p-ISSN: XXXX-YYYY
e-ISSN: XXXX-YYYY

YOLOv11-LCA: Lightweight Attention-Enhanced Object Detection for Soccer Video Analysis

Muhammad Abiya Makruf[a1], Muhammad Rafly Arjasubrata[a2], Tjokorda Agung Budi Wirayuda[a3]

[a]School of Computing, Telkom University
Bandung, Indonesia
[1]muhammadabiyamakruf@student.telkomuniversity.ac.id (Corresponding author)
[2]muhammadraflya@student.telkomuniversity.ac.id
[3]cokagung@telkomuniversity.ac.id

Abstract

Accurate object detection in soccer video analysis is essential for enabling downstream tasks such as tracking, re-identification, and tactical understanding. However, detecting small and fast-moving objects especially the soccer ball remains a challenging problem due to motion blur, frequent occlusions, and the ball's minimal visual footprint in broadcast footage. This study proposes YOLOv11-LCA, a lightweight yet effective modification of the YOLOv11 object detector that integrates the Low-Complexity Attention Module (LCAM) to enhance the network's sensitivity to small-scale features. The model was evaluated using the SoccerNet Tracking dataset and benchmarked against the original YOLOv11 architecture across three model scales (nano, small, and medium). Experimental results show that YOLOv11-LCA consistently outperforms the baseline in terms of precision, recall, and mean Average Precision (mAP), achieving notable improvements in the detection of the ball class while maintaining real-time inference capabilities. These findings demonstrate that incorporating low-overhead attention modules can significantly improve small object detection in complex sports environments without sacrificing efficiency.

Keywords: *Object Detection, Soccer Video Analytics, Low-Complexity Attention Module, YOLOv11, Small Object Detection*

1. Introduction

The analysis of sports video footage has become increasingly crucial for teams, broadcasters, and fans alike. Soccer being the world's most popular sports, generates vast amounts of video data rich with potential insights [1], [2]. Computer vision techniques offer the potential to automate the extraction of valuable information, moving beyond traditional manual annotation methods which are often time-consuming and subjective [3]. Accurate detection and localization of key elements on the field such as players, referees, goalkeepers, and ball serve as a fundamental prerequisite for a wide range of downstream tasks in soccer analytics. These include multi-object tracking [4], [5], player re-identification [6], action spotting [7], [8], highlight generation [9], and tactical analysis [10].

Despite significant advancements in deep-learning based object detection methods such as R-CNN [11], Single-Shot Detector (SSD) [12], and more advance YOLO-based methods [13], applying these methods to soccer videos presents unique and substantial challenges [5], [14]. Players often exhibit rapid and unpredictable movements, leading to motion blur and complex occlusions, both between players and with the ball. Furthermore, players belonging to the same wear visually similar kits, making appearance-based discrimination difficult, especially from typical broadcast camera viewpoints where players might occupy only small portion of the frame [4], [5]. Perhaps the most significant challenge, however, lies in the detection of the soccer ball itself. Relative to the overall image resolution in wide-angle broadcast shots, the ball is often very small object, sometimes only a few pixels in diameter [15], [16]. Its appearance can change drastically due to high speed (causing motion blur), varying lighting conditions, and frequent occlusions by players. Standard object detection architectures, while powerful, often struggle to reliably detect such small, fast-moving objects with high accuracy [15].

To address these inherent challenges, particularly the robust detection of the small soccer ball, this paper proposes a novel detection framework based on a modified YOLOv11 architecture

[17]. Our primary contribution is the integrated of an attention-based mechanism designed specifically to enhance the feature representation of small objects within the complex soccer scene. By enabling the network to focus more effectively on discriminative features relevant to small targets like ball, while still accurately detecting larger objects such as players, referee, and goalkeepers, we aim to significantly improve detection performance, especially for the challenging ball class.

The goal of this work is to establish a strong and reliable baseline detector tailored for soccer domain. The proposed model is designed to output accurate bounding boxes for key classes, providing a crucial first step for more advanced soccer video analysis systems. We demonstrate the effectiveness of our approach through rigorous evaluation, showcasing its potential to serve as a foundational component for future research in areas such as automated tracking, re-identification, and sophisticated event detection in soccer.

2. Reseach Methods

2.1. Data Acquisiton and Preperation

This study utilizes the SoccerNet Tracking dataset [6], [7] as the primary data source. This large-scale dataset provides high-definition broadcast video footage (1920×1080 resolution) from multiple professional soccer matches, along with detailed bounding box annotations and tracking identifiers for key entities such as players, referees, goalkeepers, and the ball. For the purpose of developing and evaluating an object detection model, we deliberately focused on a single complete match (Match ID 4) to ensure consistency while retaining a representative variety of game scenarios, object scales, and camera dynamics typically found in broadcast settings.

From Match ID 4, we extracted annotated frames along with their bounding box coordinates and corresponding class labels. To align with the focus of this research on frame-level object detection rather than multi-object tracking, trajectory information and instance tracking IDs were excluded from the pipeline. The selected object categories are player, referee, goalkeeper, and ball. All bounding boxes and labels were then converted to the YOLO label format, which includes the class identifier followed by normalized bounding box coordinates and dimensions. All video footage was captured from a fixed broadcast camera positioned at midfield, which dynamically pans to follow the game. Although the camera remains centrally located, these movements introduce minor variations in object scale and appearance, particularly affecting small and fast-moving objects such as the ball, as illustrated in **Figure 1**.

