ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

YOLO-DDE: A Method for Small Ship Detection in SAR Images

Shentao Wang[1], Byung-Won Min[2], Yue Hong[1], Rui Li[1]

[1] College of Yonyou Digital & Intelligence, Nantong Institute of Technology, Nantong 226000, China

[2] Division of Information and Communication Convergence Engineering, Mokwon University, Daejeon 35242, Korea

Abstract

Aiming at the problems of small targets, weak feature expression and strong background interference in satellite SAR images, this paper proposes a new target detection method based on the improved YOLOv11s architecture to improve the detection performance. The algorithm adds multi-scale convolutional blocks (MSCB) and adopts deformable attention mechanism (DAttention) to expand the receptive field of the backbone network. In addition, an efficient multi-scale attention module (EMA) is integrated in the small target detection layer of the feature fusion network to improve the accuracy of feature fusion. These improvements significantly enhance the network's ability to detect small-sized targets. The effectiveness of the proposed method is evaluated on a public SAR image dataset: High Resolution SAR Image Dataset (HRSID). Experimental results show that compared with YOLOv11, the proposed method YOLO-DDE improves the average precision (AP50) by 2.1% on the HRSID dataset.

Keywords

Target Detection; YOLOv11; SAR Imagery; Adaptively Spatial Feature Fusion.

1. Introduction

With the deepening of globalization and the continuous evolution of the geopolitical landscape, the strategic importance of the ocean in national planning has become increasingly prominent. In particular, under the ongoing advancement of the Belt and Road Initiative and the continued implementation of the “Maritime Power” strategy, the ocean has not only served as a vital space for resource development but has also progressively evolved into a core platform supporting national security, foreign cooperation, and economic growth. In the context of increasingly frequent maritime trade and escalating regional complexities, the ability to accurately grasp maritime dynamics and rapidly identify key targets has become one of the critical tasks in safeguarding national maritime interests and promoting regional stability[1].

As the primary carriers of maritime activity, the number, type, distribution, and movement patterns of vessels encapsulate vast information reflective of maritime conditions. These data are indispensable for strategic assessment and situational awareness. Traditionally, ship detection has relied on manual patrols, coastal radars, and vision systems based on optical imagery. While these methods remain effective in certain scenarios, they commonly face limitations such as restricted coverage, susceptibility to weather and lighting conditions, lack of real-time performance, and high operational costs[2]. For instance, under nighttime or adverse weather conditions, the imaging quality of optical systems often degrades significantly, directly impairing the continuity and reliability of surveillance. Meanwhile, manual patrols and nearshore radars are constrained by line-of-sight and human resource limitations, rendering them inadequate in large-scale, high-frequency, and all-weather detection tasks.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

Against this backdrop, Synthetic Aperture Radar (SAR) has emerged as a robust technological tool for maritime target monitoring, owing to its all-weather, all-day imaging capabilities[3]. Ships typically appear as bright spots in SAR images, creating distinguishable contrasts against the sea surface background and thus offering potential for automatic detection and recognition. However, the unique characteristics of SAR images also introduce a range of challenges. For example, the prevalent speckle noise can significantly obscure object contours; geometric distortions and sea clutter may mask critical features and reduce detection accuracy. Additionally, ships exhibit diverse forms and scales under different environments-particularly in long-range, complex background, or densely populated scenes-making it difficult to accurately delineate ship boundaries[4].

Moreover, the deformable nature of ships, compounded by the dynamic and complex maritime environment, contributes to the inherent uncertainty of the detection task. Under harsh sea conditions, factors such as wave surges, vapor interference, and coastal reflections may lead to false detections or missed targets. In densely populated maritime areas, occlusion and fusion among vessels can blur edges, making accurate segmentation even more challenging. Traditional detection approaches based on handcrafted feature extraction often depend on fixed templates, which struggle to adapt to the diversity of target shapes and varying scenes. Consequently, achieving efficient and reliable ship detection in SAR imagery-particularly under conditions of complex interference and unstable features-has become a pressing technical problem in fields such as ocean remote sensing, border security, and maritime administration[5].

