第 49 卷第 2 期
七月 2025

Vol. 50 No. 1
Jan - July 2025

Diqiu Kexue - Zhongguo Dizhi Daxue Xuebao/Earth Science
Journal of China University of Geosciences
https://dqkx-periodicals.com

DOI: 10.5281/zenodo.15845503

Improving Small Object Detection in Remote Sensing with YOLO11-CBAM and Deep Learning

Nima Garshasebi*

Department of Computer Engineering, K. N.
Toosi University of Technology Tehran, Iran

*Corresponding author: nima.garshasebi9776@gmail.com

Published: 09 July 2025
Accepted: 30 June 2025
Received: 29 May 2025

Abstract: In this study, we present an advanced object detection system designed specifically for remote sensing images, leveraging the YOLOv11 framework enhanced with a Convolutional Block Attention Module (CBAM). Detecting objects in remote sensing imagery poses significant challenges due to the wide variation in object scales, complex backgrounds, densely packed objects, and arbitrary orientations. Traditional detection models often struggle under these conditions, particularly in accurately identifying small objects. To address these limitations, we propose two key improvements to YOLOv11: (1) the integration of CBAM, which enhances feature extraction by focusing on critical regions through channel and spatial attention mechanisms, thereby suppressing irrelevant background information, and (2) the modification of the detection head by introducing an additional layer specifically optimized for small object detection, improving the model's ability to handle multi-scale objects. We evaluated our proposed model on the DOTA dataset, a widely recognized benchmark for aerial image object detection. Experimental results demonstrate a significant improvement in performance, achieving a mean Average Precision (mAP50) of 76.68%, which outperforms both the baseline YOLOv11 and several state-of-the-art models. Furthermore, ablation studies confirm the individual contributions of CBAM and the enhanced detection head to the overall performance. These findings highlight the effectiveness of combining attention mechanisms with multi-scale feature learning to advance object detection in remote sensing applications, offering a robust solution for real-world scenarios such as urban planning, environmental monitoring, and disaster management.

Keywords: Attention layer, Cbam, Deep learning, Object detection, Satellite images.

Introduction

Computer vision has emerged as a transformative technology across diverse domains, revolutionizing how we analyze and interpret visual data. In medicine, it plays a pivotal role in identifying diseases such as cardiovascular conditions [1], various cancers [2], and other ailments by analyzing medical imaging with unprecedented accuracy. In psychology, it aids in emotion recognition by decoding facial expressions [3], offering insights into human behavior and mental states. Firefighters benefit from computer vision through advanced fire detection systems [4] that pinpoint flames in complex environments, enhancing response times and safety. Beyond these, its applications extend to aerial imagery analysis, enabling tasks such as environmental monitoring, disaster response, and urban planning. This widespread adoption underscores the versatility of computer vision, making it a cornerstone of modern technological advancements. In this context, the detection of objects in remote sensing has become a fundamental aspect of geospatial analysis, with significant implications for civilians, leveraging the power of computer vision to address real-world challenges.Accurate identification and localization of objects in satellite images are essential for various applications, including disaster relief, environmental surveillance, city planning, and strategic defense operations [5]-[11]. Detecting objects in remote sensing is trickier than in regular object detection because of how satellite and aerial images work. Problems like objects varying in size, appearing at odd angles, being packed closely together, and having messy backgrounds make it harder to spot them accurately [12], [13]. The DOTAv1 [14] dataset, a widely recognized benchmark in remote sensing object detection, addresses these challenges. It features a diverse collection of objects, including vehicles, ships, and aircraft, which are often arranged in dense clusters and exhibit varying orientations. In addition, the dataset includes images with complex backgrounds, such as urban landscapes, forests, and water bodies, which introduce significant noise and interference. These factors make DOTAv1 an ideal testbed for evaluating the robustness and adaptability of object detection algorithms in real-world scenarios. One of the most pressing challenges in remote sensing object detection is the arbitrary orientation of objects. Unlike natural images, in which objects typically appear in upright positions, objects in satellite imagery can appear at any angle. This necessitates the development of models capable of handling rotational invariance, a feature that is not inherently present in many traditional object detection frameworks. Furthermore, the dense arrangement of objects in remote sensing images often leads to occlusion and overlapping, making it difficult for models to accurately localize and classify individual instances. Another critical challenge is the variation in object scale. Remote sensing images often contain objects that vary significantly in size, ranging from small vehicles to large ships or aircraft. This multi-scale nature requires models to simultaneously detect objects at different resolutions, which is particularly challenging in low-resolution imaging. The presence of complex backgrounds, such as textured terrains or cluttered urban environments, can introduce significant noise, further degrading detection performance.Deep learning-based approaches have revolutionized object detection, achieving state-of-the-art performance on natural image datasets such as COCO [15] and ImageNet [16]. Among these, You Only Look Once (YOLO) [17] has gained popularity owing to its real-time detection and high accuracy. However, its performance degrades in remote sensing imagery owing to challenges such as arbitrary orientations, dense object distributions, and complex backgrounds. To address these limitations, attention mechanisms, particularly the Convolutional Block Attention Module (CBAM) [18], have been introduced to enhance feature representation by sequentially applying channel and spatial attention. Integrating CBAM into object detection frameworks improves accuracy and robustness by enabling models to focus on salient features while suppressing background noise. In this study, we propose a modified version of the YOLO 11 [19] architecture, enhanced with CBAM, to address the unique challenges of remote sensing object detection. Our approach leverages the strengths of YOLO 11, such as its real-time detection capabilities and efficient architecture, while integrating CBAM to improve feature representation and handle complex backgrounds. We evaluate our model on the DOTAv1 dataset, focusing on its ability to detect objects with arbitrary orientations, dense arrangements, and varying scales. The experimental results demonstrate that our modified model achieves significant improvements in detection accuracy, particularly in challenging scenarios where traditional models often fail. This study makes the following contributions:

