysis, the research technology had high recognition accuracy, but still faced parameter complexity [4]. Bai Y et al. proposed a confident spatiotemporal attention transfer action recognition technique for action recognition in extremely low resolution videos. This technology reduced attention differences through transfer learning strategies, extracted information from high-resolution data, and considered the confidence level of monitoring signals to achieve more reliable transmission. This technology significantly improved the accuracy of action recognition in low image quality, but its adaptability to different scenes was limited and further optimization was necessary in the later stage [5]. Kumar R et al. conducted a comprehensive research on human behavior recognition in artificial intelligence to optimize the posture recognition. An improved Convolutional 3D (C3D) network was designed to construct a recognition model. The training results showed that this technology had good

application effects. However, the technology still faced excessive expenses that need to be solved [6].

At present, posture and action recognition technology has broad applications in the field of sports. By extracting and recognizing human limb feature information, the effectiveness of sports can be improved. Yan G et al. optimized the accuracy of keyframe extraction in sports videos and proposed an accurate extraction algorithm. This algorithm was based on the color components and adopted non-uniform quantization color tone and saturation space methods to convert physical education teaching videos into one-dimensional histograms, achieving video shot segmentation and training. The method could effectively extract keyframes of aerobics video actions [7]. Kong L et al. conducted research on the insufficient analysis of posture and action in volleyball scenes, and proposed a novel spatiotemporal relationship modeling method. This method captured the interactive motion relationships between players through adaptive graph CNN and attention time convolutional networks. Finally, on the volleyball dataset, and the technology demonstrated good performance in sports action recognition. However, this technology faced excessive expenses and required further optimization [8]. To improve the insufficient recognition of sports activities, Wu F et al. introduced various sports projects and compared existing sports analysis frameworks. They discussed the challenges and unresolved problems and developed a toolbox that supports multiple action recognition to promote sports analysis. Finally, experiments were conducted in multiple sports projects. The results showed that the sports analysis framework had good stability and recognition ability [9]. Sun X et al. conducted research on human tennis motion recognition. A 3D convolutional action recognition technique was proposed to improve its recognition accuracy. The research introduced an expected behavior prediction module. This module predicted subsequent motion based on observed motion, mitigated motion blur and generated predictions through adversarial time series models. In addition, the research also utilized TensorFlow to enhance insights into player performance and action prediction. The experimental results indicated that the improved technology had good application effects [10]. The relevant technical summary table is shown in Table 1.

Table 1: Summary of related technical research

Reference number	Research contents	Research results and shortcomings
Reference [3]	Khan M A et al. proposed a human action recognition technique for video detection	This technology can effectively detect human movements with excellent accuracy, but requires high computational power
Reference [4]	Zhou L et al. proposed a human motion recognition technique for multi-body part recognition	This technology has high recognition accuracy, but faces the problem of parameter complexity
Reference [5]	Bai Y et al. proposed a human motion recognition technique for solving low resolution problems	The technology has good application effects, but it still faces excessive expenses
Reference [6]	Kumar R et al. proposed an action detection technique for high-precision human posture recognition	The technology has good action recognition performance, but the cost is too high and needs to be solved
Reference [7]	Yan G et al. proposed an action recognition technique for keyframe extraction, which is applied in sports training scenes	Technology can effectively extract keyframes, but limited performance in complex scenes
Reference [8]	Kong L et al. proposed a spatiotemporal action recognition technique to address the insufficient posture detection in volleyball	This technology has good performance, but the cost is relatively high
Reference [9]	Wu F et al. proposed an action analysis framework based on deep learning to enhance the performance of sports recognition action detection	The framework has good stability and recognition ability, but its adaptability to different scenes needs improvement
Reference [10]	Sun X et al. proposed a tennis action recognition technique.	Technology has excellent motion detection capabilities, but there are shortcomings in extracting complex scene features

From the above research, posture and action recognition is crucial in fields such as security, education, and transportation. Especially with the continuous innovation of artificial intelligence technology, its application scenes have also expanded to sports training scenes and achieved good results. However, at present, existing action recognition technologies still face many effective solutions, such as poor feature extraction in complex scene targets and inaccurate positioning of extraction boxes in current human motion object detection technology. In action recognition, there are problems such as high computational overhead, slow system operation, and effective recognition scenes. These issues not only limit the development of action recognition technology, but also affect its application effectiveness in various fields. Therefore, to solve the insufficient accuracy, poor feature extraction, and excessive recognition cost in human posture recognition in sports training scenes, an efficient posture recognition technology is proposed. The novelty of this technology lies in two aspects. The first takes advanced object recognition algorithms (You Only Look Once version 5s, YOLOv5s) to construct a human posture object recognition model, which adopts attention mechanism and loss function optimization to improve the accuracy of action object detection. The second takes C3D to construct a human posture and action recognition model. The introduced residual networks, asymmetric convolutional layers, and attention mechanism layers are used to enhance feature extraction, thereby further strengthening the feature extraction. This study introduces attention mechanism to enhance feature extraction,

improve recognition accuracy, optimize model performance, and provide technical references for sports training guidance and action recognition technology improvement.

2 Methods and materials

2.1 Modeling of human posture target recognition based on improved YOLOV5s and attention mechanism

With the continuous improvement of deep learning technology, it has exerted a key role in human posture and action recognition in sports scenes [11]. Therefore, to effectively recognize human posture in sports scene training, a human posture target recognition technology based on improved YOLOV5s is proposed. Meanwhile, the Convolutional Block Attention Module (CBAM) is adopted to improve YOLOv5s. Traditional YOLOV5s mainly suffer from insufficient feature extraction, including local and global feature fusion, as well as insufficient multi-scale feature recognition. In addition, traditional loss functions can lead to significant errors when calculating the position of bounding boxes, thereby affecting action target recognition. Therefore, the study adopts an improved YOLOV5s for human posture target recognition. In human posture target recognition, the improved YOLOV5s algorithm consists of four parts, as shown in Figure 1.

