Lightweight CA-YOLOv7-Based Badminton Stroke Recognition: A Real-Time and Accurate Behavior Analysis Method

Yuchuan Lin[1]*
Public Physical Education Department, Fujian Agriculture and Forestry University, FuZhou FuJian, 35000, China

***Abstract*—With the rapid development of sports technology, accurate and real-time recognition of badminton stroke postures has become essential for athlete training and match analysis. This study presents an improved YOLOv7-based method for badminton stroke posture recognition, addressing limitations in accuracy, real-time performance, and automation. To optimize the model, pruning techniques were applied to the backbone structure, significantly enhancing processing speed for real-time demands. A parameter-free attention module was integrated to improve feature extraction without increasing model complexity. Furthermore, key stroke action nodes were defined, and a joint point matching module was introduced to enhance recognition accuracy. Experimental results show that the improved model achieved a mAP@0.5 of 0.955 and a processing speed of 44 frames per second, demonstrating its capability to deliver precise and efficient badminton stroke recognition. This research provides valuable technical support for coaches and athletes, enabling better analysis and optimization of stroke techniques.**

***Keywords*—Badminton shot; pose recognition; YOLO V7; size adaptive input; model pruning; attention mechanism**

I. Introduction

The core of badminton lies in the precise and skillful execution of movements, such as high shots, smashes, and picks. While the sport itself has a low entry barrier, mastering its complex technical aspects requires systematic training and professional guidance. However, badminton instruction faces significant challenges, including limited teacher resources, the inability to provide personalized guidance to each student, and uneven teaching standards. These factors collectively constrain the efficiency and quality of students' learning of technical movements.

Fortunately, technological advancements offer new possibilities for addressing these challenges [1] [2]. In particular, the development of artificial intelligence, computer vision, and deep learning technologies has begun to play an increasingly important role in sports training and education [3] [4]. By leveraging these technologies, it is possible to develop an intelligent badminton technique action recognition system that can automatically recognize and analyze athletes' movements, provide objective feedback for improvement, and assist both amateur enthusiasts and students in learning badminton techniques independently and efficiently.

Building on these advancements, this article proposes an improved deep learning model—a badminton posture recognition system based on YOLOv7—designed to enhance the quality and efficiency of badminton technique instruction through precise motion capture and analysis. This system aims to improve the recognition accuracy and efficiency of existing models by incorporating data augmentation, anchor box fine-tuning, and the Coordinate Attention (CA) mechanism. The proposed system will serve as a valuable tool for both instructors and students, providing better support for teaching and self-training in badminton. By improving technical proficiency and reducing the risk of injuries caused by incorrect movements, this system seeks to offer scientific and systematic support for learning badminton techniques, ultimately enhancing both the quality and effectiveness of badminton training.

The research will explore the integration of these advanced methods into a cohesive system, aiming to address the current limitations in badminton training and teaching. Through thorough experimentation and analysis, this study seeks to contribute to the development of more effective and accessible tools for badminton instruction, potentially influencing broader applications in other sports as well.

II. Literature Review

This paper will collect existing work on pose recognition in sports to highlight the shortcomings of existing research.

A. Research on Deep Learning in Sports Action Recognition

In the realm of sports, particularly in small ball racket sports like badminton, tennis, and table tennis, motion recognition technology primarily utilizes two approaches: the first involves using the acceleration sensors in wearable devices to collect and classify data for motion detection, while the second applies deep learning technology to extract and learn features from video images for action recognition.

Numerous studies have highlighted the potential of these technologies. For instance, Ang et al. [5] developed a tennis visualization system that organizes and classifies match information, helping users better understand tennis events. Johnson et al. [6] introduced a method that enhances badminton serving accuracy through 3D tracking technology. Building on this, Dierickx et al. [7] further improved the accuracy of trajectory detection. Situmeang et al. [8] combined multiple visual analysis techniques to study athletes' movement characteristics and countermeasures. Jing et al. [9] utilized a support vector machine to recognize common swing actions in badminton game videos, whereas Wang et al. [10]

*Corresponding Author

significantly enhanced action recognition accuracy with a double-layer classifier algorithm.

While deep learning demonstrates high accuracy and comprehensive feature analysis in sports action recognition, it also presents challenges, such as the need for large data sets and the complexity of data collection. In small sample data sets, these techniques are prone to overfitting and require careful processing and optimization in practical applications.