1. We improve the YOLO 11 architecture by integrating CBAM, enhancing the model's ability to handle complex backgrounds and refining feature representation.
2. Evaluation on DOTAv1: We conduct extensive experiments on the DOTAv1 dataset, demonstrating the effectiveness of our approach in detecting objects with arbitrary orientations and dense arrangements.
3. Performance Analysis: We provide a detailed analysis of our model's performance, including comparisons with the original YOLO 11 and other state-of-the-art models, highlighting the improvements achieved by our modifications.

By addressing the unique challenges of remote sensing object detection, our work aims to advance the state of the art in this field

and contribute to the development of more robust and accurate detection algorithms for real-world applications.

2- Related Work

In the following two sections, advancements in object detection algorithms are reviewed, covering both traditional techniques and deep learning-based methods. In the last 20 years, object detection has improved a lot, focusing mainly on recognizing objects in images or videos and figuring out exactly where they are and how big they are. Unlike image classification, which focuses solely on labeling objects, object detection involves both classification and accurate localization. The development of object detection can be divided into two major phases: the period dominated by traditional algorithms (1998–2014) and the era of deep learning-driven approaches (2014 to the present).

2-1-Traditional object detection algorithms

Before 2012, object detection methods primarily relied on handcrafted feature extraction due to the absence of advanced image representation techniques. These methods generally involved identifying object regions, extracting features, and classifying objects. Key approaches from this era include Scale-Invariant Feature Transform(SIFT) [20], Histogram of Oriented Gradients (HOG) [21], Speeded Up Robust Features (SURF) [22], and Oriented FAST and Rotated BRIEF (ORB) [23]. Although these algorithms laid the foundation for object detection, their reliance on handcrafted features and high computational demands limited their performance in complex scenarios, leading to the emergence of deep learning-based approaches.

2-2-Deep Learning based object detection algorithms

With the advent of deep learning in the period of deep learning-based object detection algorithms (2014–present) [24]-[27] significant breakthroughs emerged. The adoption of deep learning techniques significantly improved feature extraction and representation, enhancing the ability to tackle complex detection challenges. Consequently, traditional object detection algorithms were progressively replaced by deep learning-based approaches. Within this paradigm, two main methodologies emerged: anchor-based techniques, encompassing both one-stage and two-stage models, and anchor-free approaches. In anchor-based object detection, commonly used two-stage algorithms such as R-CNN [28], Fast RCNN [29], Faster RCNN [30], FPN [31] and Mask RCNN [32] are applied. Object detection algorithms typically follow a two-stage approach: region proposal generation and object detection. In the first stage, the algorithm identifies potential object-containing regions or bounding boxes within the input image. This is often achieved using a Region Proposal Network (RPN), which efficiently highlights regions of interest for further evaluation. In the second stage, the candidate regions undergo object detection, where each region is classified to determine whether it contains an object, and its position and boundaries are refined accordingly. This phase typically employs convolutional neural networks (CNNs) combined with classification and regression components to improve detection precision. Single-stage object detection techniques streamline the conventional two-stage approach by removing the need for a separate region proposal step.