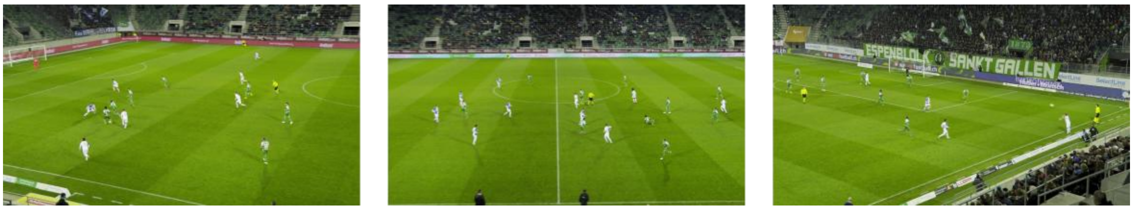


Figure 1. Comparison between different angle in the SoccerNet dataset.

The dataset was partitioned into 70% training, 10% validation, and 20% test sets. Each subset consists of continuous 30-second video segments sampled at 25 frames per second, with all data exclusively sourced from Match ID 4 to maintain domain consistency. Specifically, 17 video segments were used to construct the training and validation sets, where the final 10% of frames from each segment were assigned to validation, and the remaining 90% to training resulting in approximately 70% of the full dataset. The test set, comprising 15 separate 30-second segments, was also sampled from Match ID 4. A detailed breakdown of object class distributions across the three subsets is presented in **Table 1**.

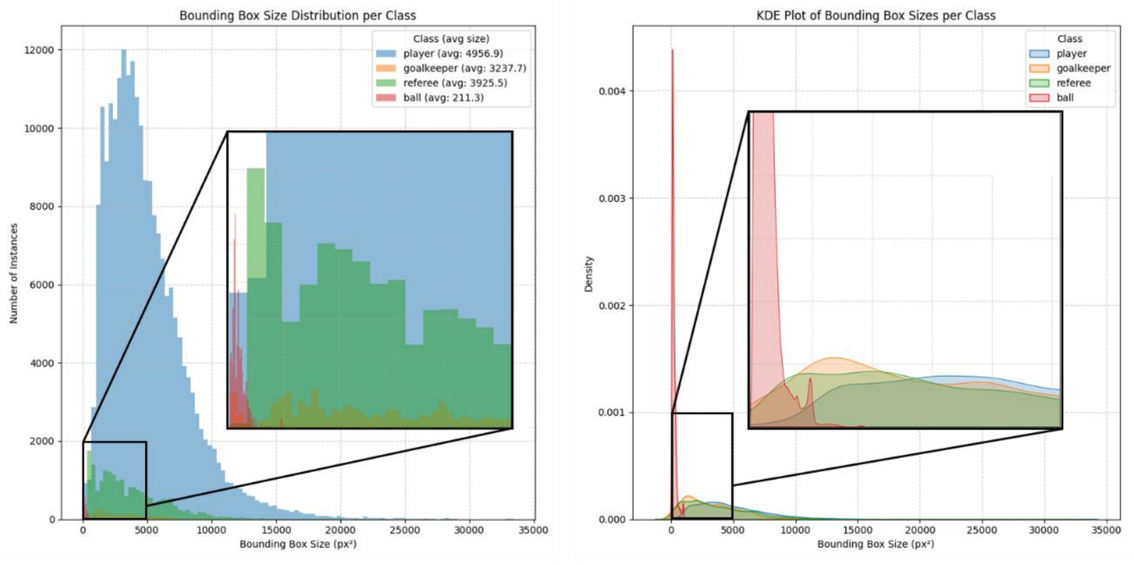
Table 1. Distribution of Annotated Object Instances Across Train, Validation, and Test.

Class	Train	Validation	Test
Player	156201	19280	39825
Referee	14499	1927	4007
Ball	8656	1117	1957
Goalkeeper	4499	760	1314
Total	183855	23084	47103

2.2. Bounding Box Size Distribution on SoccerNet Dataset

To better understand the spatial characteristics of the annotated objects, we conducted an analysis of bounding box sizes across all object classes using both histogram and Kernel Density Estimation (KDE) plots. The size of each bounding box was calculated as the area in pixels (width × height), providing an approximate measure of the visual footprint of each object in the frame.

Figure 2a presents a histogram showing the distribution of bounding box sizes per class. It is evident that the majority of player instances dominate the dataset, with an average bounding box size of approximately 4956.9 px². Goalkeepers and referees also exhibit relatively large bounding boxes, with mean areas of 3237.7 px² and 3925.5 px², respectively. In contrast, the ball class demonstrates a significantly smaller average size of just 211.3 px², and appears highly skewed toward the smallest size ranges in the distribution.



(a) (b)

Figure 2. Bounding box size distributions: (A) Histogram plot all classes, (B) KDE plot per class.

To further illustrate the density and spread of object sizes, a KDE plot is shown in **Figure 2b**. This visualization reveals that the distribution of bounding box sizes for players, referees, and goalkeepers is distributed over a wider range and generally peak between 2000–6000 px². However, the ball class exhibits a sharp density spike below 500 px², confirming its small scale and high concentration in the lower-bound region of the distribution as shown in **Figure 2b.**

These findings reinforce the inherent challenge of detecting the ball in soccer videos, particularly when using object detection architectures that are typically biased toward larger object scales. The extreme scale disparity between the ball and other objects justifies the need for targeted enhancements, such as attention mechanisms, to better capture and represent small object features.

