Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 2, 2026, 145-153.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Improving Ball Detection in Volleyball Using Deep Learning

Mohammad Jadidi Miandashti, Kourosh Kiani, and Razieh Rastgoo*

Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 23 February 2025
Revised 06 May 2025
Accepted 15 August 2025

DOI:10.22044/jadm.2025.15807.2695

Keywords:
Volleyball Detection, Deep Learning, Convolutional Neural Network, Attention Mechanism

**Corresponding author: rrastgoo@semnan.ac.ir (R. Rastgoo).*

Abstract

In recent years, the application of deep learning techniques has revolutionized various domains, including the realm of sports analytics. The analysis of ball tracking and trajectory in sports has become an increasingly vital area of research, driven by advancements in technology and the growing demand for data-driven insights in athletic performance. In volleyball, a sport characterized by rapid movements and strategic play, the ability to accurately track the trajectory of the ball is crucial for both training and competitive analysis. This paper proposes novel deep learning models for accurate volleyball ball detection and tracking. By incorporating attention mechanisms into the YOLOv8 and YOLOv10 architecture, our models significantly improve performance, particularly in challenging situations involving occlusions and fast movements. The proposed models were compared to baseline and other models across several metrics. Specifically, they achieved precision (94.2% and 94.7%, respectively) and recall (88.1% and 87.6%, respectively) as well as real-time processing speeds, making them suitable for various sports analytics applications.

1. Introduction

Object tracking, a fundamental task in computer vision, finds critical applications in diverse domains including surveillance systems [1], robotics [2], and human-computer interaction [3-15]. Notably, sports video analysis presents a significant demand for robust object tracking methodologies, driven by advancements in Artificial Intelligence, computer vision, and deep learning techniques [16-21] and the proliferation of high-performance cameras [22]. The majority of viewers of ball-based sports focus on the ball's position and movement [23].

Ball tracking, a specialized case of single-object tracking and small object detection, is a critical task in computer vision [24]. Given the rapid and unpredictable movements of balls in sports, these systems must be capable of high-speed, real-time tracking [25]. The challenges associated with tracking small objects, such as limited pixel information and potential occlusions, make this task particularly demanding. Moreover, the dynamic nature of sports environments, with varying lighting conditions and background clutter, further complicates the tracking process [23].

Sports video analysis has gained significant traction due to its numerous applications and potential commercial benefits [25]. Ball tracking is a critical component of most sports analysis systems, as it enables detailed analysis of game play [24] Systems like Hawk-Eye have revolutionized sports by providing precise 2D and 3D trajectories of balls, aiding referees and enhancing the overall viewing experience [24]. However, real-time ball tracking in dynamic sports environments presents significant challenges, such as varying lighting conditions, occlusions, and high-speed motion [26]. Advanced computer vision techniques, including deep learning, are increasingly being employed to address these challenges.

The challenges and limitations in ball sports are primarily associated with the characteristics of the

ball itself. These characteristics, including speed, shape, and size, significantly influence the choice of algorithms and approaches. Additionally, player occlusion presents a common challenge in model and algorithm design. In volleyball, numerous challenges exist, such as occlusions, misdetections, varying lighting conditions, similar colors between the ball and the environment, camera quality, varying frame rates, and the fast pace of the game [26]. These challenges must be addressed through novel approaches. Figure 1 illustrates the ball in various states, including simple, occluded, motion blur, and blended with the background. This paper aims to address these challenges by proposing a novel approach for ball tracking, leveraging advanced deep learning techniques to achieve robust and accurate performance. Our main contributions can be listed as follows:

1. We propose two novel volleyball detection models by integrating an attention mechanism into the YOLOv8 and YOLOv10 architectures, leading to improved accuracy in detecting volleyballs across various object scales.
2. By applying Convolutional Block Attention Module (CBAM), we can effectively extract spatial and channel-wise information which is crucial for small object detection.

The remainder of this paper is organized as follows. Section 2 reviews recent relevant works. Section 3 presents the proposed model, followed by an evaluation of its performance on a real-world dataset, comparing the results with those obtained from alternative approaches. Finally, Section 4 concludes the paper and outlines directions for future research.

Figure 1. Ball in different modes such as simple mode, occlusion, motion blur, and cluttered with background.

2. Related Works

In recent years, there has been a surge in research and development of artificial intelligence applications within the sports domain [27]. Various approaches, including conventional machine learning and deep learning, have been employed to detect and track balls in sports videos. While conventional machine learning models have shown promising results, deep learning techniques have emerged as a powerful tool, offering significant improvements in accuracy and robustness [26].

Previous studies have explored a range of techniques, such as color-based tracking [28], Template Matching [25,29], Mean-Shift algorithm [30], Hough transform [31], Background Estimation [32] to localize and track balls. However, these methods often struggle in challenging scenarios, such as occlusions, varying lighting conditions, and rapid ball motion [33]. To address these limitations, researchers have increasingly turned to deep learning, leveraging Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to extract relevant features and predict ball trajectories [33]. More details of recent works in these two categories (conventional machine learning and deep learning) will be presented in the following sub-sections.

