DRC²-Net: A Context-Aware and Geometry-Adaptive Network for Lightweight SAR Ship Detection

Abdelrahman Yehia , Naser El-Sheimy * , Ashraf Helmy , Ibrahim Sh. Sanad , Mohamed Hanafy

Posted Date: 7 October 2025

doi: 10.20944/preprints202509.2458.v1

Keywords: SAR; ship detection; CNNs; YOLOX-tiny; receptive field; attention mechanism; deformable convnets

 doi:10.20944/preprints202509.2458.v1

Article

DRC²-Net: A Context-Aware and Geometry-Adaptive Network for Lightweight SAR Ship Detection

Abdelrahman Yehia[1], Naser El-Sheimy[2,*], Ashraf Helmy [3], Ibrahim Sh. Sanad[1], and Mohamed Hanafy[1]

[1] Department of Electrical and Computer Engineering, Military Technical College, Cairo 11766, Egypt
[2] Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
[3] Data Reception, Analysis and Receiving Station Affairs Division, National Authority for Remote Sensing and Space Sciences (NARSS), Cairo 1564, Egypt
* Correspondence: elsheimy@ucalgary.ca

Abstract

Synthetic Aperture Radar (SAR) ship detection remains challenging due to background clutter, target sparsity, and fragmented or partially occluded ships, particularly at small scales. To address these issues, we propose the Deformable Recurrent Criss-Cross Attention Network (DRC^2-Net), a lightweight and efficient detection framework built upon the YOLOX-Tiny architecture. The model incorporates two SAR-specific modules: a Recurrent Criss-Cross Attention (RCCA) module to enhance contextual awareness and reduce false positives, and a Deformable Convolutional Net-works v2 (DCNv2) module to capture geometric deformations and scale variations adaptively. These modules expand the Effective Receptive Field (ERF) and improve feature adaptability under complex conditions. DRC²-Net is trained on the SSDD and iVision-MRSSD datasets, encompassing highly diverse SAR imagery including inshore and offshore scenes, variable sea states, and complex coastal backgrounds. The model maintains a compact architecture with 5.05M parameters, ensuring strong generalization and real-time applicability. On the SSDD dataset, it outperforms the YOLOX-Tiny baseline with AP@50 of 93.04% (+0.9%), AP_s of 91.15% (+1.31%), AP_m of 88.30% (+1.22%), and AP_l of 89.47% (+13.32%). On the more challenging iVision-MRSSD dataset, it further demonstrates superior scale-aware detection, achieving higher AP across small, medium, and large targets. These results confirm the effectiveness and robustness of DRC^2-Net for multi-scale ship detection in complex SAR environments, consistently surpassing state-of-the-art detectors.

Keywords: SAR; ship detection; CNNs; YOLOX-tiny; receptive field; attention mechanism; deformable convnets

1. Introduction

Synthetic Aperture Radar (SAR) is a high-resolution active sensing technology capable of operating under all-weather, day-and-night conditions. By exploiting microwave backscatter rather than ambient light, SAR is inherently robust to atmospheric interference such as clouds, fog, and precipitation, making it highly effective for maritime surveillance and target detection in complex environments. Nevertheless, cluttered backgrounds, geometric distortions, and the sparse distribution of ship targets in SAR imagery pose significant challenges for conventional object detection frameworks, often resulting in reduced accuracy and limited generalization. These constraints underscore the need for lightweight, context-aware detection models that are specifically tailored to the unique spatial and statistical properties of SAR data. Effective frameworks must balance real-time efficiency for deployment in resource-constrained environments with robustness to small-scale or partially occluded targets and ambiguous backscatter signatures. Addressing these challenges is essential to enable accurate, persistent, and automated monitoring across maritime, environmental, and defense-related applications [1–3].

 doi:10.20944/preprints202509.2458.v1

Traditional ship detection methods, such as the Constant False Alarm Rate (CFAR) algorithm [3], have been widely employed due to their adaptive thresholding capability in clutter-rich maritime environments. While CFAR is effective in controlled or relatively simple scenarios, its performance often degrades in practical SAR applications. The algorithm relies on manually defined features and expert-set parameters, which increase processing time and limit scalability. In complex maritime conditions—characterized by varying sea states, heterogeneous backgrounds, and low signal-to-clutter ratios—CFAR frequently suffers from reduced accuracy and weak generalization [4]. This limitation stems from its dependence on accurate clutter modeling and continuous threshold calibration, both of which must dynamically adapt to changing environments to reduce false alarms and missed detections. With the growing complexity of SAR data and the increasing demand for real-time, high-precision maritime surveillance, traditional approaches such as CFAR alone are insufficient [5]. To address these issues, several enhanced CFAR variants and hybrid detection frameworks have been proposed, as briefly discussed in [2,6].

Recent advances in deep learning have greatly advanced SAR ship detection, with convolutional neural networks (CNNs) [7,8] demonstrating strong ability to learn hierarchical representations directly from raw data. Two main categories of CNN-based object detection architectures are commonly employed. The first, known as two-stage detectors, follows a coarse-to-fine strategy: region proposals are generated initially, followed by classification and bounding box regression in a second stage. Representative models include Faster R-CNN [9], Libra R-CNN [10], and Mask R-CNN [11]. These methods typically achieve high detection accuracy but incur significant computational cost, which limits their suitability for real-time applications. The second category, single-stage detectors, performs classification and localization jointly in a unified pipeline. Examples include the YOLO family [12], SSD [13], and FCOS [14]. Owing to their end-to-end training design, single-stage detectors generally offer superior speed and simplicity, albeit sometimes at the expense of slightly reduced accuracy compared with two-stage approaches.

In SAR ship detection, key challenges stem from scale variation, occlusion, and directional backscattering, which complicate feature extraction. Background clutter, including speckle noise and sea surface texture, often leads to false alarms, particularly in lightweight models. Although deeper CNNs theoretically provide larger receptive fields, only a limited central region [15] significantly influences prediction. The fixed and spatially rigid receptive fields of CNNs make it difficult to adapt to ships of varying scales and orientations, a problem further amplified in coastal, port, and inland scenes where object–background confusion is common. These limitations highlight the need for tunable, multi-scale, and context-aware detection mechanisms.

To address these issues, recent works have explored diverse strategies. Zhao et al. [16] proposed the Attention Receptive Pyramid Network (ARPN), integrating Receptive Fields Block (RFB) and CBAM [17] to enhance global–local dependencies and suppress clutter. Tang et al. [18] introduced deformable convolutions with BiFormer attention and Wise-IOU loss to improve adaptability in complex SAR scenes. Zhou et al. [19] developed MSSDNet, a lightweight YOLOv5s-based model with CSPMRes2 and an FC-FPN module for adaptive multi-scale fusion. Cui et al. [20] enhanced CenterNet with shuffle-group attention to strengthen semantic extraction and reduce coastal false alarms. More recently, Sun et al. [21] proposed BiFA-YOLO, which employs a bidirectional feature-aligned module for improved detection of rotated and small ships. Overall, these studies emphasize that effective SAR ship detection requires models capable of balancing local detail sensitivity with global contextual awareness, particularly in cluttered and multi-scale maritime environments.

SAR ship datasets contain a high proportion of small targets with limited appearance cues such as texture and contour, making them challenging to detect. Detection performance is often hindered by the scarcity of features extracted from small ships and the mismatch between their scale and the large receptive fields or anchor sizes of conventional detectors. As mainstream frameworks typically downsample images to obtain semantic-rich features, critical information for small targets may be lost, leading to frequent missed detections [3,4,22].

 doi:10.20944/preprints202509.2458.v1

To address these issues, several lightweight attention-augmented approaches have been proposed. Hu et al. [23] introduced BANet, an anchor-free detector with balanced attention modules that enhance multi-scale and contextual feature learning. Zhou et al. [24] proposed a multi-attention model for large-scene SAR images, enhancing detection performance in complex background environments. Guo et al. [25] further extended CenterNet with multi-level refinement and fusion modules to strengthen small-ship detection and suppress clutter with minimal overhead.

