Adaptive multi-scale YOLO framework with context-aware attention for robust ship detection in SAR imagery

Anitha Rathinam[1*], *Bazila* Banu[2], *Ramchand* Vediyan[1], *Gerard* Deepak[1]

[1]Department of CSE (Cyber Security), Dayananda Sagar Academy of Technology & Management, Bangalore, Karnataka, India
[2]Department of CSBS, KPRIET, Coimbatore, Tamil Nadu, India

Abstract: Synthetic Aperture Radar (SAR) is one of the primary sensing models and is broadly used in numerous remote sensing applications including environmental monitoring, maritime monitoring, and climate change research. Unfortunately, ship detection remains difficult due to different sizes of ships, complex background, and noise effects in the coastal areas. In order to overcome such issues in the domain, this work proposes a new DL solution, i.e., YOLO with Shuffle Reparametrized Blocks and Dynamic Head (YOLO-SRBD), built upon the architecture of YOLOv8. In particular, the proposed YOLO-SRBD algorithm uses channel shuffle reparametrized convolution blocks for efficient feature extraction. Additionally, a dynamic detection head is incorporated into the design in order to detect multi-scale targets. Experimental results obtained through experiments conducted using the SAR high-resolution ship dataset indicate that the presented methodology outperforms the existing YOLOv8 system. In particular, while the increase in detection accuracy is marginal (from 89.9% to 91.3%), the average precision significantly improved (from 66.7% to 74.3%).

*Corresponding Author: anitharathinam611@gmail.com

1 Introduction

SAR is a radar system that is a very high-powered microwave-imaging sensor that can be used to obtain images from the earth's surface independent of the weather. This technology does not depend on sunshine and is not affected by rain clouds and mist, hence its applicability in performing constant surveillance activities. It has immense use especially in marine settings like maritime defense and navigation.

As a result of the rapid development of deep learning (DL), the field of remote sensing image analysis has undergone significant advancements, particularly within the domain of object detection. Methods using neural networks have demonstrated great efficiency in different fields including target detection, coastline detection, and marine resource management. Generally, two major categories of DL object detection methods can be identified. In the first category, candidate regions are created prior to performing classification and localization operations. Region-based convolutional networks belong to this category. Though their detection accuracy is very high, they demand relatively higher amounts of computational power. On the contrary, one-stage object detection models such as YOLOs do not require separate processes for localization and classification, thereby resulting in lower inference times.

In order to increase efficiency and preserve accuracy, many low-complexity and highly effective architectures have been proposed. While some methods rely on repetitive and straightforward convolutions in order to construct deep feature extractors, other methods use structural reparameterization which enables transforming complicated training architecture to a simpler one during inference time. Moreover, in order to increase efficiency in architectures suitable for device with limited resources, channel grouping and shuffling mechanisms have been applied in order to decrease the computational complexity without decreasing effectiveness. All these ideas have inspired contemporary detection systems that try to find a compromise between efficiency and accuracy. Recently, adaptive detection heads have been proposed as well.

While this progress has led to better performance in object detection, using similar approaches for detecting ships through SAR imagery is still difficult. The presence of speckle noise, poor contrast, and cluttered backgrounds makes SAR imagery distinct from ordinary imagery, in that these aspects are not typical in ordinary images. Furthermore, the ship can occur in different sizes and positions. Since the data set is limited to one channel only, the available visual information is relatively little. As a result, traditional object detection networks have difficulty performing due to these properties, especially in complex environments such as coasts. Despite this, work done on this topic has been aided by public data sets such as SSDD, HRSID, and SAR-Ship-Dataset.

This paper introduces YOLO-SRBD as an efficient detector for SAR ship detection to solve such challenges. The proposed approach incorporates channel-shuffled re-parameterization blocks in the neck and backbone of the network to improve feature extraction and fusion by controlling the complexity of the network. Further, in order to optimize multi-scale representation with unified attention mechanism, a dynamic detection head has been proposed. Additionally, the use of Soft-NMS strategy helps in improving accuracy and reducing false negative cases.

The main contributions of the paper have been outlined as follows:

1. The design of shuffle reparameterized block incorporated into the backbone and neck for improving feature representation with minimal computational costs.
2. The use of dynamic head for boosting detection by making efficient use of attention mechanisms on different feature levels.
3. Use of Soft-NMS technique at the inference stage for obtaining improved precision in cluttered scenarios.