Based on the above analysis, this paper proposes an improved model based on YOLOv11s. The proposed network integrates a Multi-Scale Convolution Block (MSCB), a Deformable Attention mechanism (DAttention), and an Efficient Multi-scale Attention (EMA) module. The original C2f module in YOLOv11s is replaced by the MSCB module to enhance multi-scale feature extraction capabilities. The DAttention module is introduced to expand the receptive field of the backbone network. Additionally, the EMA module is integrated into the small-object detection layer of the feature fusion network to improve feature fusion accuracy. The main contributions of this work are summarized as follows:

We propose a Multi-Scale Convolution Block (MSCB), which combines depthwise separable convolution with multi-scale convolutional kernels. This design enhances the network's capability to extract multi-scale features. To further improve detection performance, the original C3f2 module in YOLOv11 is replaced with the MSCB module.

We incorporate a Deformable Attention mechanism (DAttention) into the model, which introduces deformable attention and dynamic sampling points to effectively expand the receptive field of the backbone network.

We introduce an Efficient Multi-scale Attention (EMA) module into the network. This module restructures part of the channel dimension into the batch dimension and groups the channels into multiple sub-features. It significantly improves the model's ability to detect small objects.

2. Related Work

2.1 Object Detection

As one of the fundamental tasks in computer vision, object localization and classification not only involves identifying the categories of targets present in an image but also requires precise determination of their spatial positions, typically represented by bounding boxes. Compared with traditional image classification tasks, this process is inherently more complex, as it encompasses both classification and regression components.

With the rapid advancement of deep learning,significant progress has been made in visual target detection techniques. Depending on the detection pipeline, existing approaches can generally be categorized into two main types: two-stage frameworks and single-stage frameworks, each offering distinct detection mechanisms and practical advantages [6].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

Single-stage detectors offer a notable advantage in terms of computational efficiency. These methods reformulate the object detection task as an end-to-end regression problem, directly predicting both object categories and locations from the feature maps of the input image, eliminating the need for region proposal generation. Representative algorithms in this category include the Single Shot MultiBox Detector (SSD)[7] and the You Only Look Once (YOLO)[8] family. In particular, YOLO partitions the input image into a fixed grid, with each grid cell responsible for predicting objects within its region, significantly improving detection speed. Owing to their streamlined architecture and lower computational overhead, single-stage detectors are especially well-suited for deployment in time-sensitive applications such as autonomous driving, mobile device recognition, and video surveillance.

As the YOLO series continues to evolve, improvements in detection accuracy have been achieved while maintaining high inference speed. Starting from YOLOv5, substantial enhancements have been made to the network architecture, feature fusion strategies, and loss function design. The most recent version, YOLOv11, incorporates a multi-scale feature fusion mechanism, attention modules, and a lightweight design. These upgrades not only strengthen the model's capability in detecting small-scale objects but also improve its robustness in complex backgrounds, further narrowing the performance gap between single-stage and two-stage detectors in terms of accuracy.

2.2 YOLOv11

YOLO (You Only Look Once) is a deep learning-based real-time object detection algorithm that achieves fast detection by dividing the input image into a grid, offering both high detection speed and accuracy. The architecture of YOLOv11 generally comprises the following core components:

Backbone: The backbone serves as the fundamental feature extractor of the network. It utilizes a convolutional neural network (CNN) to extract multi-scale features from the input image. In YOLOv11, an improved CSPDarknet architecture is employed to capture hierarchical feature maps more effectively.

Neck: The neck serves as a critical intermediary component linking the backbone and the detection head. Its main role is to refine and integrate multi-scale feature maps extracted by the backbone, enhancing the network's ability to capture and represent semantic information across varying spatial resolutions.

Head: The head is responsible for the final prediction stage. It adopts a decoupled structure to independently perform predictions on bounding boxes, class probabilities, and object confidence scores, based on the multi-scale features provided by the neck.

Loss Functions: The loss function is a critical component used to evaluate the discrepancy between the model's predictions and the ground truth. It typically consists of three parts:

Classification Loss, which measures the difference between predicted class probabilities and true labels;

Localization Loss, which quantifies the positional error between predicted and ground truth bounding boxes;

Objectness Loss, which assesses the confidence of the model in correctly identifying the presence of an object.

These loss components are jointly optimized, with appropriate weighting, to enhance the overall detection performance of the model.