2.1. Conventional machine learning methods

The algorithm proposed by Xinguo Yu et al. [28] includes four steps for ball tracking in soccer video. The first step is to estimate ball size then detect ball candidates. After that the algorithm generates candidate trajectories and processes ball trajectory. Finally, while the proposed algorithm achieves a high accuracy of about 81% for ball location detection, this still indicates a significant margin for error, particularly in complex scenarios where the ball is not clearly visible. Another approach to track soccer ball was presented by Norihiro Ishii et al. [29], who employed Template Matching to enhance the detection process when the ball's movement is minimal. The authors proposed an algorithm that narrows the search area for the ball based on its state. Additionally, an adaptive approach employed to select the most suitable ball extraction method based on the current frame and the ball's movement. On the other hand, if the ball is in front of a player, the detection algorithm may mistakenly identify parts of the player as the ball, leading to incorrect tracking results. Bodhisattwa Chakraborty et al. [32] proposed a trajectory-based approach to detect and track a ball in basketball. The initial step involves segmenting moving objects from the background using a three-frame difference method and an edge detection technique is utilized to further distinguish the moving objects from the background. Morphological operations are also applied to the segmented objects. However, the framework also faces challenges due to motion blurring, especially at critical moments like when the ball is released from a player's hand. Sukadev Meher et al. [34] proposed a trajectory-based ball detection and tracking algorithm

specifically designed for volleyball videos to classify various game states. Their proposed model utilizes an approximate median filtering algorithm to remove the background, generating potential ball detections using the Hough transform and shape and size features. An interpolation method based on the Kalman filter was used to detect and fill in missing ball detections due to occlusions along the trajectory. Although the proposed method aims for real-time processing, the complexity of the algorithm may hinder its performance in live settings where quick decisions are necessary. In another study, Kurowski et al. [25] proposed a model for ball tracking in short volleyball rallies. The proposed method comprises two stages: training and ball tracking. It leverages background subtraction based on a Gaussian Mixture Model (GMM). In the second stage, the resulting foreground images are analyzed to detect the ball. Template matching based on a quarter-ball template is applied in this model. The method relies heavily on background subtraction techniques, which may not always be effective in varying lighting conditions typical in sports halls. Overall, such methods are often highly sensitive to environmental factors like lighting changes, occlusions, background clutter, and camera motion, leading to inaccurate tracking or failures. These methods also exhibit limited robustness to scale variations, shape deformations, and fast motion.

2.2. Deep learning methods

According to [33], Huang proposed a method for identifying volleyball trajectories using a graph convolutional neural network. The author employed the YOLOv4 model to identify regions of interest with high confidence. Subsequently, to extract deep and high-level features, the model was combined with a graph convolutional network. Finally, the DeepSORT [35] tracking algorithm was utilized to estimate the volleyball trajectory. One of the primary challenges highlighted is the low detection accuracy. In another study, Han et al. proposed a model named HMMATrack [36] for ball tracking using a neural network and a custom architecture based on improving multi-scale feature enhancement and multi-level collaborative matching to enhance tracking performance from various perspectives. In the neural network part, two object detection models, CenterNet and MNet, were used. While the authors proposed methods to enhance detection and tracking, increasing the resolution of input images to improve detection can lead to exponential increases in computational requirements. Wang and Chen have introduced a model named TrackNetV3 [37] for badminton shuttlecock tracking. TrackNetV3 comprises two primary modules: trajectory prediction and refinement. The trajectory prediction module leverages a predicted background as auxiliary information to accurately locate the shuttlecock amidst visual distractions. The tracking neural network within TrackNetV3 adopts a U-Net architecture, incorporating convolutional layers and skip connections. As demonstrated in [38], Jorge Armando et al. proposed a semi-supervised model for soccer ball detection and tracking. The proposed framework leverages the YOLOv7 convolutional neural network and incorporates the focal loss function. To track ball trajectories and perform in-depth analysis, DeepSORT was used for object tracking. Vanyi Chao et al. [39] proposed a novel approach to detect volleyball trajectory in videos. The core of the proposed method is the MaxViT Sequential model, which is designed to track high-speed, tiny balls in sports broadcasting videos. Additionally, the proposed method achieved accuracy of 85%. While the model aims to address issues like blurriness and occlusions, it still relies on the quality of the input frames. If the video quality is significantly low, the model's performance may be adversely affected. Overall, achieving real-time performance with these models remains a challenge, often limited by computational cost and latency. In robotics researches, Setiawardhana et al. [40] proposed a method to capture images and videos of the ball's movement in real-time via goalkeeper robot. The captured images are labeled and organized into a dataset that includes various lighting conditions and ball positions to train the YOLOv8 model for object detection. Once the ball is detected, the method uses a simple Convolutional Neural Network (CNN) to predict the ball's arrival position based on the data.

3. Proposed methods

The proposed methodology incorporates architectural modifications in YOLOv8 [41] and YOLOv10 [42] models, specifically utilizing convolutional block attention modules, to enhance ball detection and tracking performance. More details of the modules embedded in the proposed model will be presented in the following sub-sections.

3.1. Convolutional Block Attention Module

The objective of employing attention mechanisms is to enhance the representation of salient regions by directing the network's focus to significant features and suppressing irrelevant information.