2.3. Proposed Model Architecture

The core of our proposed approach is an enhanced object detection architecture derived from YOLOv11, referred to as YOLOv11-LCAM. This architecture is specifically designed to improve detection performance on small objects such as the soccer ball, while maintaining the computational efficiency required for real-time applications. The overall architecture flow is illustrated in **Figure 5**.

The YOLOv11 architecture developed by Ultralytics is chosen as the baseline model because it represents a recent advancement in the YOLO-based object detectors, primarily designed for efficient and real-time detection [13]. Its core architecture follows the standard paradigm of a backbone for hierarchical feature extraction, a neck for multi-scale feature fusion, and a head for generating final prediction [13]. A cornerstone of the YOLOv11 design is the optimized C3k2 block, which serves as a key building block within both the backbone and neck. The C3k2 block is more computationally efficient adaptation of the Cross Stage Partial (CSP) concept, known for its ability to improve gradient flow and reduce computational redundancy [26]. By incorporating smaller kernel size ("k2"), the C3k2 block effectively extracts and process features critical for detection across various scales while maintaining high processing speed, making it well suited for real-time application like soccer video analysis. The architecture of the YOLOv11 is depicted in **Figure 3**.

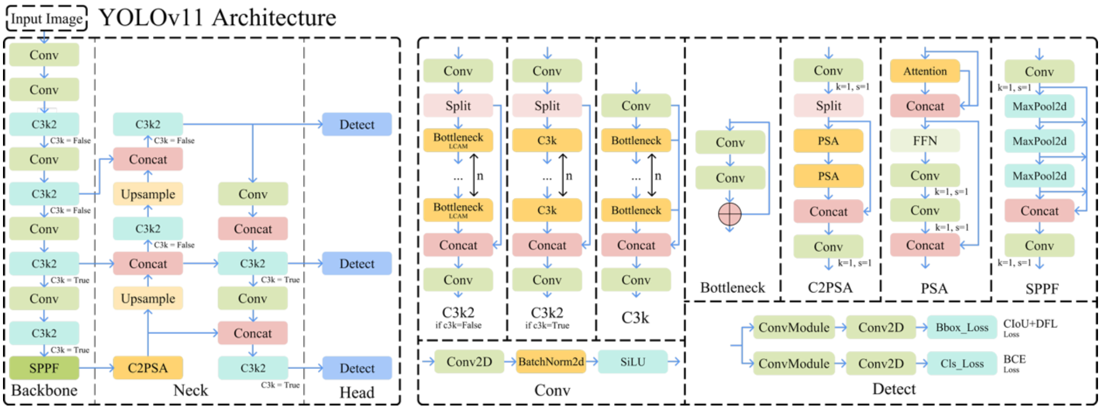


Figure 3. Architecture flow of YOLOv11.

Meanwhile, attention mechanisms have proven highly effective in enhancing object detection by enabling networks to selectively focus on the most informative features for a given task [18], [19]. For challenging scenarios involving small or particularly occluded objects, like the soccer ball, the ability to emphasize subtle yet critical visual cues while suppressing irrelevant background noise is paramount. To address this, we propose integrating a lightweight attention mechanism based on the Low-Complexity Attention Mechanism (LCBHAM), which is specifically designed to improve feature representation of small object detection with minimal computational overhead [18]. LCBHAM sequentially applies both channel and spatial attention to refine the feature maps, allowing the network to first determine "what" is important across channels and then pinpoint "where" the critical spatial information resides which depicted in **Figure 4a**.

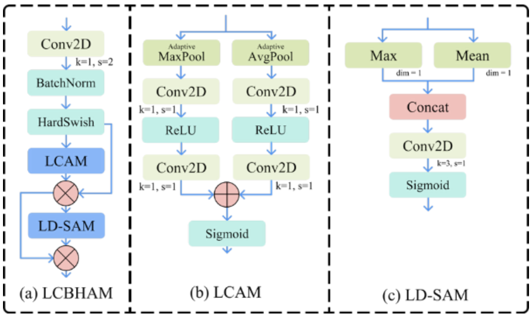


Figure 4. Architecture flow of Low-complexity Attention Mechanisms.

LCBHAM is comprised of two main sub-modules: the Low-Complexity Channel Attention Module (LCAM) for channel attention and the Lightweight Detail Spatial Attention Module (LD-SAM) for spatial attention [18]. The LCAM, responsible for channel-wise feature recalibration, process an input feature map X by applying Global Average Pooling and Global Max Pooling to generate channel descriptors. These descriptors are then passed through separate lightweight 1×1 convolutional layers followed by ReLU activation. The results are summed element-wise, and a Sigmoid function is applied to produce the attention map M_c, with the architecture detail shown in **Figure 4b**. The formulation for M_c is:

$$F_{avg}(X) = ReLU(Conv2d_{k=1,s=1}(AvgPool(X)) \tag{1}$$

$$F_{max}(X) = ReLU(Conv2d_{k=1,s=1}(MaxPool(X)) \tag{2}$$

$$M_c(X) = \sigma(F_{avg}(X) + F_{max}(X)) \tag{3}$$

Following channel refinement by LCAM, the LD-SAM refines spatial features. It takes the channel-attended feature map as input, applies channel-wise Average Pooling and Max Pooling with a kernel size of 1 along the channel dimension to produce two 2D spatial descriptor maps. These maps are concatenated and then passed through a single 2D convolution with a kernel size of 3, followed by a Sigmoid activation function, to generate the spatial attention map M_s. The architecture flow of the LD-SAM is illustrated in **Figure 4b** with the formulation for M_s is:

$$M_s(X) = \sigma(Conv2d_{k=3}([AvgPool_{k=1}(X);\ \ MaxPool_{k=1}(X)])) \tag{4}$$

