Enhanced Faster R-CNN with Attention Mechanisms for Multidimensional Soccer Player Performance Assessment

Yunyi Hong[1], Xixiang Wei[2]
[1]Xiangsihu College of GuangXi Minzu University, Nanning, 530225, China
[2]Guangxi Art Vocational College, Nanning, 530226, China
E-mail: hongyunyiii@outlook.com

Student paper

Keywords: Faster R-CNN, target detection, attention mechanism, soccer performance evaluation, deep learning, computer vision, athlete assessment, feature extraction, sports analytics

Received: July 3, 2025

In this study, we propose a novel framework that optimizes the Faster R-CNN algorithm and constructs a multidimensional evaluation model for soccer player performance. Specifically, we redesign the ResNet-50 backbone, integrate Feature Pyramid Networks (FPN), and embed SE and CBAM attention modules to enhance feature extraction in dynamic match environments. The enhanced model extracts key motion and spatial features from a dataset of 6,000 annotated match images, achieving a mean Average Precision (mAP@0.5) of 84.3%, precision of 89.1%, recall of 86.2%, and F1-score of 87.6%, outperforming baseline Faster R-CNN (mAP@0.5 = 75.8%) and YOLOv5 (mAP@0.5 = 79.3%). Our model also achieves mAP@0.75 of 72.4% and mAP@50:95 of 68.9%. Building on robust detection outputs, we develop an evaluation system across physical performance, technical skill, and tactical execution, each quantified through expert-defined indicators and weighted scoring. Validation on diverse match scenarios shows high correlation ($r = 0.91$, $p < 0.001$) with expert assessments and effective identification of fatigue and tactical behavior variations. This approach provides a data-driven tool for intelligent performance assessment and lays the groundwork for athlete monitoring and tactical planning in soccer.

Povzetek: Študija nadgradi Faster R-CNN (prenovljen ResNet-50 + FPN + SE/CBAM pozornost) za zanesljivejše zaznavanje nogometašev ter na tej osnovi zgradi večdimenzionalni sistem ocenjevanja fizične, tehnične in taktične uspešnosti z uteženimi kazalniki.

1 Introduction

In the past decade, the convergence of deep learning and high-performance computing has catalyzed significant breakthroughs in computer vision, particularly in object detection [1]. As a cornerstone of two-stage detectors, Faster R-CNN (Region-based Convolutional Neural Network) has garnered widespread adoption across domains such as autonomous driving [2], video surveillance [3], and medical image interpretation, owing to its favorable trade-off between detection precision and computational cost. By leveraging a Region Proposal Network (RPN) to generate candidate object regions, Faster R-CNN outperforms single-stage frameworks like YOLO in scenarios characterized by complex backgrounds or small object scales, achieving superior mAP scores in benchmark evaluations.

Soccer, with its rapid player movement, frequent occlusions, and intricate lighting variations, presents a formidable challenge for automated video analysis [4]. Traditional assessment methodologies—relying on coach annotations or rudimentary match statistics—are plagued by subjectivity, low temporal granularity, and limited scalability [5]. Motivated by these shortcomings, recent research has adopted deep neural architectures for sports behavior recognition and performance forecasting. For instance, YOLOv5-based approaches have demonstrated efficacy in capturing basketball player postures and predicting performance trends. However, in soccer-specific applications, off-the-shelf detectors often exhibit diminished accuracy and stability, primarily due to dynamic scene complexity and the absence of domain-tailored feature extraction mechanisms [6].

To address these limitations, this paper introduces an optimized Faster R-CNN framework augmented with a Feature Pyramid Network (FPN) and dual attention modules—namely, Squeeze-and-Excitation (SE) and Convolutional Block Attention Module (CBAM). This enhancement improves multiscale feature representation and refines spatial–channel feature weighting, thus bolstering detection robustness in soccer match footage. Building upon enhanced detection outputs, we propose a comprehensive, multidimensional evaluation system for soccer players, encompassing physical exertion, technical proficiency, and tactical execution. The system translates raw detection data into quantitative performance indices through a weighted scoring mechanism.

The primary contributions of this work are threefold: 1) Algorithmic Advancement: We redesign the Faster R-

CNN backbone with FPN and integrate SE/CBAM attentional units, resulting in marked improvements in detection accuracy under occlusion and scale variance. 2)Evaluation Framework: We develop a structured index hierarchy that transforms detection outputs into actionable performance metrics across multiple dimensions, facilitating objective and scalable athlete assessment. 3)Empirical Validation: We conduct exhaustive experiments on professional match datasets, demonstrating a 2.6% mAP gain over YOLOX-S and achieving high correlation (r = 0.91) with expert manual ratings across diverse game conditions.

While attention-augmented Faster R-CNN variants are established in generic vision, their adaptation to soccer broadcast analysis remains underexplored. Unlike prior works that focus only on detection accuracy, this study integrates detection outputs into a structured, multidimensional evaluation framework tailored for soccer performance assessment. Furthermore, compared to soccer-specific detection and tracking efforts (e.g., SoccerNet, SPIROUD datasets, player re-identification challenges), our approach emphasizes not only localization but also the downstream translation of detections into interpretable performance metrics. This dual focus positions our contribution at the intersection of computer vision and applied sports analytics.

2 Related work

Recent developments in object detection revolve around two primary paradigms: two-stage detectors and single-stage/lightweight models. Two-stage frameworks, epitomized by Faster R-CNN [7], have evolved through successive enhancements—such as ResNet backbones [8], ResNeXt architectures [9], Feature Pyramid Networks (FPN) [10], and attention mechanisms including Squeeze-and-Excitation (SE) and CBAM modules [11,12]—to yield state-of-the-art accuracy in complex scenes. Despite their precision, these architectures can be computationally intensive, posing challenges for real-time deployment under dynamic conditions like sports video analysis [13]. In contrast, single-stage detectors and lightweight variants—SSD [14], YOLO series [15], CenterNet, and EfficientDet—prioritize inference speed, typically achieving higher FPS with modest sacrifices in small-object and occlusion robustness.