Attention blocks not only indicate where to focus but also improve the representation of these salient regions. In this paper, a convolutional block attention module is utilized to achieve this goal [43].

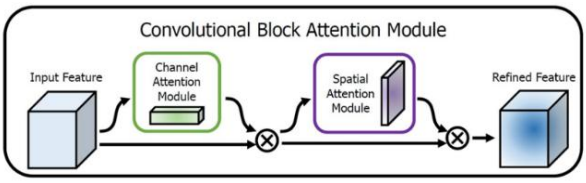


Figure 2. Convolutional Block Attention Module architecture [25].

Given the intermediate feature map, $F \in \mathbb{R}^{C\times H\times W}$, as input, this module sequentially computes a 1D channel attention map, $M_c \in \mathbb{R}^{C\times 1\times 1}$, and a 2D spatial attention map, $M_s \in \mathbb{R}^{1\times H\times W}$. The channel attention map is a vector containing weights that indicate the significance of each feature channel, while the spatial attention map is a 2D map containing weights that indicate the spatial importance or pixel-wise significance of an image. During the multiplication and calculation process, the significant values in both the channel and spatial dimensions are propagated, resulting in a refined output, F''. The computation process of the attention maps is as follows:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where F is Given an intermediate feature map as input, M_c and M_s are channel attention map and spatial attention map, respectively.

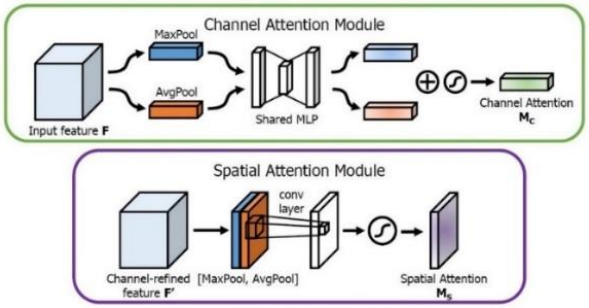


Figure 3. Channel attention and Spatial attention modules architecture [25].

Channel Attention Module. The channel attention map is generated by exploiting the inter-channel relationships of features. To capture spatial information, both average pooling and max pooling are simultaneously employed. This approach significantly enhances the network's representational power by leveraging the benefits of both pooling operations, rather than relying on either independently, and extracts more discriminative features for each object. The computation process of the channel attention maps is as follows:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$
$$= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max})))$$

where the sigmoid function is represented by σ, $W_0 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$. W_0 and W_1 are the MLP weights. Also, ReLU is used as the activation function.

Spatial Attention Module. Unlike the channel attention module, the spatial attention module focuses on "where" an informative and feature-rich part is located, complementing the channel attention. To compute spatial attention, average pooling and max pooling operations are initially applied along the channel axis and concatenated to form an effective feature descriptor.

$$M_s(F) = \sigma(f^{(7\times7)}([AvgPool(F); MaxPool(F)])) \tag{4}$$
$$= \sigma(f^{(7\times7)}([F^s_{avg}; F^s_{\max}]))$$

where $f^{7\times7}$ is a convolutional layer with a 7x7 kernel size is applied to generate the spatial attention map. As illustrated in Figure 4, this module can be implemented as a residual network by adding the input to the output.

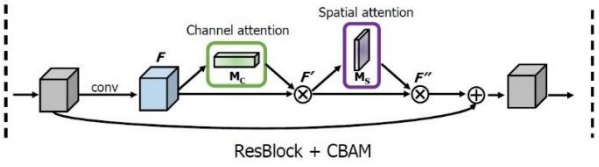
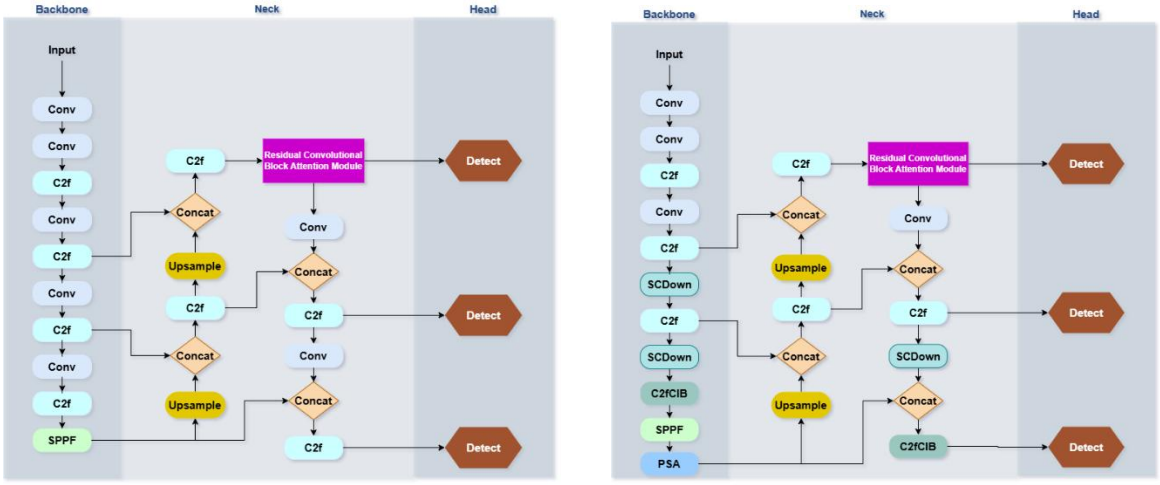


Figure 4. Residual Convolutional Block Attention Module [43].