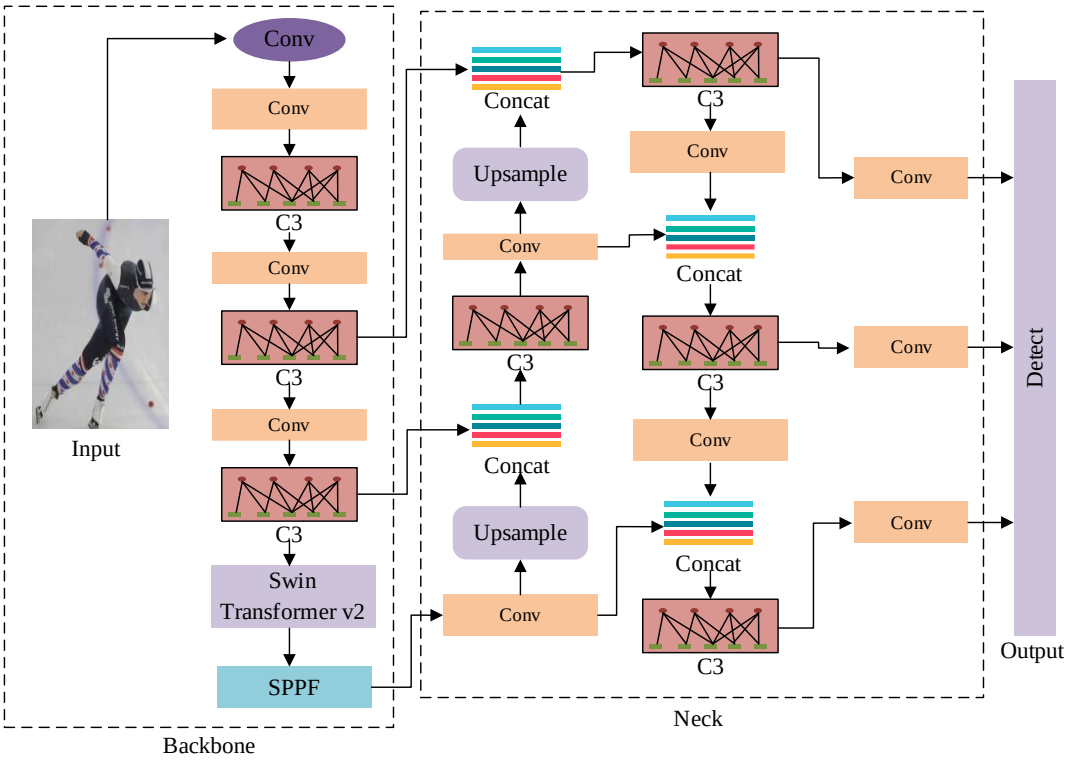


Figure 1: Structure of improved YOLOV5s

From Figure 1, the traditional YOLOV5s mainly consists of four parts: input, backbone, neck, and output. In the improved YOLOV5s, a simplified combination structure of cross stage partial connections (CSP Bottleneck 3, C3) and Focus is adopted in the backbone section. An improved Spatial Pyramid Pooling Fast (SPPF) layer is introduced to fuse local and global features in the model. SPPF is embedded between the backbone output and the detection head, enabling the model to extract richer feature information on feature maps of different scales, thereby improving the accuracy of object detection. In addition, in the backbone section, Swin Transformer v2 network is introduced to replace the final connection layer, which can partition long sequence features into blocks, reduce network parameters, and improve network stability [12]. Neck is mainly responsible for feature fusion. The research adopts Path Aggregation Network (PANet) to extract features and enhance the multi-scale localization ability of targets. The output suppression adopts an improved Complete Intersection over Union Loss (CIOU) function to achieve the target output [13]. In addition, to improve the YOLOV5s in recognizing posture features of sports training process, a CBAM is introduced in the Neck tail research. The role of CBAM is to enhance the attention to important features at the target input end, thereby improving the model's ability to extract multi-scale features and enhance target recognition performance. Figure 2 displays the attention mechanism structure [14].

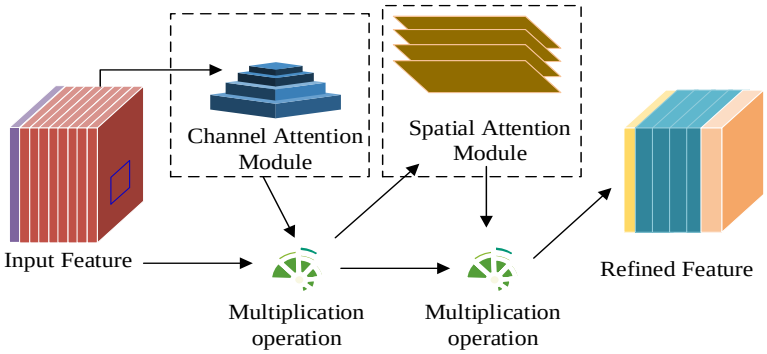


Figure 2: Attention mechanism structure

From Figure 2, it involves channel attention and spatial attention, where the former is shown in equation (1).

$$M_c(F)=\sigma\left(MLP_c\left(\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F_{ij}\right)\right) \quad (1)$$

In equation (1), H signifies the height of the feature map, and W signifies the width of the feature map. σ signifies the sigmoid activation function. F_{ij} signifies the input feature map. c signifies the index. MLP_c indicates a two-layer Fully Connected Network (FCN) in channel attention. The channel attention part is mainly responsible for capturing the channel relationships of input feature maps and adjusting weights using the sigmoid function [15]. The spatial attention part is shown in equation (2).

$$M_S(F)=\sigma(MLP_s\left(Avgpool\left(F_r\right)\right) \quad (2)$$

In equation (2), F_r represents the input feature map. Avgpool is the average pooling layer. MLP_s represents the two-layer FCN in spatial attention. Finally, in the output section, YOLOv5 generally uses three types of loss functions for object recognition determination, including position prediction, bounding box position, and confidence loss functions. Equation (3) displays the total loss [16].

$$L_{loss}=\lambda_1 L_{cls}+\lambda_2 L_{obj}+\lambda_3 L_{loc} \quad (3)$$

In equation (3), L_{cls}, L_{loc}, and L_{obj} represent three loss functions: prediction, bounding box position, and confidence. L_{loss} represents the weighted sum of three loss functions. The weighted parameters corresponding to the function are λ_3, λ_2, and λ_1. A binary cross entropy function is taken as the confidence loss, which mainly determines whether the recognition target exists, as shown in equation (4) [17].

$$L_{obj}=-y\log p-(1-y)\log(1-p)=\begin{cases}-\log p, y=1\\-\log(1-p), y=0\end{cases} \quad (4)$$

In equation (4), p represents the predicted positive class. y signifies the true label. To determine the boundary position loss, the CIOU function is used in this study. CIOU introduces three additional penalty terms on the basis of traditional IoU, including overlapping area, center point distance, and aspect ratio consistency. It can better distinguish between predicted and real boxes, which is suitable for detecting different sports targets [18]. The CIOU function is shown in equation (5).

$$L_{loc}=1-IoU+\frac{{p_p}^2\left(b,b^{gt}\right)}{c^2}+\alpha v \quad (5)$$

In equation (5), c signifies the diagonal distance of the error. IoU represents the Intersection over Union (IoU) ratio. b^{gt} signifies the center point of the real box. b is the center point of the prediction box. p_p signifies the Euclidean distance between these two points. α signifies the weight parameter. v represents the combination of width and height consistency parameters used to determine the target width and height parameters. α is displayed in equation (6).

$$\alpha=\frac{v}{(1-IoU)+v} \quad (6)$$

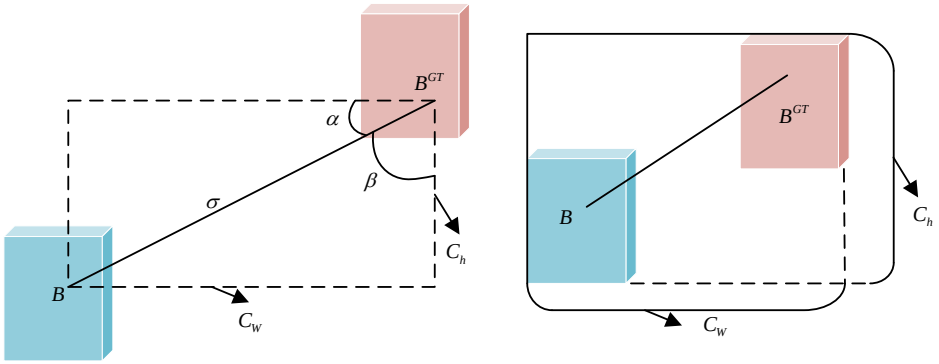
The parameter v is continued to be calculated, as displayed in equation (7).

