Yang Q et al. (2025) AUTOMATIC EVALUATION AND IMPROVEMENT OF SOCCER SERVING TECHNIQUES UTILIZING COMPUTER VISION. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 25 (99) pp. 402-417.
DOI: https://doi.org/10.15366/rimcafd2025.99.026

ORIGINAL

AUTOMATIC EVALUATION AND IMPROVEMENT OF SOCCER SERVING TECHNIQUES UTILIZING COMPUTER VISION

Mingliang Song[1], Xiao Chen[2,3], Qishun Yang[4*], Zhengdong Mi[1], Qin Qin[1], Dan Feng[1]

[1] Wuhan Auto Valley Stadium Operation Investment Development Co., LTD, Wuhan, 430000, China.
[2] School of Sports Training, Tianjin University of Sport, Tianjin, 301617, China
[3] Department of Physical Education, Zhongnan University of Economics and Law, Wuhan, 430000, China.
[4] School of Physical Education, International Equestrian Academy, Wuhan Business University, Wuhan, 43000, China
E-mail: yangqishun0422@sina.com

Recibido 11 de Marzo de 2024 **Received** March 11, 2024
Aceptado 11 de Octubre de 2024 **Accepted** October 11, 2024

ABSTRACT

Soccer serving technology has an important role and value in the game, and excellent serving technology can create a variety of offensive opportunities. Computer vision technology, especially target detection technology, has a wide range of applications in soccer. This paper is simplified based on YOLOv5, and the structure of CSPDarkNet53 is streamlined into Mobile Net structure, which reduces the number of parameters of the model and improves the detection speed of the model. Aiming at the problems of target occlusion and uneven illumination conditions, different attention mechanisms are embedded into the network model respectively, which improves the detection ability of the model on the target. The improved YOLOv5 model is tested for performance on a publicly available dataset, and the experimental results show that the model proposed in this paper has better detection performance.

KEYWORDS: Computer Vision; YOLOv5; Soccer Serving

1. INTRODUCTION

Soccer, as a global sport, has a broad and far-reaching influence. Soccer is a sport that transcends national boundaries and races, and it promotes exchanges and integration among different cultures around the world. Soccer

can become a symbol of countries and regions, inspire people's sense of belonging and pride in the communities they belong to, help unite people's hearts and promote social harmony. The soccer industry chain covers a wide range of fields, including television broadcasting, sponsorship, ticket sales, peripheral products, etc., so the role of soccer in boosting the local economy and employment is not to be underestimated. The development of soccer technology and tactics continues to drive the progress of the sport. Artificial Intelligence (AI) technology and Internet of Things (IoT) technology have brought more technological means to the game of soccer (Chen, Han, et al., 2023; Chen, Li, et al., 2023; Li & Cao, 2021), for example, the introduction of Video Assistant Referee (VAR) technology has provided a more accurate basis for the match referee to award penalties, and the application of data analytics and sports science has led to a more scientific and refined aspect of the team's training and matches. These technological and tactical advances have not only improved the quality of the game, but also provided players and coaches with more room for exploration and innovation. Soccer serving technique has an important role and value in the game, especially in corner kicks, free kicks and kickoffs. Excellent serving technique can create various attacking opportunities. For example, in corner kicks and free kicks, accurate passes or direct shots can directly threaten the opponent's goal and create good scoring opportunities for the team. In a match, an excellent serve can change the whole situation of the game. Whether it is a precise corner kick, a stunning direct free kick, or a perfect kick-off, it can be the turning point of the game. On the defensive end, serving technique is also important. A solid kick-off can help the team to release the ball quickly and reduce the pressure from the opponent; a high ball scramble to win can also lay the foundation for a solid defense. Soccer serving technique plays a crucial role in the game, not only a mere technical action, but also a key link that affects the whole game. Excellent serving technique can bring many unexpected gains for the team, and become one of the key factors to achieve victory. Computer vision technology has a wide range of applications in the field of soccer (Huang et al., 2023), such as the VAR system that has revolutionized the sport. The system carries out video replay and analysis of controversial scenes in the game through computer vision technology, which helps the referee make more accurate judgments, reduces the possibility of misjudgment, and improves the fairness and professionalism of the game referee. Computer vision technology can be used to capture various data in the game in real time, including team positions, players' running trajectories, passing routes and so on. These data can be used for game tactics analysis, player performance evaluation, and injury prevention. Through computer vision technology, players' physical quality can be evaluated more scientifically, including speed, flexibility, explosive power, etc., thus helping coaches to better develop personalized training plans. Target detection technology as a branch of computer vision has a high application value in the field of soccer. Target detection methods are usually implemented using deep convolutional neural networks, such as Faster