To enhance the YOLOv11 architecture capability for accurately detection small and challenging objects like soccer ball, we integrate these low-complexity attention mechanisms strategically. One primary modification involves replacing the standard C3k2 blocks within the neck with modified C3k2LCAM blocks. These enhanced blocks incorporate the LCAM into their internal structure, allowing for channel-wise feature recalibration early in the feature extraction process and during multi-scale feature fusion in the neck. Furthermore, to leverage both channel and spatial attention at critical feature interaction points, some the deepest convolutional block in the neck is replaced by the LCBHAM, which combines the function of LCAM and the LD-SAM. This integration strategy aims to improve the network's sensitivity to the fine-grained features characteristic of small objects, helping to better distinguish the ball from cluttered background and overcome challenges like motion blur and occlusion, while maintaining the computational efficiency necessary for real-time performance. The modified YOLOv11 architecture is illustrated at **Figure 5**.

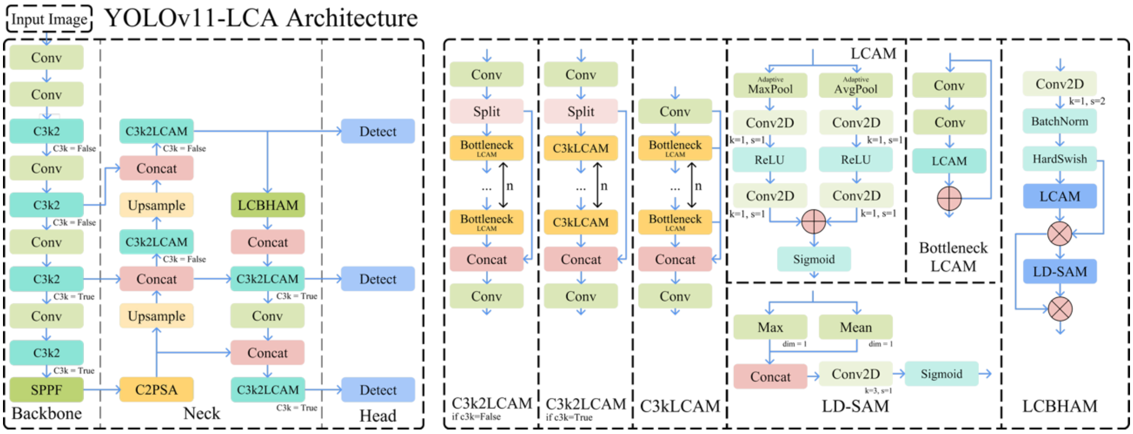


Figure 5. Architecture flow of the proposed method.

We hypothesize that the integration of LCAM and the LCBHAM enables the YOLOv11-LCA model to better preserve and amplify subtle features critical for detecting small objects. By focusing on both local details and long-range dependencies in a computationally efficient manner, the model is expected to improve detection accuracy for the ball class while maintaining real-time inference speed. This improved detection capability establishes a more robust baseline for downstream soccer video analysis tasks such as object tracking, player re-identification, and action spotting.

2.4. Training Scenario and Configuration

To evaluate the effectiveness of the proposed modifications, we conducted experiments under two primary training scenarios: (1) training of the baseline YOLOv11 architecture variants and (2) training of the modified YOLOv11-LCA architecture variants incorporating the Low-Complexity Attention Module. A total of distinct model configurations were trained for comparative analysis where the first three is the original YOLOv11 with the nano, small and the medium counterpart that representing the baselines, and the other three is their respective LCA-enhanced counterparts. Both model families were trained from scratch without any pretrained weights. This approach ensures a fair comparison, as the structural changes introduced in the modified architectures prevent the direct use of pre-trained parameters and necessitate training from initialization.

Training was performed using the SoccerNet Tracking dataset, specifically the annotated from Match ID 4 as described in Section 2.1. A consistent set of training hyperparameters was applied to all six model configurations to ensure a controlled and comparable evaluation environment. These parameters, detailed in **Table 2**, were selected to balance training stability and computational feasibility within a realistic research setup.

Table 2. Training Parameters For Training All of The YOLO Model.

Parameter	Value
Epoch	50
Input Size	640×640
Batch Size	16
Initial Learning Rate	0.01
Momentum	0.937
Optimizer	Adam

All training experiments were conducted on a hardware with specific resources optimized for deep learning tasks. The primary computational power used throughout the training process was provided by an NVIDIA RTX 3080 GPU, which offer sufficient resource to handle workload for various YOLOv11 scales efficiently and with the Ultralytics version of 2.3.96, enabling reproducible training results across all tested scenario.

This standardized training setup, utilizing a specific dataset split, consistent hyperparameters, and dedicated hardware, was designed to provide a robust basis for directly comparing the performance baseline of the baseline YOLOv11 model against their LCA-enhanced versions. This results obtained from these training runs form the foundation for the quantitative evaluation and analysis presented in the subsequent sections, particularly in assessing the impact of the LCA on small object detection performance in complex soccer environments.

2.5. Model Evaluation

To quantitively assess the performance of each object detection model configuration, we employ four standard evaluation metrics widely used in the field: Precision, Recall, and mean Average Precision averaged over multiple IoU thresholds from 0.50 to 0.95 (mAP50-95). These metrics collectively provide a comprehensive view of the model’s accuracy, completeness, and localization capabilities, which are particularly important for challenging small objects like the soccer ball.