B. Human Posture Recognition Method

The advent of Convolutional Neural Networks (CNNs) has led researchers to adopt deep learning-based techniques for human pose recognition. This approach essentially involves constructing convolutional layers to leverage large datasets, thereby extracting effective feature information to represent the human body. These features are then used in conjunction with efficient classification models for supervised training to achieve accurate predictions. The input data typically consists of an image or a video sequence, and after training the deep learning model, it can identify key body parts in the image, such as the head, arms, and knees.

Currently, there are two main research approaches to deep learning-based human pose recognition: direct regression based on keypoint coordinates and regression based on keypoint heatmaps. In 2014, the release of the MPII dataset [11], which contains approximately 25,000 images and covers over 40,000 human instances, with each instance including 16 keypoints, significantly advanced research in this field. In 2016, Wei et al. designed the CPM deep network [12], which extracts receptive fields of different sizes through repeated convolution operations and combines contextual information from the image to recognize human poses. CPM also introduced the concept of intermediate supervision, effectively addressing the issues of network depth and vanishing gradients, thus improving pose recognition accuracy. Additionally, the Stacked Hourglass Network (SHN) [13] introduced multi-resolution heatmap regression and multi-scale receptive field mechanisms, which enhanced the CNN structure for feature extraction, yielding excellent results. That same year, the release of the MSCOCO dataset [14] further enriched research resources, increasing the number of human keypoints to 18, making it more precise and comprehensive than MPII.

Pose recognition in complex multi-person scenes involves two approaches: top-down and bottom-up. The top-down approach first detects each person in the image through object detection, then performs keypoint recognition on each detected person. This method is straightforward but its performance heavily depends on the accuracy of human detection, and its processing time increases linearly with the number of people. Representative models include HRNet [15] and RMPE [16]. The bottom-up approach, on the other hand, first identifies all keypoints and then connects them according to the human pose model. This method does not slow down with an increase in the number of people, but matching keypoints in complex scenes can be challenging. Representative algorithms include OpenPose [17] and DeepCut [18].

Despite the achievements in theoretical innovation and technological development both domestically and internationally, human pose recognition technology still faces several challenges and difficulties that require further research and optimization.

C. Research Gaps

Despite significant progress in badminton stroke posture recognition technology, several challenges remain in practical applications:

1) The demand for high-precision recognition: Accurate recognition of badminton stroke posture is crucial, as even minor differences in movement can significantly affect the effectiveness of a stroke. However, existing technologies still require improvements in accuracy and robustness, especially in complex environments.

2) Challenges with real-time performance: In actual training or matches, there is a critical need for real-time posture recognition, enabling coaches and athletes to receive immediate feedback and adjust strategies accordingly. Current systems continue to face limitations in processing speed and providing real-time feedback.

3) Issues with automation and adaptability: Most current posture recognition systems require specific setups, such as particular camera angles and lighting conditions, which restricts the system's applicability and flexibility.

The above three points are the key to improving badminton hit recognition and also the focus of this study.

III. Detection Model

A. Framework

Badminton stroke posture recognition technology is primarily utilized in sports training and match analysis. By recognizing athletes' stroke postures, coaches can more effectively guide athletes in adjusting their techniques and optimizing training outcomes. This technology captures detailed aspects of athletes' movements, thereby aiding in the analysis of the precision and efficiency of stroke techniques. Currently, this field predominantly relies on video analysis and sensor technology, combined with machine learning and deep learning methods, to achieve motion capture and data analysis.

To address the limitations of existing methods, a badminton stroke posture recognition method based on an improved YOLO V7 model is proposed. The enhanced YOLO V7 model features a fast detection architecture composed of the Backbone module (lite-Darknet), Bottleneck module, Head module, and threshold matching module, as illustrated in Fig. 1. By introducing a parameter-free attention module, the model's feature extraction capability is enhanced, allowing it to adapt to varying environments and conditions. Additionally, model pruning techniques are employed to improve processing speed, and key nodes are defined, with a joint point matching module integrated to enhance matching accuracy. These advancements significantly increase the automation, real-time performance, and accuracy of posture recognition,

thus better meeting the needs of badminton training and match analysis.