RCNN and YOLO. Tracking algorithms that relate the detected target of interest to the target in the previous frame are commonly used, such as Kalman filtering based on the Hungarian algorithm and algorithms based on correlation filtering, such as KCF (Henriques et al., 2014), CN (Danelljan et al., 2014). For player tracking, literature (Olagoke et al., 2020) proposed a multi-camera data fusion method in order to solve the problem of insufficient content in the single camera lens frame, using a multi-camera system to capture the player's movement more comprehensively. Literature (Dicle et al., 2013) proposes an interference term-aware color model and a target-aware depth model, and combines the two to track the upper and lower body of the target respectively, which improves the effective overlap rate of the algorithm. In literature (Naik & Hashmi, 2023), the authors use the Deep SORT algorithm to track players in soccer videos, but the improvement of Deep SORT and the experimental results of the algorithm on soccer dataset are missing. Literature (He et al., 2022) proposes a transformer-based duplicate detection eliminator D3, firstly, when duplicate detection occurs, the D3 eliminator mitigates the process by generating enhanced labeled frame losses, and then matches them using the Rally-Hungarian method, which speeds up the convergence rate and saves training time during model training. Literature (Hurault et al., 2020) introduces a self-supervised detection algorithm to detect small targets in soccer videos for the small targets of players in soccer match videos, while the step of manually labeling the dataset is eliminated because of the use of self-supervised algorithms. In terms of tracking, a pedestrian re-labeling mechanism using triplet loss is proposed. Literature (Dhassi & Aarab, 2018) combined local and global features of the target and combined particle filtering to estimate the object motion. There has been a lot of academic research on sports target detection and tracking algorithms for soccer games. However, the soccer game scene is complex, and there are common unfavorable conditions such as target deformation, occlusion, background clutter and light source changes, player detection and tracking in soccer and subsequent team class discrimination. To address the above difficulties, this paper proposes an improved YOLOv5 network model. The feature extraction backbone network part, the bottleneck part for feature fusion and the header for detection prediction of the original YOLOv5 network model are improved respectively, which increases the network model's ability to extract targets, especially small targets, through fine feature engineering and the introduction of the attention mechanism. The model has stronger detection ability and robustness, effectively improving the problem of wrong and missed detection due to occlusion and deformation, and reflecting better comprehensive performance.

2. Improved YOLOV5 Model

2.1 The Basic Yolov5 Model

The most critical part of the target detection process is the construction

of the detection network, and the network structure directly affects the detection performance. In this paper, the original YOLOv5 detection model is chosen as the baseline standard, which is used to compare with the improved model. From the above, it can be seen that the baseline YOLOv5 model consists of three parts: backbone network backbone, neck and prediction head. Among them, the backbone network backbone uses CSPDarknet-53, the neck part uses the combined structure of PANet and SPP, and the head part includes three convolutional layers, which output the positional information, confidence and category of the labeled boxes, respectively. The network structure of baseline YOLOv5 is shown in Figure 1: As can be seen from the figure, the input accepted by the YOLOv5 network is an RGB image of 640×640 size with three color channels, and the input is passed through the backbone network backbone for feature extraction. The convolution kernel is utilized to generate feature vectors and the process of constant down sampling is used to obtain feature maps of different sizes and channel counts of 80×80×255, 40×40×255 and 20×20×255. The structure of the backbone consists of the Conv module and the C3 module. The Conv module is composed of the convolution operation, the BN batch normalization process and the SiLU activation function cascade, and the SiLU function is expressed as follows:

$$SiLU(x) = x * \sigma(x) = x * \frac{1}{1+e^{-x}} \tag{1}$$

Compared to the ReLU function, the SiLU function is smoother, not monotonically increasing in the domain of definition, and the image is not horizontal when the input is less than zero, but still has a gradient, and the minimum of the function acts as a soft-bottom for the weights, inhibiting the learning of large weights.