The network structure of YOLOv11 is shown in Figure 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

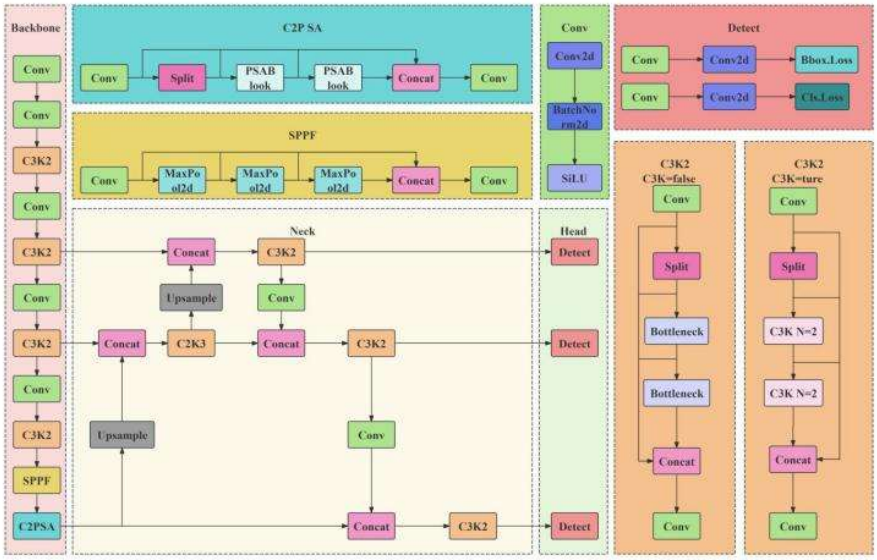

Figure 1. YOLOv11 Model Architecture

3. Method

In the context of SAR ship detection tasks, although YOLOv11 achieves a favorable trade-off between detection speed and accuracy, it still exhibits several limitations when confronted with the unique challenges posed by SAR imagery. Firstly, YOLOv11's ability to extract multi-scale features remains insufficient for effectively detecting small-scale ship targets, which are prevalent in SAR images, leading to frequent missed detections. Secondly, due to the presence of strong speckle noise and complex backgrounds-such as ocean waves, ports, and coastal interference-YOLOv11 shows limited capability in distinguishing targets from cluttered scenes, often resulting in false positives. Moreover, its fixed-grid prediction mechanism can introduce localization errors when ship targets are irregularly oriented or located near grid boundaries. As a general-purpose detection framework originally designed for optical images, YOLOv11 also lacks adaptation to the distinctive structural and scattering characteristics of SAR data, which undermines its feature representation effectiveness. Although the model integrates lightweight multi-scale feature fusion modules and attention mechanisms, these components are still inadequate when dealing with the significant scale variation of ship targets and the high complexity of SAR scenes, thereby limiting the overall robustness and accuracy of the model. Therefore, further architectural enhancements are necessary to adapt YOLOv11 more effectively to SAR-specific detection challenges.

3.1 MSCB

In this study, we introduce a Multi-Scale Convolution Block (MSCB)[9] designed to enhance the network's ability to extract and represent features across multiple spatial scales, particularly for small object detection in complex imaging scenarios such as SAR data. The MSCB architecture is inspired by the inverted residual block of MobileNetV2, but it is specifically tailored to improve multi-scale feature extraction efficiency. It begins by expanding the input channels through a point-wise convolution, followed by the application of parallel depth-wise convolutions with multiple kernel sizes (e.g., 1×1, 3×3, and 5×5) to capture local and global contextual information. To mitigate the limitation of depth-wise convolution in modeling inter-channel dependencies, a channel shuffle operation is introduced, allowing effective information exchange between feature groups. Finally, a second point-wise convolution restores the original number of channels and enables further feature transformation. By integrating this block, the network achieves a better balance between computational efficiency and representational capacity, making it well-suited for scenarios with significant scale variation and high background complexity. The MSCB serves as a foundational component to enhance subsequent attention mechanisms and feature fusion modules, ultimately improving overall detection performance.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

3.2 DAttention

In this work, we incorporate a Deformable Attention (DAttention)[10] mechanism into the network architecture to enhance its adaptability to spatially variant features and improve the detection of targets under complex imaging conditions, such as those found in SAR images. Unlike conventional attention mechanisms that compute dense attention over fixed, regular grids, DAttention dynamically learns sparse and content-adaptive sampling locations, allowing the model to focus on semantically relevant regions while reducing computational overhead. This is particularly beneficial in scenarios where object shapes, orientations, or spatial distributions vary significantly. DAttention integrates the strengths of both convolutional and self-attention approaches: it retains spatial inductive biases while enabling long-range dependency modeling. By assigning learnable offsets and attention weights to a small number of key positions, it effectively captures contextual information with minimal redundancy. The integration of DAttention into the backbone network expands its receptive field and improves its capacity to capture fine-grained and spatially diverse features, thereby enhancing overall detection robustness-especially in cluttered backgrounds and for small or irregularly shaped targets.