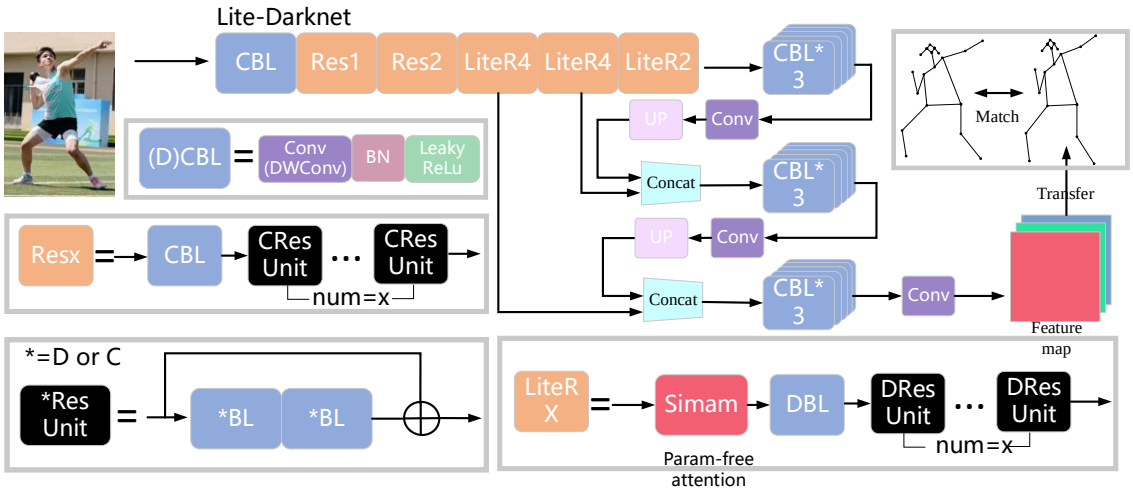

Fig. 1. Model architecture diagram.

To further improve the model's adaptability, the size of the input image is first adjusted to 640×640 pixels through scale normalization. The Backbone module extracts feature from the input images, identifies areas where node features are located, and proceeds to fuse these features across layers via the Bottleneck module. The semantic features and location features are then combined, and the simplified Head layer classifies the image. Finally, the classified features are sent to the node matching module, which accurately determines the connections between human nodes, thereby generating the detection result.

B. Node Feature Definition

Human body nodes need to be set before model training. In this paper, 19 nodes are selected and marked with 1-19 labels respectively, as shown in Fig. 2.

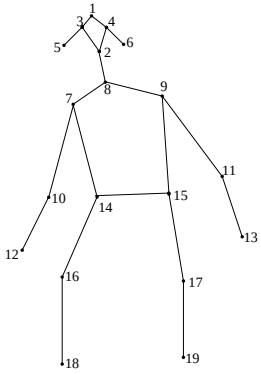

Fig. 2. Human body node diagram.

In Fig. 2, 1 is forehead, 2 is nose, 3 is left eye, 4 is right eye, 5 is left ear, 6 is right ear, 7 is left shoulder, 8 is neck. 9 is the right shoulder, 10 is the left elbow, 11 is the right elbow, 12 is the left wrist, 13 is the right wrist, 14 is the left hip, 15 is the right hip, 16 is the left knee, 17 is the right knee, 18 is the left ankle, 19 is the right ankle. The marked data set of 19 nodes was input into the improved YOLO V7 network for training, and the inference model was obtained. The inference model could detect 19 nodes of the human body in the image and preliminarily match them to get the result of human body pose.

C. Node Matching Template

There may be some errors in the preliminary matching result, for example, the node is matched to the wrong target body when multiple people overlap, so the node matching template needs to be used for accurate matching. According to the actual situation, this paper sets up a single human body posture of the node template, part of the Fig. 3, mainly including human swing, standing, single leg lift, squat and other posture.

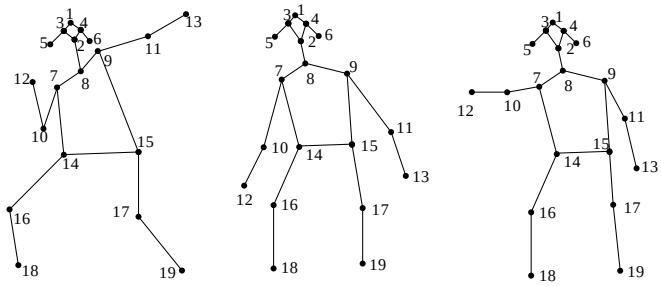

Fig. 3. Single human body posture of the node template.