Despite progress with compact detectors that cut redundancy and incorporate attention for scale adaptability, reliable SAR ship detection remains difficult. Lightweight models, in particular, struggle with clutter, noise, and scale variation due to limited context modeling and rigid receptive fields. These gaps motivate the design of specialized, domain-tailored frameworks. To address these limitations, this paper proposes **DRC^2-Net**, a compact and context-aware enhancement of YOLOX-Tiny. The proposed framework integrates lightweight semantic reasoning and adaptive spatial modules to strengthen feature representation, improve geometric adaptability, and enhance detection robustness in complex maritime scenes, all while maintaining high efficiency. The key contributions of this work are summarized as follows:

- **Enhanced Semantic Context Modeling:** Long-range spatial dependencies are captured by integrating a recurrent attention mechanism after the SPPBottleneck in the backbone. This placement enables semantic reasoning over fragmented, elongated, or partially visible ship structures, improving robustness against weak or ambiguous contours in complex maritime scenes.
- **Adaptive and Flexible Receptive Fields:** A novel DeCSP module embeds deformable convolutions into the bottleneck paths of three CSP layers in the neck, enabling dynamic, content-aware sampling. This design adapts to irregular ship scales and shapes while recovering shallow and boundary information often overlooked by conventional FPN-based fusion.
- **Lightweight and Generalizable Detection Framework:** The proposed DRC^2-Net extends YOLOX-Tiny with targeted architectural enhancements while maintaining its lightweight nature ($\sim$5.05M parameters). Evaluations on SSDD and iVision-MRSSD demonstrate strong generalization across varying resolutions, target densities, and clutter conditions, ensuring real-time performance suitable for maritime surveillance and edge deployment.

The remainder of this paper is organized as follows. Section 2 introduces the YOLOX-Tiny baseline and reviews the theoretical foundations of recurrent attention and deformable convolution. Section 3 presents the proposed DRC^2-Net architecture, emphasizing its attention-aware and geometry-adaptive modules. Section 4 describes the experimental setup, datasets, evaluation metrics, and ablation studies conducted on SSDD and iVision-MRSSD. Finally, Section 5 summarizes the main findings and discusses potential avenues for future research.

2. A Lightweight Backbone

Accurate ship detection in SAR imagery requires broad contextual reasoning to suppress false alarms caused by sea clutter, together with detailed semantic discrimination to reliably localize weak or fragmented targets. Due to the frequent presence of coarse-resolution ships and highly dynamic maritime environments, traditional detectors often struggle to achieve an optimal balance between precision and efficiency. Current research increasingly focuses on lightweight, anchor-free frameworks tailored to the unique properties of SAR data. Such designs combine adaptive spatial sampling with long-range dependency modeling, enabling real-time operation in resource-limited settings while maintaining strong detection reliability [3].

2.1. YOLOX-Tiny Architecture

As introduced in the original "YOLOX: Exceeding YOLO Series in 2021" paper [26], the YOLOX family comprises six progressively larger variants: Nano, Tiny, S, M, L, and X, each balancing speed and accuracy to suit different deployment needs. In this work, we adopt YOLOX-Tiny as the baseline architecture due to its compact design and favorable trade-off between inference speed and detec-

 doi:10.20944/preprints202509.2458.v1

tion accuracy. YOLOX adopts a center-based, anchor-free detection paradigm that localizes objects directly using key points, eliminating the reliance on predefined anchor boxes. This approach simplifies the detection pipeline, reduces computational complexity, and avoids the burden of extensive hyperparameter tuning [27].

The YOLOX network structure is composed of four main components: the input layer, the backbone for feature extraction, the neck for multi-scale feature fusion, and the prediction head. The overview of the YOLOX-Tiny model is illustrated in Figure 1.

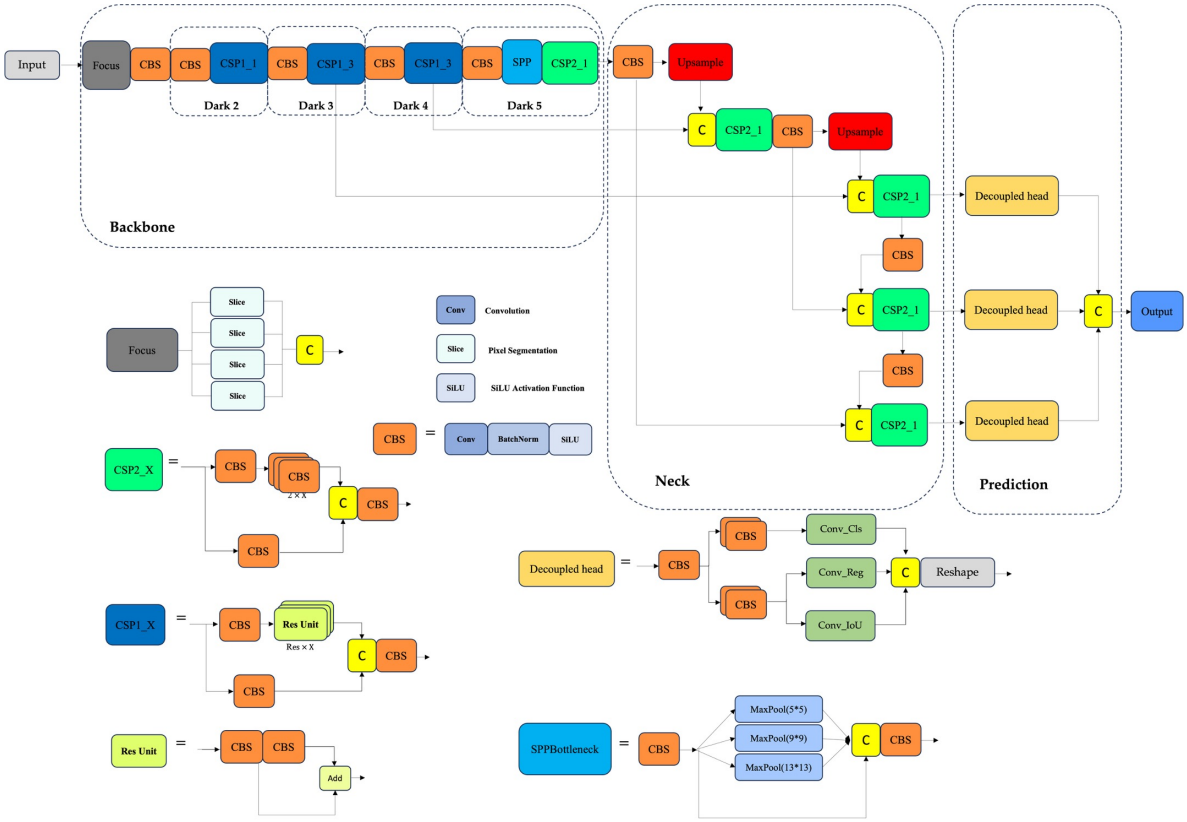


Figure 1. Overview of the YOLOX-Tiny model [26].

YOLOX framework uses CSP-Darknet as the backbone, leveraging Cross Stage Partial Networks (CSPNet) [28] to enhance gradient flow and reduce computational redundancy, and Spatial Pyramid Pooling (SPP) [29] for multi-scale feature extraction. CSP-Darknet offers a robust balance between accuracy and efficiency, making it a preferred choice in modern object detection models. YOLOX-Tiny is well-suited for real-time applications on resource-constrained platforms. The neck utilizes a Path Aggregation Network (PANet) [30], which fuses features through both top-down and bottom-up pathways. The top-down flow, inspired by Feature Pyramid Networks (FPN) [31], enriches semantic information, while the bottom-up path enhances spatial detail and localization precision, resulting in stronger representations across scales. The decoupled head design in YOLOX separates the object detection process into two distinct branches: classification and regression. This structural decoupling allows the model to independently optimize feature extraction for identifying object categories and for precisely localizing their spatial positions and dimensions. By minimizing task interference and distributing computational focus, the decoupled approach enhances both detection accuracy and operational efficiency, particularly important in complex SAR environments where fine-grained semantic discrimination and precise localization are critical.