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}\right)^2 \quad (7)$$

In equation (7), w and h signify the true width and height of the target box, and w^{gt} and h^{gt} signify the predicted width and height of the target box. The CIOU function mainly determines the position error of the bounding box from the aspects of border aspect ratio, overlapping area, and center distance point. The parameter v only distinguishes the aspect ratio of the border and cannot fully reflect the specific difference between the width and height of the border [19]. Therefore, the study improves the CIOU function by further considering the influence of border height and width. The improved CIOU is displayed in equation (8).

$$L_E = (1-IoU) + p_p{}^2\left(b, b^{gt}\right)/C^2 + p_p{}^2\left(w, w^{gt}\right)/C_w^2 + p_p{}^2\left(h, h^{gt}\right)/C_h^2 \quad (8)$$

In equation (8), C_w and C_b signify the width and height of the bounding rectangle. In the improved CIOU function, additional consideration is given to the width and height loss of the border to further optimize the convergence effect of the model. Finally, the study adopts a Smoothed Intersection over Union (SIoU) function as the prediction loss function. Based on CIOU, the stability and convergence of the model have been improved by introducing a smoothing function to suppress the gradient variation problem. The SIoU function fully considers shape, distance, angle, and IoU loss in object detection. The angle and distance loss are displayed in Figure 3 [20].



(a) SIoU angle loss (b) SIoU distance loss

Figure 3: Schematic diagram of angle loss and distance loss

In the angle loss, C_h is used to represent the height difference between the center points of these two boxes. σ_t represents its center point distance of these two boxes. In the distance error, B and B^{GT} represent the real box and the predicted box. The target distance is determined by the width C_w and height C_h of the bounding rectangle between the real box and the predicted box.

2.2 Modeling of human posture and action recognition in sports training based on improved C3D

Based on the improved YOLOV5s human posture target recognition technology, human posture in sports training scenes is recognized. Next, the study introduces a human posture action recognition technology based on an improved C3D network. The technical improvements include introducing asymmetric convolutional layers and residual networks to address network degradation problems, and adding an improved CBAM mechanism to enhance feature attention. The asymmetric convolutional layer in C3D networks reduces the computational complexity and resource consumption of convolution operations by asymmetrically splitting larger convolution kernels into two smaller ones. When the number of network layers increases to extract deeper human posture features, network degradation occurs, that is, as the number of network layers increases, the performance of the network actually decreases. The study takes a residual learning module to directly add input to output, thereby avoiding gradient vanishing. C3D network is a 3D convolutional model. Compared with traditional 2D convolutional networks, C3D network can directly extract spatiotemporal features, capturing continuous actions and complex features. The C3D network framework is shown in Figure 4.

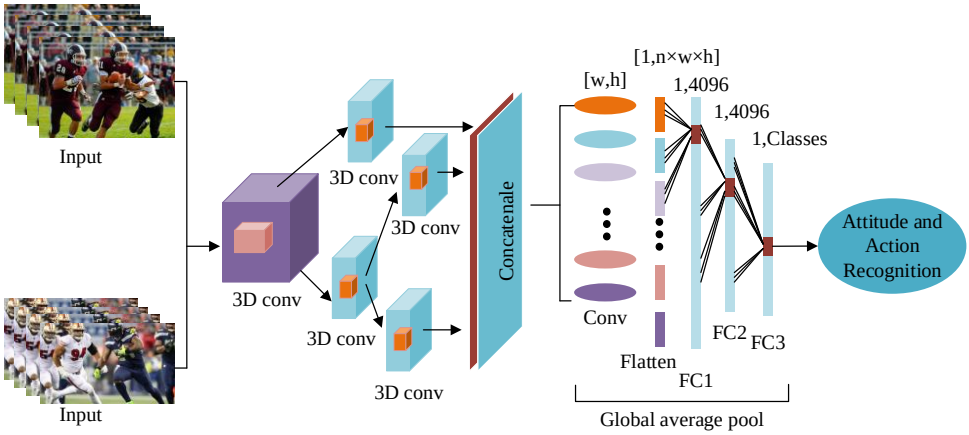
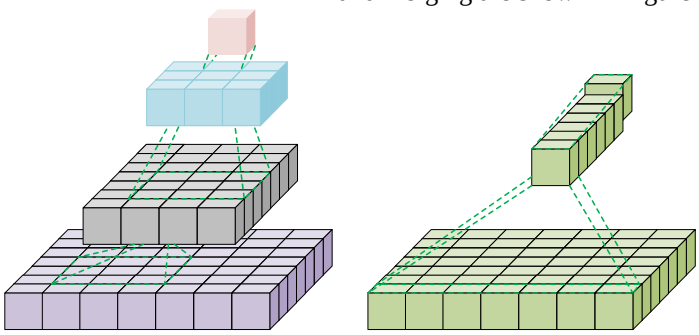


Figure 4: C3D network architecture framework

According to the C3D structure in Figure 4, it mainly consists of 3D pooling, 3D convolution, fully connected components, etc [21]. The 3D convolution part enhances the temporal feature extraction and adjusts the over-fitting problem caused by FCN through Dropout regularization. However, the FCN of C3D still faces many parameters. Therefore, the global average pooling replaces the original connection layer [22]. The 3D convolution calculation is shown in equation (9).

$$x_{ij}^{tt} = f\left(\sum_{n}\sum_{x=0}^{X_{i-1}}\sum_{y=0}^{Y_{i-1}}\sum_{z=0}^{Z_{i-1}} w_{ijn}^{xyz} x_{(i-1)n}^{(\alpha+x)(\beta+y)(\gamma+z)} + b_{ij}\right) \quad (9)$$

In equation (9), $f(\bullet)$ represents the activation function. x_{ij}^{tt} represents the convolution value of the j -th convolution kernel in the i -th representation layer at position (α,β,γ). Z_i, Y_i and X_i respectively signify the depth, height, and width of the 3D convolution kernel in the i -th representation layer. w_{ijn}^{xyz} signifies the convolution weight value of the feature map at position (x,y,z). b_{ij} represents the bias value. Due to the increased temporal information analysis, 3D convolution has a more complex network structure compared with 2D networks [23]. To lower the computational complexity, an asymmetric 3D convolutional layer is introduced into the C3D network, which utilizes a splitting and merging strategy to improve the convolutional layer structure and enhance feature extraction. The convolution layer splitting and merging are shown in Figure 5.