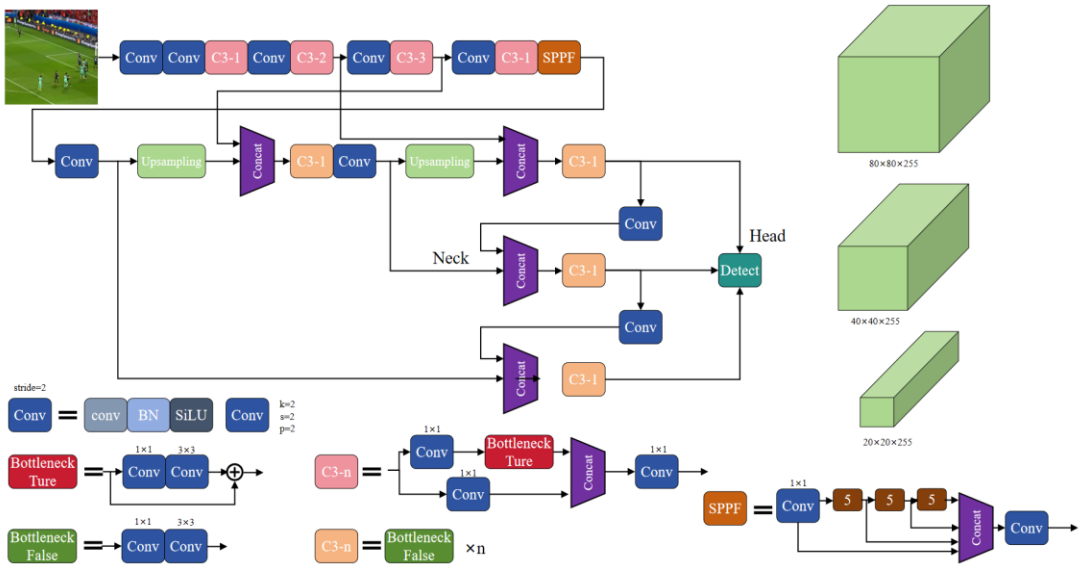

Figure 1: YOLOv5 Network Architecture

$C3-n$ stands for $C3$ module, where n stands for the number of bottlenecks in $C3$ module. When the parameter of bottleneck is False, the $C3$ module is a 1×1 $Conv$ module cascaded with a 3×3 $Conv$ module; when the parameter of bottleneck is True, the $C3$ module is a residual connection, which adds the original input to the output after two $Conv$ modules. The parameters of $C3$ are related to the bottleneck; when the parameter of $C3$ is False, a bottleneck with a parameter that is also False is used, and vice versa (indicated by the same color scheme in the figure). SPPF (Spatial Pyramid Pooling-Fast) is an upgrade of SPP Spatial Pyramid Pooling. SPPF is mathematically equivalent to the output of SPP, but SPPF cascades three 5×5 pooling layers, reducing the amount of computation and increasing the speed.

2.2 Improved Yolov5 Algorithm

In this paper, we apply the YOLO algorithm to real-time video streaming, in addition, the YOLO algorithm may perform poorly in the face of scenarios such as dense targets, occlusion, and small target size. Therefore, we need to improve the YOLO structure to obtain higher accuracy and faster detection speed.

2.2.1 Streamlining the Backbone Network

By analyzing the network structure of YOLOv5, it can be seen that the YOLOv5 network chooses CSPDarknet53 as the backbone network, which achieves high detection accuracy, but its number of parameters is large and the computational volume is too large, which cannot satisfy the real-time requirements of target detection in the video scene of a soccer match. Therefore, in order to reduce the number of parameters of the neural network, improve the detection speed of soccer game video, and ensure the real-time and smoothness of detection, this paper streamlines the backbone network of YOLOv5 into a lightweight Mobile Net, while keeping the other parts of the YOLOv5 network unchanged. The new network structure is shown in Figure 2. It can be seen that Mobile Net is a lightweight network structure, Mobile Net's biggest feature is to use depth separable convolution instead of ordinary convolution kernel, which is suitable for mobile computing power limited scenarios, the use of CPU computational inference, the requirement of lightweight and fast deployment requirements, in order to ensure the detection of accuracy at the same time to accelerate the speed of inference. Common convolutional operations to extract features of an image usually need to process both spatial and channel features of the input image. Depending on the input features, the number of convolution kernels used needs to be adjusted, which tends to lead to a significant increase in the number of network model parameters. Depth-separable convolution consists of two parts: depth-by-depth convolution and point-by-point convolution.