2.2. Attention Mechanisms in SAR Ship Detection: (RCCA)

Detecting maritime targets in SAR imagery—particularly small, low-contrast, or partially visible ships—remains difficult. Small targets offer limited features that are easily lost due to receptive field

 doi:10.20944/preprints202509.2458.v1

mismatches, where anchors or kernels are disproportionately large. Discrete anchor scales further conflict with the continuous variation in ship size and orientation, reducing recall for targets between intervals. Incomplete targets, affected by sensor limits or background clutter, provide fragmented features that hinder accurate detection and classification [3]. CNN architectures, while effective in extracting semantic abstractions through deep hierarchical layers, often suffer from spatial resolution loss due to successive down-sampling. Consequently, small ships—occupying only a few pixels in SAR images—may lose discriminative features in deeper stages, leading to missed detections and reduced fine-grained recognition accuracy [2]. To address these limitations, attention mechanisms have been widely adopted for adaptively enhancing spatial and channel-wise features [32]. Transformer-based self-attention provides strong contextual modeling but remains computationally prohibitive for real-time or edge scenarios. Lightweight alternatives such as SE, SimAM [33], and CBAM are more practical but show clear drawbacks for SAR imagery: SE neglects spatial cues, while SimAM and CBAM depend on static pooling that limits long-range context. More recent approaches like Monte Carlo Attention (MCA) [34] attempt to capture global dependencies via stochastic sampling, yet face instability in cluttered maritime backgrounds. These challenges underscore the need for efficient, spatially aware attention mechanisms tailored to SAR ship detection.

To enhance pixel-wise representational capacity, the Criss-Cross Attention (CCA) mechanism was developed to efficiently capture contextual information along horizontal and vertical directions. Unlike traditional self-attention mechanisms, which incur high computational cost, CCA selectively aggregates features across rows and columns, significantly reducing complexity [35]. As illustrated in Figure 2, given an input feature map $H \in \mathbb{R}^{C\times W\times H}$, the attention module begins by applying two 1×1 convolutional layers to produce the query (Q) and key (K) feature maps, where $\{Q, K\} \in \mathbb{R}^{C'\times W\times H}$ and $C' < C$ for dimensionality reduction. After generating the query and key maps Q and K, an attention map $A \in \mathbb{R}^{(H+W-1)\times(W\times H)}$ is computed via an affinity operation. For each spatial location u in Q, a feature vector $Q_u \in \mathbb{R}^{C'}$ is extracted. Correspondingly, a set $\Omega_u \in \mathbb{R}^{(H+W-1)\times C'}$ is constructed by collecting feature vectors from K that lie along the same row and column as u. Each element $\Omega_{i,u} \in \mathbb{R}^{C'}$ in this set represents a context vector. The affinity score between Q_u and $\Omega_{i,u}$ is then calculated as:

$$d_{i,u} = Q_u \cdot \Omega_{i,u} \tag{1}$$

where $d_{i,u} \in D$ represents the correlation score between the query feature Q_u and the corresponding key feature $\Omega_{i,u}$, for $i = 1, \ldots, H + W - 1$. The resulting correlation matrix $D \in \mathbb{R}^{(H+W-1)\times(W\times H)}$ captures the attention strength between each spatial location and its horizontal and vertical context.

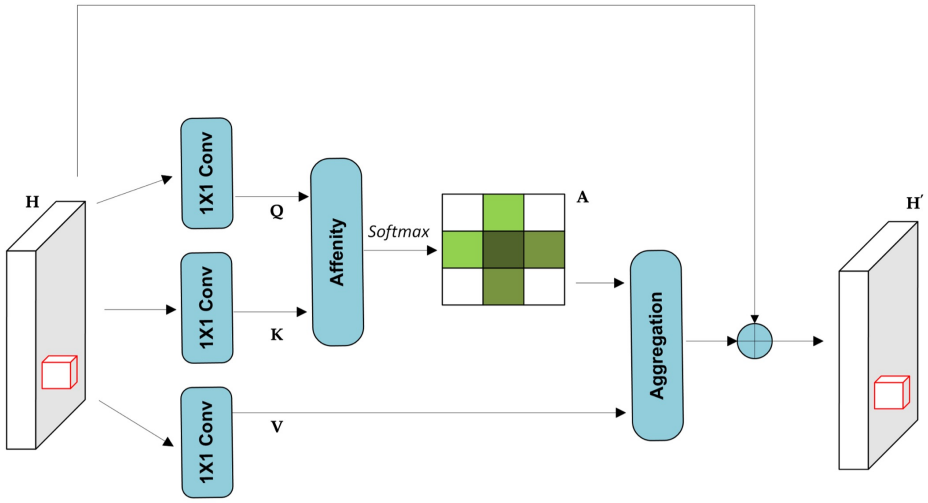


Figure 2. The Criss-Cross Attention module [35].

 doi:10.20944/preprints202509.2458.v1

A `SoftMax` operation is then applied along the attention dimension of D to normalize the values and produce the final attention map A. To adapt the features, another 1×1 convolution is applied to the input feature map H, producing $V \in \mathbb{R}^{C \times W \times H}$. For each spatial position u, a feature vector $V_u \in \mathbb{R}^C$ is extracted, along with a contextual set $\Phi_u \in \mathbb{R}^{(H+W-1) \times C}$ comprising vectors from the same row and column. The final contextual representation is computed using an aggregation operation that fuses this information with the original feature at u.The final contextual representation is computed as follows:

$$H'_u = \sum_{i=0}^{H+W-1} A_{i,u} \cdot \Phi_{i,u} + H_u \tag{2}$$

where H'_u is a feature vector in $H' \in \mathbb{R}^{C \times W \times H}$ at position u, and $A_{i,u}$ is a scalar value at channel i and position u in the attention map A. The contextual information $\Phi_{i,u}$ is integrated with the local feature H_u to enrich the pixel-wise representation.

This mechanism enables a broader spatial receptive field and selectively aggregates relevant context via the spatial attention map. As a result, the enhanced features become more semantically expressive and robust, which is particularly beneficial for pixel-level tasks such as semantic segmentation [36]. While the CCA module enables efficient capture of horizontal and vertical dependencies, its single-pass operation may be insufficient to fully model the complex spatial relationships often encountered in SAR ship detection, where targets may appear fragmented or rotated within cluttered scenes.

To overcome this limitation, RCCA extends CCA by introducing iterative refinement across R loops. In the first loop, the input feature map H yields an updated representation H', with the same shape. A second pass then reprocesses H' to generate H'', effectively integrating contextual information from all pixels. By sharing parameters across loops, RCCA enhances global semantic reasoning while maintaining a lightweight footprint.

As illustrated in Figure 3, setting $R = 2$ allows the module to aggregate full-image contextual information from all pixels, resulting in dense, context-rich feature representations. Let A and A' denote the attention maps in loop 1 and loop 2, respectively. With the help of a propagation function f, we can describe the information flow between any position u in H'' and any position θ in H.

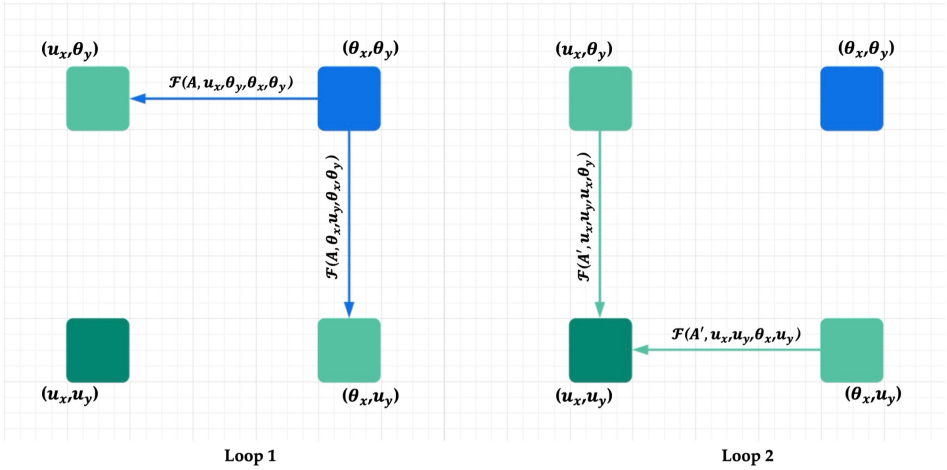


Figure 3. Information propagation when the loop number is 2 [35].