In the exact matching operation, we first need to establish the threshold node descriptor set of the preliminary matching result graph and the template graph and realize the threshold node accurate matching by comparing the distance of the threshold node descriptor in the two-point sets. The calculation of the node descriptor in the template diagram is shown in Eq. (1):

$$R_i = (r_{i1}, r_{i2}, \ldots, r_{in}) \quad (1)$$

Where, the R_i presentation template key descriptor in figure collection, r_{in} presentation template the key points in the graph. Preliminary matches the key descriptor in figure calculation as shown in Eq. (2):

$$S_i = (s_{i1}, s_{i2}, \dots, s_{in}) \quad (2)$$

Where, the S_i said preliminary joint point descriptor set, match the picture s_{in} said preliminary match key points in the figure. Any two-descriptor similarity measure computation as shown in Eq. (3):

$$d(R_i, S_i) = \sqrt{\sum_{j=1}^{n} \left(r_{ij} - s_{ij}\right)^2} \quad (3)$$

$d(R_i, S_i)$ need satisfaction:

$$\frac{S_j}{S_p} < \delta \quad (4)$$

The S_j for distance R_i point recently, S_p for distance R_i near point, δ for distance threshold, when the threshold value is less than the set value to get the final figure.

D. Size Adaptive Normalization

In order to make the image have a unified scale, it is necessary to carry out size adaptive operation. At the same time, in order to deal with the problem of image feature intensity decline after operation and stabilize the detection accuracy of the model, it is necessary to enhance the image first. When image enhancement is carried out, the standardization process is carried out first. Common standardization methods include linear stretching, mean-variance normalization, histogram equalization, etc. This paper adopts mean-variance normalization method. Variance normalized (Z - score Normalization) is put all the data to the distribution of mean, variance 1 to 0, calculated as shown in Eq. (5):

$$x_{\text{saale}} = \frac{x-\mu}{S} \quad (5)$$

Where, the x_{saale} for the value of the normalized after x to the value of the normalized μ for the average of the image pixels, S as the standard deviation of the image and the standard deviation of calculated as shown in Eq. (6):

$$S = max\left(\sigma, \frac{1.0}{\sqrt{N}}\right) \quad (6)$$

Where σ is the standard variance and N is the number of pixels in the image. Standardized normalized processing after processing, the original data is mapped to get image Z on [0, 1] interval, calculated as shown in Eq. (7):

$$Z = \frac{x_i - min(x)}{max(x) - min(x)} \quad (7)$$

Where, the x_i with the value of image pixels, Max (x) and min (x) respectively, of the maximum and the minimum of image pixels. After normalized image to zoom in, fixed to 640 x 640 pixels, calculation process is as follows: according to the target size and the size of the original image, computing needs zoom ratio, can according to the proportion of long or short while zooming. For example, suppose the target size is 640×640 pixels, the scaling ratio needs to be calculated based on the width and height of the original image, as shown in Eq. (8) and (9):

$$\text{scale}_1 = max\left(\frac{w}{W}, \frac{h}{H}\right) \quad (8)$$

$$\text{scale}_2 = min\left(\frac{w}{W}, \frac{h}{H}\right) \quad (9)$$

Where, the scale_1 to long side scaling, scale_2 short while scaling w for target image (this article is 640 pixels), long W Z long, for the image h for target image of high (this article is 640 pixels), H is the height of the Z image. Will be calculated according to the original image to zoom scaling operation, during operation (such as bilinear interpolation), using interpolation algorithm is used to keep the image quality.

E. Parameterless Attention Mechanism

The core idea of Simam is based on the local self-similarity of images. In an image, there is usually a strong similarity between adjacent pixels, while the similarity between distant pixels is weak. Simam uses this property to generate attention weights by calculating the similarity between each pixel in the feature map and its neighbors. The following energy function is defined for each neuron:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \quad (10)$$

Where, $\hat{\mathrm{t}} = w_t t + b_t$, $\hat{\mathrm{x}}_i = w_t x_i + b_t$ is the target neurons and other neurons $x\ i$ in figure x characteristics of the single channel of the linear transformation, i is the index on the spatial dimension, T y_t and y_o is the target neurons and other neurons x_i two different values, M is the number of neurons on one channel. w_t and b_t is linear transformation of weights and bias. All values are in the scalar type, when t equals y_o, x_i equals y_t, minimum energy function. To minimize the above formula is equivalent to the training of neurons within the same channel linear separability between t with other neurons. For simplicity, we take binary labels and add regular terms, and the final energy function is defined as follows:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + (1 - (w_t t + b_t))^2 + \lambda {w_t}^2 \quad (11)$$

The analytical solution of the above equation is:

$$w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \quad (12)$$

$$b_t = -\frac{w_t(t+\mu_t)}{2} \quad (13)$$

Since all neurons on each channel follow the same distribution, it is possible to first calculate the mean and variance of the input features in the H and W dimensions to avoid double computation:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (14)$$

The whole process can be expressed as:

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \quad (15)$$

By calculating the mean and variance, the weight of each pixel in the feature map is adjusted adaptively, so that important pixels are amplified, and unimportant pixels are suppressed. Since SimAM does not require additional convolutional operations or full connection layers, it has low computational overhead and is suitable for embedding in various convolutional neural networks. In YOLOv7 target

detection algorithm in the whole network structure is divided into three parts, respectively is the network part of feature extraction, feature fusion network part and YOLO head testing part finally. Since the feature extraction network will output feature images of three different scales, three attention mechanisms need to be added, followed by a three-layer feature extraction module LiteR2. Attention mechanism is added in the network to determine the image characteristics of the channel number, the number of channels can be determined by the upper the output of the network structure, network structure exist three output in feature extraction, from top to bottom respectively is 80 x 80 x 512, 1024, 20 x 40 * 40 * 20 x 1024, Therefore, the number of channels corresponding to the three attention mechanism modules is 512, 1024, and 1024 respectively.

IV. Experiment and Verification

In this section, we will conduct ablation experiments based on self-made test data sets to verify the effectiveness of the proposed method.

A. Data Set Construction and Processing

In this study, an enhanced OpenPose model was selected to detect the skeletal key points of badminton technical movements. Given the limitations of existing datasets such as FineGym and UCF101 in supporting badminton-specific scenarios, a new badminton action sample dataset was independently constructed. This dataset specifically focuses on collecting videos of forehand high shots in badminton, which will be used for subsequent feature extraction and classification tasks.

The forehand high shot (using a right-handed racket as an example) is a fundamental and crucial technique in badminton, as it can effectively force the opponent to the back of the court and create an advantageous position. This technique can be systematically broken down into the following four steps, as shown in Fig. 4.

1) Preparation Phase

- Position and Posture: Stand with feet shoulder-width apart, left foot forward, right foot back, and body positioned sideways to the net, forming a stable support base.
- Grip and Arm Position: Hold the racket in the right hand with a moderate grip. Bend the arms naturally, point the racket in the direction of the ball, and slightly raise the left hand to aid in body balance.
- Sight: Focus the eyes on the incoming ball, preparing to execute the stroke.

2) Lead-Up Action

- Racket Head Movement: From the ready position, lift the right elbow upward and pull the racket head back and upward to a position directly above the head, with the racket face oriented forward.
- Body Coordination: Move the body slightly backward in sync with the racket head to increase the power of the shot. Gradually shift the weight from the right foot to the left, maintaining balanced movement.
- Racket Face Adjustment: Rotate the racket face slightly inward when overhead, ensuring readiness to strike the ball at the correct angle.

3) Swinging to the Ball

- Hit Point Selection: Select a hitting point slightly above shoulder height on the right side of the body, allowing better control over the flight and power of the ball.
- Swing: Swing the racket from the bottom of the lead position forward and upward, swiftly and forcefully, ensuring that the racket face is aligned with the ball at the moment of contact.
- Wrist Utilization: At the end of the swing, accelerate the racket head through rapid wrist action to generate more power and control, directing the ball toward the opponent's far court.

4) Follow-Through and Recovery

- Finishing Position: After striking the ball, rotate the body gently to the left and forward, aiding in returning to a stable state and preparing for the next action.
- Weight Adjustment: Quickly redistribute the weight after the shot to be ready to move or return to a defensive position, in anticipation of the opponent's return.
- Racket Face Adjustment: Ensure that the racket face is realigned toward the net after the shot, facilitating preparation for the next stroke.

Following these four detailed steps allows for effective execution of forehand high shots, thereby not only improving the quality of the stroke but also gaining better control over the rhythm and strategic layout of the court.

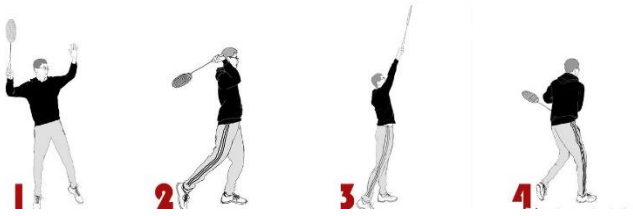

Fig. 4. Forehand high ball technique breakdown diagram.

B. Collection of Experimental Data Sets

In this study, a high-quality badminton stroke pose recognition dataset was constructed by carefully selecting 20 right-handed participants, including 10 males and 10 females, aged between 22 and 27 years. The participant group consisted of half national-level badminton players and half badminton enthusiasts with a certain level of skill. All participants passed a series of motor stability and reliability tests before inclusion, and their heights were controlled between 173 cm and 178 cm to minimize variations in technical movements due to differences in body shape.