3.2. Architecture

In the first proposed method, the primary architecture is based on the YOLOv8 object detection model. This model employs convolutional block, C2f module, SPPF module, and detect block. The C2f module is designed to improve the flow of information and gradients within the network, leading to more effective feature extraction. SPPF Allows the network to accept images of varying sizes and makes the model more robust to different input sizes. It pools the feature maps at different scales and then concatenates them. SPPF is a faster version of SPP. Figure 5-(a) presents a detailed illustration of this architecture. As depicted in Figure 5-(a), the modifications were made to the upper part or head of the model. To avoid increasing computational costs, we implemented only one additional block (the purple block).

In the second proposed method, the YOLOv10 object detection model architecture is employed. Given its newer version compared to YOLOv8, it has demonstrated improved performance on various metrics and reduced the number of parameters. However, it exhibits weaker performance in detecting small objects compared to YOLOv8. Figure 5-(b) illustrates second method architecture. similar to YOLOv8, this model utilizes attention block to enhance feature extraction, implemented as residual convolutional attention blocks.



(a) (b)

Figure 5: (a) The first modification of YOLOv8 architecture with Residual Convolutional Block Attention Module, (b) The second modification of YOLOv8 architecture with Residual Convolutional Block Attention Module.

4. Evaluation

In this section, details of the implementation environment, hyperparameters, dataset, and experimental results and discussion are presented.

4.1. Implementation environment

Given the significance of the Graphics Processing Unit (GPU) in deep learning model experimentation and performance evaluation, the proposed methods were trained on the Kaggle environment equipped with a GeForce Tesla P100 GPU featuring 16GB of graphics memory and 29GB of RAM.

4.2. Hyperparameters

Both proposed models were trained using identical hyperparameters for 100 epochs with the SGD optimizer, a learning rate of 0.01, and a momentum of 0.937. The batch size was set to 16.

4.3. Dataset

In this paper, we utilize the VolleyVision dataset [44], a collection of 25,239 images extracted from volleyball match videos. The dataset's strength lies in its large size, encompassing both indoor and beach volleyball matches. Through data augmentation, the dataset was further expanded to include diverse images with varying lighting conditions and duplicates. The dataset was divided into three subsets: 17,679 images for training, 5021 for validation, and 2,539 for testing.

4.4. Experimental results

4.4.1. Baseline Comparison

Evaluation results of the proposed models, compared to the baseline model, on the VolleyVision validation set (5021 images) are presented in Tables 1 and 2, using metrics such as precision, recall, mAP50 and mAP50-95. The results highlight the effectiveness of our approach.

Table 1. Comparison of first proposed model and baseline model evaluation on the VolleyVision.

	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
YOLOv8n	92.9	82.9	89.8	57.9
YOLOv8n+RCBAM	**93.5**	**83.1**	**90.2**	**58.4**
YOLOv8s	93.8	86.2	92.2	61.4
YOLOv8s+RCBAM	**94**	**86.6**	**92.8**	**62**
YOLOv8m	93.7	86.2	92.4	61.7
YOLOv8m+RCBAM	**94**	**88.5**	**93.4**	**63.5**
YOLOv8l	**94.5**	87	93.3	63.4
YOLOv8l+RCBAM	94.2	**87.3**	**93.6**	**63.7**
YOLOv8x	**93.9**	**88.7**	**94.1**	**64.3**
YOLOv8x+RCBAM	93.6	88.1	93.8	64.2

As shown in Tables 1 and 2, the proposed models exhibited relatively better performance in the evaluation results, indicating the effectiveness of the attention block in enhancing feature extraction compared to the baseline model. For example, YOLOv8m with RCBAM increases by 0.3%,

2.3%, 0.1% and 2.2% in precision, recall, mAP50 and mAP50-95, respectively. In the first proposed method, only YOLOv8x with an attention block demonstrated slightly lower performance by 0.3%, 0.6%, 0.3% and 0.01% in precision, recall, mAP50 and mAP50-95, respectively. While in the second proposed method, the YOLOv10b with an attention block evaluation results decreases in recall by 1.1% and the YOLOv10l with an attention block evaluation results decreases in precision by 0.3%.

Table 2. Comparison of second proposed model and baseline model evaluation on the VolleyVision.