Information can directly flow from θ to u when θ lies along the Criss-Cross path of u. However, when $\theta = (\theta_x, \theta_y)$ is **not** in the Criss-Cross path of $u = (u_x, u_y)$, the propagation is indirect:

- In **Loop 1**, θ transmits information to two intermediate positions: (u_x, θ_y) and (θ_x, u_y) (light green points), both of which lie on the Criss-Cross path of u.

 doi:10.20944/preprints202509.2458.v1

- In **Loop 2**, these intermediate positions then relay the information to $u = (u_x, u_y)$ (dark green point).

This two-step message-passing mechanism enables θ to influence u even if it does not lie directly in its Criss-Cross path. As a result, RCCA captures long-range spatial dependencies and semantic context more effectively across the image domain.

2.3. Deformable Convolution Networks

In SAR-based maritime surveillance, ships often exhibit irregular or elongated geometries and may appear fragmented within cluttered sea or coastal environments. These hard-to-detect samples, combined with sparse and misaligned target distributions, pose significant challenges for lightweight detectors. While CNNs can extract hierarchical features, their grid-aligned and spatially rigid receptive fields—along with the limited scope of the effective receptive field—restrict adaptability to such complex cases. This mismatch between fixed sampling locations and actual ship structures frequently reduces detection accuracy in challenging SAR scenarios [15].

Unlike standard convolutions with fixed sampling grids, *deformable convolutions* dynamically adjust sampling positions based on local features, enabling the receptive field to adapt to the geometry of SAR ship targets. This flexibility enhances the model's ability to extract relevant information from distorted or obliquely shaped ships, significantly improving robustness in cluttered or ambiguous maritime conditions.

To overcome the limitations of fixed-grid sampling in conventional convolutional layers, Deformable Convolutional Networks (DCNs) introduce a learnable offset mechanism that dynamically adjusts the sampling positions based on local content, as shown in Figure 4. This enhances the model's ability to align with the actual structure of ship targets, which often vary in shape, scale, and orientation.

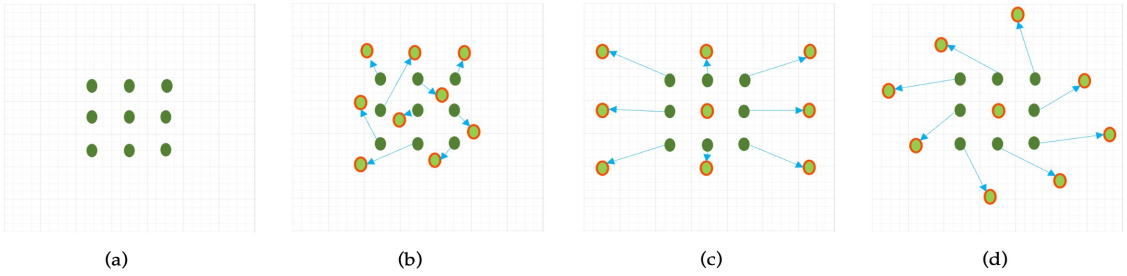


Figure 4. Illustration of sampling locations in 3x3 standard and deformable convolutions. (a) Standard 3×3 convolution; (b) Deformable convolution with learned offsets enabling adaptive kernel shapes; (c, d) Specialized variants of deformable convolution [37].

In a standard convolution, the output feature at location p_0 is computed as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} \omega(p_n) \cdot X(p_0 + p_n) \tag{3}$$

where X denotes the input feature map, ω represents the learnable weights, p_n is the relative offset from p_0 in the convolutional grid $\mathcal{R}$ (e.g., for a 3×3 kernel, $\mathcal{R} = \{(-1,-1), ..., (1,1)\}$).

In contrast, deformable convolution introduces learnable offsets Δp_n to each sampling location, enabling the network to shift the receptive field adaptively based on the input content. As illustrated in Figure 5, these offsets are predicted via parallel convolutional layers and organized into a $2N$-channel offset map, where N is the number of sampling locations. Due to the fractional nature of Δp_n, bilinear interpolation is applied to compute precise feature values at the deformed positions [38].

This adaptive sampling improves representation of irregular ship shapes and orientations commonly seen in SAR imagery.

The deformable convolution operation is thus expressed as:

 doi:10.20944/preprints202509.2458.v1

$$y(p_0) = \sum_{p_n \in \mathcal{R}} \omega(p_n) \cdot X(p_0 + p_n + \Delta p_n) \tag{4}$$

In deformable convolution, each sampling location is adjusted by a learnable offset Δp, allowing the receptive field to shift flexibly according to the input features. Since Δp often results in fractional coordinates, the corresponding feature values are obtained via bilinear interpolation. However, this interpolation may inadvertently sample from irrelevant or noisy regions, potentially degrading the quality of the extracted features. To mitigate this issue, DCNv2 [39] introduces a modulation scalar $\Delta m_k \in (0, 1)$ at each sampling location. This scalar acts as an adaptive attention weight, suppressing uninformative or noisy spatial regions by assigning lower values to less relevant sampling points.

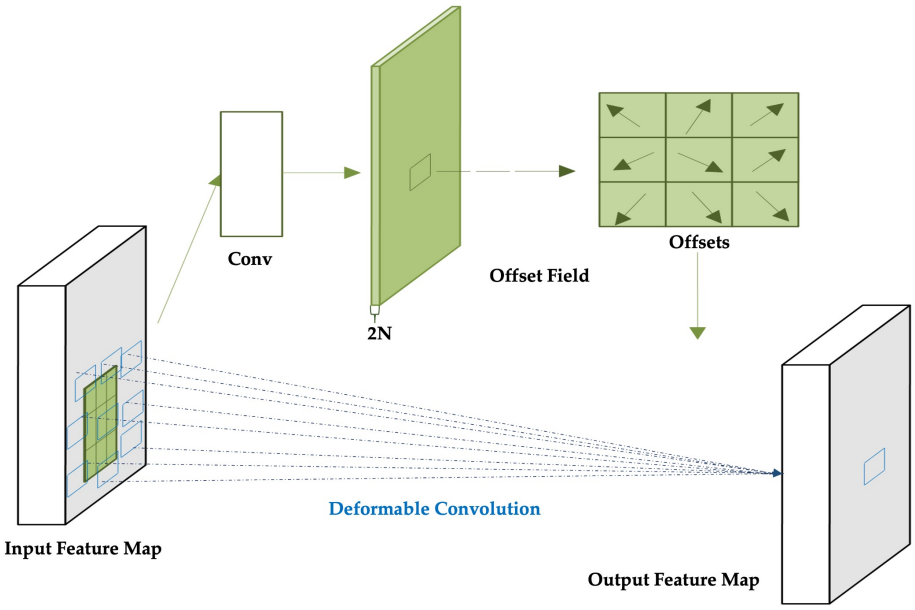


Figure 5. Illustration of a 3×3 deformable convolution neural network. The offset field is derived from the input feature map and has the same spatial resolution as the input [38].

The deformable convolution with modulation is mathematically defined as:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} \omega(p_n) \cdot X(p_0 + p_n + \Delta p_n) \cdot \Delta m_k \tag{5}$$

where:

- p_0 is the current location in the output feature map,
- $\mathcal{R}$ denotes the regular grid of the convolution kernel,
- $\omega(p_n)$ is the weight for the n-th location in the kernel,
- Δp_n is the learnable offset for position p_n,
- Δm_k is the modulation scalar applied to the sampled value.

The modulation coefficient Δm_k is an adaptive attention weight that suppresses irrelevant spatial regions by assigning lower values to uninformative sampling points.