3.3 EMA

To improve the detection performance for small-scale objects and enhance the quality of feature representations, we incorporate the Efficient Multi-Scale Attention (EMA) module into the network's feature fusion stage[11]. Unlike traditional attention mechanisms that typically compress channel dimensions, EMA maintains richer semantic detail by partially reshaping the channel dimension into the batch dimension. The module is composed of parallel processing branches: a 1×1 convolution path designed for capturing detailed local information, and a 3×3 convolution path aimed at gathering wider contextual cues. By leveraging grouped feature division, EMA distributes spatial features across multiple subspaces, facilitating more effective modeling of both local and global dependencies. Additionally, it introduces a cross-spatial interaction mechanism that merges attention features through matrix-based operations, allowing fine-grained pixel-level context learning. This design enables the model to better concentrate on informative regions, especially in cluttered environments. As a result, EMA contributes to more precise detection of small targets while retaining high computational efficiency, making it well-suited for real-time visual recognition applications.The EMA structure is shown in Figure 2.

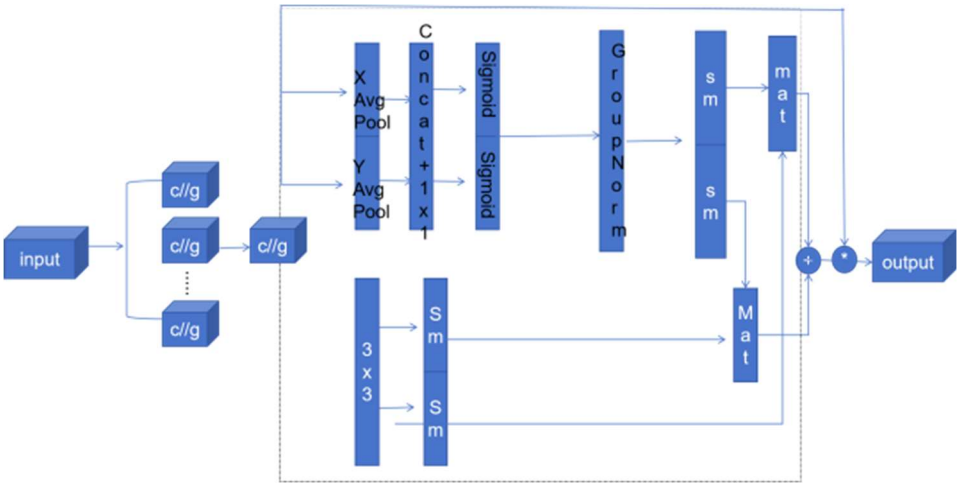

Figure 2. EMA structure

4. Experiment

4.1 Dataset and Experimental Settings

The High-Resolution Ship Instance Dataset (HRSID)[12] is a publicly available dataset for ship detection and instance segmentation in optical remote sensing imagery. Developed by Wuhan

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

University, the dataset contains a large number of high-resolution images featuring diverse and complex background scenes. It provides precise annotations of ship instances, making it well-suited for evaluating the performance of detection and segmentation algorithms in challenging environments.

In order to ensure the effectiveness of the improvement measures based on the YOLOv11 algorithm, all experimental environments in this study have the same configuration and model training parameters. The specific training parameter settings are shown in Table 1.

The experiments were conducted on a high-performance computing platform configured with Ubuntu 22.04 as the operating system and Python 3.9 as the programming environment. The hardware setup includes an Intel(R) Xeon(R) Platinum 8481C CPU and 90 GB of RAM, ensuring sufficient computational capacity for training and evaluation. For deep learning tasks, the PyTorch 2.1.0 framework was utilized. Additionally, a NVIDIA GeForce RTX 4090 GPU was employed to accelerate model training and inference, providing strong support for large-scale computations and complex neural network operations.

Table 1. The training parameters

Parameter Type	Value or Type
imgsize	640
epochs	500
batch	32
optimizer	Stochastic gradient descent (SGD)

4.2 Experimental Results

This experiment uses the High-Resolution Ship Instance Dataset as the experimental dataset.The ablation experiment results are shown in Table 2.