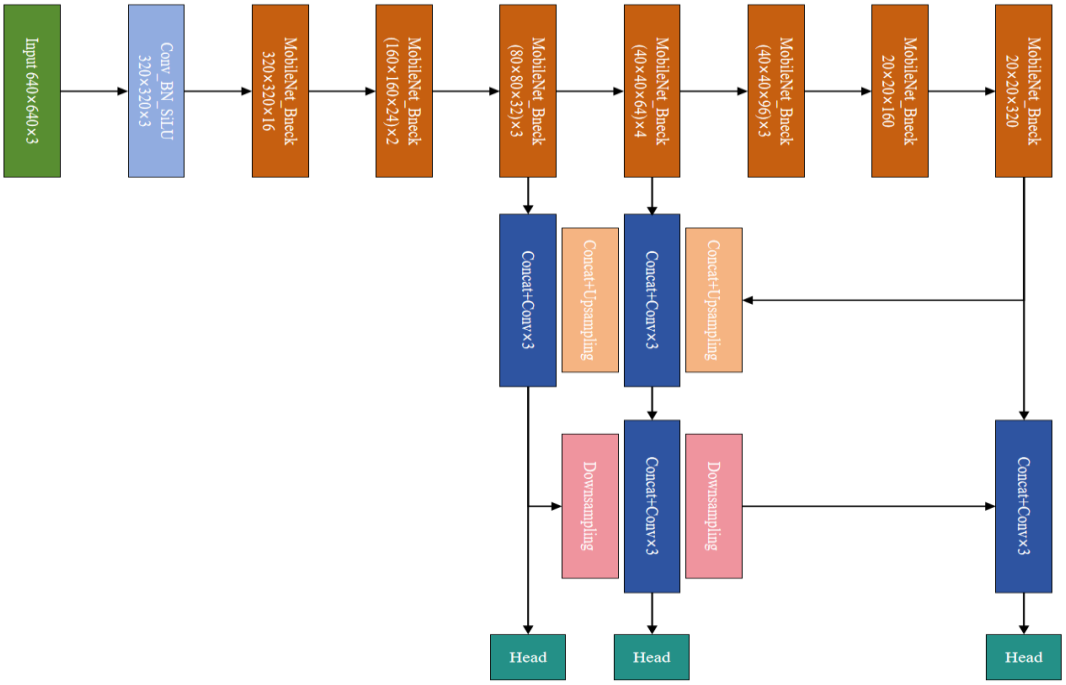

Figure 2: MobileNet-YOLOv5 Network Architecture

An ordinary convolution kernel accepts as input a feature map of size $D_F \times D_F \times M$, where D_F is the length and width of the feature map and M is the number of channels of the feature map, and outputs a feature map of size $D_F \times D_F \times N$, where D_F is the length and width of the feature map and NNNN is the number of channels of the feature map. Assuming that the size of the convolution kernel is D_K ,the amount of computation generated by one convolution is:

$$Compute_{cost} = D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \tag{2}$$

Depth separable convolution first convolves each channel of the input feature map with a convolution kernel of size $D_K \times D_K$, called depth-by-depth convolution. This is shown in Figure 3:

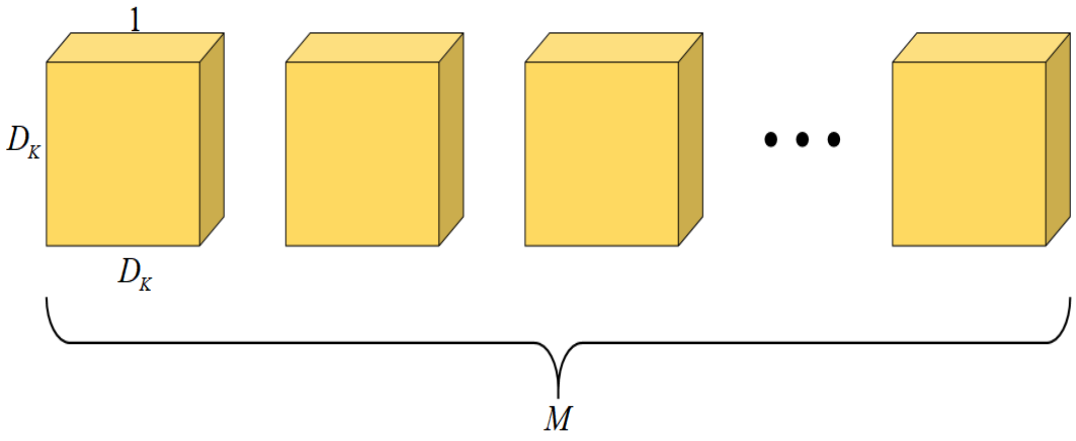

Figure 3: Channel-by-Channel Convolution Schematic

Then the convolution kernel of size 1×1 is used for point-by-point convolution to linearly combine the convolved feature maps together. As shown in Figure 4.

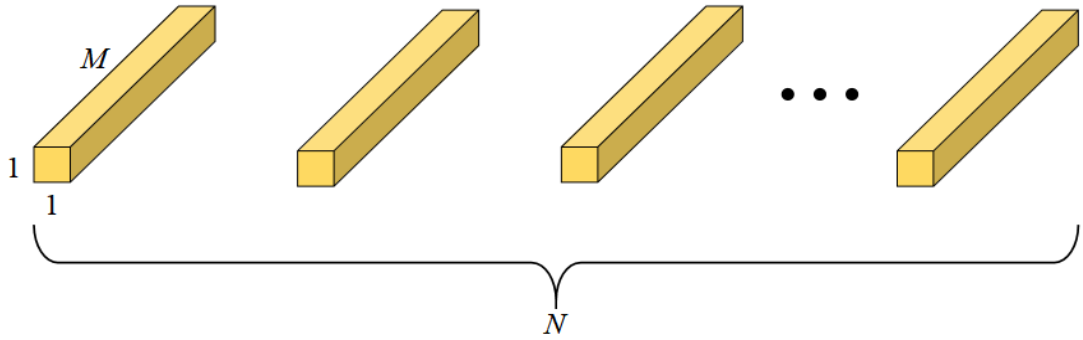

Figure 4: Schematic of point-by-point convolution

The computational effort required for deeply separable convolution is:

$$Compute_{cost_{dsc}} = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \quad (3)$$

Dividing Equation 3 by Equation 2 yields.

$$\frac{Compute_{cost_{dsc}}}{Compute_{cost}} = \frac{1}{N} + \frac{1}{K^2} \quad (4)$$

When the number of convolution kernels is large enough, the first half of the ratio, $\frac{1}{N}$, is negligible, so that the speed of deeply separable convolution is roughly proportional to the square of the convolution kernel size. For example, with a convolution kernel size of 3 × 3, depth-separable convolution is nine times faster than normal convolution. This and subsequent experimental results show that the detection speed of the improved YOLOv5 is greatly improved.