These algorithms directly predict the class probabilities and object position coordinates, enabling faster detection speeds. Notable examples of one-stage detection algorithms are YOLOv1–11 [33], SSD [34], the development of these algorithms shows how object detection keeps getting better, becoming more efficient and useful over time.

Anchor-free object detection algorithms eliminate the dependency on predefined anchors, a key characteristic of traditional anchor-based methods. By doing so, they reduce computational complexity and minimize the number of hyperparameters, leading to improved model efficiency. Recently, anchor-free approaches have focused on detecting key points for object localization instead of relying on anchors, streamlining model architecture and further decreasing computational overhead. Noteworthy anchor-free object detection algorithms include CornerNet [35], CenterNet [36], and FSAF [37].

2-3-Dataset Description

To train and evaluate our enhanced model, we utilized the DOTA dataset, a comprehensive collection of aerial images designed for object detection tasks. The dataset was initially released by Wuhan University in 2017 and consists of three versions: DOTAv1.0, DOTAv1.5 and DOTAv2.0. For this study, we employed the DOTAv1.0 version, which comprises 2,806 aerial images with resolutions ranging from 800 × 800 pixels to 4000 × 4000 pixels. These images include 188,282 annotations across 15 distinct categories of remote sensing targets. The DOTA images are sourced from multiple platforms, including Google Earth, GF-2 and JL-1 satellites (provided by the China Centre for Resources Satellite Data and Application), and aerial images from CycloMedia B.V. The dataset contains both RGB images and grayscale images. The RGB images are obtained from Google Earth and CycloMedia, while the grayscale images are derived from the panchromatic band of the GF-2 and JL-1 satellite images. Compared to other publicly available remote sensing datasets, DOTAv1.0 stands out due to its extensive collection of images featuring small and multi-scale targets, as well as its sufficient sample size for each category. These characteristics make it an ideal choice for training our network, enabling the model to achieve robust performance in detecting a wide variety of objects in aerial imagery. These categories, such as baseball diamond, storage tank, tennis court, basketball court, ground track field,

harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, and swimming pool, are illustrated in figure 1.Figure 2 illustrates sample aerial images from the DOTA v1.0 dataset, showcasing different environments such as transportation hubs, urban areas, and sports facilities. The images demonstrate the multi-scale nature of the dataset, where objects appear in various orientations, lighting conditions, and occlusions, making object detection more challenging. The diversity of these objects, along with variations in scale, orientation, and background complexity, makes the DOTAv1.0 dataset an ideal choice for training and evaluating robust object detection models.

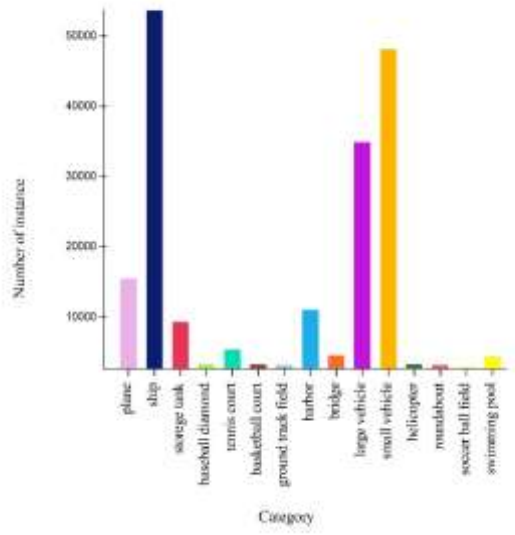

Fig. 1. Distribution of object categories in the DOTA v1.0 dataset, illustrating the imbalance in class frequencies and highlighting the challenge of detecting underrepresented categories.

Fig. 2.Examples of DOTA aerial images showcasing diverse object categories, varying scales, arbitrary orientations, and complex backgrounds, emphasizing the need for robust object detection models.

3-Methods

3-1-Proposed YOLO 11 Architecture with Modifications

The YOLO framework has significantly influenced the field of object detection by introducing an end-to-end neural network architecture capable of simultaneously performing object localization and classification. The original YOLO models utilize a unified approach, enabling efficient inference while maintaining high detection accuracy for real-time applications [15]. Building upon this foundation, YOLO 11 integrates novel enhancements over its predecessors, particularly focusing on improving small object detection and addressing challenges in complex environments, such as aerial imagery datasets. The modifications introduced in the proposed model are designed to enhance feature extraction, improve attention mechanisms, and optimize detection across multiple scales. The following sections detail the architecture and key changes.