Precision measures the accuracy of the model’s positive predictions. It is defined as the ratio of the number of correctly predicted positive instances (True Positives) to the total of the number predicted positive instance (True Positives + False Positives). Precision indicates how reliable the model when it predicts that an object is present. The formula of Precision is written as:

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

Where:

- TP (True Positive): A correctly detected object (prediction has an IoU ≥ 0.50 with the a ground truth box of the correct class)
- FP (False Positive): An incorrect detection (prediction has an IoU < 0.50 with any ground truth box, or detects a non-existent object, or detect the wrong class)

Recall, also known as sensitivity, measures the mode's ability to detect all relevant objects within the dataset. It is calculated as the ratio of the number of correctly predicted positive instances (True Positives) to the total number of actual positive instances (True Positive + False Negatives). Recall indicates what proportion of the actual objects in the image the model successfully identified. The equation of the Recall can be formulated as:

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

Where:

- TP (True Positive): A correctly detected object (prediction has an IoU ≥ 0.50 with the a ground truth box of the correct class)
- FN (False Negative): An actual object that the model failed to detect

The original YOLOv11 models serves as baselines to evaluate the effectiveness of the proposed YOLOv11-LCA architecture. This comparison conducted across multiple scales and various metrics, is essential to determine whether integrating lightweight attention mechanisms leads to measurable improvements in detection performance, particularly for challenging small objects like the soccer ball, which suffers from a limited visual footprint and frequent occlusions. Ultimately, we aim to validate the hypothesis that such modules can enhance model sensitivity and precision for small-object detection without sacrificing overall performance or computational efficiency.

3. Result and Discussion

In this section, we present the experimental results and provide a detailed analysis to assess the effectiveness of the proposed YOLOv11-LCA architecture in comparison with the original YOLOv11. The discussion focuses on both quantitative and qualitative evaluations, as well as model efficiency considerations, with special attention given to the performance on small objects particularly the ball class, which poses significant detection challenges due to its size and motion characteristics.

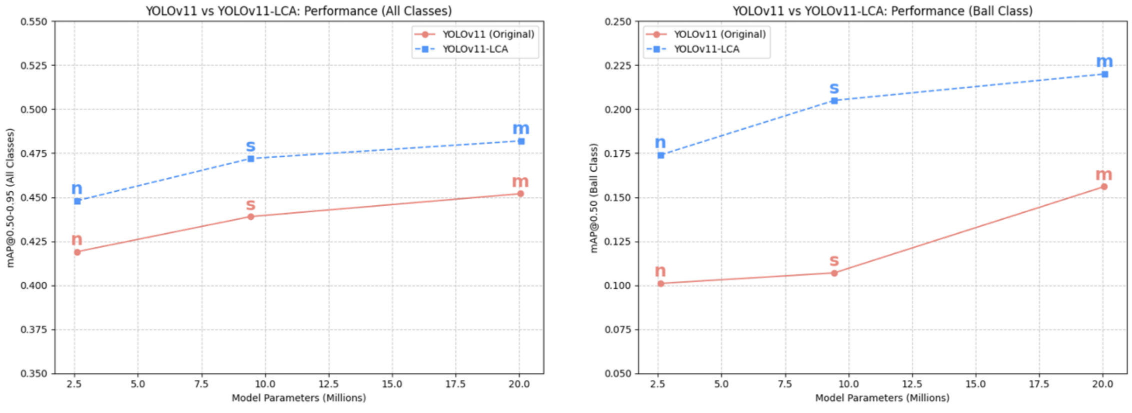
3.1. Quantitative Evaluation

Quantitative evaluation involved comparing the original YOLOv11 models against the modified with low-complexity attention counterparts. As summarized in **Table 3**, the YOLOv11-LCA variants consistently demonstrated superior performance across all model scales for overall "All Classes" detection. The advantage was evident in key metrics such as mAP50-95 and Recall. For example, YOLOv11m-LCA achieved an mAP50-95 of 0.482 for "All Classes", compared to 0.452 for the baseline YOLOv11m, showing the successful implementation of the LCA mechanism to improve the YOLOv11 performance for detection on the soccer match.

Table 3. Performance Comparison of Original Yolov11 and The Proposed Method Yolov11-LCA Across Different Scales For "All Classes" and "Ball" Detection Metrics.

Model	**All Classes**			**Ball**		
	mAP$_{50-95}$	*Precision*	*Recall*	*mAP$_{50-95}$*	*Precision*	*Recall*
YOLOv11n	0.419	0.793	0.716	0.030	0.585	0.070
YOLOv11n-LCA	0.448	0.761	0.824	0.057	0.590	0.139
YOLOv11s	0.439	0.783	0.735	0.033	0.486	0.094
YOLOv11s-LCA	0.472	0.848	0.770	0.063	0.613	0.151
YOLOv11m	0.452	0.821	0.743	0.048	0.558	0.137
YOLOv11m-LCA	0.482	0.846	0.778	0.077	0.577	0.169

The most substantial improvement due to LCA integration were observed in the detection of the challenging “ball’ class, a key focus of this study. **Table 3** reveals that the attention-enhanced models yielded remarkable gains in ball detection accuracy across all scales. For instance, the YOLOv11n-LCA gains remarkable gains in ball detection accuracy, even outperform the baseline YOLOv11m model across all metrics. Ball recall also saw considerable enhancement such as an increase from 0.094 to 0.151 for the small-scale model, highlighting the improvement of the small object detection on the modified YOLOv11. **Figure 6** illustrating “Ball” class performance increase on the modified model, showing a notably wider performance margin between the LCA and the baseline model, underscoring the low-complexity attention efficiency for small and often occluded object like the soccer ball.