Data acquisition was conducted through two methods: field acquisition and network video acquisition. Field data were collected using DJI Action2 and Nikon D70s cameras from May to June 2024 on a standard badminton court, with the ADIBO Smart Badminton server A200 used to ensure consistent service conditions, as shown in Fig. 5. In addition, qualified badminton match videos were screened from online video platforms to enhance the diversity and coverage of the dataset. To ensure data consistency and repeatability, the position of the serve was precisely controlled, and participants were required to hit the ball within a specified area, effectively reducing data bias caused by variations in participant positioning.

As a result of these methods, a comprehensive badminton stroke pose dataset was successfully constructed, including data from both professional players and enthusiasts. The aim is to promote the scientific analysis of badminton technical training and competition by accurately identifying and analyzing stroke techniques. This dataset serves as a valuable resource not only for the technical analysis of badminton but also as an experimental foundation for subsequent research in computer vision and machine learning.

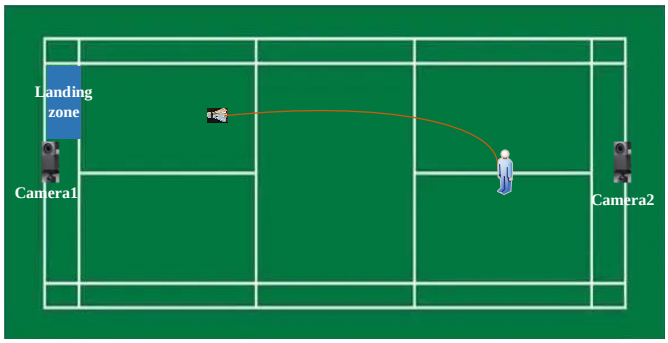

Fig. 5. Data acquisition method.

To further refine the dataset, point coordinates were normalized to reduce the influence of body center variations, and the body was rotated to a fixed angle to minimize the impact of different viewing angles.

The experimental setup is outlined in Table I. The PyTorch deep learning framework was utilized for the experiment. Stochastic Gradient Descent (SGD) was selected as the optimizer, with a batch size of 16. The cross-entropy loss function was employed as the loss function. The learning rate decay strategy implemented a reduction to one-tenth of the original rate at each specified interval. For the JDTD datasets, the initial learning rate was set at 0.1. The first dataset was trained for 50 epochs in total, with the learning rate reduced at the 30th and 40th epochs. The second dataset underwent 65 epochs of training, with the learning rate decaying at the 45th and 50th epochs.

TABLE I. ALGORITHM EXPERIMENT ENVIRONMENT

Environment	Version	Environment	Version
System	Ubuntu20.04	**CPU**	I7-8700k
Graphics card	GTX1080Ti	**Internal memory**	32G
Python	3.7	**PyTorch**	1.2 .0
CUDA	11.1	**CUDNN**	8.0 .4

C. Ablation Experiment

In order to verify the effectiveness of the improved algorithm in this paper, the detection experiment results of YOLO V7 original model, YOLO V7+ adaptive input module, YOLO V7 pruning model + adaptive input module, YOLO V7 model + adaptive input module + node matching module were studied on the same data set, as shown in Table II.

As can be seen from Table II, although the detection speed of the model is not improved after the addition of the adaptive module, the adaptability of the model to images of different sizes is greatly improved, and the detection accuracy is improved. When the pruning model is used, the detection speed and accuracy of the model are greatly improved when the IOU threshold is low, but the detection accuracy is not improved when the IOU threshold is high. After the node matching module is added, the detection speed and accuracy of the model are improved, which proves the effectiveness of the model improvement.

TABLE II. ABLATION RESULTS

Model	mAP@0.5	mAP@0.75	FPS
YOLO V7 original model	0.946	0.81	35
YOLO V7+ adaptive input module	0.947	0.83	35
YOLO V7 Pruning model + adaptive input module	0.952	0.83	39
YOLO V7 model + adaptive input module + node matching module	0.955	0.86	44

Note: mAP@0.5 indicates that the IOU threshold is 0.5, and FPS indicates the number of images processed per second

Fig. 6 illustrates the partial results of badminton stroke posture recognition using the trained model. As shown in the figure, the model demonstrates effective detection for both single-player and multi-player scenarios with minimal occlusion. For images with some occlusion, a node matching template is employed to enhance the model's detection accuracy. This issue arises when multiple players overlap significantly or exhibit similar postures, resulting in the detected nodes being unable to accurately distinguish the target player, thereby leading to detection errors.