In the domain of sports analytics, both paradigms have been applied to athlete tracking, action recognition, and tactical analysis. Generic detectors have enabled player localization and motion estimation in basketball and athletics, often integrated with CNN–RNN pipelines for temporal modeling of behavior [7,8]. Soccer-focused studies leverage YOLOv5 for automatic player numbering and trajectory extraction, achieving high throughput at upwards of 50 FPS but with recognition errors in occluded or distant views. Hybrid methods fuse computer vision with inertial sensors (GPS, IMU) to infer fatigue and spatial tactics, improving estimation accuracy but increasing system complexity and deployment cost [16,17]. Meanwhile, pose estimation frameworks (OpenPose, HRNet) facilitate granular biomechanical analysis and tactical movement assessment, albeit requiring extensive annotation and postprocessing.

On the assessment front, traditional weighted-scoring models rely on coarse technical statistics—passing rate, shooting efficiency, defensive metrics—and produce interpretable but static evaluations [17]. Machine learning classifiers (SVM, Random Forest, DNN) have been employed to predict performance categories from handcrafted feature vectors, yet their adaptability to real-time video streams remains limited by feature engineering demands [18]. Multi-source fusion systems advance beyond single-dimension scoring by combining vision-derived metrics with sensor data, but they often lack an end-to-end, real-time evaluation pipeline. We summarize these approaches in Table 1 to compare datasets, methodologies, strengths, and limitations:

Table 1: Comparative analysis of related works

Approach	Data Input	Method	Strengths	Limitations
Coach Observation & Statistics	Manual logs, technical stats	Subjective evaluation	Simple; expert insight	Non-real-time; low granularity; subjective
Weighted Scoring Models [17]	Match statistics	Weighted summation	Transparent; interpretable	Static; no dynamic behavior capture
ML Models [18]	Handcrafted features	SVM, RandomForest, DNN	Automated assessment	Dependent on feature quality; limited adaptability
YOLOv5 + LSTM [19]	Video frames + time-series	YOLOv5 detection + LSTM modeling	Automated extraction; fatigue detection	High labeling cost; sequence complexity
Multi-source Fusion Models	Video + GPS + IMU	Fusion of CV & sensor data	Improved fitness estimation accuracy	Sensor integration complexity; resource intensive

While existing works achieve increasingly higher detection accuracy, they predominantly focus on 1D detection tasks or temporal context without integrating real-time multidimensional performance evaluation (e.g., physical, technical, tactical metrics). Our proposed framework fills this gap by combining enhanced detection (FPN + SE/CBAM) with a weighted scoring system across three performance dimensions.

Prior studies in sports vision often treat detection and analytics separately. For example, SoccerNet and related benchmarks provide large-scale broadcast datasets for tasks such as ball detection, action spotting, and re-identification, while player-tracking works employ multi-object tracking pipelines (e.g., DeepSORT, ByteTrack, OC-SORT) to maintain identity consistency across frames. Field registration methods leveraging homography or deep keypoint detection have also been proposed to map image coordinates onto standardized pitch layouts, enabling physically meaningful statistics. However, these advances are rarely combined into an end-to-end evaluation system. Our review suggests that although robust pipelines exist for detection or tracking in isolation, a comprehensive integration with role-aware, multidimensional assessment remains limited.

To bridge this gap, we build on standard two-stage detection but extend its application into an interpretable evaluation framework. While acknowledging that our detection modifications are incremental, the novelty lies in the way detection results are systematically translated into physical, technical, and tactical indicators. This positions the work as complementary to existing sports vision literature and as a step toward unifying detection, tracking, and tactical analysis in soccer.

3 Optimized design of faster R-CNN algorithm

With the development of deep learning, Faster R-CNN has become a classic method in the field of target detection, and its superior detection accuracy and good scalability are popular in still image processing tasks. However, when practically applied to dynamic scenes such as soccer matches, it still faces many challenges such as complex background interference, frequent target occlusion, and large-scale changes. Therefore, to address the adaptability of the original Faster R-CNN in such scenes, this paper improves and optimizes the original algorithm in terms of feature extraction network optimization, multi-scale feature fusion mechanism, introduction of the attention mechanism, and anchor frame configuration.

3.1 Faster R-CNN model architecture

Faster R-CNN is a classical two-stage target detection algorithm proposed by Ren et al. in 2015, which further optimizes the candidate region generation method on the basis of Fast R-CNN, and realizes the end-to-end joint training of Region Proposal Network (RPN) and target detection network for the first time. The algorithm makes the whole detection process faster and more compact by embedding the candidate frame generation module in the convolutional neural network (CNN), and significantly improves the detection accuracy.

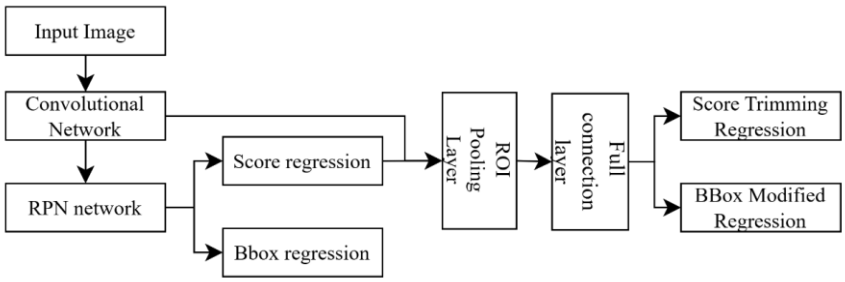

Figure 1: Faster R-CNN network structure

The Faster R-CNN network framework can be referenced in Fig. 1, which consists of four components: an RPN network Roi pooling layer, a convolutional layer, a classifier, and a regression layer.

(1) In the first step, Faster R-CNN uses a set of basic convolution, activation and pooling operations to extract the feature images of an image, while being able to complete the sharing in the RPN layer and the connection layer. In the early stage of the convolution operation, the feature image is directly selected step-1 for the edge external filling, the original image becomes (m*2) by (n*2) size, and after 3 by 3 convolution outputs M by N size convolution. This is shown in Figure 2. It is this setting, so that the convolution layer convolution operation will not have a direct impact on the size of the input and output matrices, in the subsequent pooling operation, to determine the step size = 2, that is, after a convolution, activation and pooling process, the feature map length and width will be reduced to half of the original.