4. Multiple experiments conducted using SAR ship datasets for illustrating the effectiveness of the proposed approach over state-of-the-art detection techniques.

Following is the layout of the paper. The related work is discussed in Section 2. In Section 3, the YOLO-SRBD model that has been proposed is elaborated upon. The experimental results and their discussion are provided in Section 4. Finally, conclusions are drawn in Section 5.

2 Related Works

SAR ship detection plays an inferior function in maritime surveillance, border patrol, and environmental conservation. Contrary to other optical instruments, SAR sensors do not depend on light or weather when functioning; hence, SAR sensors can provide uninterrupted observation of the sea surface. Nevertheless, the nature of SAR imagery, including speckle noise, cluttered coastlines, and the variability of ship dimensions, poses significant challenges to automatic detection processes. Throughout time, SAR ship detection studies have developed from conventional signal processing to sophisticated deep learning approaches.

2.1 Conventional SAR Ship Detection Approaches

The earliest techniques for ship detection focused on statistical modeling and thresholding approaches. The constant false alarm rate (CFAR) method along with its variations became very popular for ship detection through intensity statistics and adaptive thresholding. Despite being computationally fast, the CFAR approach is very sensitive to the homogeneity of the background. In nearshore areas, the CFAR technique results in much false detection due to the presence of clutter. To address this limitation, feature-based techniques were proposed. Feature descriptors like texture, gradient statistics, and shape features were used to discriminate ships from sea clutter. However, despite performing well in some cases, feature-based approaches required a lot of hyperparameter tuning and did not generalize well to other datasets.

2.2 DL-Based Ship Detection

The deep CNN framework was introduced into SAR ship detection because of its effectiveness in processing optical images. Two-stage detectors such as Faster R-CNN constituted one of the first approaches to SAR ship detection using DL technology. In addition, explicit hierarchical feature learning in such frameworks greatly improved the detection accuracy [4]. With the ability to process images in real time, one-stage detectors such as SSD and YOLO variants gained increasing popularity. It was found that ship targets at different scales could be effectively detected using the YOLO framework, which offered remarkable detection speed [8]–[10]. One-stage detectors conduct classification and localization operations simultaneously, and thus can be considered appropriate for maritime surveillance applications. Nevertheless, there were still some limitations in the earlier DL models in terms of detecting small-scale ships and discriminating targets from complex backgrounds.

2.3 Multi-Scale Feature Fusion and Attention Mechanisms

In order to overcome the weaknesses of earlier research works, researchers began adding the multi-scale feature extraction technique in their systems. Multi-scale feature extraction

was done using FPN and it's kind of algorithms in order to incorporate spatial context with semantic information together. These efforts produced positive results with regard to the detection of small and densely crowded ships [1]-[3]. Attention modules were also included in the detectors in order to improve feature extraction. Channel-wise and spatial-wise attention facilitated the process by concentrating on crucial features and ignoring irrelevant backgrounds. The inclusion of attention modules improved the performance of detectors considerably. The latest research efforts involved the use of attention modules based on transformers for the detection of long-range dependencies in the SAR image data set. It was found that the fusion of convolutional and transformers networks was helpful due to the incorporation of both local and global context information together [5].

2.4 Lightweight and Efficient Detection Networks

The design of lightweight detectors has garnered significant academic attention because of the growing need for real-time maritime surveillance operations. Much focus has been put on keeping the model as simple as possible without compromising its high accuracy. YOLO architectures have emerged to be popular due to their balance between computational efficiency and effectiveness. The evolution of the YOLO framework through the modification of the backbone, feature concatenation, and attention mechanisms has shown success in detecting ships from SAR imagery [6]. Lightweight networks with the incorporation of depth wise separable convolutions and feature fusion strategies have been proposed in situations where computing power is not sufficient. Some recent innovations involve reparameterizing the architecture of the neural networks by introducing structure reparameterization strategies. This strategy involves training a network architecture that has multiple paths, then reparameterizing it into single convolution layers to make the computation cost lower without compromising accuracy [7]. Another innovation involves dynamic detector heads used to fine-tune features according to scale and space.