3-2-Backbone

The backbone in YOLO 11 is responsible for feature extraction from input images, generating multi-scale feature maps through a series of convolutional layers and residual blocks. As shown in Figure 5, the architecture incorporates an enhanced feature extraction pipeline to improve performance in challenging environments. A CBAM layer is integrated after the second C3k2 block. The CBAM mechanism sequentially applies channel and spatial attention, refining feature maps to focus on relevant regions while suppressing background noise. The CBAM module enhances the model's ability to prioritize meaningful spatial information, which is particularly crucial for aerial imagery datasets, where small objects often appear in cluttered backgrounds. The integration of CBAM after the C3k2 block ensures that early-stage feature representations are effectively filtered before deeper layers process them. Unlike traditional feature extraction methods, which treat all spatial regions equally, CBAM dynamically enhances the most informative features, leading to improved localization and classification accuracy [16]. Additionally, the proposed backbone increases the number of channels in early convolutional layers to capture fine-grained details more effectively. This modification helps compensate for the high object density and small object scales present in aerial imagery datasets like DOTA.

3-3-Neck

The neck aggregates and fuses multi-scale features from the backbone before passing them to the detection head. In the proposed architecture, a key improvement is the introduction of an enhanced neck structure (highlighted in orange in Figure 4) that incorporates an additional detection layer specifically designed for super small objects. This modification includes an extra convolutional pathway, refining feature representations before feature fusion. Unlike the previous YOLO architectures, which rely solely on Feature Pyramid Networks (FPN) or Path Aggregation Networks (PANet), the improved design in figure 4 introduces an additional feature extraction step before concatenation. This ensures that fine-grained details from high-resolution feature maps are preserved, which is particularly beneficial for detecting objects with minimal pixel coverage. Another improvement in the neck design is the inclusion of adaptive feature fusion, which adjusts feature importance dynamically based on object scale. Traditional multi-scale aggregation methods, such as FPN, use a fixed hierarchy of feature maps, which may not be optimal for aerial images. By refining feature interactions through a dynamic weighting strategy, the proposed model achieves better alignment of feature representations across different object scales.

3-4-Head

The detection component is tasked with producing bounding box coordinates and class likelihoods for every identified object. A notable improvement comes from incorporating multi-scale detection with a CBAM adjustment, allowing the model to generate predictions across four different feature scales. This enhancement greatly boosts the model's capacity to identify objects of diverse sizes, particularly emphasizing the detection of extremely small ones.

3-5-Performance and Efficiency

The proposed modifications, including CBAM and the additional detection layer, enhance detection accuracy without significantly increasing computational complexity. Early experiments indicate improved mAP scores, particularly for small objects, while maintaining resource efficiency. Compared to the original YOLO 11 architecture, which lacked specific attention mechanisms and super-small object detection capabilities, these enhancements address these limitations and make the model more robust for datasets like DOTA. Figure 5 illustrates the key components of the modified YOLO 11 architecture, highlighting the integration of CBAM and the additional detection layer that improves feature refinement and small object localization.