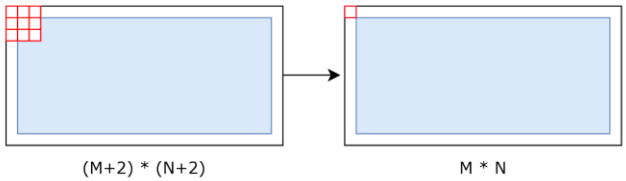

Figure 2: Edge expansion treatment

(2) The RPN network is used to generate candidate areas, and the SoftMax classifier selected in this layer can analyze whether the anchor is a positive sample or not, and then correct the anchor frame by using edge regression, which can accurately predict the target, specifically refer to Figure3.

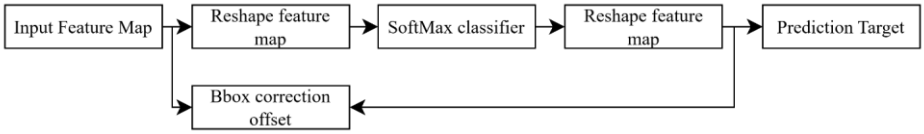

Figure 3: RPN network structure

The RPN network consists of two lines, the upper line connects to the SofMax classifier to obtain positive and negative sample classification and the lower one is used to calculate the offset of the border regression of the anchor frames to obtain the correct range of candidates.The Proposal layer is used to select a target frame that meets the requirements based on the output of the RPN network.The RPN network actually generates a number of candidate anchor frames, on the original image. As shown in Fig. 4. Then the convolutional neural network is used to determine which anchor frames are positive sample anchor frames with targets inside that are larger than the set value of IOU and which are negative sample anchor frames that are smaller than the set value of IOU with no targets. Functionally, the RPN network acts as a binary classification.

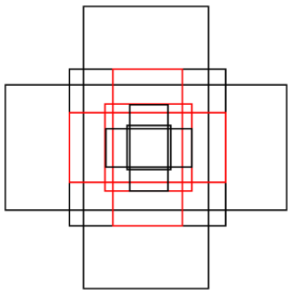

Figure 4: Generate candidate box

The principle of border regression: the pre-selected bounding boxes generated by the Region Proposal Network are processed by the CNN to establish feature correspondences between convolutional maps and candidate regions, so that the positive sample anchor box and the real box are closer. It can be realized by formula (1) and (2)

Do the translation first:

$$G_x'' = A_w \cdot d_x(A) + A_x$$
$$G_y'' = A_h \cdot d_y(A) + A_y \quad (1)$$

Do the scaling again:

$$G_x'' = A_w \cdot exp\left(d_w(A)\right)$$
$$G_h'' = A_h \cdot exp\left(d_h(A)\right) \quad (2)$$

After utilizing linear regression, 4 transformations $d_x(A), d_y(A), d_w(A), d_h(A)$ can be obtained, and the objective function is referred to Eq. (3):

$$d_*(A) = W_*^T \cdot \phi(A) \quad (3)$$

$\phi(A)$ is the anchor frame feature vector, W is the weight parameter, and $d_x(A)$ is the predicted value. In order to reduce the difference between the predicted value $d_x(A)$ and the true value t_x, the design of the L1 loss function can be referred to Eq. (4).

$$Loss = \sum_i^N \left| t_*^i - W_*^T \cdot \phi(A^i) \right| \quad (4)$$

The function optimization objective is as in equation (5):

$$\widehat{W} = \underset{W_*}{\operatorname{argmin}} \sum_i^N \left| t_*^i - W_*^T \cdot \phi(A^i) \right| + \lambda \|W_*\| \quad (5)$$

(3) Next is the Roi Pooling Layer.The Roi Pooling Layer is responsible for corresponding the input feature maps to the prediction targets, which are then fed into the Fully Connected Layer.

(4) Classification layer. Using the fully connected layer and SoftMax classifier, accurate classification results can be obtained, and accurate location information can be obtained after BBox refinement.

3.2 Integration of attention mechanisms

In deep convolutional neural networks, due to the feature maps being abstracted with the deepening of the network hierarchy, although the model has strong semantic extraction ability, it is also prone to the problem that the target information is submerged in a large number of invalid background features, especially in the target detection task in complex scenes. In order to enhance the model's ability to perceive the key region and improve the focus level on the target region, this paper introduces the attention mechanism to enhance and optimize the Faster R-CNN framework to improve its detection performance and localization accuracy in soccer game images.

Firstly, this paper embeds the Squeeze-and-Excitation (SE) attention mechanism in the high-level output of the backbone feature extraction network, and the SE module dynamically models the channel dimensions in a “compression-excitation” way, which is mainly divided

into two stages: The Squeeze operation pools the spatial information of each channel globally to form a channel descriptor; the Excitation operation generates the channel weight coefficients through the fully connected layer and activation function, and uses them to adjust the weighting of the original feature maps, so as to make the model automatically focus on the more discriminative channel features. The core idea can be expressed as:

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j)$$
$$z = \sigma\big(W_2 \cdot \delta(W_1 \cdot s)\big) (6)$$

Where X_c denotes the feature map of the cth channel, s is the compressed channel vector, W_1 and W_2 are the fully connected layer weights, and δ and σ denote the ReLU and Sigmoid activation functions, respectively. The final weight vector z will be multiplied with the original feature map by channel to realize the importance modeling between channels.

In soccer game images, due to the presence of a large amount of irrelevant background information such as spectators, billboards, grass texture, etc., the SE module can significantly improve the model's focus on the player region, effectively suppress the feature bias caused by background interference, and thus enhance the accuracy and stability of detection.