3. Methodology

3.1. Overall Network Structure of DRC^2-Net

DRC^2-Net is a lightweight, context-aware detector built on the YOLOX-Tiny framework, noted for its balance of speed, compactness, and accuracy in real-time applications. While YOLOX-Tiny serves as a solid baseline, it encounters limitations in SAR imagery, especially when detecting sparse or partially visible ships within challenging maritime environments characterized by speckle noise and background

 doi:10.20944/preprints202509.2458.v1

artifacts. To address these challenges, this work introduces dual-modular enhancements applied to both the backbone and neck. The backbone preserves the four-stage design (Dark2–Dark5), producing feature maps C2–C5, with C3–C5 used as P3–P5 for detection, forming a hierarchical progression from high-resolution to high-semantic features. To strengthen deep semantic reasoning, the RCCA module is placed after the SPP bottleneck in Dark5, capturing horizontal and vertical dependencies through iterative refinement. This improves discrimination of ship targets while preserving the network's lightweight efficiency.

In the neck, a bidirectional feature fusion strategy is adopted to enhance multi-scale ship representation. The top-down path first propagates deep semantic features to generate the feature pyramids P3, P4, and P5, which capture coarse but highly informative contextual cues. To preserve the fine-grained spatial details essential for detecting small or partially occluded ships, a bottom-up enhancement path subsequently aggregates shallow features upward, producing the refined outputs N3, N4, and N5. To further improve geometric adaptability within this fusion process, DCNv2 are strategically embedded into key CSP blocks. Unlike standard convolution, DCNv2 introduces *modulated* learnable offsets. This mechanism predicts not only spatial adjustments to the sampling grid but also a modulation mask that weights the contribution of each sampled value. Consequently, the network dynamically adjusts its receptive field in both position and intensity based on the local geometry of the input features, leading to superior adaptation to the diverse and complex shapes of maritime targets.

As illustrated in Figure 6, the input SAR image is first processed by the **backbone** (dark2–dark5), which extracts hierarchical features represented as C2–C5. From these, multi-scale pyramids P3–P5 are constructed, capturing small, medium, and large target information. These pyramids are then fed into the **neck**, where bidirectional fusion generates intermediate maps N3–N5, enriching feature interactions across scales. Finally, three **decoupled detection heads** operate on N3–N5 to predict classification scores and bounding-box (BBox) regression. This end-to-end pipeline—from backbone encoding, through pyramid feature generation, to neck fusion and multi-head prediction—illustrates how the proposed framework transforms raw SAR imagery into accurate and scale-aware ship detections. Strategic enhancements within the backbone and neck further strengthen semantic continuity and geometric adaptability, while maintaining the lightweight nature of the design.

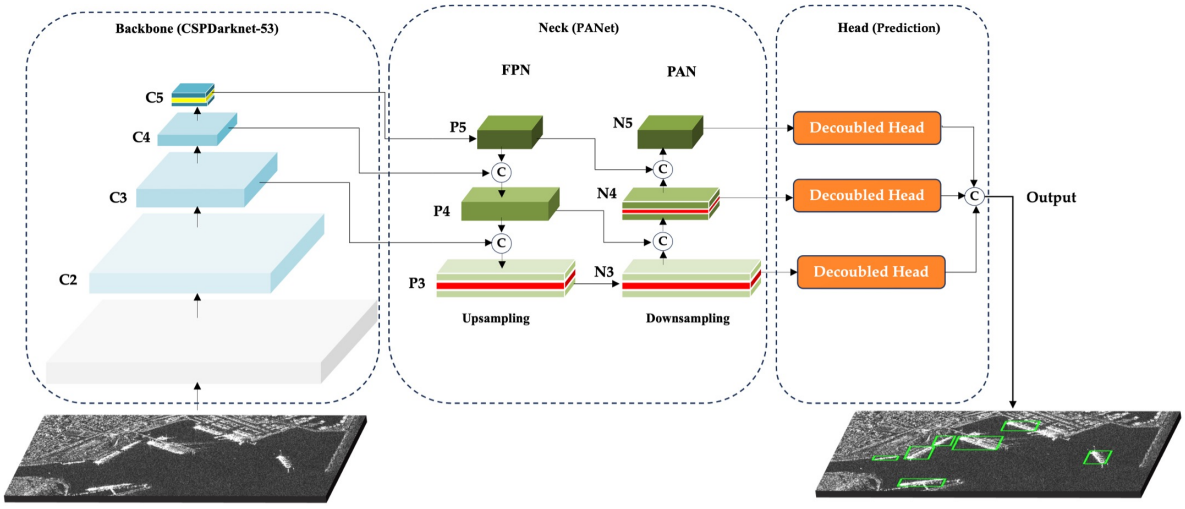


Figure 6. The overall structure of the proposed DRC^2-Net.

3.2. RCCA Integration for Sparse Maritime Contextual Enhancement

In CNNs, spatial resolution diminishes with depth, reducing the ERF and causing loss of fine-grained detail. This is especially problematic for hard-to-detect samples such as small or partially visible ships, where complex backscatter and multi-path reflections obscure object boundaries and increase false negatives. To mitigate this, DRC^2-Net integrates the RCCA module at the deepest backbone stage. As shown in Figure 7, RCCA is placed immediately after the SPPBottleneck block in

 doi:10.20944/preprints202509.2458.v1

dark5, where semantic abstraction is high but spatial precision is weakened. By iteratively aggregating horizontal and vertical context, RCCA expands the ERF and restores continuity across distant regions while preserving essential spatial cues. This placement enables the network to better distinguish fragmented or low-contrast ships from background clutter with minimal overhead.

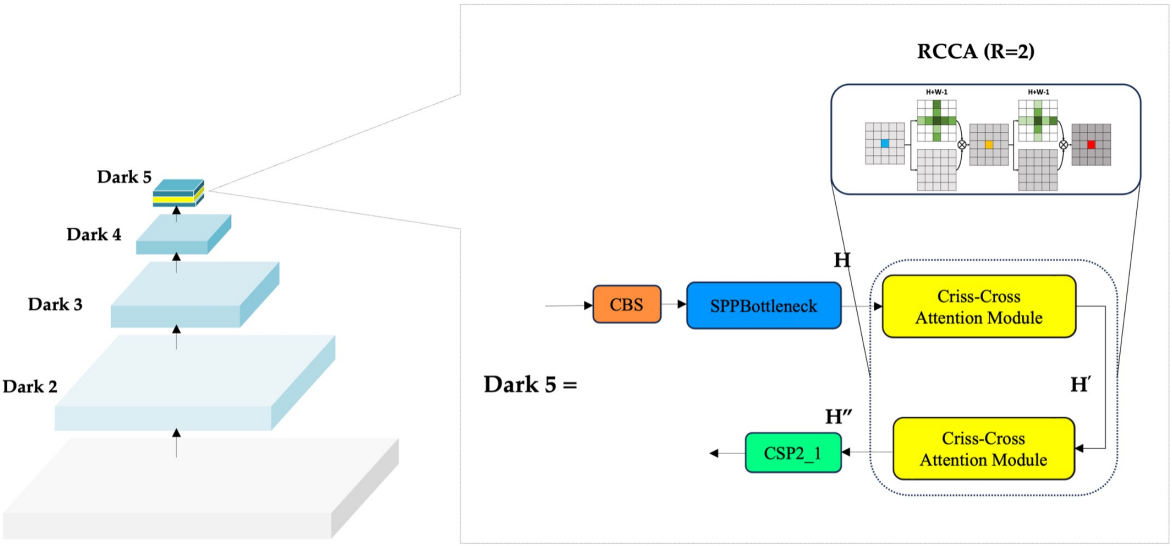


Figure 7. The details of the proposed RCCA module.

RCCA enhances semantic continuity by approximating global self-attention through a lightweight two-pass refinement process ($R = 2$). This iterative design improves upon single-pass CCA by enabling dense contextual aggregation across all pixels without adding parameters or significant computational cost. Its ability to capture elongated and irregular ship geometries makes it particularly effective for SAR imagery. Unlike conventional attention mechanisms constrained to local or channel-specific focus, RCCA models structured global context. Integrating it at the deep semantic stage of DRC^2-Net strengthens spatial reasoning, especially in challenging maritime scenes where ships may be rotated or visually fused with background clutter—conditions where fixed-window attention often fails.