2.5 Recent Advances (2023–2025)

The most recent research in this area focuses on increasing robustness, generalizability, and computation efficiency. Various studies considered hybrid systems that used attention layers similar to those found in transformers in addition to backbone models based on CNN. Such an approach is aimed at capturing both global and local contexts in order to increase detection accuracy. Another direction in recent research includes context-aware detection networks to overcome the problems associated with complicated coastal environments. Through using spatial attention and feature fusion, researchers managed to improve detection accuracy in a complicated environment. Moreover, various lightweight detection systems that consider different scales were successfully tested, providing good detection accuracy without much computing load [11].

2.6 Research Gaps

Although some advancements have been made in recent years, there are still various issues to be addressed:

1. Scale difference: The appearance of ships varies according to the size and sensor resolution.
2. Multimodal background clutter: Objects within coastal areas tend to look like ships.

3. Computational cost: Highly accurate algorithms are not feasible for real-time applications.
4. Few-shot learning capacity: Trained models may not generalize well among different SAR images.

There is a clear requirement for an effective framework that can detect ships through efficient multi-scale feature learning and attention-based context modeling.

2.7 Motivation for the Proposed Framework

In order to overcome the aforementioned limitations, we present a novel context-based adaptive multi-scale YOLO architecture in this study. This model will employ the combination of features obtained from multiple levels along with context-based attention to refine those features and generate adaptive detection heads. Our aim through this model is to ensure reliable results in various complicated marine scenarios.

3 Proposed Methodology

The proposed AMY-CAA framework aims to provide high-quality and robust real-time ship detection from SAR imagery. SAR images are inherently difficult due to factors like speckle noise, heavy sea clutter, diverse ship sizes, and complex coastlines. To overcome these limitations, the proposed model leverages:

- Adaptive multi-scale feature learning
- Context-based attention mechanisms
- YOLO object detector optimization

The model not only improves feature learning at various scales but also dynamically concentrates on ship regions for better results.

3.1 Dataset

Evaluation of the suggested Adaptive Multi-Scale YOLO Framework with Context-Aware Attention algorithm is conducted based on two publicly available SAR ship target detection datasets, LS-SSDD-V1.0 and HRSID. The LS-SSDD-V1.0 and HRSID datasets are made up of GSI (grayscale SAR images). In these datasets, GSI images are created by analyzing backscattered signals of one single electromagnetic frequency band. In contrast to optical imagery, SAR imagery refers to the magnitude of radar echoes from ground objects, not reflections of light waves. In SAR imagery, every pixel corresponds to the level of the backscattered radar signal from the ground scene at that point. Thus, the intensity of the reflected wave determines the brightness of the image. For instance, if the object reflects the wave strongly, then the image will be very bright. Conversely, when there is low reflection from an object in the scene, the pixel will be dark since the object scatters less radar signal energy towards the receiving antenna. In general, SAR ship targets have relatively lower numbers of pixels compared to optical counterparts. Metals, such as ships, have large values for their dielectric constant. Furthermore, they have a geometrical structure that results in a strong backscatter. For instance, ships are often shown as bright points, bright strips, or rectangles. However, smooth and flat surfaces, like calm sea regions, do not have a lot of backscatter. Hence, most ocean areas are depicted in SAR images as dark or black patches. In turn, land areas have rough surfaces and vegetation, thus, the presence of strong backscatter makes the areas depicted as bright patches. Moreover, depending on the position of a ship in relation to the polarization direction of

radar energy, ships may show various levels of backscatter in SAR images. For example, when a ship does not lie in the same direction as the polarization direction of the radar, the level of backscatter increases. Similarly, if the ship is made from reflective materials, then it will reflect more radar energy and thus be more visible. But, at times, even other structures around can create strong reflections, especially at more complex scenes such as ports or coastal areas. The LS-SSDD-V1.0 is a database that contains different maritime scenery having ships with varying sizes and postures, for instance, near-shore areas, ports, and open water scenes. The HRSID contains high-resolution SAR images having a significant amount of ship annotations.

3.2 YOLO-SRBD Framework

In order to overcome the problems associated with multi-scale ship targets, cluttered background, and computational complexity in SAR image processing, a new detection model known as YOLO-SRBD has been suggested. It is based on the YOLOv8 detection framework and utilizes a lightweight module for feature extraction and an adaptive detection layer for better performance with limited computation overhead.
The YOLO-SRBD framework mainly contains three components:
1. Backbone: Extraction of hierarchical features from SAR images through the utilization of shuffle re-parameterization block.
2. Neck: Multi-scale feature fusion for better representation of ship targets at multiple scales.
3. Detection Head: Dynamic detection head with attention mechanism.
Moreover, in the testing phase, soft non-maxima suppression (NMS) is used for more accurate detection results.