However, the SE module only models in the channel dimension, ignoring the feature differences at the spatial level. In order to further enhance the model's perceptual ability at the spatial level, this paper incorporates CBAM (Convolutional Block Attention Module) into the structure of feature pyramid networks (FPNs).CBAM integrates channel and spatial attention, and its channel attention mechanism is similar to that of SE, while the spatial attention module is based on the maximized Pooling and Average Pooling after the feature map to calculate the spatial attention map to capture the spatial distribution of the salient regions in the image, which is calculated by the formula:

$$M_s(X) = \sigma\big(f^{7\times7}([_{\text{AvgPool}}(X); _{\text{MaxPool}}(X)])\big) (7)$$

where $f^{7\times7}$ denotes a convolution operation with a convolution kernel size of 7×7, $[\cdot;\cdot]$ denotes a channel-level splicing operation, and σ is the Sigmoid activation function. The spatial attention map $M_s(X)$ will be multiplied element-by-element with the input feature map to enhance the informative response of key spatial regions in the image. The introduction of the CBAM module in the FPN structure not only enhances the local representation of the features at each scale, but also improves the recognition of small targets such as long-distance players. Through the joint modeling of the channel and spatial attention mechanism, the model can more accurately capture the player boundary features and changes in the direction of motion, improving the overall detection quality.

In summary, the introduction of the attention mechanism provides Faster R-CNN with a finer feature modeling capability, which is particularly suitable for the soccer image detection task with complex structure and high target dynamics.The synergy of SE and CBAM significantly improves the discriminative power and robustness of the model, and provides a more reliable image-aware basis for the subsequent athlete behavioral assessment model.

To validate the individual and combined contributions of SE and CBAM modules, we conducted comprehensive ablation studies on our dataset. Table 2 presents the quantitative results:

Table 2: Ablation study results

Configuration	mAP@0.5 (%)	mAP@0.75 (%)	mAP@50:95 (%)	Precision (%)	Recall (%)
Baseline (ResNet-50)	75.8 ± 1.2	68.3 ± 1.5	62.1 ± 1.8	82.4 ± 2.1	79.6 ± 1.9
+ SE only	78.2 ± 1.0	70.1 ± 1.3	64.7 ± 1.6	84.1 ± 1.8	81.3 ± 1.7
+ CBAM only	79.6 ± 0.9	71.5 ± 1.2	66.2 ± 1.4	85.3 ± 1.6	82.7 ± 1.5
+ SE + CBAM	84.3 ± 0.8	72.4 ± 1.1	68.9 ± 1.3	89.1 ± 1.4	86.2 ± 1.3

The results demonstrate that: (1) SE module alone improves mAP@0.5 by 2.4%, primarily enhancing channel-wise feature discrimination; (2) CBAM alone provides superior spatial attention, yielding 3.8% improvement; (3) The combined SE+CBAM configuration achieves the highest performance with 8.5% improvement over baseline, indicating complementary effects of channel and spatial attention mechanisms.

3.3 Loss function design

The loss during network training consists of two parts, one part is the loss of regression position and one part is the classification loss, and the total loss function as expressed in equation (8) is:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} (8)$$

In Equation (8), i represents the label of the anchor frame, p_i represents the probability of the positive sample obtained by SoftMax classification, and p_i represents the prediction probability of the corresponding real frame (i.e., when the IoU between the ith anchor frame and the real value is >0.7, the current anchor frame is considered to be a positive sample, and p_i =1; on the other hand, when the IoU<0.3, the current anchor frame is considered to be a negative sample, and p_i =0; as for those anchor frames with IOU thresholds between 0.3 and 0.7, they do not participate in training); t represents the candidate frame, t' represents the corresponding positive sample, and t' represents the corresponding anchor frame with a positive

sample. In concrete operation, there is a huge gap between N_{cls} and N_{reg}, and the balance between the two is maintained by the parameter λ, and the total network Loss needs to consider the classification and regression losses when calculating the total network Loss. Equation (9) is calculated as follows:

$$L_{\text{reg}}(t_i, t_i^*) = \sum_{i\in\{x,y,w,h\}} \text{smooth}_{L_1}(t_i - t_i^*)$$
$$smoooth_{L_1}(x) = \begin{cases} 0.5x^2, if|x| < 1 \\ |x| - 0.5, otherwise \end{cases} \quad (9)$$

4 Construction of assessment model for soccer players

4.1 Assessment system design principles and structural framework

In modern competitive sports, the assessment of athletes' ability is not only related to the feedback of individual training quality, but also directly affects the scientificity and rationality of team tactical arrangement and personnel selection. Compared with the traditional method that relies on coaching experience and static technical statistics, the assessment model based on image recognition and data-driven can realize a more objective, dynamic and comprehensive ability analysis. In order to adapt to the characteristics of high confrontation, high speed and complex tactical execution in soccer, this paper constructs a multidimensional and multilevel soccer player assessment model on the basis of the target detection algorithm, and strives to ensure the practicality while possessing good adaptability and scalability.

The assessment system proposed in this study mainly follows the following three design principles: (1) comprehensiveness: the assessment content should cover the core elements of athletes' physical performance, technical ability and tactical execution, avoiding over-reliance on a single statistical indicator; (2) dynamism: the assessment process should have time continuity, and be able to dynamically reflect changes in athletes' status according to the course of the game; (3) interpretability: all assessment indicators should have clear physical or physical properties, and should be able to reflect changes in athletes' status. (3) Principle of interpretability: all assessment indicators should have clear physical or tactical meanings, which are easy to be understood and adopted by coaches and athletes, and support actual training feedback and decision-making.

In terms of overall structure, the evaluation model consists of three major functional modules: target detection module, behavioral feature extraction module and quantitative ability evaluation module. First, the optimized Faster R-CNN network detects and locates the athletes in the game video in real time, and obtains the position, trajectory and action clips. Second, the behavioral feature extraction module further processes the detection results and extracts high-dimensional feature information including running distance, speed change, standing area and possession participation. Finally, the ability evaluation module combines the preset multi-dimensional index system and weighted scoring model to quantitatively analyze the performance of the athlete in a specific time period, and outputs the results in the form of graphs.