(a)Schematic diagram of merging convolution kernels

(b)Schematic diagram of asymmetric split convolution kernel

Figure 5: Schematic diagram of convolution splitting and merging

According to Figure 5, it asymmetrically splits a 7×7 convolution kernel into 7×1 or 1×7. The smaller the size of the convolution n×m, the lower the resource consumption. In addition, convolutions such as 1×1 and 3×3 can also be merged to form larger convolutions to meet the feature extraction. Next, the C3D also needs to consider the degradation problem that arises from extracting deeper human posture features. To solve this problem, an improved Group Normalization (GN) residual network is introduced to perform regularization. Compared with batch normalization, GN can reduce the network dependence on batches and improve the training accuracy of the network [24]. Then, GN normalizes the input data. The variance σ_i and mean μ_i of the image features are shown in equation (10).

$$\begin{cases} \sigma_i = \sqrt{1/m\sum_{k\in S_i}(x_k-\mu_i)^2+\varepsilon} \\ \mu_i = \dfrac{1}{m}\sum_{k\in S_i} x_k \end{cases} \quad (10)$$

In equation (10), m represents the quantity of input images. x_k signifies the k -th input information in set S_i. ε represents a constant value that suppresses the denominator to zero. Next, after completing the variance and mean calculations, the input data is regularized to obtain new input data x_i , as shown in equation (11).

$$x_i = 1/\sigma_i(x_i-\mu_i) \quad (11)$$

Next, the processed data x_i is scaled and translated to get the final normalized value y_i , as displayed in equation (12).

$$y_i = \gamma x_i + \beta \quad (12)$$

In equation (12), β and γ represent the scaling and translation parameters, respectively. The input set data S_i is displayed in equation (13).

$$\mathbf{S}_i = \{k \,|\, k_N = i_N, |\frac{k_C}{C/G}| = \frac{i_C}{C/G}\} \quad (13)$$

In equation (13), N signifies batch size. C signifies the channel dimension. G signifies the quantity of groups. k signifies the quantity of data inputs. $k_N = i_N$ represents the input σ_i and μ_i values oriented towards the (C,H,W) axis. H and W signify the height and width dimensions of the input image, respectively. In addition, the research considers that the input network's action keyframes contain many redundant frames, which will directly affect the performance of posture action recognition [25]. Therefore, an improved CBAM is applied to enhance the feature map attention in

the classification process. Compared with traditional CBAM, it adds temporal dimension feature analysis. The improved CBAM is displayed in Figure 6 [26].

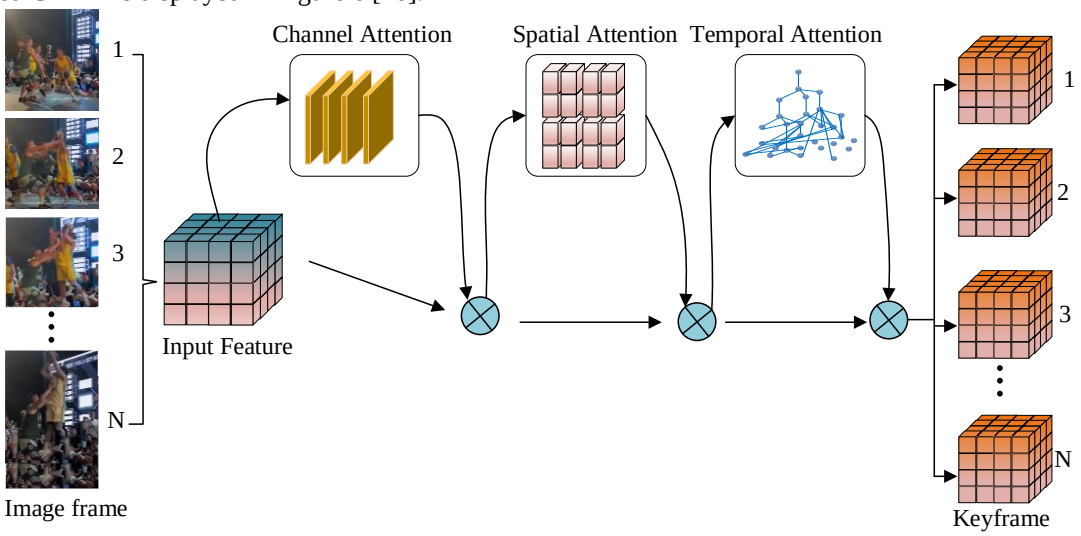


Figure 6: Improved CBAM attention mechanism

From Figure 6, the feature map extracted by the input 3D convolution is processed through three attention modules: channel, space, and time, to obtain the attention feature map F'. The calculation is shown in equation (14) [27].

$$F' = \mathrm{T}_{\mathrm{C}}\left(F\right) \bullet F \tag{14}$$

In equation (14), $\bullet$ represents element wise multiplication. F is the original feature map. The two-dimensional spatial attention feature map $\mathrm{T}_{\mathrm{S}}\left(F'\right)$ is obtained through the above processing, where T_{S} represents the spatial attention. The refined feature map F'' is obtained by element wise multiplication, as expressed in equation (15).

$$F'' = \mathrm{T}_{\mathrm{s}}\left(F'\right) \bullet F' \tag{15}$$

Finally, to accurately identify keyframes in the image data, the time attention module T_{M} is taken to continue processing the feature map and obtain the final key feature map, as displayed in equation (16).

$$F''' = \mathrm{T}_{\mathrm{M}}\left(F''\right) \bullet F'' \tag{16}$$

3 Results

3.1 Effect analysis of target recognition model

Next, the performance of the proposed sports training human posture recognition technology is analyzed. The experimental environment takes PyTorch deep learning framework, the system is WINDOWS 10, the system runs 32GB, the processor is INTEL i7-8700 CPU @ 3.20GHz, the graphics card is NVIDIA GeForce GTX 1070, and the experimental dependency library is OpenCV 4.4 Cuda 10.1. The experiment selects the Kinetic400 dataset and UCF101 dataset for training, where the Kinetic400 dataset contains 600 categories of various human posture actions. The UCF101 dataset contains 101 types of action categories, with a duration of 27 hours, covering various sports training content. In addition, the study also considers environmental hardware limitations and adds a self-made training dataset, which includes 45 categories of sports and 5,411 segments containing common sports like high jump, long jump, running, basketball, and table tennis. The experiment is recorded using DJI Pocket3 gimbal equipment. To ensure camera clarity, the video is recorded at 1080P resolution and 60fps, with a total of 600 videos. The specific recording is carried out in a clear and bright environment, with recording conditions including different lighting, different perspectives, and background conditions. Before experimental training, it is necessary to preprocess the sports video data and reverse processing. Specifically, in the preprocessing of sports video datasets, the video frames are first subjected to data augmentation processing, including random cropping, flipping, color adjustment, etc., to increase data diversity and improve model generalization ability. Subsequently, the video frames are standardized to normalize pixel values to a specific range, facilitating model training and convergence. Then, the improved YOLOv5s is used for video object recognition, with a training set ratio of 70% and a validation set ratio of 30%. The parameter settings are displayed in Table 2.