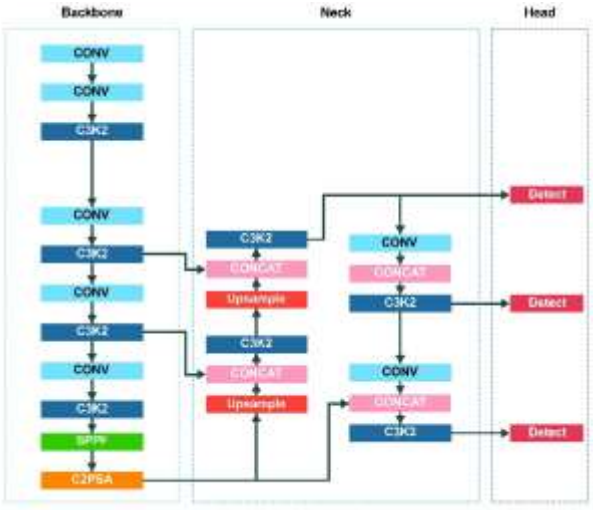

Fig. 3. OLO 11 original architecture

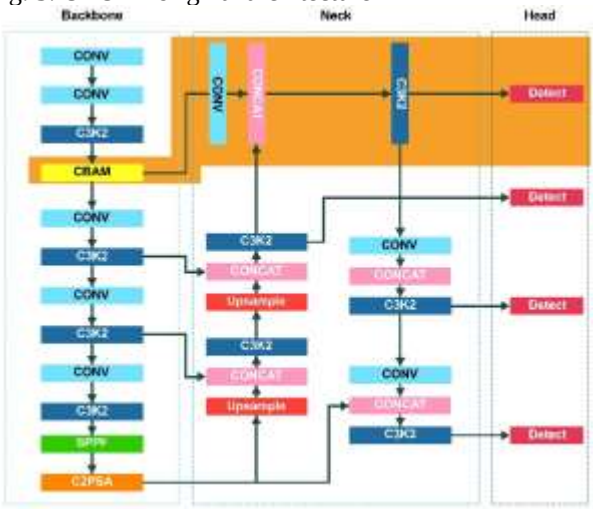

Fig. 4.Improvement YOLO 11

4-Experiments and Results

The training process was conducted using the DOTA-v1.0 dataset, which contains 15 common categories, 2,806 images, and 188,282 instances. The dataset was split into training, validation, and test sets with proportions of 1/2, 1/6, and 1/3, respectively. This split ensures a balanced distribution of data for training, validation, and evaluation. The model was trained using the AdamW optimizer with an initial learning rate of 0.001. The training process included 100 epochs with a batch size of 16 and an image size of 640 × 640 pixels. To improve convergence, warmup epochs were applied for the first 10 epochs. Additionally, the close_mosaic parameter was set to 10, disabling mosaic augmentation in the final 10 epochs to stabilize training. To prevent overfitting, data augmentation techniques were enabled, including mosaic augmentation (before the final 10 epochs), random cropping, and flipping. The model was trained on an NVIDIA A100 GPU, leveraging its high computational power to accelerate the training process. The training process was configured to save checkpoints every 10 epochs and store the final model weights for evaluation.

4-1-Baseline Comparisons

The new YOLO 11 model outperformed both the original YOLO 11 and YOLOv8, especially when it came to spotting tiny objects in aerial photos. As shown in Table 1, our YOLO 11 model achieved a mean Average Precision (mAP50) of 76.68\%, surpassing the original YOLO 11 by 1.22\% and YOLOv8 by 2.48\%. Additionally, the mAP50-95 score improved to 60.35\%, indicating better precision across different IoU thresholds. The recall also increased to 73.12\%, reflecting a higher capability of detecting objects correctly. These results highlight the effectiveness of our modifications, including the integration of CBAM and the additional detection layer, in improving detection performance without significantly increasing computational complexity.

Table 1: comparing the proposed model against YOLOv11, YOLOv8 on the DOTA dataset

Model	*mAP50*	*mAP50-95*	*Recall*
Yolo 11	*75.45*	*59.43*	*70.75*
Yolo 8	*74.19*	*58.58*	*68.98*
Our model	*76.68*	*60.35*	*73.12*

This study evaluates the performance of YOLO 11, YOLOv8, and OURYOLO on the DOTA v1.0 dataset, focusing on three key metrics: mAP50, mAP95, and Recall. figures 5, 6, and 7 illustrate the comparative performance of these models over the training epochs. Figure 5 presents the mAP50 metric, which measures detection accuracy at an IoU threshold of 0.50. The results indicate that all models exhibit an increasing trend in precision over the training process, with significant improvements occurring within the first 40–50 epochs. Notably, OURYOLO outperforms both YOLO 11 and YOLOv8, particularly in the later stages of training, demonstrating superior object detection capability at this threshold. Figure 6 depicts mAP95, a stricter evaluation metric that averages precision across multiple IoU thresholds ranging from 0.50 to 0.95. While all models follow a similar learning trajectory, the absolute values are lower compared to mAP50 due to the increased IoU constraints. The results indicate that OURYOLO consistently achieves a higher mAP95, highlighting its effectiveness in detecting small objects with greater localization accuracy. This performance advantage suggests that architectural enhancements in OURYOLO contribute to improved feature extraction and object representation. Figure 7 illustrates the recall metric, which reflects the proportion of ground-truth objects correctly detected by the models. The recall curves indicate that OURYOLO maintains higher recall values in the middle stages of training and stabilizes at a competitive level in the final epochs. In contrast, YOLO 11 experiences a slight decline in recall towards the end of training, suggesting potential overfitting or reduced generalization.