(a) (b)

Figure 6. Performance comparison of mAP50-95 for the YOLOv11 and the proposed method YOLOv11-LCA across model scales. (a) Performance comparison for all classes. (b) Performance comparison for ball class.

Crucially, these significant performance enhancements were achieved with minimal computational overhead detailed in **Table 4**, which outlines the model parameters, GFLOPs, and layer counts. The integration of LCAM resulted in only a negligible 1% increase in model parameters and GFLOPs across all the scales. For example, the YOLOv11m-LCA model showed 20.077M parameters which only marginally higher than the baseline YOLOv11m 20.056M. While the number of layers did increase with the addition of the attention modules, the minimal impact on overall computational load suggests that the real-time inference capabilities, vital for practical soccer video analysis, are well-reserved in the YOLOv11-LCA architecture.

Table 4. Comparison of Model Complexity, Detailing Parameters, Gflops, And Layer Count of Yolov11 and Yolov11-LCA Variants.

Model	**Params**	**GLOPs**	**Layers**
YOLOv11n	2,590,620	6.444	181
YOLOv11n-LCA	2,593,326	6.450	220
YOLOv11s	9,429,340	21.555	181
YOLOv11s-LCA	9,440,110	21.575	220
YOLOv11m	20,056,092	68.202	231
YOLOv11m-LCA	20,077,614	68.241	288

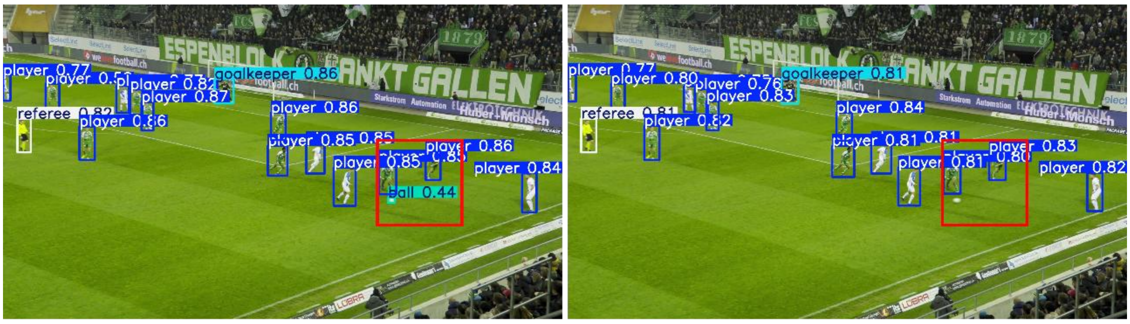
A consolidated view of the top-performing configurations is presented in **Table 5**, which highlights the best model for each metric and object class. This summary further substantiates the benefits of the LCAM integration, with YOLOv11-LCA securing all of the leading scores. Notable the YOLOv11m-LCA as it’s the larger model of the three. For the critical “Ball” class, YOLOv11m-LCA consistently led in Recall, mAP50, and mAP50-95, while YOLOv11s-LCA achieved the highest Precision. These collective findings affirm that the LCAM enhancement significantly improves detection capabilities, particularly for small and challenging objects, without imposing a substantial computational burden.

Table 5. Best Performing Models Across All Metric And Class.

Class	Metric (Highest)	Model Name	Value
All	Precision	yolo11s-LCA	0.848
	Recall	yolo11m-LCA	0.778
	mAP50	yolo11m-LCA	0.794
	mAP50–95	yolo11m-LCA	0.482
Player	Precision	yolo11m-LCA	0.936
	Recall	yolo11m-LCA	0.969
	mAP50	yolo11m-LCA	0.975
	mAP50–95	yolo11m-LCA	0.645
Goalkeeper	Precision	yolo11s-LCA	0.984
	Recall	yolo11m-LCA	0.989
	mAP50	yolo11m-LCA	0.994
	mAP50–95	yolo11s-LCA	0.601
Referee	Precision	yolo11m-LCA	0.895
	Recall	yolo11m-LCA	0.985
	mAP50	yolo11m-LCA	0.988
	mAP50–95	yolo11m-LCA	0.610
Ball	Precision	yolo11s-LCA	0.613
	Recall	yolo11m-LCA	0.169
	mAP50	yolo11m-LCA	0.220
	mAP50–95	yolo11m-LCA	0.077