	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
YOLOv10n	92.7	80.6	88.4	56.8
YOLOv10n+RCBAM	**92.8**	**80.9**	**88.5**	**57**
YOLOv10s	93.6	84.6	91.4	60
YOLOv10s+RCBAM	**94.1**	**84.7**	**91.7**	**60.4**
YOLOv10m	94.3	85.7	**92.6**	61.6
YOLOv10m+RCBAM	**94.7**	**86**	92.5	**61.8**
YOLOv10b	94.2	**87.6**	93.2	62.3
YOLOv10b+RCBAM	**94.3**	86.5	93.2	**62.5**
YOLOv10l	**94.5**	86	92.8	62
YOLOv10l+RCBAM	94.2	**86.9**	**93**	**62.1**
YOLOv10x	93.8	86.5	92.7	61.7
YOLOv10x+RCBAM	**94**	**87.2**	**92.9**	61.7

4.4.2. Performance Evaluation

Tables 3 and 4 present the inference time of the first proposed model and its baseline, as well as the evaluation time of the second proposed model and its baseline on the validation set, along with the number of trainable parameters. The impact of the graphics processor on the execution time is evident. These evaluations were conducted using a GeForce Tesla P100 GPU. The execution time encompasses pre-processing, detection, and post-processing stages.

Table 3. Inference Time and Parameter Comparison of Baseline and modified YOLOv8 Models.

	Inference time (ms)	Parameters
YOLOv8n	3.1	3005843
YOLOv8n+RCBAM	3.5	3055541
YOLOv8s	5.7	11125971
YOLOv8s+RCBAM	6.3	11323573
YOLOv8m	11.7	25856883
YOLOv8m+RCBAM	12.5	26284149
YOLOv8l	19.8	43607379
YOLOv8l+RCBAM	20.9	44395701
YOLOv8x	28	68124531
YOLOv8x+RCBAM	29.5	69355669

The added attention block did not significantly affect the speed or parameter count of our models, as shown in Tables 3 and 4. The performance gains justify the minor speed trade-off.

Tables 5 and 6 present a comparison of the proposed models with two other models, HMMATrack and GE-YOLOv4, in terms of FPS, precision and AP50, respectively. It is worth noting that some metrics were not available for all models and are therefore omitted.

Table 4. Inference Time and Parameter Comparison of Baseline and modified YOLOv10 Models.

	Inference time (ms)	Parameters
YOLOv10n	3	2694806
YOLOv10n+RCBAM	3.4	2744504
YOLOv10s	6.2	8035734
YOLOv10s+RCBAM	6.5	8233336
YOLOv10m	12	16451542
YOLOv10m+RCBAM	12.7	16895352
YOLOv10b	16.9	20412694
YOLOv10b+RCBAM	17.8	21201016
YOLOv10l	20.3	25717910
YOLOv10l+RCBAM	21.3	26506232
YOLOv10x	29.9	31586006
YOLOv10x+RCBAM	31.2	32817144

Table 5. performance comparison of first proposed model and other models.

	FPS	Precision (%)	AP50 (%)
YOLOv8n+RCBAM	285	93.5	90.2
YOLOv8s+RCBAM	158	94	92.8
YOLOv8m+RCBAM	80	94	93.4
YOLOv8l+RCBAM	47	94.2	93.6
YOLOv8x+RCBAM	33	93.6	93.8
HMMATrack/Mnet	28.2	82.7	N/A
GE-YOLOv4	34	N/A	66.2

Based on the results and comparison of the proposed models with two other models, it can be concluded that the proposed models exhibit higher execution speed and better accuracy. Furthermore, the proposed models demonstrate a quantitatively better performance in detecting small objects such as volleyballs, making them suitable for real-time systems.

Table 6. performance comparison of second proposed model and other models.

	FPS	Precision (%)	AP50 (%)
YOLOv10n+RCBAM	294	92.8	88.5
YOLOv10s+RCBAM	153	94.1	91.7
YOLOv10m+RCBAM	78	94.7	92.5
YOLOv10b+RCBAM	56	94.3	93.2
YOLOv10l+RCBAM	46	94.2	93
YOLOv10x+RCBAM	32	94	92.9
HMMATrack.Mnet	28.2	82.7	N/A
GE-YOLOv4	34	N/A	66.2

The comparisons show that CBAM enhances YOLOv8 and YOLOv10 for small object detection by making the network focus on the most informative feature channels and precise spatial locations. This attention mechanism refines feature representations, making small objects more distinguishable from the background.

4.4.3. Statistical Analysis

To assess the comparative object detection performance of the proposed models, a statistical analysis was conducted between the baseline YOLOv8s model and YOLOv8s with a Residual CBAM (Convolutional Block Attention Module). Both models were evaluated across 10 randomly sampled, identical subsets (each containing 254 images) drawn from the comprehensive VolleyVision test set. A paired samples t-test was employed for each metric to account for the dependency introduced by evaluating both models on the same image subsets. According to Table 7, According to Table 7, for mAP50, the paired samples t-test indicated a statistically significant difference between the two models. YOLOv8s with Residual CBAM (Mean mAP50: 0.9208) demonstrated superior performance compared to the baseline YOLOv8s (Mean mAP50: 0.9132), with a mean improvement of approximately 0.0076 (95% CI [0.004, 0.0112]). The statistical results were $t(9)$=4.8131, p=0.001. Similarly, for mAP50-95, a statistically significant difference was also observed. YOLOv8s with Residual CBAM (Mean mAP50-95: 0.6165) again outperformed YOLOv8s (Mean mAP50-95: 0.6071) by a mean of approximately 0.0094 (95% CI [0.004, 0.0148]). The corresponding statistical values were $t(9)$=3.9335, p=0.0034.