2.2.2 Integration of Attention Mechanisms

There are background complex interferences and non-detectable objects such as field-side billboards and spectators in soccer game videos, which affect the detector's effectiveness in target detection. We try to introduce the attention mechanism into the YOLOv5 network structure, which has the advantage of enhancing the detection network's attention to the target and ignoring irrelevant background factors, thus improving the detection accuracy of the target. To this end, this paper surveys the current state-of-the-art attention mechanisms, such as SE attention and CBAM attention, and combines the two attention mechanisms with the YOLOv5 network structure respectively. Through extensive experimental validation, it is demonstrated that the YOLOv5 network incorporating the attention mechanisms has improved detection

accuracy in soccer match video scenarios. SENet, whose full name is Squeeze-and-Excitation Network, belongs to the channel attention mechanism, which focuses on the correlation between the channel dimensions of an image, and investigates how much each channel affects the network. Convolutional neural networks use a large number of convolutional kernels to extract spatial information and channel information separately in the local receptive domain, and fuse the two to construct target feature information. The size of the convolutional kernel is usually 3×3 or 5×5, and as the convolutional layers are stacked, the receptive domain of the deep network becomes larger, and there are many multilevel dependency terms, and the edge portion of the image is not sufficiently feature-extracted, and the positional information of different distances is transmitted centrally in the network layer, which generates the long-distance dependency problem. SENet proposes a mechanism that allows the network to recalibrate the features, through which it learns to use global information to selectively emphasize informative features and suppress less useful ones, thus realizing the reconfiguration of the dimensionality of the convolutional feature channel. SENet consists of four main steps: F_{tr}, F_{sq}, F_{ex} and F_{scale}.

$$F_{tr}: X \to U, X \in R^{H' \times W' \times C'}, U \in R^{H \times W \times C} \tag{5}$$

F_{sq} is the Squeeze operation, which aims to compress global spatial information into a channel descriptor, essentially a channel-by-channel information statistic using global averaging pooling. in general, U are compressed in the $H \times W$ spatial dimension to obtain the statistic z. z is computed using the following formula:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=i} uc(i,j) \tag{6}$$

F_{ex} is the excitation operation, also called adaptive recalibration. After the compressed information is obtained by Squeeze operation, another channel-by-channel dependency extraction is performed to satisfy the nonlinear and non-mutually exclusive relationship between the feature channels. The formula is shown below:

$$s = F_{ex}(z,\ W) = \sigma\big(g(z,W)\big) = \sigma\big(W_2 \delta(W_1 z)\big) \tag{7}$$

where δ stands for ReLU activation function and σ for sigmoid activation function. F_{scale} uses the activation function to resize the converted output U. Each channel of U is multiplied by the corresponding weight to obtain the output $\tilde{x}_C$ of the SENet module. $\tilde{x}_C$ is calculated using the following formula:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{8}$$

There are two ways to add an attention module: one is to add it at the end of the backbone of the original network structure, before the SPPF, and the other is to replace all the $C3$ modules in the backbone with an attention mechanism module. CBAM is a hybrid domain attention mechanism that processes both the spatial and channel domains of an image to help different feature channels and spatial features with adaptive attention, which can enhance the feature representation capability of neural networks.

The CBAM attention mechanism operates in the spatial dimension more than the SE attention mechanism, and consists of two parts, the channel attention module and the spatial attention module, stacked by order. The combination of these two modules, embedded in the convolutional layer of the neural network, utilizes the ability of the convolutional network to extract features while emphasizing meaningful regions in both the spatial and channel dimensions, realizing the fusion of cross-channel and spatial information. In the channel attention module, attention is focused on what is relevant in the context of the visual input.

The spatial information of the input features is collected through average pooling and maximum pooling, which are represented by F_{avg}^{c} and F_{max}^{c}, respectively. These spatial features will then be passed through a multilayer perceptron with only one hidden layer to get the channel attention mapping. Finally, the output feature vector is obtained by summing the corresponding elements. The formula for the channel attention is:

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^{c}\right)\right) + W_1\left(W_0\left(F_{\max}^{c}\right)\right)\right) \end{aligned} \tag{9}$$

Where σ denotes the sigmoid activation function, and $W_0 \in R^{C/r \times C}$, $W_1 \in R^{C \times C/r}$, represent the weights of the multilayer perceptron model.