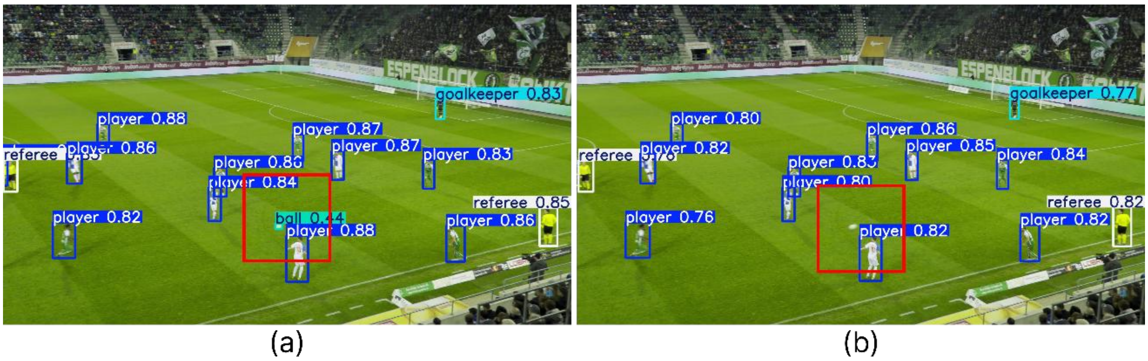
3.2. Qualitative Results

Beyond the numerical metrics, qualitative evaluation through visualization of prediction outputs on representative frames offer crucial insights into the practical performance enhancements of the proposed YOLOv11-LCA models. These visualizations consistently reveal that the modified architectures are more adept at detecting challenging instances of the soccer ball, such as those partially occluded, affected by motion blur, or appearing very small within the frame conditions where the original YOLOv11 models often failed. The integrated LCAM evidently helps the network to better focus on subtle spatial cues, enabling more accurate localization of the ball. This is demonstrated in **Figure 7a**, where the YOLOv11s-LCA model successfully identified the ball with 0.44 confidence rate amidst a cluster of players, a scenario in which the baseline YOLOv11s depicted in **Figure 7b** fails detect the ball. Similarly, **Figure 8b** showcases another instance where the YOLOv11s-LCA model correctly detect ball near a player, while the corresponding baseline model fails to register its presence, highlighting the LCA improved sensitivity.



(a) (b)

Figure 7. Example 1 illustrating the ball detection result of baseline and modified YOLOv11s models. (a) YOLOv11s-LCA shows a successful detection of the ball. (b) Baseline YOLOv11s missed the ball detection.

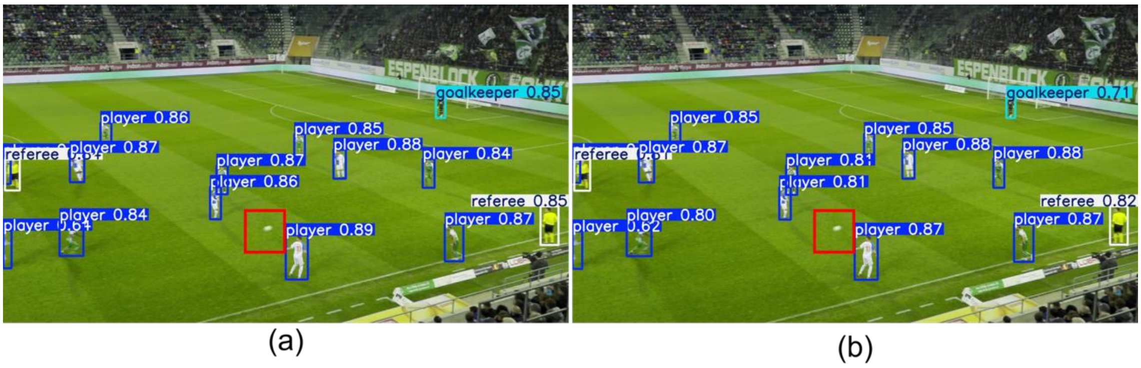


(a) (b)

Figure 8. Example 2 illustrating the ball detection result of baseline and modified YOLOv11s models. (a) YOLOv11s-LCA shows a successful detection of the ball. (b) Baseline YOLOv11s missed the ball detection.

While the LCAM integration significantly boosts detection capabilities, particularly for the ball, extreme conditions can still pose a challenge. **Figure 9** illustrates such a scenario, where both the YOLOv11m-LCA depicted in **Figure 9a** and the baseline YOLOv11m shown in **Figure 9b** fail to detect the soccer ball. In this specific frame, the ball is heavily affected by motion blur due to its high speed, rendering its feature indistinct and difficult for either model to capture. However, a noteworthy observation in **Figure 9a** is that despite the missed ball detection, the YOLOv11m-LCA model generally assigns higher confidence scores to other correctly identified objects on the field, compared to the baseline model in **Figure 9b**. For example, the goalkeeper is detected with a confidence of 0.85 by the YOLOv11m-LCA versus 0.71 by the baseline, and similar trends are visible for several players.

These qualitative examples collectively underscore the benefits and limitations of the proposed approach. The consistent success of the YOLOv11-LCA models in **Figure 7** and **Figure 8** in detecting balls missed by baseline model highlights the practical value of the low-complexity attention mechanism for improving robustness in typical, yet challenging, game situations. The scenario in **Figure 9**, while demonstrating a current boundary for detection, also subtly indicate that the LCAM implementation to the YOLOv11 can contribute to more confident feature representations for other, less severely degraded object. Overall, the visual evidence shows inline with the quantitative findings, illustrating that YOLOv11-LCA provides a more reliable detection framework for soccer video analysis, particularly for the critical task of ball detection, even if extreme visual degradation remains an open challenge.



(a) (b)

Figure 9. Example illustrating a challenging scenario where both baseline and proposed YOLOv11m model fails to detect balls. (a) YOLOv11m-LCA fails to detect ball due to its motion blur. (b) Baseline YOLOv11m fails to detect fast moving ball.

4. Conclusion

This study introduced YOLOv11-LCA, a modified object detection architecture tailored for enhanced soccer video analysis, with a particular focus on improving the detection of small objects like the soccer ball. By strategically integrating the Low-Complexity Attention Module

(LCAM) into the YOLOv11 backbone, our approach significantly boosts the network's ability to capture fine-grained spatial and channel-wise features crucial for small object recognition, without imposing a substantial computational overhead.