These results collectively suggest that YOLOv8s with Residual CBAM consistently and significantly outperforms the baseline YOLOv8s across both mAP50 and mAP50-95 metrics on this dataset. This highlights the enhanced overall detection accuracy and localization capabilities contributed by the Convolutional Block Attention Module, confirming its positive impact on detection performance.

Table 7. Paired Samples T-Test Results for YOLOv8s vs. YOLOv8s with Residual CBAM.

	mAP50	mAP50-95
p value	0.001	0.0034
t statistic	4.8131	3.9335
95% CI	0.004 – 0.0112	0.004 – 0.0148

4.4.4. Visual Detection Analysis

We present a qualitative analysis through visual examples of ball detections on predicted video frames. Figures 6 showcases our models' robustness and effectiveness.

To complement our visual detection examples and gain deeper insight into the internal workings of our proposed models, we have further included visualizations of mean feature maps. Specifically, Figure 7 illustrates the average activation patterns across channels within the C2f and Convolutional Block Attention Module (CBAM) modules. The C2F module, positioned earlier in the feature extraction pipeline, is crucial for efficiently aggregating features and creating rich representations. Following this, the CBAM adaptively refines these features by highlighting salient spatial regions and emphasizing critical channels. This sequence provides a qualitative understanding of how the C2f module first generates robust feature representations, which are then further refined by CBAM's attention mechanism.

5. Conclusion

In this paper, we introduced two proposed models for volleyball ball detection and tracking using attention blocks, which demonstrated superior performance compared to the baseline model and two other state-of-the-art models. A comprehensive comparison between the proposed models and other models was also conducted.

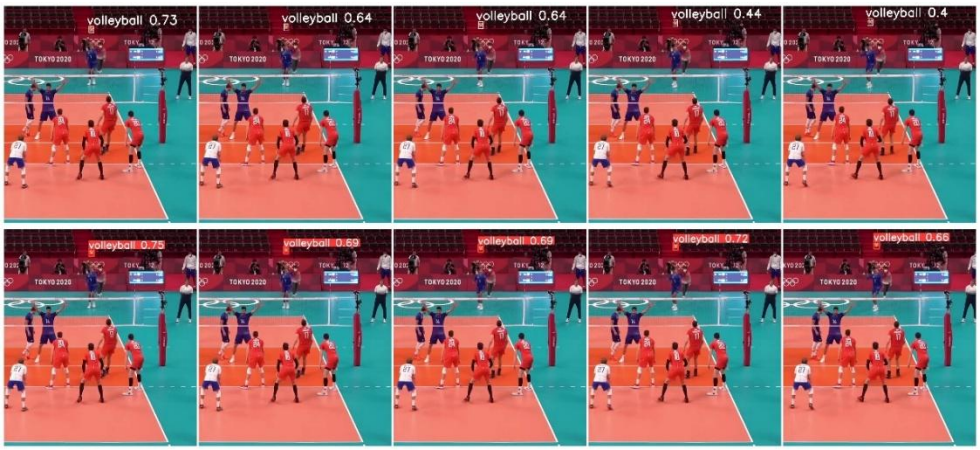


Figure 6. Volleyball prediction results comparison on 5 frames in order: YOLOv8n with Kalman filter (top row) vs YOLOv8n with residual CBAM (bottom row).

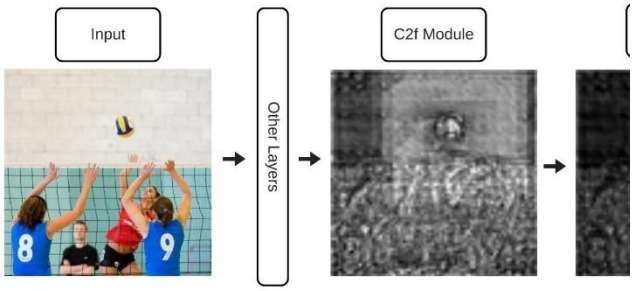


Figure 7. C2f module and Residual Convolutional Block Attention Module feature map example.

Given the high speed of the YOLO detection model and the integration of convolutional attention blocks, which introduced minimal speed degradation but enhanced feature extraction for small objects and ultimately improved accuracy, the proposed models can be effectively employed in real-time systems. As a future work, we aim to utilize new attention blocks and effective fusion in the middle layers of prominent object detection models in order to extract important features, especially in small object projects. Also, for achieving better results, collecting a dataset appropriate to the work domain and considering all model conditions is recommended.

References

[1] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed. "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77 103116, 2021.

[2] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke. "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1 pp. 1-66, 2023.

[3] K. Ranade, T. Khule, and R. More. "Object Recognition in Human Computer Interaction: A Comparative Analysis," *arXiv:2411.04263*, 2024.

[4] N. Esfandiari, K. Kiani, R. Rastgoo, "Development of a Persian Mobile Sales Chatbot based on LLMs and Transformer," *Journal of AI and Data Mining*, vol. 12, no. 4, pp. 465-472, 2024.

[5] N. Esfandiari, K. Kiani, R. Rastgoo, "Transformer-based Generative Chatbot Using Reinforcement Learning," *Journal of AI and Data Mining*, vol. 12, no. 3, pp. 349-358, 2024.