The spatial attention module focuses on the spatial distribution of useful information and complements the channel attention module. Maximum pooling and average pooling operations are performed on the channel dimensions and the outputs are spliced together to obtain spatial attention. The formula for spatial attention is:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}\left(\left[AvgPool(F); MaxPool(F)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F_{avg}^{s}; F_{\max}^{s}\right]\right)\right) \end{aligned} \tag{10}$$

where σ stands for sigmoid activation function, and $f^{7\times7}$ stands for convolution kernel of size 7×7.

3. Yolo Model Training

3.1 Preparation of The Dataset

The dataset selected for the experiments in this paper is the Soccer DB dataset collected and labeled by Cui et al (Jiang et al., 2020), in which the images are collected from 346 soccer matches with a total of 668.6 hours of soccer game videos. Among them, 270 matches were selected from the six major European leagues in the 2014 to 2017 seasons, and 76 matches were selected from the Chinese Super League (CSL) in the 2017 to 2018 seasons as well as some of the matches in the last three World Cups, which can be considered to be broadly representative. In order to increase the difficulty of the dataset, the authors first crawled 24475 images from the Internet covering a variety of soccer match scenes and used them as data to initially train a detector, which was used to perform target detection on the match footage, annotate the information of target labeling frames contained in the images, and then select the frames with the poorest detection effect as the final target detection dataset. The images used for the target detection part of the final dataset total 45732 images, each with a size of 1280×720, and a total of about 700,000 target labeled frames, which are classified into three categories, soccer ball, player and goal. The specific distribution is shown in Table 1:

Table 1: Distribution of labeled boxes in the data set

CATEGORY	QUANTITY
PLAYER	643581
FOOTBALL	45160
GOALMOUTH	13355
TOTAL	702096

The dataset is divided into training and test sets in the ratio of 8:2, and the dataset is labeled using the COCO dataset labeling format, which generates a text file with the same name for each image, in which the information about the target contained in that image is stored. Each row stores information about a target label box, divided into several columns, the first column indicates the category to which the target belongs, and the second to fifth columns indicate the coordinates of the center point of the label box and the length and width of the label box, respectively.

3.2 Experimental Setup and Parameters

The environment used for the experiments in this paper is: the operating system is Ubuntu 16.04, the processor is Intel® Core™ i7-9750H CPU @ 2.60GHz, and the graphics card is NVIDIA GeForce RTX 2080 Ti with 11GB of video memory. the deep learning framework used is PyTorch.

3.3 Evaluation Indicators

For the detection results of the model, they can be classified into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) based on the combination of the target's true category and predicted category. TP means that the IoU of the labeled frame is greater than 0.5, FP means that the IoU is less than 0.5 or the labeled frame is duplicated, and FN means that the labeled frame does not detect the target or the IoU is greater than 0.5 but the detection category is incorrect. TN means that the labeled frames are not detected at the location where there is no target and does not affect the target detection performance, so it will not be used. The confusion matrix for the detection result categorization is shown in Table 2:

Table 2: Confusion Matrix of Detection Results

REFERENCE	**PROJECTION**	
	Positive	Negative
POSITIVE	Ture Positive (TP)	False Negative (FN)
NEGATIVE	False Positive (FP)	Ture Negative (TN)

Precision is defined as:

$$precision = \frac{TP}{TP+FP} \tag{11}$$

Recall is defined as:

$$recall = \frac{TP}{TP+FN} \tag{12}$$

The commonly used criterion for evaluating the effectiveness of target detection is the average precision AP. AP, originally proposed by the PASCOL VOC target detection contest, refers to the area under the PR curve with precision as the horizontal coordinate and recall as the vertical coordinate. The specific algorithm for AP is shown in the following equation:

$$P = \frac{1}{11}\sum_{r\in[0,0.1,\cdots,1]} p_{int\,erp}(r) \tag{13}$$

Where，

$$p(r)\underset{\tilde{r}:\tilde{r}\geq r}{max}(\tilde{r})_{interp} \tag{14}$$

represents the precision for each check rate value, and the largest precision value on the right side of the point is taken as the precision value of the point. An image often contains multiple categories of targets. mAP (mean

Average Precision) is obtained by summing the mean accuracies of all categories of targets and then taking the mean. mAP is calculated by the formula:

$$mAP = \frac{\sum_{c=1}^{C} AveP(c)}{C} \quad (15)$$

where C represents the number of categories, c represents a specific category, and $Ave(c)$ represents the average accuracy of the category c.

4. Experimental Results and Analysis

4.1 Backbone Network Comparison

The model complexity and performance comparison between MobileNet-YOLOv5 designed in this paper and the original YOLOv5 after training on the dataset are shown in Table 3.