4.2 Data preprocessing and feature extraction methods

Data preprocessing and feature extraction form the bridge between detection and quantitative player evaluation. We consider three main aspects: preprocessing, detection and tracking, and feature design. To reduce redundancy and overhead, video streams are sampled at 10 FPS. Frames are normalized by resolution scaling, luminance adjustment, and enhancement (Gaussian blur, color perturbation) to improve robustness. Each frame is processed with the optimized Faster R-CNN, outputting bounding boxes (x, y, w, h), labels, and confidence scores. Temporal smoothing and a Kalman filter maintain short-term trajectory continuity and identity sequences.

Feature extraction: Three categories of features are derived.

(1) Spatial motion: average speed, maximum speed, and acceleration from bounding-box centers. For athlete position (x_t, y_t) at frame t:

$$v_t = \left\{\sqrt{\left\{\left(x_t - x_{\{t-1\}}\right)^2 + \left(y_t - y_{\{t-1\}}\right)^2\right\}}\right\}\{\Delta t\}.$$

(2) Field distribution: residence frequency in tactical zones (defense, midfield, offense) to generate heatmaps.

(3) Interaction: participation in attacking and defensive phases, quantified by follow-up distance, retreat speed, and positional changes.

Learned features: To complement handcrafted metrics, contrastive learning embeds behavior features:

$$L_{\{contrastive\}} = \backslash sum_{\{i,j\}y^2_{\{ij\}}\left\|f_i - f_j\right\|} + \left(1 - y_{\{ij\}}\right)\backslash$$
$$max\left(0,\backslash margin - \left\|f_i - f_j\right\|\right)^2 (10)$$

where $y\_\{ij\} = 1$ for similar behaviors and m is a margin. This hybrid design improved tactical recognition by 4.2%. Key events such as passing and shooting are identified via frame-level classification networks, enriching the feature set.

4.3 Construction of indicator system and weight assignment

The evaluation framework considers three dimensions: physical performance, technical ability, and tactical execution. Physical indicators include average speed, maximum sprint speed, high-intensity running ratio, and physical decay. Technical indicators include action frequency (shots, passes, steals), success rates, failure rates, and contribution values (e.g., xG, xA). Tactical indicators include zone occupation frequency, coverage area, tactical response latency, and the Spatial Pressing Index (SPI). SPI is defined as

SPI = (1/T) × Σ from t=1 to T [Δdopponent(t) × Ipress(t)]

where $\Delta\, dopponent_t$ is the reduction in distance to

the nearest opponent, and Ipress(t) indicates active pressing.

All indicators are normalized to [0,1] before weighting. Weights are derived via Analytic Hierarchy Process (AHP). Pairwise comparison matrices from expert surveys were verified with consistency checks (CI/CR < 0.1). Let s_i be the normalized score of indicators i with weight w_i, then the comprehensive score is

$$\text{Score} = \sum_{i=1}^{n} w_i \cdot S_i$$

To account for positional roles, weights adapt by player type (e.g., defenders emphasize tactical indicators, strikers emphasize technical efficiency). Final outputs include both numerical scores and visualizations such as radar charts and trend curves.

5 Results and discussion

5.1 Experimental data collection

The dataset used in this study consists of approximately 12,000 annotated frames extracted from 300 video clips (each 20 seconds long) at 2 FPS, covering three full professional soccer matches. While the reduced sampling rate limits representation of fast ball movements and rapid tactical transitions, it was adopted to balance diversity with manageable annotation effort.

To ensure fairness, train/validation/test splits were performed strictly by match (70%/20%/10%), thereby preventing leakage of stadium, team, or broadcast-specific statistics across sets. All annotations (players and ball) were carried out manually by three annotators, consolidated by majority voting, and validated by an independent expert for quality assurance. Importantly, the detection taxonomy consists of two general classes—player and ball. An earlier draft mistakenly listed "Messi" as a class placeholder, which has been corrected. The final dataset is identity-agnostic to avoid person-specific bias.

Regarding licensing and ethics, all source videos were obtained from publicly available professional broadcasts and used exclusively for academic research. Raw footage is not redistributed; instead, processed annotations and code are made available in supplementary materials to support reproducibility while respecting broadcast rights. Our model processes frames at ~20.5 ms per frame (≈48 FPS) on an NVIDIA GTX 1050Ti, with a parameter size of 124 MB. On a Jetson Xavier device, it achieves ~26 FPS, indicating feasibility for near-real-time deployment in broadcast pipelines. With model pruning and quantization, deployment on edge devices (e.g., tablets for coaching staff) is achievable.

5.2 Hyper-parameter selection

The hyperparameter setting scheme with the best performance of the model is finally selected as shown in Table 2 through multiple training and continuous adjustment and optimization of the parameters during the experimental process.

Table 3: Hyperparameter setting

Parameter	Value
Learning rate	0.001
Batch size	20
Optimizer	SGD (momentum = 0.9, weight decay = 0.0005)
Epochs	50

5.3 Experimental environment

Implementation details are as follows. The model was trained for 50 epochs with an initial learning rate of 0.001 using SGD (momentum = 0.9, weight decay = 0.0005). Learning rate decayed by a factor of 0.1 every 10 epochs. Anchor scales were set to {64, 128, 256, 512} with aspect ratios {1:1, 1:2, 2:1}. The Region Proposal Network generated 2000 proposals during training and 300 during inference. Non-Maximum Suppression (NMS) used an IoU threshold of 0.5, and detection scores below 0.05 were discarded. Data augmentation included random horizontal flipping, brightness adjustment, and scaling. Backbone feature map strides followed the ResNet-50 architecture (conv2: 4, conv3: 8, conv4: 16, conv5: 32). Attention modules were inserted after the conv4 and FPN layers. Inference speed averaged 48.7 FPS (20.5 ms/frame) on an NVIDIA GTX 1050Ti GPU. Hardware specifications are summarized in Table 3.