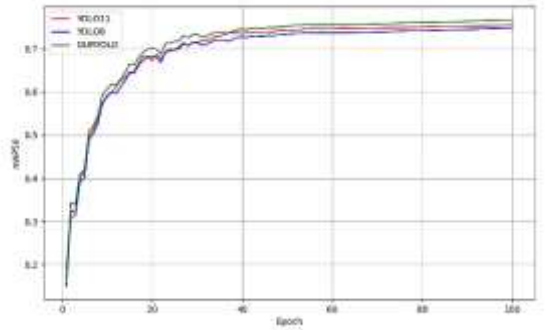

Fig. 5. Comparison of mAP50 scores across different models, demonstrating the performance improvements achieved by the proposed YOLO11-CBAM architecture.

Overall, the results confirm that OURYOLO achieves the best performance across all three metrics, surpassing YOLO 11 and YOLOv8 in terms of precision and recall. The improvements are particularly notable in mAP95, which underscores the model's ability to detect small objects in aerial images with high localization accuracy. Furthermore, the training curves indicate that all models converge after approximately 40–50 epochs, suggesting that further training beyond this point yields minimal gains. These findings demonstrate that OURYOLO is a highly effective model for small object detection in aerial imagery, making it a strong candidate for real-world applications requiring precise localization. Future improvements may focus on optimizing the model architecture to further enhance recall while maintaining high precision.

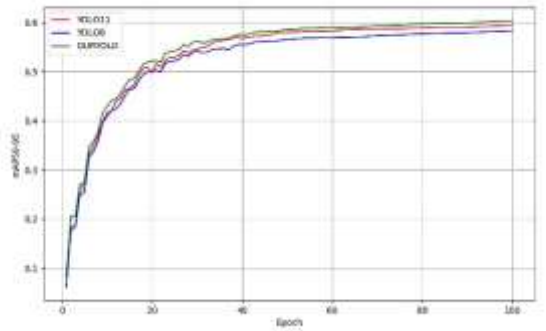

Fig. 6.Comparison of mAP50-95 scores, illustrating the precision of different object detection methods, particularly in detecting small objects in aerial imagery

Figure 8 illustrates the objects detected by the YOLO11 algorithm on the DOTA dataset. Figure 9 showcases the results of our proposed algorithm on the same images, where the detected objects are clearly marked. Additionally, the differences between

figures 8 and 9 are also highlighted. To comprehensively evaluate the model's performance, Precision-Recall (P–R) curves were plotted for each category, as illustrated in figure 10. The area under these curves, referred to as Average Precision (AP), was calculated, with higher AP values indicating superior detection performance. Furthermore, two critical parameters in evaluating deep learning models are the Intersection-over-Union (IoU) threshold and the confidence threshold.

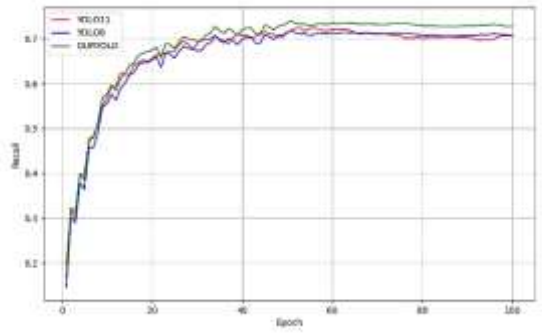

Fig. 7.Comparison of recall values, highlighting the detection sensitivity of various models, particularly in challenging aerial scenarios.

Fig. 8. Detected object by YOLO11.

Fig. 9. Detected object by our YOLO11.

To comprehensively evaluate the model's performance, Precision-Recall (P–R) curves were plotted for each category, as illustrated in figure 10. The area under these curves, referred to as Average Precision (AP), was calculated, with higher AP values indicating superior detection performance. Furthermore, two critical parameters in evaluating deep learning models are the Intersection-over-Union (IoU) threshold and the confidence threshold.

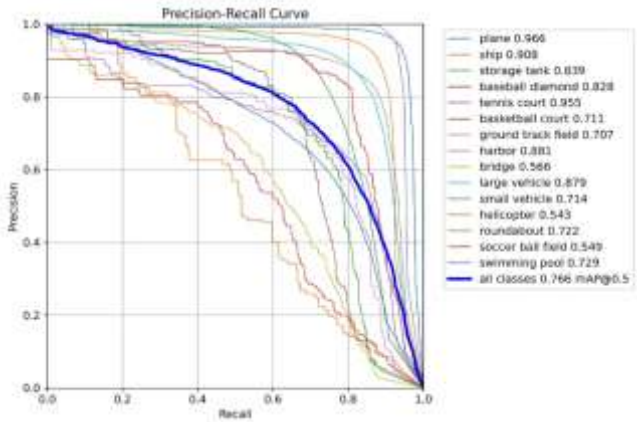

Fig. 10. Percision-Recall curve.