Table 2: Target recognition model parameter settings

Experimental parameters	Numerical value
Initial learning rate	0.001
Weight attenuation coefficient	0.0005
Learning rate momentum	0.937
Loss coefficient	0.05

Training threshold	0.2
Target aspect ratio	4.0

In Table 2, the YOLOv5s learning rate setting mainly considered the influence of the Adam optimizer. Adam had a relatively low sensitivity to learning rate. When the learning rate was set to 0.001, it could achieve good training results in most cases. Therefore, considering the overall situation, the learning rate is set to 0.001. Next, the research conducts training on sports object detection, introducing YOLOv7, Single Shot MultiBox Detector (SSD) algorithm, and Fast Region-based CNN (Faster R-CNN) as testing benchmarks in the experiment. Meanwhile, the study introduces Precision, Recall, IoU, Frames Per Second (FPS), Mean Average Precision under 50% IoU threshold (mAP-0.5), and mAP-0.50:0.95 as evaluation indicators. Firstly, the precision-recall performance of different target recognition techniques is compared, as shown in Figure 7.

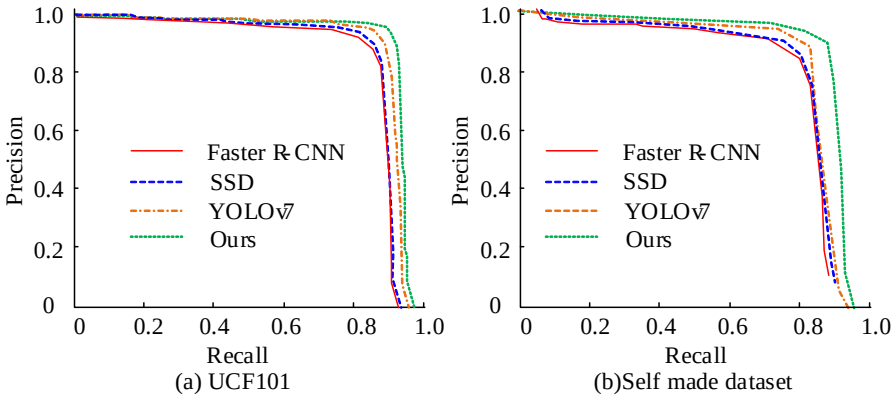


Figure 7: Precision-Recall curve test

Figure 7 (a) displays the test results on the UCF101. According to the test results, the research method outperformed YOLOv7, SSD, and Faster R-CNN. For example, the precision of the research model was 0.988, which was significantly better than YOLOv7's 0.976, SSD's 0.965, and Faster R-CNN's 0.959. In the self-made dataset of Figure 7 (b), compared with the UCF101 dataset, the recognition performance of all four techniques was decreased in more complex sport scenes. Overall, the Precision-Recall area of the research method is still the largest. Compared with YOLOv7, SSD, and Faster R-CNN, the research model has improved target recognition performance by 4.52%, 5.25%, and 7.25%, respectively. Next, the loss performance test is shown in Figure 8.

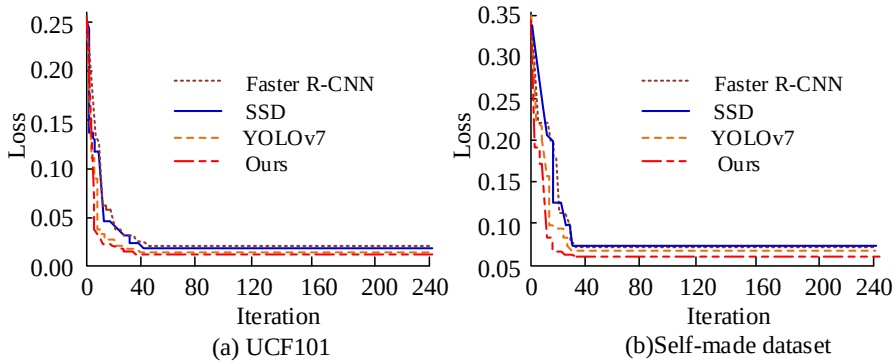


Figure 8: Loss performance test

Figure 8 (a) displays the training loss results on the UCF101 dataset. The overall convergence of the research model was better, approaching convergence at iteration 40 with a loss value of 0.021, demonstrating the best overall performance. The second best performance was YOLOv7, with a convergence loss value of 0.028, while SSD and Faster R-CNN were 0.035 and 0.038. Figure 8 (b) displays the loss test results on the self-made dataset. Overall, Faster R-CNN has a similar loss to SSD in the self-made dataset, but SSD has slightly better convergence. The final Faster R-CNN and SSD loss values were 0.078 and 0.077. Overall, the research model performs the best, with a final loss of 0.061. Next, based on self-made data, the comprehensive performance of different target recognition methods is compared, as shown in Figure 9.

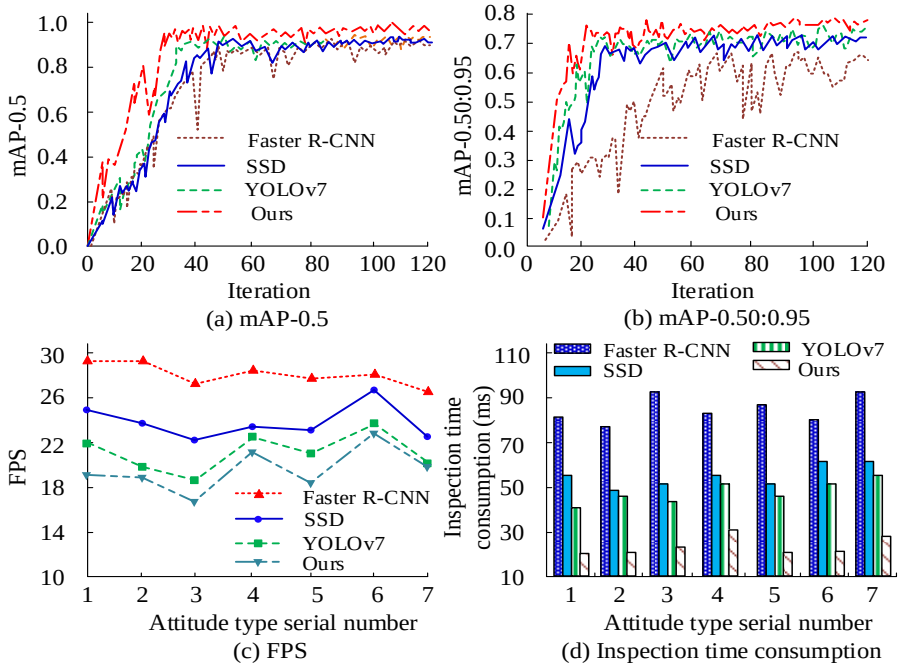


Figure 9: Comprehensive testing of target recognition model

Figures 9 (a) to 9 (d) show mAP-0.5, mAP-0.5:0.95, FPS, and detection time. Firstly, in the mAP-0.5 test, different models showed significant differences. For example, Faster R-CNN performed the worst, with obvious fluctuations and slow convergence during training. The final mAP-0.5 value was 0.845, while SSD and YOLOv7 performed better, with final results of 0.867 and 0.878, respectively. The research model performed best, with a final value of 0.942. In the mAP-0.5:0.95 test, Faster R-CNN showed significant fluctuations, indicating poor stability in accuracy for detecting objects of different categories. Overall, the research model performed the best, with an mAP-0.5:0.95 value of 0.872 at convergence. In addition, in the FPS value and detection time tests, the research model adopted a lightweight design to strengthen data analysis, which significantly improved its image processing ability per second. The average FPS was 28.5FPS, and the overall performance was the best. Meanwhile, in the time-consuming detection of multiple postures, the research model showed the best performance with an average time of 15.2ms. Figure 10 displays the object detection performance.