[6] A.M. Ahmadi, K. Kiani, R. Rastgoo, "A Transformer-based model for abnormal activity recognition in video," *Journal of Modeling in Engineering*, vol. 22, no. 76, pp. 213-221, 2024.

[7] F. Bagherzadeh, R. Rastgoo, "Deepfake image detection using a deep hybrid convolutional neural network," *Journal of Modeling in Engineering*, vol. 21, no. 75, pp. 19-28, 2023.

[8] M. Talebian, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Fingerprint Verification," *Journal of AI and Data Mining*, vol. 12, no. 2, pp. 241-248, 2024.

[9] R. Rastgoo, K. Kiani, S. Escalera, "ZS-GR: zero-shot gesture recognition from RGB-D videos," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 43781-43796, 2023.

[10] R. Rastgoo, K. Kiani, S. Escalera, "A deep co-attentive hand-based video question answering framework using multi-view skeleton," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1401-1429, 2023.

[11] H. Zaferani, K. Kiani, R. Rastgoo, "Real-time face verification on mobile devices using margin distillation," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 44155-44173, 2023.

[12] S. Zarbafi, K. Kiani, R. Rastgoo, "Spoken Persian digits recognition using deep learning," *Journal of Modeling in Engineering*, vol. 21, no. 74, pp. 163-172, 2023.

[13] N. Esfandiari, K. Kiani, R. Rastgoo, "A conditional generative chatbot using transformer model," *Neural Computing and Applications,* 2025.

[14] R. Rastgoo, K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," *Journal of Modeling in Engineering*, vol. 17, no. 58, pp. 103-111, 2019.

[15] F. Alinezhad, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Gender Recognition in Mobile Devices," *Journal of AI and Data Mining*, vol. 11, no. 2, pp. 229-236, 2023.

[16] S. Shekarizadeh, R. Rastgoo, S. Al-Kuwari, M. Sabokrou, "Deep-disaster: unsupervised disaster detection and localization using visual data," *26th International Conference on Pattern Recognition (ICPR)*, pp. 2814-2821, 2022.

[17] N. Majidi, K. Kiani, R. Rastgoo, "A deep model for super-resolution enhancement from a single image," *Journal of AI and Data Mining*, vol. 8, no. 4, pp. 451-460, 2020.

[18] R. Rastgoo, V. Sattari-Naeini, "Gsomcr: Multi-constraint genetic-optimized qos-aware routing protocol for smart grids," *Iranian Journal of Science and Technology, Transactions of Electrical, Engineering*, vol. 42, no. 2, pp. 185-194, 2018.

[19] R. Rastgoo, V. Sattari-Naeini, "Tuning parameters of the QoS-aware routing protocol for smart grids using genetic algorithm," *Applied Artificial Intelligence*, vol. 30, no. 1, pp. 52-76, 2016.

[20] R. Rastgoo, V. Sattari Naeini, "A neurofuzzy QoS-aware routing protocol for smart grids," *22nd Iranian Conference on Electrical Engineering (ICEE)*, pp. 1080-1084, 2014.

[21] K. Kiani, R. Rastgoo, A. Chaji, S. Escalera, "Image Inpainting Enhancement by Replacing the Original Mask with a Self-attended Region from the Input Image," *Journal of AI and Data Mining*, vol. 13, no. 3, pp. 379-391, 2025.

[22] C.B. Santiago, A. Sousa, M.L. Estriga, L.P. Reis, and M. Lames, "Survey on team tracking techniques applied to sports." *In 2010 International Conference on Autonomous and Intelligent Systems, AIS*, pp. 1-6, 2010.

[23] P.R. Kamble, A.G. Keskar, and K.M. Bhurchandi, "Ball tracking in sports: a survey." *Artificial Intelligence Review*, vol. 52, pp.1655-1705, 2019.

[24] M. Takahashi, K. Ikeya, M. Kano, H. Ookubo, and T. Mishina, "Robust volleyball tracking system using multi-view cameras." *In 2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2740-2745, 2016.

[25] P. Kurowski, K. Szelag, W. Zaluski, and R. Sitnik, "Accurate ball tracking in volleyball actions to support referees." *Opto-Electronics Review,* vol. 26, no. 4, pp. 296-306, 2018.

[26] P.R. Kamble, A.G. Keskar, and K.M. Bhurchandi, "A deep learning ball tracking system in soccer videos," *Opto-Electronics Review,* vol. 27, no. 1, pp. 58-69, 2019.

[27] R. Beal, T.J. Norman, and S.D. Ramchurn. "Artificial intelligence for team sports: a survey," *The Knowledge Engineering Review,* vol. 34, e28, 2019.

[28] X. Yu, H.W. Leong, C. Xu, and Q. Tian. "Trajectory-based ball detection and tracking in broadcast soccer video," *IEEE Transactions on multimedia*, vol. 8, no. 6 pp. 1164-1178, 2006.

[29] N. Ishii, I. Kitahara, Y. Kameda, and Y. Ohta. "3D tracking of a soccer ball using two synchronized cameras," *In Advances in Multimedia Information Processing*, vol.8, pp. 196-205. 2007.