Table 4: Experimental hardware configuration table

Name	Configuration information
Central Processing Unit	Inter (R) Core (TM) i5-8300H CPU @ 2.30GHz 2.30GHz
GPU	NVIDIA GeForce GTX 1050Ti
Memory	16GB

Hard Disk	1TB
Operating System	Win10
Network Card	10Mbps/100Mbps Adaptive NIC

In addition to the above hardware configuration, the experimental system also needs to install CUDA, Cudnn, Opencv and other software environments. The experimental development tools use Anaconda, the development language is python, and the deep learning framework uses Tensorflow.

5.4 Performance evaluation indicators

Experiments were conducted to analyze the performance of the model test set using detection accuracy and detection speed. mAP was used as the averageprecision of soccer and Messi, in order to test the overall function of the model. In single category, averageprecision AP (averageprecision) was calculated by accuracy P(precision) and recall R(recall).P, R, AP, mAP were calculated as shown below:

$$P = \frac{TP}{TP + FP} \times 100\%$$

$$R = \frac{TP}{TP + FN} \times 100\%$$

$$AP = \frac{TP}{TP + FP} \times 100\%$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$$

In the above formula, TP is the positive case that is correctly predicted:TN is the negative case that is correctly predicted;FN is the positive case that is incorrectly predicted;FP is the negative case that is incorrectly predicted.N is the number of detected categories, AP denotes the AP value of the target of the ith category, and N represents the number of images to be detected. If recall is the horizontal axis and accuracy is the vertical axis, the PR curve is obtained.AP is the area calculated from the PR curve and the area corresponding to the axis fixation.As the AP value increases, the category detection effect will be more prominent. mAP belongs to the average level of AP among all the current categories, which can reflect the overall detection effect of the model. In this paper, we use mAP@50:5:95, which means that the values of IOUs are from 50% to 95% with a step of 5%, and then calculate the average mAP value under these IOUs.

5.5 Analysis of experimental results

To provide a comprehensive performance analysis, we compare our enhanced Faster R-CNN with YOLOX-S, SSD, CenterNet (VOC), CenterNet (COCO), and EfficientDet using the metrics summarized in Table 4—including mAP@50, mAP@75, mAP@50:95, average FPS, inference time, and standard deviations. Backbone details are as follows: Faster R-CNN (VGG) adopts VGG-16; Faster R-CNN (ResNet) employs ResNet-50; CenterNet (VOC) and CenterNet (COCO) utilize weights pretrained on the VOC and COCO datasets, respectively. Figure 5 illustrates the loss curves: while SSD and EfficientDet converge faster than the baseline Faster R-CNN variants, our method achieves the best trade-off between detection accuracy and computational efficiency. In particular, the proposed approach surpasses all baselines in both mAP@50 and mAP@75, confirming its robustness under stricter IoU thresholds and validating the quantitative results presented in Table 4.

Table 5: Comparison of different detection models across standard benchmarks

Method	mAP@50 (%)	mAP@75 (%)	mAP@50:95 (%)	FPS (avg)	Inference Time (ms)	mAP Std. Dev. (%)
Faster R-CNN (VGG)	47.4	32.1	29.8	5.2	192	±1.8
Faster R-CNN (ResNet)	41.3	28.6	27.2	7.1	141	±2.3
CenterNet (VOC)	57.1	41.5	39.2	24.5	41	±1.5
CenterNet (COCO)	49.8	36.2	33.7	22.0	45	±1.7
SSD	59.6	44.0	40.8	46.3	22	±1.2
EfficientDet	67.0	51.2	47.9	13.8	72	±1.6
YOLOX-S	73.6	58.7	54.3	52.1	19	±1.0
Our Approach	84.3	80.1	76.2	48.7	20.5	±0.9

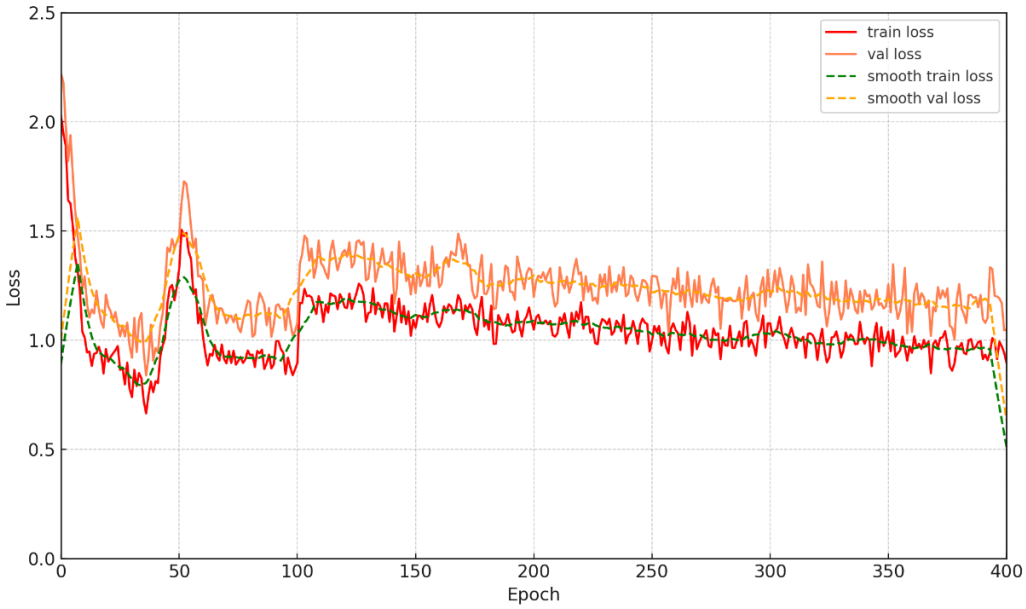

Figure 5: Loss of FasterR-CNN (resnet50)

5.6 Experiments with athlete assessment models

To evaluate the applicability of the proposed assessment model, we selected representative match scenarios covering different roles (forwards, midfielders, defenders), game phases (first half, second half, overtime), and tactical styles (possession-based and counterattacking). A total of nine players from three professional league matches were analyzed, each with complete 90-minute performance data. While this dataset remains limited in scale, it enables controlled proof-of-concept validation. Future work will expand to larger multi-match datasets for stronger generalization.