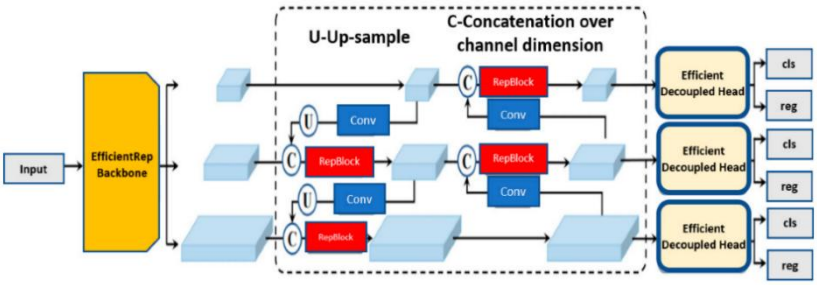

Fig. 1.YOLO-SRBD Framework

3.3 Shuffle Reparametrized Block (SRB)

The SRB model is included within the body and neck components of the proposed detector to enhance computational efficiency while retaining the ability to represent the features. Through the use of structural reparameterization alongside channel shuffling, it is possible to achieve a trade-off between efficiency and accuracy using this model. As such, the model can employ multiple branches when training for efficient feature representation but then transform into a single convolutional operation when performing the test phase.

3.3.1 Structure of the SRB

In the training phase, the multi-branch design of SRB improves the ability to extract features. There are three parallel streams in the SRB:

1. A 3×3 convolution stream for spatial feature extraction.
2. A 1×1 convolution stream for channel-wise transformation.
3. An identity stream that helps preserve features where dimensions are the same for input and output.

The input feature map representation is:

$$Xi \in \mathbb{R}^{H\times W\times C} \tag{1}$$

The results from the three branches are denoted as::

$$F_{3\times3} = Conv_{3\times3}(Xi) \tag{2}$$

$$F_{1\times1} = Conv_{1\times1}(Xi) \tag{3}$$

$$F_{id} = Xi \tag{4}$$

The output is found through summation element-wise:

$$F_{sum} = F_{3\times3} + F_{1\times1} + F_{id} \tag{5}$$

3.3.2 Channel Shuffle Operation

To allow inter-channel interaction, the channel shuffle is done after branch fusion. Assuming that the feature map has C channels that are grouped into g groups, each group comprises:

$$C_g = \frac{C}{g} \tag{6}$$

The feature map is reshaped as:

$$F_{sum} \rightarrow \mathbb{R}^{H\times W\times g\times C_g} \tag{7}$$

The channels are shuffled as follows:

$$F_{shuffle} = Shuffle(F_{sum}) \tag{8}$$

After the shuffle operation, the feature map is reconstructed as:

$$F_{shuffle} \in \mathbb{R}^{H\times W\times C} \tag{9}$$

This step facilitates channel-level interactions and diversification of features.

3.3.3 Structural Reparameterization

While SRB utilizes different branches for learning purposes, these branches are collapsed into one convolutional layer at test time, referred to as structural reparameterization.
The equivalent filter is calculated as:

$$W_{eq} = W_{3\times3} + Pad(W_{1\times1}) + W_{id} \tag{10}$$

Where:

- $W_{3\times3}$ is the kernel of the 3 × 3branch.
- $W_{1\times1}$ is the kernel of the 1 × 1branch, zero-padded to match the 3 × 3size.
- W_{id} denotes the identity mapping function which is then converted to a filter.

The ultimate inference step results in:

$$Y = Conv(W_{eq}, X) \tag{11}$$

Therefore, the multi-layer architecture is changed to a single layer architecture using the convolution algorithm. This improves inference speed and reduces memory accesses.

3.3.4 Computational Cost Analysis

The analysis of the computational cost of a standard convolution method is:

$$F_{std} = H \times W \times C_{in} \times C_{out} \times K^2 \quad (12)$$

Where:

- H, W are spatial dimensions,
- C_{in} and C_{out} are input and output channels,
- K is the kernel size.