Figure 10: Research model for sports training posture target recognition

Based on the detection performance of 10 sports training targets, the research method can effectively identified target objects and determined human posture positions. Especially when there were static and dynamic characters in the scene, the research method could accurately determine the real sports training object, laying the foundation for subsequent human posture and action recognition.

3.2 Analysis of human posture and action recognition in sports training

At present, deep learning technology has extensive applications in the field of sports action recognition, providing important guidance for the evaluation and analysis of font actions in sports scenes. However, current action recognition technology faces problems such as low accuracy and high computational resource consumption. Therefore, a new sports training human posture recognition technology that can be used for real-time sports scene analysis is proposed. Next, this technology is subjected to specific experimental analysis to verify its feasibility. The experimental environment remains consistent. The study introduces Two-Stream Convolutional Network (Two-Stream) and C3D network as testing benchmarks. The learning rate is 0.025, the weight decay is 0.0003, and the spatial length parameter is 56×56. Recognition accuracy, Receiver Operating Characteristic curve (ROC), Peta Floating Point Operations (PLOPs), Accuracy, and mAP are taken as testing metrics. Firstly, the research conducts ablation tests, including removing residual networks, removing asymmetric convolutional layers, and removing attention mechanisms to evaluate the training performance of the model on the UCF101 dataset. Among them, A1: Remove the residual network, A2: Remove the asymmetric convolutional layer, and A3: Remove attention mechanism. B1: Simultaneously remove the residual network and asymmetric convolutional layer, B2: Simultaneously remove the residual network and attention mechanism, and B3: simultaneously remove asymmetric convolutional layers and attention mechanisms. The specific ablation experiment is shown in Figure 11.

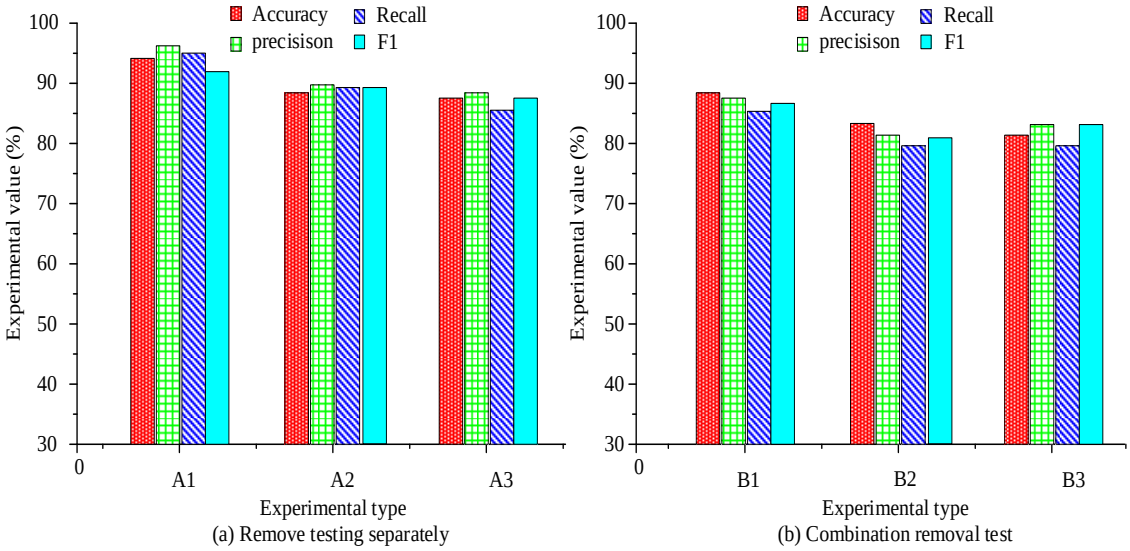


Figure 11: Results of ablation experiment

Figure 11 (a) shows a separate removal test. According to the results, removing the attention mechanism (A3) had a significant impact on the overall training accuracy, with accuracy, precision, recall, and F1 score of 90.12%, 91.25%, 88.21%, and 90.21%, respectively. Figure 11 (b) shows the combined removal test. Simultaneously removing the residual network and attention mechanism (B2) and simultaneously removing the asymmetric convolutional layer and attention mechanism (B3) had a significant impact on the overall model. In B2, its accuracy, precision, recall, and F1 score were 84.25%, 82.05%, 80.21%, and 81.25%, respectively. According to the test results, removing residual network, attention mechanism, and asymmetric convolution from the model can all affect the training performance of the network, which also verifies that improving the model can improve action recognition accuracy. Based on a self-made dataset, multiple sports training scenes are identified, as shown in Figure 12.

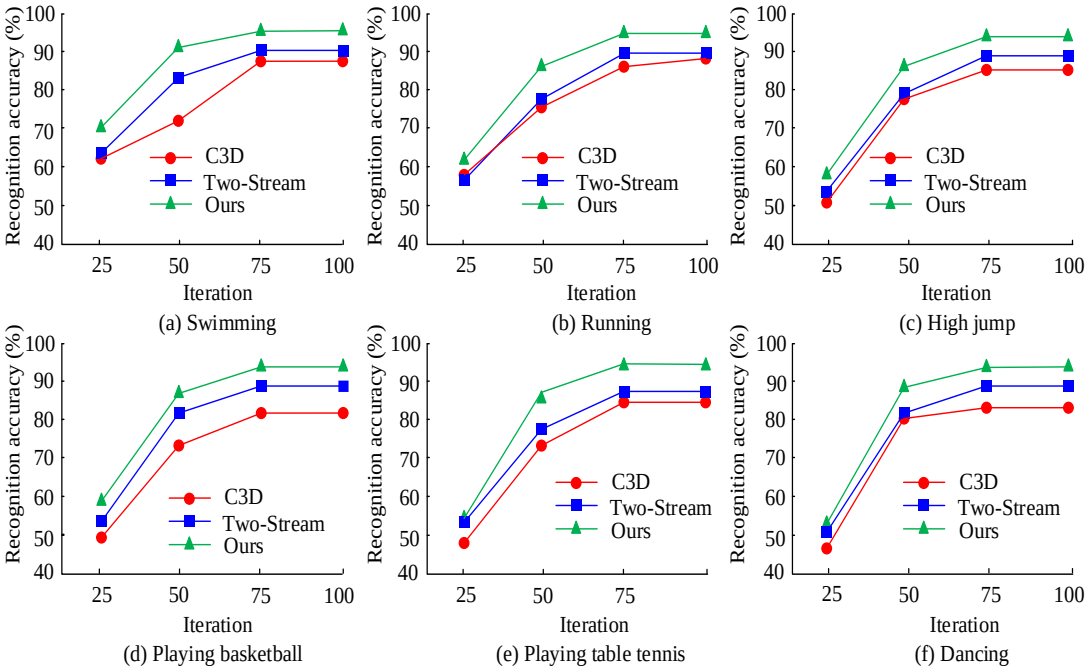


Figure 12: Comparison of accuracy of posture and action recognition in sports training

Figures 12 (a) to 12 (f) show the recognition results of actions such as swimming, running, high jump, playing basketball, playing table tennis, and dancing, respectively. In different sports scenes, the overall performance of the research model was the best, followed by Two-Stream, and finally C3D. In the table tennis scene that tests the ability of action recognition, the accuracy of the research model was 92.54%, which was better than Two-Stream's 87.54% and C3D's 83.25%. The ROC curve test results are displayed in Figure 13.