5-Ablation Experiments

To analyze the impact of our modifications, we conducted ablation experiments to evaluate the contribution of each component. Initially, integrating CBAM into YOLO 11 resulted in an improvement in mAP50 from 75.45\% to 75.93\%, showcasing the module's effectiveness in enhancing feature extraction. Building on this, we introduced an additional detection layer tailored for small object detection, further boosting mAP50 to 76.68\%. This progression highlights the synergistic effect of CBAM and the newly added module in refining detection performance. The results validate our approach in addressing small object challenges in aerial imagery, confirming that attention mechanisms and specialized detection layers can significantly enhance object detection capabilities.

Table 2: PERFORMANCE COMPARISON OF YOLOV11 VARIANTS, SHOWING THE IMPACT OF CBAM AND OUR PROPOSED MODIFICATIONS ON MAP AND RECALL.

Model	*mAP50*	*mAP50-95*	*Recall*
Yolo 11	*75.45*	*59.43*	*70.75*
Yolo11+CBAM	*75.93*	*59.72*	*71.32*
Our model	*76.68*	*60.35*	*73.12*

6- Conclusion

In this study, we presented an improved YOLO 11 architecture incorporating CBAM and additional detection layers to enhance small object detection in remote sensing images. Our modifications led to a significant increase in detection accuracy, achieving an mAP50 score of 76.68\% on the DOTA dataset, outperforming the baseline YOLO 11 model. By integrating CBAM, we improved feature selection by emphasizing important regions while reducing background noise. Furthermore, the additional detection layer allowed for better localization of small-scale objects, addressing a critical challenge in aerial imagery analysis. These enhancements demonstrate the potential of incorporating attention mechanisms and multi-scale detection strategies for robust object detection in complex environments. Future work will explore further optimization techniques to enhance model efficiency and accuracy while reducing computational costs.

References

[1] A. Sandooghdar and F. Yaghmaee, "Deep Learning Approach for Cardiac MRI Images," Journal of Information Systems and Telecommunication, vol. 10, no. 37, pp. 61–67, Dec. 2022, doi: 10.52547/JIST.16121.10.37.61.

[2] E. Gholami, S. Reza, K. Tabbakh, and M. Kheirabadi, "Diagnosis of Gastric Cancer via Classification of the Tongue Images using Deep Convolutional Networks."

[3] M. Navraan, N. M. Charkari, and M. Mansoorizadeh, "Automatic facial emotion recognition method based on eye region changes,"

Journal of Information Systems and Telecommunication, vol. 4, no. 4, pp. 221–231, Sep. 2016, doi: 10.7508/JIST.2016.04.003.

[4] K. Rezaee, S. J. Mousavirad, M. Rasegh Ghezelbash, and J. Haddania, "Accurate fire detection system for various environments using Gaussian mixture model and HSV space," Journal of Information Systems and Telecommunication, vol. 1, no. 1, pp. 47–54, Dec. 2013, doi: 10.7508/JIST.2013.01.007.

[5] W. Han et al., "Methods for Small, Weak Object Detection in Optical High-Resolution Remote Sensing Images: A survey of advances and challenges," IEEE Geosci Remote Sens Mag, vol. 9, no. 4, pp. 8–34, Dec. 2021, doi: 10.1109/MGRS.2020.3041450.

[6] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 9, pp. 3446–3456, Sep. 2010, doi: 10.1109/TGRS.2010.2046330.

[7] N. Proia and V. Pagé, "Characterization of a bayesian ship detection method in optical satellite images," IEEE Geoscience and Remote Sensing Letters, vol. 7, no. 2, pp. 226–230, Apr. 2010, doi: 10.1109/LGRS.2009.2031826.

[8] J. Xu, X. Sun, D. Zhang, and K. Fu, "Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized hough transform," IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 12, pp. 2070–2074, 2014, doi: 10.1109/LGRS.2014.2319082.

[9] F. Yang, Q. Xu, F. Gao, and L. Hu, "Ship detection from optical satellite images based on visual search mechanism," International Geoscience and Remote Sensing Symposium (IGARSS), vol. 2015-November, pp. 3679–3682, Nov. 2015, doi: 10.1109/IGARSS.2015.7326621.