Comprehensive experiments on the SoccerNet dataset unequivocally demonstrate that YOLOv11-LCA consistently surpasses the original YOLOv11 models across key evaluation metrics, including Precision, Recall, and mAP50–95. The most notable advancements were observed for the challenging ball class which typically suffers from low accuracy due to its diminutive size, motion blur, and frequent occlusions achieving up to a 200% relative improvement in mAP50–95. Importantly, these enhancements for small objects were accompanied by improved performance on larger classes such as players and goalkeepers.

Crucially, these accuracy gains were achieved while retaining the real-time processing capabilities inherent to YOLOv11, positioning YOLOv11-LCA as a practical and scalable solution for demanding real-world soccer analytics applications. These findings affirm that lightweight attention mechanisms serve as effective architectural enhancements for object detection in complex, dynamic visual environments. Future work will explore extending this robust detection framework to advanced tasks such as multi-object tracking, player re-identification, and automated event spotting.

References

[1] M. F. Hashmi, B. T. Naik, and A. G. Keskar, "BDTA: events classification in table tennis sport using scaled-YOLOv4 framework," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 6, pp. 9671–9684, Jun. 2023, doi: 10.3233/JIFS-224300.

[2] K. Seweryn, G. Chęć, S. Łukasik, and A. Wróblewska, "Improving Object Detection Quality in Football Through Super-Resolution Techniques," Jan. 2024.

[3] X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin, "Feature Combination Meets Attention: Baidu Soccer Embeddings and Transformer based Temporal Detection," Jun. 2021.

[4] M. Gran-Henriksen, H. A. Lindgaard, G. Kiss, and F. Lindseth, "Deep HM-SORT: Enhancing Multi-Object Tracking in Sports with Deep Features, Harmonic Mean, and Expansion IOU," Jun. 2024.

[5] C. Yang *et al.*, "A survey on soccer player detection and tracking with videos," *Vis Comput*, vol. 41, no. 2, pp. 815–829, Jan. 2025, doi: 10.1007/s00371-024-03367-6.

[6] A. Deliege *et al.*, "SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2021, pp. 4503–4514. doi: 10.1109/CVPRW53098.2021.00508.

[7] A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, and M. Van Droogenbroeck, "Scaling up SoccerNet with multi-view spatial localization and re-identification," *Sci Data*, vol. 9, no. 1, p. 355, Jun. 2022, doi: 10.1038/s41597-022-01469-1.

[8] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking Objects as Points," 2020, pp. 474–490. doi: 10.1007/978-3-030-58548-8_28.

[9] J. Komorowski, G. Kurzejamski, and G. Sarwas, "FootAndBall: Integrated Player and Ball Detector," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, SCITEPRESS - Science and Technology Publications, 2020, pp. 47–56. doi: 10.5220/0008916000470056.

[10] M. Manafifard, H. Ebadi, and H. Abrishami Moghaddam, "A survey on player tracking in soccer videos," *Computer Vision and Image Understanding*, vol. 159, pp. 19–46, Jun. 2017, doi: 10.1016/j.cviu.2017.02.002.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[12] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[14] K. Host and M. Ivašić-Kos, “An overview of Human Action Recognition in sports based on Computer Vision,” *Heliyon*, vol. 8, no. 6, p. e09633, Jun. 2022, doi: 10.1016/j.heliyon.2022.e09633.

[15] 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, Mar. 2019, doi: 10.1016/j.opelre.2019.02.003.

[16] G. Jin, “Player target tracking and detection in football game video using edge computing and deep learning,” *J Supercomput*, vol. 78, no. 7, pp. 9475–9491, May 2022, doi: 10.1007/s11227-021-04274-6.

[17] N. R. Sulake, “A Comprehensive Guide to YOLOv11 Object Detection,” Oct. 2024, [Online]. Available: https://www.analyticsvidhya.com/blog/2024/10/yolov11-object-detection/

[18] Y. Zhang, H. Zhang, Q. Huang, Y. Han, and M. Zhao, “DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects,” *Expert Syst Appl*, vol. 241, p. 122669, May 2024, doi: 10.1016/j.eswa.2023.122669.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[20] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[21] Y. Zhang, X. Wang, M. S. Shakeel, H. Wan, and W. Kang, “Learning upper patch attention using dual-branch training strategy for masked face recognition,” *Pattern Recognit*, vol. 126, p. 108522, Jun. 2022, doi: 10.1016/j.patcog.2022.108522.

[22] Q. Hou, D. Zhou, and J. Feng, “Coordinate Attention for Efficient Mobile Network Design,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2021, pp. 13708–13717. doi: 10.1109/CVPR46437.2021.01350.

[23] T. Liu *et al.*, “Spatial Channel Attention for Deep Convolutional Neural Networks,” *Mathematics*, vol. 10, no. 10, p. 1750, May 2022, doi: 10.3390/math10101750.

[24] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to Attend: Convolutional Triplet Attention Module,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2021, pp. 3138–3147. doi: 10.1109/WACV48630.2021.00318.

[25] R. Mo, S. Lai, Y. Yan, Z. Chai, and X. Wei, “Dimension-aware attention for efficient mobile networks,” *Pattern Recognit*, vol. 131, p. 108899, Nov. 2022, doi: 10.1016/j.patcog.2022.108899.

[26] R. Xiao, H. Wang, L. Wang, and H. Yuan, “C3Ghost and C3k2: performance study of feature extraction module for small target detection in YOLOv11 remote sensing images,” in *Second International Conference on Big Data, Computational Intelligence, and Applications (BDCIA 2024)*, S. S. Agaian, Ed., SPIE, Mar. 2025, p. 139. doi: 10.1117/12.3059792.