3.3. DCNv2-Enhanced Neck: Adaptive Geometry Modeling in Multi-Scale Fusion

To enhance spatial adaptability in multi-scale feature fusion, DRC^2-Net integrates DCNv2 into the CSP modules of the neck. In contrast to standard convolutions that utilize a fixed grid, DCNv2 introduces learnable offsets which dynamically adjust the sampling positions based on the input's local geometry. This allows the network's receptive field to adaptively align with the diverse shapes and orientations characteristic of maritime targets.

This enhancement is implemented through a custom *Deformable CSP* (DeCSP) layer, which preserves the original CSP architecture's split–transform–merge strategy. Specifically, the standard 3×3 convolutions within the bottleneck blocks are replaced with DCNv2 layers, forming a *DCN-Bottleneck*. By embedding these DCN-Bottlenecks within the CSP structure, the network gains a superior capacity to capture rotated and distorted ship features, all while maintaining a low computational overhead. Architecturally, deformable convolutions are integrated at three critical points in the neck: C3–N3 and C3–N4 in the bottom-up path, and C3–P3 in the top-down path, where accurate multi-scale feature alignment is essential. Instead of altering entire residual branches or replacing all convolutions—which provided only marginal benefits in preliminary trials—we adopt a selective design: only the **3×3 convolution** inside the bottleneck block is substituted with a **DCNv2** layer, forming a modular **DeCSP** block as illustrated in Figure 8. This modularity ensures that the original CSP structure can be preserved or extended with minimal architectural disruption.

 doi:10.20944/preprints202509.2458.v1

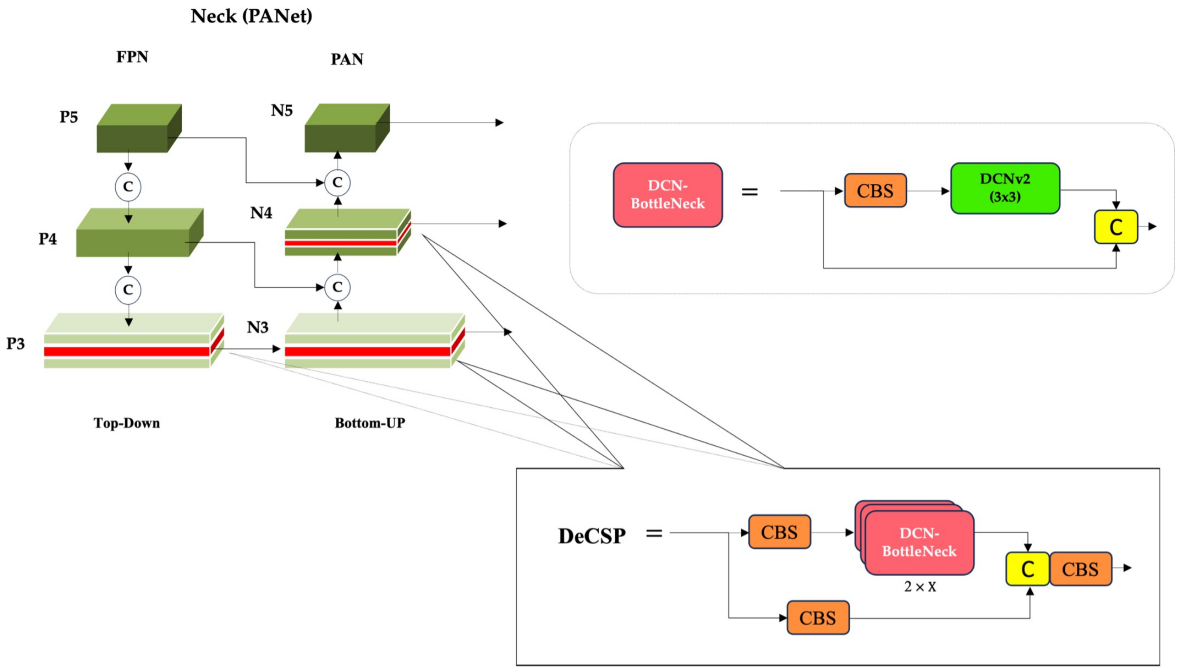


Figure 8. The details of the proposed DeCSP Module.

The two bottom-up insertions enhance early semantic fusion by adapting receptive fields to local geometric variations, while the top-down insertion reinforces high-level refinement, capturing global shape consistency. Together, these placements complement each other by balancing low-level adaptability with high-level contextual reasoning. By combining this geometric flexibility with the efficiency of CSPDarknet, the design strengthens multi-scale feature fusion and significantly improves robustness to hard samples in SAR imagery.

This dual strategy leverages semantic attention and spatial adaptability in a complementary manner, effectively addressing both contextual ambiguity and geometric deformation. Importantly, the enhancements preserve the original YOLOX-Tiny detection head, ensuring that DRC²-Net retains its real-time inference speed and compact size.

4. Experiments

4.1. Dataset Description

To evaluate the proposed method, we employ two publicly available datasets.

The **SSDD** dataset [40] contains a wide variety of maritime scenes, including port terminals, offshore waters, and open seas, with ship types ranging from small fishing boats to large container ships and tankers. Each image is standardized to a size of 512 × 512 pixels, with spatial resolutions spanning 3 m, 5 m, 8 m, and 10 m. The dataset is divided into training and testing sets in an 8:2 ratio. Annotations follow the *PASCAL VOC* format, ensuring compatibility with common object detection pipelines and enabling consistent evaluation across experiments.

The **iVision-MRSSD** dataset [41] consists of 11,590 SAR image tiles, each of size 512 × 512 pixels. The images are acquired from six distinct satellite sensors, providing a wide range of spatial resolutions and imaging geometries. This multi-sensor diversity ensures coverage of varied maritime conditions, from open seas to cluttered coastal zones. The dataset includes both inshore (3,605) and offshore (7,985) scenes, as well as negative examples (images without ships) to improve background discrimination. On average, each image contains 2.4 ships, enabling both single-target and multi-target detection evaluation. The dataset is split into training (70%), validation (10%), and testing (20%) subsets, supporting robust model development and fair benchmarking.

A detailed comparison of the two datasets is provided in Table 1.

 doi:10.20944/preprints202509.2458.v1

Table 1. Basic parameters of the SSDD and iVision-MRSSD datasets.

Attribute	SSDD	iVision-MRSSD
Data sources	RadarSat-2, TerraSAR-X, Sentinel-1	Capella Space, ICEYE, TerraSAR-X, Paz, Alos-PALSAR, Sentinel
Polarization modes	HH, VV, VH, HV	Single, Dual, Quad
Bands	X and C	C, L, X
Resolution (m)	1–15	Multiple spatial resolutions
Category	Ship	Ship
Number of images	1,160	11,590
Image size (pixels)	$28 \times 28 - 256 \times 256$	512×512
Number of ships	2,456	27,885

4.2. Implementation Settings

Since the SSDD dataset contains a relatively limited number of images, we adopt a **transfer learning** strategy by initializing the network with weights pre-trained on large-scale datasets. This enables the model to acquire general visual representations, thereby facilitating faster convergence and improved performance on the SAR-specific ship detection task. The baseline architecture is **YOLOX-Tiny**, configured with a depth multiplier of 0.33 and a width multiplier of 0.375, resulting in a compact network with approximately5.05M parameters.

All experiments were carried out on a Linux platform using PyTorch 2.0 and CUDA versions 11.x–12.x, with an NVIDIA Tesla T4 GPU (16 GB). The input resolution was fixed at 512×512 pixels. Training was performed for 96 epochs, organized into four cycles of 24 epochs each, with a 5-epoch warm-up. Early stopping was applied with a patience of 12 epochs to mitigate overfitting.