[30] K. Zhao, W. Jiang, X. Jin, and X. Xiao. "Artificial intelligence system based on the layout effect of both sides in volleyball matches," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3075-3084, 2021.

[31] D. Budden, S. Fenn, J. Walker, and A. Mendes. "A novel approach to ball detection for humanoid robot soccer," In AI 2012: Advances in Artificial *Intelligence: 25th Australasian Joint Conference*, vol. 25, pp. 827-838. 2012.

[32] B. Chakraborty and S. Meher. "A real-time trajectory-based ball detection-and-tracking framework for basketball video," *Journal of Optics,* vol. 42, pp.156-170, 2013.

[33] G. Huang, "An Effective Volleyball Trajectory Estimation and Analysis Method with Embedded Graph Convolution," *International Journal of Distributed Systems and Technologies (IJDST),* vol. 14, no. 2, pp.1-13, 2023.

[34] B. Chakraborty and S. Meher, "A trajectory-based ball detection and tracking system with applications to shot-type identification in volleyball videos," *In 2012 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1-5, 2012.

[35] N. Wojke, A. Bewley, and D. Paulus, "Simple online and real-time tracking with a deep association metric," *In 2017 IEEE international conference on image processing (ICIP)*, pp. 3645-3649, 2017.

[36] X. Han, Q. Wang, and Y. Wang, "Ball Tracking Based on Multiscale Feature Enhancement and Cooperative Trajectory Matching," *Applied Sciences*, vol. 14, no. 4 pp. 1376, 2024.

[37] Y.J. Chen and Y.S. Wang, "Tracknetv3: Enhancing shuttlecock tracking with augmentations and trajectory rectification," *In Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pp. 1-7. 2023.

[38] J.A. Vicente-Martínez, M. Márquez-Olivera, A. García-Aliaga, and V. Hernández-Herrera, "Adaptation of YOLOv7 and YOLOv7_tiny for soccer-ball multi-detection with DeepSORT for tracking by semi-supervised system," *Sensors*, vol. 23, no. 21, pp. 8693, 2023.

[39] V. Chao, H.Q. Nguyen, A. Jamsrandorj, Y.M. Oo, K.R. Mun, H. Park, S. Park, and J. Kim. "Tracking the Blur: Accurate Ball Trajectory Detection in Broadcast Sports Videos," *In Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, pp. 41-49. 2024.

[40] I.K. Wibowo and N.A.H. Bernardt, "Prediction of Ball Position Using CNN Methods with Zed Camera on Goalkeeper Robot Application," IEEE Access, vol. 13, pp. 41559 – 41570, 2025.

[41] Ultralytics, YOLOv8, 2023. [online]. Available: https://docs.ultralytics.com/models/yolov8/

[42] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. "Yolov10: Real-time end-to-end object detection," *arXiv:2405.14458*, 2024.

[43] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, "Cbam: Convolutional block attention module," *In Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19, 2018.

[44] VolleyVision, 2024. [online]. Available: https://github.com/shukkkur/VolleyVision [Accessed: Sept. 25, 2023].

بهبود شناسایی توپ والیبال با استفاده از یادگیری عمیق

محمد جدیدی میاندشتی، کوروش کیانی و راضیه راستگو*

دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۵/۰۲/۲۳؛ بازنگری ۲۰۲۵/۰۵/۰۶؛ پذیرش ۲۰۲۵/۰۸/۱۵

چکیده:

در سال‌های اخیر، کاربرد تکنیک‌های یادگیری عمیق، حوزه‌های مختلفی از جمله عرصه تحلیل ورزشی را متحول کرده است. تحلیل مسیر و ردگیری توپ در ورزش به یک حوزه تحقیقاتی فزاینده حیاتی تبدیل شده است که ناشی از پیشرفت‌های فناوری و تقاضای رو به رشد برای یافته‌های داده‌محور در عملکرد ورزشی است. در والیبال، ورزشی که با حرکات سریع و بازی استراتژیک مشخص می‌شود، توانایی ردیابی دقیق مسیر توپ برای تحلیل‌های تمرینی و رقابتی بسیار مهم است. این مقاله مدل‌های یادگیری عمیق جدیدی را برای شناسایی و ردیابی دقیق توپ والیبال پیشنهاد می‌کند. با ترکیب مکانیسم‌های توجه در معماری YOLOv8 و YOLOv10، مدل‌های ارائه شده به‌طور قابل توجهی عملکرد را بهبود می‌بخشند، به‌ویژه در سناریوهای چالش برانگیز مانند انسداد و حرکت سریع توپ. مدل‌های پیشنهادی در مقایسه با مدل پایه و دیگر مدل ها، عملکرد بهتری در معیارهای مختلف نشان دادند. مدل‌های پیشنهادی به دقت (به ترتیب ۹۴.۲ و ۹۴.۷ درصد)، Recall (به ترتیب ۸۸.۱ درصد و ۸۷.۶ درصد) و سرعت بلادرنگ دست یافتند که آنها را می توان برای سیستم های مختلف تحلیل ورزشی استفاده نمود.

کلمات کلیدی: شناسایی توپ والیبال، یادگیری عمیق، ماژول توجه، شبکه عصبی پیچشی.