Table 3: Performance Comparison between YOLOv5 and Mobile-YOLOv5

MODEL	MAP@.5	FPS(GPU)	FPS(CPU)	NO. OF PARAMETERS	GFLOPS
YOLOV5	0.915	67.53	4.68	7.08M	16.3
OURS	0.879	78.89	7.24	3.57M	6.2

As can be seen from the above table, MobileNet-YOLOv5 reduces the mAP indicator by 2.6 percentage points compared to the original YOLOv5, indicating that the streamlined backbone network has reduced the ability to extract target features, resulting in a decrease in accuracy. In terms of the number of parameters and complexity of the model, the number of parameters of MobileNet-YOLOv5 is reduced by about half, and the amount of computation required is about 38% of the previous one, indicating that the streamlining of the backbone network effectively reduces the complexity of the model. Meanwhile, the detection speed of MobileNet-YOLOv5 is 17 percentage points higher than that of the original model (on GPU), and the detection speed on CPU is half as high. It shows that the model after streamlining the backbone network has a slight decrease in detection accuracy, but it reduces the number of parameters, decreases the model size, and improves the detection speed of the model, which is more in line with the requirements of target detection in the video scenes of soccer matches.

4.2 Comparison of detection performance for adding small targets

Since the size of the model is not categorized on the original dataset, in the actual scenario, the soccer ball can be regarded as a small target because the camera lens is far away from the pitch and the soccer ball itself is small. The detection accuracies of the two models for soccer ball targets are shown

in Table 4.

Table 4: Performance comparison before and after improving small target detection

MODEL	MAP@.5
YOLOV5	85.6
OURS	87.8

As can be seen from the above figure, after optimizing the neck part of YOLOv5, the model's detection accuracy for small targets is improved by nearly two percentage points due to the fusion of multi-scale features.

4.3 Comparison of Attention Modules

Table 5: Performance Comparison of Attention Models

SIZE	C3	SE	CBAM	MAP@ .5
SMALL	◎			0.868
		◎		0.879
			◎	0.896
	◎	◎		0.902
	◎		◎	0.879
MIDDLE	◎			0.913
		◎		0.921
			◎	0.932
	◎	◎		0.922
	◎		◎	0.916

The network model embedded with each attention mechanism is trained on the dataset and the performance of the model is recorded to obtain Table 5. where size represents the volume of the model, and s and m represent that the model is based on the structure of YOLOv5s, YOLOv5m, and YOLOv5l, respectively. c3, se, and cbam represent the individual modules of the network, respectively. The rows with only C3 checked represent the basic, unimproved YOLOv5 structure. Rows with only the Attention Mechanism module checked indicate that the C3 module is replaced by the checked Attention Mechanism module. Rows that check both the Attention Mechanism module and the C3 module indicate that the Attention Mechanism module is added before the SPPF module and does not replace the C3 module. From the above table, it can be seen that as the neural network model increases in size, the detection accuracy of the model also increases. At the same volume, embedding the attention mechanism brings about an increase in detection accuracy, with the network with the embedded attention mechanism being about two or three percentage points higher than the original YOLOv5 network, while noting that the embedded attention mechanism helps to compensate for the accuracy gap

brought about by the model size. However, there are advantages and disadvantages among the various attention mechanisms, and on the whole, the CBAM attention mechanism is more effective in improving the accuracy. In addition, from the point of view of the number of parameters, in the case of similar detection accuracy, since replacing the C3 module with the attention mechanism module reduces the number of parameters, the replacement embedding is preferred.

4.4 Comparative Experiments

In this paper, we use Faster R-CNN and Center Track network models, respectively, and use the same dataset as in literature (Komorowski et al., 2019) to train the two models, respectively, and by doing a comparison of the detection results, we get an evaluation of the detection performance of the models as shown in Table 6:

Table 6: Performance Comparison of Different Models

	ISSIA-CNR	SOCCER PLAYER	FPS
FASTER R-CNN	87.3	92.6	8
FOOT AND BALL	92.2	88.3	36
CENTER TRACK	90.2	90.3	38

From the table, we can see that the above-mentioned two literatures perform better than Faster R-CNN on the ISSIA-CNR dataset, in which the literature (Komorowski et al., 2019) outperforms Faster R-CNN by 4.7 percentage points, and the literature outperforms Faster R-CNN by 2.7 percentage points. However, on the Soccer Player dataset, the accuracies of the two instead decrease compared to the Faster R-CNN. Center Track performs about the same on both datasets, and it is speculated that it may have better generalization. The network structure used in this paper achieves a detection accuracy of up to 97.9% on a larger and more comprehensive dataset, which is five percentage points higher than the accuracy of literature (Komorowski et al., 2019) on ISSIA-CNR, suggesting that the method used in this paper has a better effect under the application of soccer game videos.