Table 2. Ablation experimen Results

	mAP50/%	mAP50-95/%	GFLOPs/G	Parameters/M,
YOLOv11s	89.7	63.2	21.5	9.43
YOLOv11s+MSCB	90.7	63.8	20.6	9.25
YOLOv11s+DAttention	90.5	64.1	21.6	9.51
YOLOv11s+EMA	91.0	64.4	21.7	9.43
Ours	**91.8**	**65.0**	**21.2**	**9.48**

Table II presents the results of ablation experiments evaluating the contribution of each proposed module-MSCB, DAttention, and EMA-to the performance of the baseline YOLOv11s model. The evaluation metrics include mAP@50 (mAP50), mAP@50-95 (mAP50-95), GFLOPs, and the number of parameters .

The baseline YOLOv11s achieves an mAP50 of 89.7% and mAP50-95 of 63.2%. After integrating the MSCB module, which enhances multi-scale feature extraction, the model sees improvements in both accuracy metrics, reaching 90.7% mAP50 and 63.8% mAP50-95, with a slight reduction in computation to 20.6 GFLOPs.

Incorporating the DAttention module further boosts detection performance, achieving 90.5% mAP50 and 64.1% mAP50-95, indicating its effectiveness in capturing spatially adaptive features through deformable attention mechanisms.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

The addition of the EMA module also leads to performance gains, with mAP50 rising to 91.0% and mAP50-95 to 64.4%, alongside a modest increase in computational cost to 21.7 GFLOPs.

Finally, the proposed full model (Ours), which combines all three modules, achieves the highest performance, with 91.8% mAP50 and 65.0% mAP50-95. Notably, this performance gain is achieved with only 21.2 GFLOPs and 9.48M parameters, demonstrating that the integrated design effectively improves detection accuracy while maintaining computational efficiency.

These results confirm that each module contributes positively to the model's performance, and their combination yields a synergistic effect, particularly enhancing the detection of small and challenging targets.

5. Conclusion

In this work, we propose an enhanced object detection framework built upon the YOLOv11s architecture, aiming to improve the detection of small and complex targets in remote sensing imagery. To overcome the limitations of conventional models in handling multi-scale objects and cluttered backgrounds, we introduce three key modules: the Multi-Scale Convolution Block (MSCB), the Deformable Attention mechanism (DAttention), and the Efficient Multi-Scale Attention (EMA) module.

Extensive ablation studies conducted on the HRSID dataset validate the individual effectiveness of each component and demonstrate the complementary benefits achieved through their integration. Compared to the baseline YOLOv11s, the proposed model achieves notable performance gains in both mAP@50 and mAP@50–95 metrics, while maintaining a lightweight structure and low computational overhead. Specifically, our final model achieves 91.8% mAP@50 and 65.0% mAP@50–95, outperforming all baseline configurations.

These results confirm that the proposed enhancements significantly improve detection accuracy, particularly for small and densely distributed targets, without compromising model efficiency. This makes the framework well-suited for real-time deployment in complex remote sensing scenarios. Future work will explore extending the model to multi-modal inputs and incorporating temporal dynamics for video-based object detection applications.

References

[1] Zhang C, Liu P, Wang H, et al. A review of recent advance of ship detection in single-channel SAR images[J]. Waves in Random and Complex Media, 2023, 33(5-6): 1442-1473.

[2] Luo R, Chen L, Xing J, et al. A fast aircraft detection method for SAR images based on efficient bidirectional path aggregated attention network[J]. Remote Sensing, 2021, 13(15): 2940.

[3] Chang S, Deng Y, Zhang Y, et al. An advanced echo separation scheme for space-time waveform-encoding SAR based on digital beamforming and blind source separation[J]. Remote Sensing, 2022, 14(15): 3585.

[4] Yang, X., Sun, H., Fu, K., et al. Ship detection from remote sensing images using domain adaptive YOLO network. Remote Sensing, 2020, 12(3): 307.

[5] Marino, A., Dalla Mura, M., & Poggi, G. A novel approach for ship detection in SAR imagery based on wavelet domain and fuzzy logic. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(5): 2703–2714.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202507_11(7).0027

[8] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[9] Rahman M M, Munir M, Marculescu R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 11769-11779

[10] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4794-4803.

[11] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[12] Wei S, Zeng X, Qu Q, et al. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation[J]. Ieee Access, 2020, 8: 120234-120254.