The improved model, which incorporates an attention mechanism alongside data augmentation and anchor box adjustment, is based on YOLOv7. This enhanced model is compared with the baseline model to evaluate the performance improvements brought by the addition of the attention mechanism as shown in Table III and Table IV. In terms of posture recognition accuracy, there is noticeable improvement across all phases of the stroke. For instance, the accuracy of detecting the "Swinging to the Ball" phase increased from 86.12% to 88.26%, while the accuracy for the "Follow-Through and Recovery" phase approached 90%. Similarly, recall rates saw significant increases: the "Preparation Phase" improved by 1.73%, the "Lead-Up Action" by 2.06%, the "Swinging to the Ball" by 1.4%, the "Follow-Through and Recovery" by 3%, and overall detection recall exceeded 90% for some phases. These comparisons indicate that the performance of the YOLOv7 model has been significantly enhanced following the integration of the attention mechanism.

The mAP (mean Average Precision) for the model is detailed in Table V.

In this study, the average accuracy of the badminton stroke posture recognition model was improved from an initial 85.25% to 88.50%. This 3.25% increase in accuracy was achieved through a combination of data augmentation, anchor box adjustment, and the incorporation of an attention mechanism.

TABLE III. COMPARISON OF MODEL ACCURACY AFTER CA MECHANISM IS ADDED

Algorithm	Preparation Phase	Lead-Up Action	Swinging to the Ball	Follow-Through and Recovery
YOLOv7+ Data expansion +anchor fine-tuning	86.12%	83.51%	85.52%	88.31%
YOLOv7+ Data expansion +anchor fine-tuning +CA	88.26%	85.58%	86.92%	89.30%

TABLE IV. COMPARISON OF MODEL RECALL RATE AFTER ADDING SIMAM MECHANISM

Algorithm	Preparation Phase	Lead-Up Action	Swinging to the Ball	Follow-Through and Recovery
YOLOv7+ Data expansion +anchor fine-tuning	86.12%	83.51%	85.52%	88.31%
YOLOv7+ Data expansion +anchor fine-tuning +CA	87.85%	85.57%	86.92%	91.30%

TABLE V. COMPARISON OF MODEL RECALL RATE AFTER ADDING SIMAM MECHANISM

Algorithm	mAP
YOLOv7	85.25%
YOLOv7+ Data expansion	87.01%
YOLOv7+ Data expansion +anchors fine tuning	87.50%
YOLOv7+ Data expansion +anchors trimming +CA	88.50%

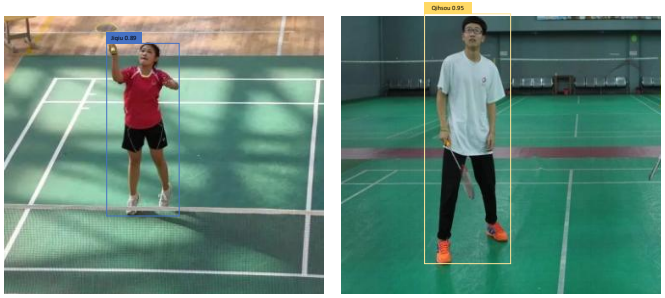

Fig. 6. Identify results after adding the simam attention mechanism.

D. Comparison of Other Algorithms

In the previous discussion, data augmentation, anchor box adjustment, and the integration of an attention mechanism were applied to the original YOLOv7 model. These modifications ultimately demonstrated an improvement in the accuracy of YOLOv7 for badminton stroke posture recognition. The enhanced YOLOv7 algorithm is compared with other target detection algorithms to demonstrate the superiority of the proposed approach.

In the field of target detection, Faster R-CNN, a two-stage detection algorithm, is widely utilized, including in transmission line inspections. Similarly, CenterNet, an anchor-free detection method, has achieved significant results in 3D target detection. Other network structures, such as DenseNet and BiFPN, are also extensively used in target detection. These network structures are applied to badminton stroke posture recognition, and the resulting average accuracies are compared to validate the advantages of the proposed algorithm.

- YOLOv3: Utilizes the Darknet53 network structure, combining residual learning and multi-scale output to optimize deep networks and enhance feature extraction capabilities.
- Faster R-CNN: A typical two-stage target detection algorithm that merges Fast R-CNN and RPN networks to rapidly and accurately generate candidate regions and detect targets.
- CenterNet: An anchor-free detection method that simplifies the detection process and speeds it up by detecting the target center and using heatmap technology.
- MobileNet: A lightweight deep neural network that employs depthwise separable convolution to significantly reduce computation, making it suitable for embedded systems.
- BiFPN: A bidirectional weighted feature pyramid network structure that optimizes feature fusion and improves the performance and efficiency of the detection network.