Optimization was conducted using **AdamW** with an initial learning rate of 1.25×10^{-4} (scaled by batch size) and a weight decay of 0.05. A **cosine annealing schedule** was employed for dynamic learning rate adjustment. Data augmentation strategies included Mosaic (1.0), MixUp (0.3), and horizontal flipping (0.5). Evaluation was carried out every two epochs with a confidence threshold of 0.5. To ensure reproducibility, all experiments used a fixed random seed (42) and four data loading workers.

4.3. Evaluation Indicators

To comprehensively assess the detection performance of the proposed model, we employ a set of standard evaluation metrics widely used in object detection research. These include accuracy indicators and multi-scale precision analyses tailored for SAR ship detection.

Precision (P) is defined as the ratio of correctly predicted positive samples—true positives (TP), to all samples predicted as positive, including false positives (FP). It reflects the model's ability to minimize false alarms and is particularly critical in reducing false detections under cluttered SAR backgrounds:

$$P = \frac{TP}{TP + FP} \tag{6}$$

Recall (R) measures the proportion of actual positive samples correctly identified by the model. It captures the model's capacity to detect all relevant targets:

$$R = \frac{TP}{TP + FN} \tag{7}$$

Precision and recall together provide a nuanced view of detection quality, especially important in maritime SAR scenarios where objects may be sparse or embedded in noisy backgrounds. The **F1 score**,

 doi:10.20944/preprints202509.2458.v1

defined as the harmonic mean of P and R, offers a comprehensive measure of a model's classification performance:

$$F1 = \frac{2 \times (P \times R)}{P + R} \tag{8}$$

Average Precision (AP) measures the area under the precision–recall (PR) curve, evaluating the trade-off between precision and recall across confidence levels:

$$AP = \int_0^1 P(R)\, dR \tag{9}$$

In this study, we report AP_{50} (computed at a fixed Intersection over Union (IoU) threshold of 0.5) and the more comprehensive AP, averaged over multiple IoU thresholds from 0.5 to 0.95 in 0.05 increments. Additionally, we report AP_s, AP_m, and AP_l, corresponding to the model's performance on small, medium, and large ship targets, respectively.

Intersection over Union (IoU) is a standard metric used to evaluate the accuracy of object detection models by comparing predicted bounding boxes to ground truth:

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})} \tag{10}$$

where B_{pred} is the predicted bounding box, and B_{gt} is the ground truth. A higher IoU indicates better localization accuracy.

Furthermore, computational efficiency is assessed using the number of parameters (**Params**) and the number of floating-point operations (**FLOPs**). The total parameter count is the sum across all layers. For a convolutional layer, it is given by:

$$Params = (k_h \cdot k_w \cdot C_{in}) \cdot C_{out} \tag{11}$$

where k_h and k_w denote kernel height and width, C_{in} is the number of input channels, and C_{out} is the number of output channels.

4.4. Ablation Experiments: Results and Analysis

To ensure a principled integration of the proposed attention mechanism, an initial comparative evaluation was conducted against state-of-the-art alternatives. Specifically, CBAM and SimAM were embedded at the same spatial location as RCCA to enable a fair and isolated comparison. This preliminary study confirmed the superior representational capacity of RCCA, thereby justifying its selection for subsequent integration.

Following this, a three-stage ablation study was designed to systematically quantify the impact of each architectural enhancement. Spanning eight controlled configurations, the experiments independently assessed the contributions of contextual attention and deformable convolutions across both the backbone and neck. The final experiment integrated the most effective modules into a unified DRC^2-Net design, highlighting their complementary synergy in boosting multi-scale SAR ship detection performance.

Table 2 reports the results of **Experiment 1**, which examines the effect of contextual attention mechanisms embedded within the backbone. Using the default YOLOX-Tiny (*BB-0*) as the baseline, several enhanced variants were evaluated by inserting attention modules immediately after the SPP bottleneck in the Dark5 stage. These configurations include *BB-1* (CBAM), *BB-2* (SimAM), *BB-3* (CCA), and *BB-4* (RCCA).

 doi:10.20944/preprints202509.2458.v1

Table 2. Results of the ablation studies of Experiment 1 on the backbone (best results in bold).

Backbone	mAP50	P	R	F1	AP	AP50	AP_s	AP_m	AP_l
BB-0 (Baseline)	90.89	96.74	92.49	94.57	61.32	92.12	89.97	87.23	78.95
BB-1 (CBAM)	90.58	**97.30**	92.94	94.84	59.71	92.12	89.09	88.30	84.20
BB-2 (SimAM)	90.46	96.20	92.86	94.50	60.02	92.76	89.97	**89.63**	84.20
BB-3 (CCA)	**91.59**	97.10	92.12	94.55	61.21	91.94	89.09	88.30	78.95
BB-4 (RCCA)	91.09	96.58	**93.04**	**94.78**	**61.92**	**93.04**	**89.97**	87.23	**84.21**

To systematically evaluate the effect of attention mechanisms within the YOLOX-Tiny backbone, five configurations were tested by inserting different modules after the SPP bottleneck in the Dark5 stage. The baseline (BB-0), without attention, achieved a solid reference performance with mAP@50 of 90.89%, AP of 61.32%, and F1-score of 94.57%. This configuration established the baseline semantic representation and localization capacity for SAR ship detection. Integrating CBAM (BB-1) produced the highest precision (97.30%) but reduced AP to 59.71% (−1.61% compared to baseline), indicating that while CBAM reinforces object confidence, it lacks sufficient adaptability in cluttered SAR scenes. SimAM (BB-2) reached 60.02% AP and the highest AP_m (89.63%), but only modestly improved recall (92.86%) and AP_l (+5.25%), suggesting its neuron-level importance estimation does not generalize well across scales. CCA (BB-3) enabled criss-cross feature interactions, yielding a balanced performance with AP of 61.21% and recall of 92.12%, alongside moderate improvements in AP_m, though it still underperformed for small and large targets. RCCA (BB-4), by contrast, achieved the best overall performance. Its iterative refinement across horizontal and vertical dimensions improved AP to 61.92% (+0.60%), recall to 93.04% (+0.55%), and AP_l to 84.21% (+5.26%), while maintaining strong precision (96.58%) and competitive mAP@50 (91.09%).

The results demonstrate that RCCA surpasses all tested alternatives, offering stronger contextual modeling and scale adaptability with minimal overhead. As shown in Figure 9, the Precision–Recall curve confirms this advantage: the RCCA-enhanced backbone sustains higher precision across a broad recall range, yielding superior AP@50. This visual evidence reinforces the quantitative findings, validating RCCA's role in strengthening spatial–semantic representation and supporting its integration into the final DRC^2-Net architecture.

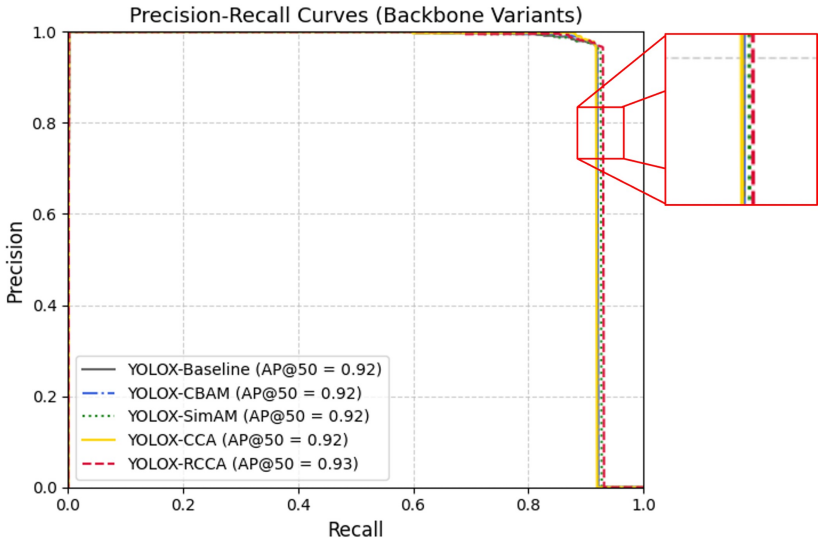


Figure 9. PR curves of different attention-enhanced backbones.