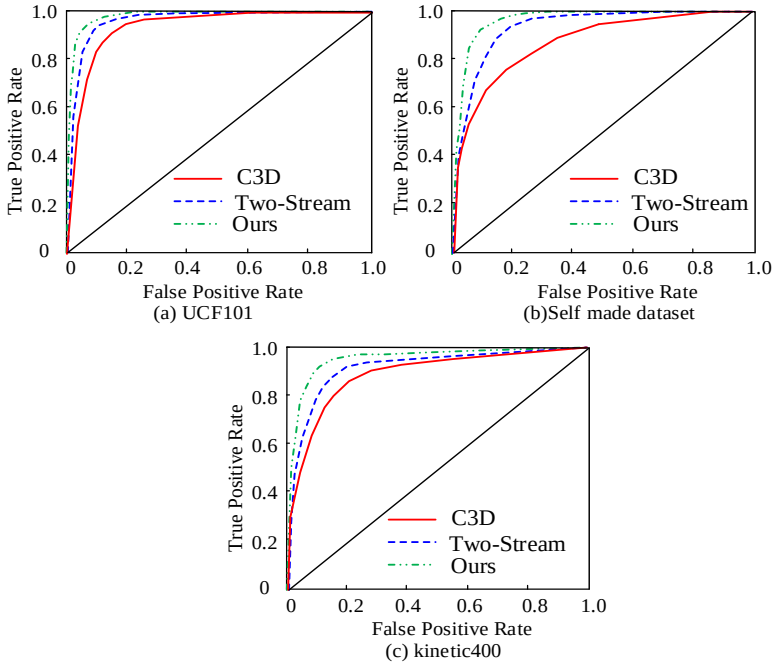


Figure 13: ROC curve test

Figures 13 (a) to 13 (c) show the ROC test results under kinetic400, self-made dataset, and UCF101, respectively. Overall, the research model had the largest area under the curve in all three datasets, ranked from best to worst as the research model, Two-Stream, and C4D. The research model had the best performance in posture and action classification. Compared with Two-Stream and C4D, the research model improved classification performance by 7.25% and 11.25%, respectively. The relationship between true positive rate and false positive rate reveals that as the false positive rate increases, the true positive rate also shows an increasing trend and remains stable. As the area between the two approaches the upper left corner, the AUC value becomes larger, indicating that

the overall performance of the research model is better. Table 3 presents a comprehensive comparison and statistical analysis of different action recognition methods.

Table 3: Comprehensive comparison of action recognition methods

Index	Algorithm type	kinetic400	UCF101	Self-made dataset
mAP	C3D	0.531	0.546	0.515
	Two-Stream	0.604	0.615	0.598
	Ours	0.701	0.708	0.701
	p*	0.0085	0.0091	0.0095
	p**	0.0152	0.0124	0.0135
PLOPs (MB)	C3D	38.456	45.456	46.458
	Two-Stream	39.456	40.258	37.152
	Ours	38.456	37.458	35.264
	p*	0.0091	0.0098	0.0092
	p**	0.0189	0.0135	0.0128
Identification accuracy	C3D	0.856	0.864	0.842
	Two-Stream	0.904	0.908	0.894
	Ours	0.956	0.974	0.945
	p*	0.0098	0.0078	0.0085
	p**	0.0355	0.0256	0.0119

Note: p* represents the comparison between the research model and C3D. p** represents the comparison between the research model and Two-Stream.

In Table 3, identification accuracy represents the ratio of the number of correct recognitions to the total number of recognitions. The better the value, the better the action recognition effect of the technology. Three datasets were selected for testing in the experiment, including mAP, PLOPs, accuracy, and action recognition accuracy. Overall, the research model performed the best in different indicator tests. For example, in the self-made dataset, the mAP of the research model was 0.701, while the mAP of Two-Stream and C3D was 0.598 and 0.515, respectively. In the PLOPs test that tests the algorithm's processing ability, taking the UCF101 dataset as an example, the C3D, Two-Stream, and research model were 46.458MB, 40.258MB, and 37.458MB, respectively. This research model consumed less overall. In addition, in statistical analysis, the research model had statistical significance compared with C3D and Two-Stream in various indicators. In the mAP test (kinetic400), the comparison between the research model and C3D had a p* value of 0.0085, while the comparison between the research model and Two-Stream had a p** value of 0.0152. It can be seen that the posture recognition method has the best application effect in sports training scenes.

4 Discussion and conclusion

In the field of sports training, human posture recognition technology has important guiding significance for standardized sports training. This technology can monitor athletes' postures in real-time, providing personalized guidance for standardized sports training. Therefore, the study introduces a network sports training human posture recognition technology that combines the improved YOLOV5s with C3D. The human posture recognition is completed from two aspects: action target recognition and posture action recognition.

In action target recognition, the research model is significantly superior to similar methods such as YOLOv7, SSD, and Faster R-CNN in Precision-Recall curve analysis. In the self-made dataset, the research model showed a 4.52% improvement in target recognition performance compared to YOLOv7, which was significantly better than SSD and Faster. The main reason is that the research model adopts a lightweight design and introduces CBAM for multi-feature extraction. In FPS, recognition time, and mAP-0.5 accuracy testing, the research model still showed excellent performance. In posture and action recognition, various sports training contents were selected for training. For example, in table tennis training, the accuracy of the model recognition reached 92.54%, better than Two-Stream's 87.54% and C3D's 83.25%. The main reason is that the research introduces asymmetric convolutional layers and residual networks to solve the model parameters and reduce the network dependence on batches. In addition, the improved attention mechanism layer enhances the capture of video action keyframes. Therefore, the research model performed well in subsequent mAP, PLOPs, and other tests, significantly outperforming similar technologies. Among them, PLOPs testing showed significant advantages of the research model. For example, on kinetic400, UCF101, and self-made datasets, the PLOP values of the research model were 38.456MB, 37.458MB, and 35.264MB, respectively, which were significantly better than similar techniques. The research model consumed less resources in data processing, and the comparison between data was statistically significant (p<0.05). This result indicates that asymmetric convolutional layers and residual networks effectively optimize network parameters and improve network computational performance. In addition, the proposed technology is compared with relevant action recognition techniques such as reference [5] and reference [6]. This research model used asymmetric convolutional layers and reduces system resource consumption through convolutional kernel splitting. Compared with references [5] and [6], the resource utilization rate of the research model was reduced by more than 20%, and the computational efficiency was improved by more than 15%.