5. Simulation Applications

In order to demonstrate the detection effect of the model more intuitively, it is also necessary to do a qualitative analysis of the model's detection performance in specific scenarios. In a throw-in scenario there is both a distant player and a near soccer ball. At the same time, the near player is in the middle of a confrontation and the target has an obscured part. The original YOLOv5 detection result is shown on the left, and the improved YOLOv5 detection result is shown on the right, as shown in Figure 5.

OURS Original YOLOv5

Figure 5: Player Detection for Out-of-Bounds Throws

As can be seen from the figure, the original YOLOv5 detector does not do a good job of solving the phenomenon of missed and wrong detections in the presence of occlusion in the frame. The red player on the left is in front of the white player, obscuring most of the white player, and the original YOLOv5 detector does not detect the white player. Moreover, the original YOLOv5 detector misdirected the white player's exposed shoulder jersey as a soccer ball, and again duplicated the detection of the soccer ball target near the soccer ball's true location. The improved YOLOv5 detector, faced with the same test, accurately detected both the player and the soccer ball target, identified the white player who was blocked by the red player, and detected the correct location of the soccer ball without over-detection or under-detection.

6. Conclusion

This bit centers around the soccer serve action using computer vision technology to detect the target of the game video, in order to achieve the automatic evaluation and improvement of the serve action. In terms of target detection, this paper takes the YOLOv5 detection network model as the theoretical basis to carry out research. Aiming at the problems of insufficient detection real-time, target occlusion, uneven lighting conditions and small target size affecting detection accuracy in the multi-target detection task of soccer game video, the YOLOv5 network structure is optimized and improved. In this paper, we simplify the backbone structure of the YOLOv5 network by streamlining the original CSPDarkNet53 structure into a Mobile Net structure with depth-separable convolution, which reduces the number of parameters of the model and effectively improves the detection speed of the model. (2) For the problems of target occlusion and uneven illumination conditions, different attention mechanisms are embedded into the network model respectively, which improves the detection ability of the model for the target. For the problem of small target size, the problem of missed and wrong detection of small targets is effectively solved by reconstructing the multi-scale output layer and fusing multi-scale features.

REFERENCES

Chen, J., Han, P., Zhang, Y., You, T., & Zheng, P. (2023). Scheduling energy consumption-constrained workflows in heterogeneous multi-processor embedded systems. *Journal of Systems Architecture*, *142*, 102938.

Chen, J., Li, T., Zhang, Y., You, T., Lu, Y., Tiwari, P., & Kumar, N. (2023). Global-and-local attention-based reinforcement learning for cooperative behaviour control of multiple uavs. *IEEE Transactions on Vehicular Technology*.

Danelljan, M., Shahbaz Khan, F., Felsberg, M., & Van de Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. Proceedings of the IEEE conference on computer vision and pattern recognition,

Dhassi, Y., & Aarab, A. (2018). Visual tracking based on adaptive interacting multiple model particle filter by fusing multiples cues. *Multimedia Tools and Applications*, *77*, 26259-26292.

Dicle, C., Camps, O. I., & Sznaier, M. (2013). The way they move: Tracking multiple targets with similar appearance. Proceedings of the IEEE international conference on computer vision,

He, R., Fu, Z., Liu, Q., Wang, Y., & Chen, X. (2022). D^ 3: Duplicate Detection Decontaminator for Multi-Athlete Tracking in Sports Videos. Proceedings of the Asian Conference on Computer Vision,

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, *37*(3), 583-596.

Huang, Z., Zhang, P., Liu, R., & Li, D. (2023). An Improved YOLOv3-Based Method for Immature Apple Detection. *IECE Transactions on Internet of Things*, *1*(1), 9-14.

Hurault, S., Ballester, C., & Haro, G. (2020). Self-supervised small soccer player detection and tracking. Proceedings of the 3rd international workshop on multimedia content analysis in sports,

Jiang, Y., Cui, K., Chen, L., Wang, C., & Xu, C. (2020). Soccerdb: A large-scale database for comprehensive video understanding. In Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports. 1-8.

Komorowski, J., Kurzejamski, G., & Sarwas, G. (2019). Footandball: Integrated player and ball detector. *arXiv preprint arXiv:1912.05445*.

Li, Y., & Cao, J. (2021). WSN node optimal deployment algorithm based on adaptive binary particle swarm optimization. *ASP Transactions on Internet of Things*, *1*(1), 1-8.

Naik, B. T., & Hashmi, M. F. (2023). YOLOv3-SORT: detection and tracking player/ball in soccer sport. *Journal of Electronic Imaging*, *32*(1), 011003-011003.

Olagoke, A. S., Ibrahim, H., & Teoh, S. S. (2020). Literature survey on multi-camera system and its application. *IEEE Access*, *8*, 172892-172922.