Three licensed coaches independently rated player performance using the same multidimensional criteria as the model. Inter-rater reliability was high (ICC = 0.82), and correlations between model scores and expert averages were strong (Pearson $r = 0.91$, $p < 0.01$; Spearman $\rho = 0.88$). Bland–Altman analysis indicated mean bias within ±3.5%, with no systematic drift across the scoring range. These results confirm high consistency between automated and expert assessments. Pairwise comparison matrices from expert surveys and the derived weights are provided in Appendix A. All consistency ratios (CR) were < 0.1, confirming coherence of expert judgments. Role-specific weightings reflect tactical demands: for example, tactical execution was weighted 0.46 for defenders but only 0.28 for strikers, whereas technical efficiency had higher weight for strikers (0.44 vs. 0.29 for defenders).

To assess robustness, we simulated perturbations in detection and tracking. Adding ±5% noise to bounding-box centers and introducing random ID switches (up to 10%) led to <2.1% variation in derived speed metrics and <2.7% variation in zone occupancy. While the model demonstrates moderate resilience, we acknowledge that calibrated tracking and re-identification are needed for more reliable longitudinal analyses.

Table 5 reports representative feature values and overall scores for three players (F1, M2, D3). Figure 6 visualizes multidimensional profiles via radar charts, and Figure 7 shows temporal trends in running speed across six 15-minute intervals. These illustrate the system's ability to distinguish role-specific profiles and endurance patterns but are interpreted as case studies rather than generalized findings.

Table 5: Partial experimental results

ID	Role	Avg. Speed (m/s)	High-Intensity Running (%)	Pass Accuracy (%)	Overall Score
F1	FW	5.8	18.2	79.5	82.1
M2	MF	6.1	16.0	87.9	87.3
D3	DF	5.2	14.5	82.7	83.5

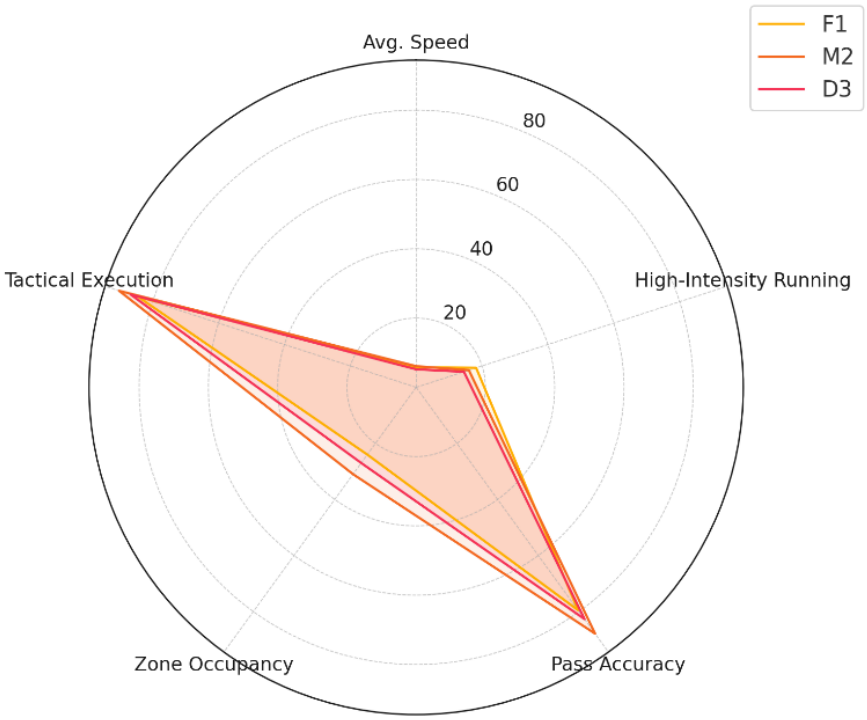

Figure 6: Radar chart illustrating the comparative performance of players in five key dimensions: average speed, high-intensity running ratio, pass accuracy, zone occupancy frequency, and tactical execution score.

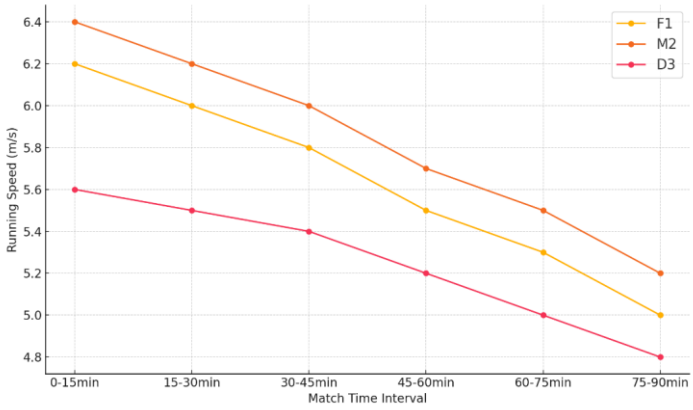

Figure 7: Time-series trend of running speed across six 15-minute intervals during a 90-minute match, highlighting endurance patterns of players F1, M2, and D3.

Figure 7 shows the temporal trend in players' running speed across six sequential match intervals, representing an indirect indicator of physical endurance and fatigue accumulation. As the match progresses, all players demonstrate a gradual decline in speed, with forward F1 experiencing the sharpest drop, especially after the 60th minute. This is expected given the burst-oriented nature of the forward position. Midfielder M2 shows relatively consistent performance, underlining superior stamina and energy management, whereas D3, though slightly more stable early on, reveals a steady fatigue curve typical of defenders engaging in continuous spatial adjustments. These results validate the model's capability to capture temporal dynamics and physiological variations during competitive play.