As shown in the Table VI, the average accuracy of the improved YOLOv7 algorithm presented in this study is 3.6% higher than that of the YOLOv7-BiFPN algorithm, 22.75% higher than that of the MobileNet algorithm, and 4.58% higher than that of the CenterNet algorithm. It also surpasses the Faster R-CNN algorithm by 3.83% and the YOLOv3 algorithm by 7.34%.

TABLE VI. COMPARISON OF MODEL ACCURACY UNDER DIFFERENT ALGORITHMS

Algorithm	mAP
YOLOv3	81.16%
Faster-RCNN	84.67%
CenterNet	83.92%
MobilNet	65.75%
YOLOv7-BiFPN	84.90%
YOLOv7+CA	88.50%

V. CONCLUSION

Badminton, as a fast-paced and technically demanding sport, requires precise analysis of player movements to enhance performance and training outcomes. Recognizing this need, this study proposes a deep learning-based algorithm for badminton player pose recognition. The proposed method addresses the challenges of accuracy and efficiency in pose

recognition by integrating data augmentation, anchor box fine-tuning, and the Coordinate Attention (CA) mechanism into the YOLOv7 algorithm. These improvements significantly enhance the model's recognition capabilities.

Experimental results demonstrate that the average accuracy of the improved YOLOv7 model has increased from 85.25% to 88.5%, outperforming other popular algorithms such as YOLOv3, Faster R-CNN, CenterNet, MobileNet, and YOLOv7-BiFPN. Moreover, the introduction of the CA attention mechanism has yielded particularly noteworthy results in recognizing specific badminton actions, such as smashes and picks, by reducing instances of missed or false detections. Looking ahead, future research will focus on further refining the dataset, exploring additional challenges in motion recognition, and considering the integration of edge computing techniques to optimize real-time processing capabilities.

ACKNOWLEDGMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Liu S, Zheng P, Xia L, et al. A dynamic updating method of digital twin knowledge model based on fused memorizing-forgetting model[J]. Advanced Engineering Informatics, 2023, 57: 102115.

[2] Fu T, Li P, Liu S. An imbalanced small sample slab defect recognition method based on image generation[J]. Journal of Manufacturing Processes, 2024, 118: 376-388.

[3] Fu T, Liu S, Li P. Digital twin-driven smelting process management method for converter steelmaking[J]. Journal of Intelligent Manufacturing, 2024: 1-17.

[4] Fu T, Liu S, Li P. Intelligent smelting process, management system: Efficient and intelligent management strategy by incorporating large language model[J]. Frontiers of Engineering Management, 2024: 1-17.

[5] Polk Tom,Yang Jing,Hu Yueqi,et al.TenniVis:Visualization for Tennis Match Analysi s[J].IEEE transactions on visualization and computer graphics.2014,20(12):2339-2348.

[6] Shayne Vial,Jodie Cochrane,Anthony J,et al.Using the trajectory of the shuttlecock as a measure of performance accuracy in the badminton short serve[J].International Jou rnal of Sports Science&Coaching,2019,14(1):91-96.

[7] Dierickx T.Badminton Game Analysis from Video Sequences[D].Ghent:University of Ghent, 2015 .

[8] WeiTa C,Sitmeang S.Badminton Video Analysis based on Spatitemporal and Stroke Features[J].International Conference on Multimedia Retrieval,2017,12(1):121-122.

[9] Heemskerk C, David L, Steve S, et al. The effect of physical education lesson intensity and cognitive demand on subsequent learning behaviour[J]. Journal of science and medicine in sport,2020,23(6):586-590.

[10] Ben AT, Elleuch I, Guermazi R. Student Behavior Recognition in Classroom using De -ep Transfer Learning with VGG-16[J]. Procedia Computer Science,2021,192:951-960.

[11] Andriluka M, Pishchulin L, et al. Human Pose Estimation: New Benchmark and State of the Art Analysis[C]// Computer Vision and Pattern Recognition (CVPR). IEEE, 2014.

[12] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[C]// Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[13] Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation[J]. arXiv e-prints, 2016.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[J]. 2014.

[15] Sun K, Xiao B, Liu D, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[16] Fang H, Xie S, Tai Y, et al. RMPE: Regional Multi-person Pose Estimation[J]. 2016.

[17] Cao Z, Simon T, Wei S E, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[18] Rajchl M, Lee M, Oktay O, et al. DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks[J]. IEEE Transactions on Medical Imaging, 2016, 36(2):674-683.