The proposed technology has good application effects in complex sports training scenes. Based on lightweight design and parametric improvement, the technology has good accuracy and generalization ability. However, there are also shortcomings in the research, as the research techniques did not take into account factors such as low light and complex environments. In addition, for real-time multi-motion scenes and very low light motion scenes, the technology still faces insufficient recognition. In the future, to adapt to complex and invisible motion scenes, infrared assisted monitoring technology can be added. Moreover, feature extraction and analysis in low light scenes can be optimized to enhance the effectiveness of this technology.

References

[1] Liu W, Jia X, Ju Y, Jiang K, Wu S, Zhong L. Fragrant: frequency-auxiliary guided relational attention network for low-light action recognition. The Visual Computer, 2025, 41(2): 1379-1394. http://dx.doi.org/10.1007/s00371-024-03427-x

[2] Sheng B, Han R, Cai H. CDFi: Cross-domain action recognition using WiFi signals. IEEE Transactions on Mobile Computing, 2024, 23(8): 8463-8477. http://dx.doi.org/10.1109/TMC.2023.3348939

[3] Khan M A, Javed K, Khan S A, Abbasi A A. Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimed Tools Appl., 2024, 83(5): 14885-14911. http://dx.doi.org/10.1007/s11042-020-08806-9

[4] Zhou L, Jiang T. Learning body part-based pose lexicons for semantic action recognition. IET Computer Vision, 2023, 17(2): 135-155. http://dx.doi.org/10.1049/cvi2.12143

[5] Bai Y, Zou Q, Chen X, Li L, Ding Z, Chen L. Extreme Low-Resolution Action Recognition with Confident Spatial-Temporal Attention Transfer. International Journal of Computer Vision, 2023, 131(6): 1550-1565. http://dx.doi.org/10.1007/s11263-023-01771-4

[6] Kumar R, Kumar S. Survey on artificial intelligence-based human action recognition in video sequences. Optical Engineering, 2023, 62(2): 23102-23123. http://dx.doi.org/10.1117/1.OE.62.2.023102

[7] Yan G, Woźniak M. Accurate key frame extraction algorithm of video action for aerobics online teaching. Mobile Netw Appl., 2022, 27(3): 1252-1261. http://dx.doi.org/10.1007/s11036-022-01939-1

[8] Kong L, Pei D, He R, Huang D, Wang Y. Spatio-temporal player relation modeling for tactic recognition in sports videos. IEEE T Circ Syst Vid, 2022, 32(9): 6086-6099. http://dx.doi.org/10.1109/TCSVT.2022.3156634

[9] Wu F, Wang Q, Bian J, Ding N, Lu F X, Cheng J, Dou D J, Xiong H Y. A survey on video action recognition in sports: Datasets, methods and applications. IEEE T Multimedia, 2022, 25: 7943-7966. http://dx.doi.org/10.1109/TMM.2022.3232034

[10] Sun X, Wang Y, Khan J. Hybrid LSTM and GAN model for action recognition and prediction of lawn tennis sport activities. Soft Computing, 2023, 27(23): 18093-18112. http://dx.doi.org/10.1007/s00500-023-09215-4

[11] Le V H. Deep learning-based for human segmentation and tracking, 3D human pose estimation and action recognition on monocular video of MADS dataset. Multimed Tools Appl., 2023, 82(14): 20771-20818. https://doi.org/10.1007/s11042-022-13921-w

[12] Dhiman C, Varshney A, Vyapak V. AP-TransNet: a polarized transformer based aerial human action recognition framework. Mach Vision Appl., 2024, 35(3): 52-58. http://dx.doi.org/10.1007/s00138-024-01535-1

[13] Ma N, Wu Z, Cheung Y, Guo Y, Gao Y, Li J H, Jiang B Y. A survey of human action recognition and posture prediction. Tsinghua Sci Technol., 2022, 27(6): 973-1001. http://dx.doi.org/10.26599/TST.2021.9010068

[14] Liu Y, Zhang H, Xu D, He K. Graph transformer network with temporal kernel attention for skeleton-based action recognition. Knowledge-Based Systems, 2022, 240(3): 108146-10862. http://dx.doi.org/10.1016/j.knosys.2022.108146

[15] Dang R, Liu C, Liu M, Chen Q. Channel attention and multi-scale graph neural networks for skeleton-based action recognition. AI Commun., 2022, 35(3): 187-205. http://dx.doi.org/10.3233/AIC-210250

[16] Kulsoom F, Narejo S, Mehmood Z, Chaudhry H N, Butt A M, Bashir A K. A review of machine learning-based human activity recognition for diverse applications. Neural Computing and Applications, 2022, 34(21): 18289-18324. http://dx.doi.org/10.1007/s00521-022-07665-9

[17] Su T, Wang H, Qi Q, Wang L, He B. Transductive learning with prior knowledge for generalized zero-shot action recognition. IEEE T Circ Syst Vid, 2023, 34(1): 260-273. http://dx.doi.org/10.1109/TCSVT.2023.3284977

[18] Zhu X, Huang Y, Wang X, Wang R. Emotion recognition based on brain-like multimodal hierarchical perception. Multimed Tools Appl., 2024, 83(18): 56039-56057. http://dx.doi.org/10.1007/s11042-023-17347-w

[19] Qin Z, Liu Y, Ji P, Kim D, Wang L. Fusing higher-order features in graph neural networks for skeleton-based action recognition. IEEE T Neur Net Lear., 2022, 35(4): 4783-4797. http://dx.doi.org/10.1109/TNNLS.2022.3201518

[20] Liu S, Li Y, Fu W. Human-centered attention-aware networks for action recognition. Int J Intell Syst., 2022, 37(12): 10968-10987. http://dx.doi.org/10.1002/int.23029

[21] Chen B, Tang H, Zhang Z, Tong G J, Li B Q. Video-based action recognition using spurious-3D residual attention networks. IET Image Processing, 2022, 16(11): 3097-3111. http://dx.doi.org/10.1049/ipr2.12541

[22] Sánchez-Caballero, A., de López-Diz, S., Fuentes-Jimenez, D. et al. 3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information. Multimed Tools Appl., 2022, 81(17): 24119-24143. https://doi.org/10.1007/s11042-022-12091-z

[23] Timmons A C, Duong J B, Simo Fiallo N, Lee T, Quynh Vo H P. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci., 2023, 18(5): 1062-1096. http://dx.doi.org/10.1177/17456916221134490

[24] Wu J. Log Data Mining of User Purchase Behavior Based on Distributed Intelligent Optimization Algorithm. Informatica, 2024, 48(20). http://dx.doi.org/10.31449/inf.v48i20.6779

[25] Li S, Liu H, Qian R, Li Y X, See J, Fei M J, Yu X Y, Lin W Y. Ta2n: Two-stage action alignment network for few-shot action recognition. AAAI, 2022, 36(2): 1404-1411. https://doi.org/10.48550/arXiv.2107.04782

[26] Ansari M A, Singh D K. ESAR, an expert shoplifting activity recognition system. Cybern Inf Technol., 2022, 22(1): 190-200. https://doi.org/10.2478/cait-2022-0012

[27] Zhou Y. Structural Damage Identification of Large-Span Spatial Grid Structures Based on Genetic Algorithm. Informatica, 2024, 48(17). https://doi.org/10.31449/inf.v48i17.6428