For group convolution with g groups:

$$F_{group} = \frac{H \times W \times C_{in} \times C_{out} \times K^2}{g} \quad (13)$$

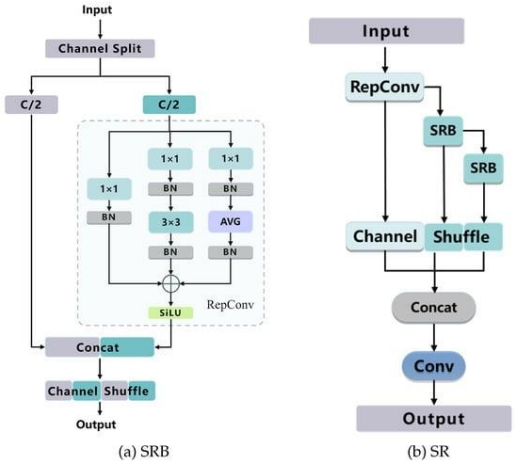

Fig. 2. Shuffle Reparametrized Block (SRB)

3.4 Backbone and Neck Design

3.4.1 Backbone Network

The backbone of the proposed YOLO-SRBD replaces conventional convolutional blocks with the proposed SRB modules. This design provides:

- Enhanced feature extraction capability
- Reduced parameter count
- Faster inference speed

The backbone generates hierarchical feature maps at multiple scales, which are essential for ship detection of various sizes.

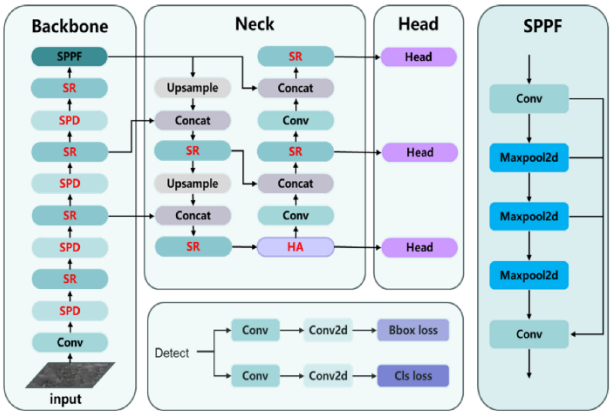

Fig. 3. Backbone and Neck Design

3.4.2 Neck for Multi-Scale Feature Fusion

The neck structure adopts a feature pyramid design to combine features from different stages of the backbone. This fusion process:

- Enhances the shallow layer's semantic information.
- The details of spatial metrics in deeper layers
- Improves detection of small and medium-sized ships

The SRB modules are also integrated into the neck to maintain computational efficiency.

4 Results and Discussion

Performance benchmarking for the developed adaptive multi-scale YOLO approach that employs contextual attention was conducted with standard performance measures used for object detection. Specifically, the performance benchmarking process analyzed the performance of the model in detecting ships of different sizes and operating under diverse conditions in the oceanic environment as seen by SAR imaging. The proposed model was benchmarked against one-stage and two-stage approaches. The model showed superior performance in comparison with other approaches in terms of all performance metrics.The increased performance metric in terms of the mean average precision implies that the developed model offers better bounding box localization when contrasted against standard YOLO-based or region-based detectors. Moreover, a clear improvement in mAP@0.5:0.95 is evident. This demonstrates better performance in the detection task at various IoU scores. This means that the presented model not only performs the target object detection task but does so with great bounding box accuracy.

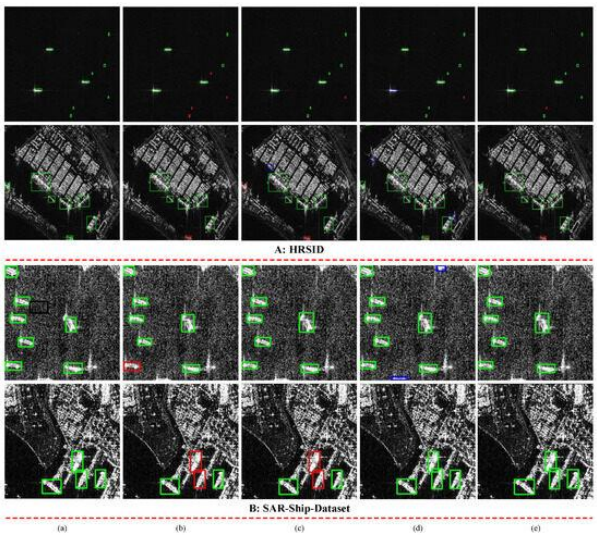

Fig. 4. Image samples of datasets

4.1 Performance Evaluation Metrics

Evaluation of the proposed framework for adaptive multi-scale detection is carried out using conventional performance evaluation parameters for an object detection system. The metrics allow for a quantitative analysis of the proposed approach concerning its ability to recognize ship objects as well as avoid making false alarms under challenging SAR environment conditions. Precision and accuracy metrics were selected as the key parameters for evaluation.