[10] Y. Yao, Z. Jiang, H. Zhang, M. Wang, and G. Meng, "Ship detection in panchromatic images: a new method and its DSP implementation," https://doi.org/10.1117/12.2234677, vol. 9901, pp. 165–170, Mar. 2016, doi: 10.1117/12.2234677.

[11] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," IEEE International Conference on Image Processing, vol. 1, 2002, doi: 10.1109/ICIP.2002.1038171.

[12] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sens Environ, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/J.RSE.2017.06.031.

[13] D. Li, Y. Ke, H. Gong, and X. Li, "Object-Based Urban Tree Species Classification Using Bi-Temporal WorldView-2 and WorldView-3 Images," Remote Sensing 2015, Vol. 7, Pages 16917-16937, vol. 7, no. 12, pp. 16917–16937, Dec. 2015, doi: 10.3390/RS71215861.

[14] G. S. Xia et al., "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3974–3983, Dec. 2018, doi: 10.1109/CVPR.2018.00418.

[15] T. Y. Lin et al., "Microsoft COCO: Common objects in context," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[16] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.

[17] J. S. D. R. G. A. F. Redmon, "(YOLO) You Only Look Once," Cvpr, vol. 2016-December, pp. 779–788, Dec. 2016, doi: 10.1109/CVPR.2016.91.

[18] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11211 LNCS, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1/TABLES/8.

[19] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.17725

[20] T. Nguyen, E. A. Park, J. Han, D. C. Park, and S. Y. Min, "Object Detection Using Scale Invariant Feature Transform," Advances in Intelligent Systems and Computing, vol. 238, pp. 65–72, 2014, doi: 10.1007/978-3-319-01796-9_7.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. I, pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177.

[22] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3951 LNCS, pp. 404–417, 2006, doi:

10.1007/11744023_32.

[23] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” Proceedings of the IEEE International Conference on Computer Vision, pp. 2564–2571, 2011, doi: 10.1109/ICCV.2011.6126544.

[24] A. Rosenfeld, “The Max Roberts Operator is a Hueckel-Type Edge Detector,” IEEE Trans Pattern Anal Mach Intell, vol. PAMI-3, no. 1, pp. 101–103, 1981, doi: 10.1109/TPAMI.1981.4767056.

[25] F. Ulupinar and G. Medioni, “Refining edges detected by a LoG operator,” Comput Vis Graph Image Process, vol. 51, no. 3, pp. 275–298, Sep. 1990, doi: 10.1016/0734-189X(90)90004-F.

[26] Y. Zhang, X. Han, H. Zhang, and L. Zhao, “Edge Detection Algorithm of Image Fusion Based on Improved Sobel Operator,” Proceedings of 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference, ITOEC 2017, vol. 2017-January, pp. 457–461, Nov. 2017, doi: 10.1109/ITOEC.2017.8122336.

[27] C. G. Harris and M. Stephens, “A Combined Corner and Edge Detector,” Alvey Vision Conference, pp. 23.1-23.6, Apr. 1988, doi: 10.5244/C.2.23.

[28] R. Girshick, J. Donahue, T. Darrell, J. Malik, U. C. Berkeley, and J. Malik, “1043.0690,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, p. 5000, Sep. 2014, doi: 10.1109/CVPR.2014.81.

[29] R. Girshick, “Fast R-CNN,” in 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Dec. 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.

[30] 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.

[31] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 936–944, Nov. 2017, doi: 10.1109/CVPR.2017.106.

[32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 2980–2988, Dec. 2017, doi: 10.1109/ICCV.2017.322.

[33] M. L. Ali and Z. Zhang, “The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection,” Dec. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/computers13120336.

[34] W. Liu et al., “SSD: Single shot multibox detector,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2/FIGURES/5.

[35] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11218 LNCS, pp. 765–781, 2018, doi: 10.1007/978-3-030-01264-9_45/TABLES/4.

[36] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” Proceedings of the IEEE International Conference on Computer Vision, vol. 2019-October, pp. 6568–6577, Oct. 2019, doi: 10.1109/ICCV.2019.00667.

[37] C. Zhu, Y. He, and M. Savvides, “Feature selective anchor-free module for single-shot object detection,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June, pp. 840–849, Jun. 2019, doi: 10.1109/CVPR.2019.00093.