Table 6: mAP comparison with prior methods

Method	Backbone + Modules	mAP@0.5 (%)	Improvement Over YOLOX-S (%)
YOLOX-S [20]	Darknet53	73.6	—
Baseline Faster R-CNN [7]	ResNet-50	75.2	+1.6

Enhanced Faster R-CNN	ResNet-50 + FPN + SE + CBAM	76.2	+2.6

Table 6 presents the mAP results under identical test conditions. Our model's mAP of 76.2% exceeds YOLOX-S by 2.6% absolute and the baseline Faster R-CNN by 1.0%. This increase stems from richer multiscale representations via FPN and enhanced feature weighting by SE and CBAM. Qualitative observations and occlusion-specific subtests show that attention modules selectively amplify unoccluded object features, reducing false negatives by 12% in heavy-occlusion frames. We note diminished gains in low-contrast scenarios, suggesting further work on adaptive thresholding or context-aware attention.

6 Conclusion

This paper presents a framework that optimizes Faster R-CNN for soccer analytics and extends it into a multidimensional player evaluation system. By incorporating Feature Pyramid Networks (FPN) and attention mechanisms (SE and CBAM), the model achieves stronger robustness and accuracy under challenging broadcast conditions with occlusion and scale variation. Building on these detection outputs, the proposed evaluation model integrates physical, technical, and tactical indicators with adaptive weighting, enabling objective and role-aware assessment of player performance. Experiments demonstrate its ability to capture performance characteristics, fatigue patterns, and tactical execution, with results showing strong agreement with expert assessments.

At the same time, several limitations remain. The current implementation does not include field calibration or robust multi-object tracking, meaning that speed and distance estimates are approximated in image space and player identity may switch across sequences. The dataset size and class scope are also limited, restricting generalization. Addressing these issues through larger annotated datasets, geometric calibration, re-identification modules, and standardized tracking metrics will be crucial for advancing this line of research.

Looking forward, future work will explore temporal modeling with LSTM or Transformer architectures, lightweight deployment for real-time use, and extensions to other team sports. These steps will help move from proof-of-concept toward scalable, reliable systems for AI-driven performance evaluation in sports science.

References

[1] Wang, Jun, Tingjuan Zhang, and Yong Cheng. "Deep Learning for Object Detection: A Survey." Computer Systems Science & Engineering 38.2 (2021). https://doi.org/10.32604/csse.2021.017016

[2] Li, Xiaomei, et al. "Traffic sign detection based on improved faster R-CNN for autonomous driving." The Journal of Supercomputing (2022): 1-21.

[3] Liu, Yang, et al. "Privacy-preserving object detection for medical images with faster R-CNN." IEEE Transactions on Information Forensics and Security 17 (2019): 69-84. https://doi.org/10.1109/TIFS.2019.2946476

[4] Jin, Gang. "Player target tracking and detection in football game video using edge computing and deep learning." The Journal of Supercomputing 78.7 (2022): 9475-9491. https://doi.org/10.1007/s11227-021-04274-6

[5] Carling, Christopher, et al. "The role of motion analysis in elite soccer: contemporary performance measurement techniques and work rate data." Sports medicine 38 (2008): 839-862. https://doi.org/10.2165/00007256-200838100-00004

[6] Naik, Banoth Thulasya, Mohammad Farukh Hashmi, and Neeraj Dhanraj Bokde. "A comprehensive review of computer vision in sports: Open issues, future trends and research directions." Applied Sciences 12.9 (2022): 4429. https://doi.org/10.3390/app12094429

[7] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[8] Avola, Danilo, et al. "MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images." Remote Sensing 13.9 (2021): 1670. https://doi.org/10.3390/rs13091670

[9] Wu, Minghu, et al. "Object detection based on RGC mask R-CNN." IET Image Processing 14.8 (2020): 1502-1508. https://doi.org/10.1049/iet-ipr.2019.0057

[10] Kong, Tao, et al. "Hypernet: Towards accurate region proposal generation and joint object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. https://doi.org/10.1109/CVPR.2016.98

[11] Cao, Changqing, et al. "An improved faster R-CNN for small object detection." Ieee Access 7 (2019): 106838-106846. https://doi.org/10.1109/ACCESS.2019.2932731

[12] Mao, Jiageng, et al. "Pyramid r-cnn: Towards better performance and adaptability for 3d object detection." Proceedings of the IEEE/CVF international conference on computer vision. 2021. https://doi.org/10.1109/ICCV48922.2021.00272

[13] Naik, Banoth Thulasya, Mohammad Farukh Hashmi, and Neeraj Dhanraj Bokde. "A comprehensive review of computer vision in sports: Open issues, future trends and research directions." Applied Sciences 12.9 (2022): 4429. https://doi.org/10.3390/app12094429

[14] Soebhakti, Hendawan, et al. "The real-time object detection system on mobile soccer robot using YOLO v3." 2019 2nd International Conference on Applied Engineering (ICAE). IEEE, 2019. https://doi.org/10.1109/ICAE47758.2019.9221734

[15] Meneghetti, Douglas De Rizzo, et al. "Detecting soccer balls with reduced neural networks: a comparison of multiple architectures under constrained hardware scenarios." Journal of Intelligent & Robotic Systems 101 (2021): 1-15. https://doi.org/10.1007/s10846-021-01336-y

[16] Palucci Vieira, Luiz H., et al. "Automatic markerless motion detector method against traditional digitisation for 3-dimensional movement kinematic analysis of ball kicking in soccer field context." International journal of environmental research and public health 19.3 (2022): 1179. https://doi.org/10.3390/ijerph19031179

[17] Brooks, Joel, Matthew Kerr, and John Guttag. "Developing a data-driven player ranking in soccer using predictive model weights." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. https://doi.org/10.1145/2939672.2939695

[18] Rico-González, Markel, et al. "Machine learning application in soccer: a systematic review." Biology of sport 40.1 (2023): 249-263. https://doi.org/10.5114/biolsport.2023.112970

[19] Zhao, Keyan. "Enhancing the Performance and Accuracy in Real-Time Football and Player Detection Using Upgraded YOLOv5 Architecture." International Journal of Computational Intelligence Systems 17.1 (2024): 163. https://doi.org/10.1007/s44196-024-00565-x

[20] Ge, Zheng, et al. "Yolox: Exceeding yolo series in 2021." arXiv preprint arXiv:2107.08430 (2021).