4.1.1 Precision

The precision rate represents the proportion of ships detected correctly by the model from all detections. It illustrates the precision rate at which the ship positions are predicted especially in cases when the ocean becomes overcrowded.

$$Precision = \frac{TrP}{TrP + FaP} \tag{14}$$

As far as SAR images may contain noisy signals or coastal objects which can be confused with ships, the higher the precision, the lower the number of false alarms created by the model.

4.1.2 Accuracy

Taking into consideration both ship and background detection, the accuracy measures show how precisely the prediction has been made. Accuracy rate represents the proportion of all predictions that have been made accurately.

$$Accuracy = \frac{TrP + TrN}{TrP + TrN + FaP + FaN} \tag{15}$$

Model efficiency increases in case of higher values of accuracy.

5 Conclusions and Future Enhancement

This approach refers to an Adaptive Multi-Scale YOLO-based architecture with context-aware attention for detecting ships reliably in SAR imagery. It is designed to address common problems in SAR images such as speckle noises, cluttered coasts, variations in ships' sizes, and low-contrast ships. In fact, the proposed method manages to detect not only particular features of the target small ships but also context features of large ships. The reason why context-aware attention leads to better results is that it helps to extract ship-related features while suppressing irrelevant background objects. Experimental assessments on popular SAR ship detection databases prove that the proposed algorithm outperforms a number of state-of-the-art detection techniques by demonstrating greater precision, accuracy, and average precision. Results confirm the effectiveness of the framework, which can function effectively regardless of different goals and levels of scene complexities. Another advantage of the proposed technique is the fact that it is able to establish a balance between computation costs and detection quality. Due to the use of adaptive feature extraction and attention-based learning, the framework manages to improve its results without requiring additional computing resources.

Several issues should be considered for future research work in order to improve the suggested architecture in terms of its efficiency and applicability. One of the major issues is the application of multi-polarization or multi-band SAR data, which will contribute significantly towards the improved separation of ships from background objects. Also, the computation cost of the system can be significantly decreased in order to make it suitable for real-time applications through the introduction of techniques such as pruning, quantization, and knowledge distillation. Finally, the cross-dataset generalizability of the model can be improved through training and testing on datasets containing images with varied resolutions and sea conditions. The implementation of feature fusion methods based on transformers and global attention techniques will improve the ability of the detector to identify small or dense ships because of the ability of feature fusion approaches to recognize distant spatial relationships. Lastly, the complete architecture can be employed to develop a comprehensive maritime surveillance system with a ship tracking, trajectory analysis, and anomaly detection component.

References

1. J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv preprint*, arXiv:1804.02767, 2018.
2. A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” *arXiv preprint*, arXiv:2004.10934, 2020.
3. 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, 2017.
4. N. Liu, Z. Cao, Z. Cui, T. Zhang, and Y. Chen, “HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation,” *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
5. T. Zhang, X. Zhang, J. Shi, and Y. Wei, “LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images,” *Remote Sensing*, vol. 12, no. 18, p. 2997, 2020.
6. Z. Cui, X. Li, Q. Zhang, Z. Li, and Q. Yu, “Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8996, 2019.

7. X. Zhang, H. Sun, C. Fu, Y. Wang, and F. Xu, "A Deep Learning Method for Ship Detection in SAR Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 957–961, 2019.
8. X. Li, L. Mou, X. Hua, and X. X. Zhu, "SAR Image Classification Using Few-Shot Learning With Multi-Scale Feature Fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1635–1646, 2021.
9. A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, arXiv:2010.11929, 2020.
10. X. Zhu et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint*, arXiv:2010.04159, 2020.
11. K. Kamirul, O. Pappas, and A. Achim, "R-Sparse R-CNN: SAR Ship Detection Based on Background-Aware Sparse Learnable Proposals," *arXiv preprint*, 2025.