Table 3 presents the results of Experiment 2, which evaluates the impact of deformable convolutional enhancements within the neck while reusing the same backbone variants from Experiment 1. Three neck configurations were explored: **NK-0** denotes the original YOLOX neck; **NK-1** integrates

 doi:10.20944/preprints202509.2458.v1

two DeCSP blocks into the bottom-up path at C3_N3 and C3_N4; and **NK-2** extends this design by adding a third DeCSP block in the top-down path at C3_P3. This setup isolates the contribution of the neck, particularly the influence of DCNv2, on multi-scale feature fusion and spatial adaptability, while maintaining consistency in the backbone structure across all variants.

Table 3. Results of the ablation studies of Experiment 2 on the neck (best results shown in bold).

Neck	mAP50	P	R	F1	AP	AP50	AP_s	AP_m	AP_l
NK-0	90.89	96.74	**92.49**	94.57	**61.32**	**92.12**	**89.97**	87.23	78.95
NK-1	**90.97**	**97.48**	91.94	**94.63**	60.53	91.94	88.79	86.17	84.21
NK-2	89.60	97.09	91.58	94.25	61.20	91.58	86.73	**90.43**	**89.47**

On the other hand, The baseline configuration (NK-0), which employs standard CSPLayers, establishes strong performance with the highest AP (61.32%) and AP_s (89.97%), confirming its suitability for small target detection. When two DeCSP blocks are introduced into the bottom-up path (NK-1), precision improves to **97.48%** and AP_l rises to **84.21%** (+5.26% over baseline), indicating that adaptive sampling strengthens localization for larger and irregular ship targets. However, the overall AP decreases slightly (60.53%), suggesting that the deformable design sacrifices some sensitivity to small-scale targets.

Extending the architecture with a third DeCSP block in the top-down path (NK-2) further boosts AP_m to **90.43%** (+3.20%) and AP_l to **89.47%** (+10.52%), demonstrating improved multi-scale refinement and geometric adaptability. Nevertheless, this configuration results in reduced mAP50 (-1.29%) and a noticeable decline in AP_s (-3.24%), confirming that excessive deformability can weaken fine-grained detection in cluttered SAR backgrounds. These findings highlight that while deformable convolutions benefit mid-to-large targets, they require balancing to maintain small-scale accuracy.

Table 4 presents the results of Experiment 3, which integrates the most effective components identified in the earlier studies—namely, the RCCA-augmented backbone (BB-4) and the 3-DeCSP neck configuration (NK-2)—into a unified architecture, referred to as DRC^2-Net. While the neck-only experiments indicated that deformable convolutions primarily benefit mid-to-large targets at the expense of small-scale accuracy, their combination with RCCA effectively balances this trade-off. In the final design, the three DeCSP modules work in harmony with RCCA, enhancing multi-scale representation without sacrificing the lightweight nature of the YOLOX-Tiny foundation.

Table 4. Results of Experiment 3 for the DRC^2-Net (best results in bold).

Model	mAP50	P	R	F1	AP	AP50	AP_s	AP_m	AP_l
YOLOX-Tiny	90.89	96.74	92.49	94.57	61.32	92.12	89.97	87.23	78.95
DRC^2-Net	**91.87**	**96.77**	**93.41**	**95.06**	**61.50**	**93.04**	**91.15**	**88.30**	**89.47**

Experiment 3 validates the complementary synergy between global contextual attention and deformable convolutional sampling. The integrated design enhances both semantic representation and spatial adaptability, yielding consistent improvements across scales. Specifically, DRC^2-Net achieves gains of +0.98% in mAP@50, +0.61% in overall AP, and +0.49% in F1-score over the baseline YOLOX-Tiny. Notably, improvements are most pronounced for small-object detection (AP_s: +1.18%) and large-object detection (AP_l: +10.52%). In conclusion, DRC^2-Net represents a focused architectural refinement of YOLOX-Tiny, where RCCA strengthens long-range contextual reasoning while DeCSP modules adaptively refine multi-scale spatial features. These enhancements deliver a lightweight yet powerful SAR ship detector that balances efficiency with robustness, making it suitable for real-time maritime surveillance in complex environments.

 doi:10.20944/preprints202509.2458.v1

4.5. Comparative Evaluation with Lightweight and State-of-the-Art SAR Detectors on SSDD

To comprehensively assess the performance of the proposed DRC^2-Net, we benchmarked it against a range of representative object detectors. These include mainstream YOLO variants such as YOLOv5 [42], YOLOv6 [43], YOLOv3 [44], YOLOv7-tiny [45], and YOLOv8n [46], as well as lightweight SAR-specific models including YOLO-Lite [47] and YOLOSAR-Lite [48].

As summarized in Table 5, DRC^2-Net achieves the highest F1-score of **95.06%**, outperforming all baseline detectors. It also attains the highest precision (96.77%) and a strong recall (93.41%), highlighting its ability to minimize false positives while maintaining sensitivity to true targets. These results demonstrate that the integration of contextual reasoning and geometric adaptability in DRC^2-Net leads to superior SAR ship detection performance across diverse conditions.

Table 5. Objective evaluation of recent lightweight detection models on the SSDD dataset (best results in bold).

Model	P (%)	R(%)	F1 (%)	Params (M)	FLOPs (G)
YOLOv5	92.11	89.84	90.96	9.12	24.04
YOLOv6	94.56	86.18	90.20	16.31	44.21
YOLOv3-tiny	94.00	90.40	92.22	12.00	18.90
YOLOv7-tiny	92.90	**94.90**	–	6.00	13.00
YOLOv8n	95.40	94.40	–	**3.00**	8.10
YOLO-Lite	96.28	90.63	93.39	7.64	–
YOLOSAR-Lite	92.30	91.20	91.75	**2.05**	**4.48**
DRC^2-Net	**96.77**	93.41	**95.06**	5.05	9.59

As summarized, the proposed DRC^2-Net achieves superior performance across both general-purpose and SAR-specific lightweight detectors. It attains the highest F1-score of **95.06%**, reflecting an optimal balance between precision (96.77%) and recall (93.41%). This is achieved with only 5.05M parameters and 9.59 GFLOPs, underscoring its efficiency. Compared to mainstream detectors such as YOLOv5 and YOLOv8n, which demonstrate strong precision above 95% but do not report F1-scores, DRC^2-Net provides a more complete and favorable performance profile. While YOLOv7-tiny yields the highest recall (94.9%), its relatively lower precision (92.9%) and absence of F1-score reporting limit a balanced assessment.

In contrast, DRC^2-Net consistently outperforms domain-specific lightweight models. YOLO-Lite achieves an F1-score of 93.39% and YOLOSAR-Lite 91.75%, yet both fall short of DRC^2-Net's accuracy while maintaining similar or larger parameter counts. With its compact size (5.05M parameters) and moderate computational cost (9.59 GFLOPs), DRC^2-Net establishes a favorable balance between detection effectiveness and efficiency. These results make it particularly suitable for real-time SAR ship detection in resource-constrained environments, setting a new benchmark for lightweight detection frameworks. The strong gains further motivate broader generalization experiments on diverse and higher-resolution datasets such as iVision-MRSSD, discussed in the following section.

4.6. Quantitative Evaluation on the iVision-MRSSD Dataset

We further evaluated the proposed model on the recently introduced **iVision-MRSSD** dataset, a high-resolution SAR benchmark released in 2023. In contrast to SSDD, iVision-MRSSD presents greater challenges due to its wide range of ship scales, dense coastal clutter, and highly diverse spatial scenarios, making it an appropriate benchmark for testing robustness in realistic maritime surveillance applications. A notable limitation of this domain is that many existing SAR ship detection models are not publicly available or lack detailed implementation specifications, hindering reproducibility. To ensure a fair and meaningful comparison, we therefore adopt uniform experimental settings wherever feasible and report the best available metrics as documented in the respective original publications. As shown in Table 6, recent lightweight detectors such as YOLOv8n (58.1%), YOLOv11n (57.9%), and YOLOv5n (57.5%) achieve the highest overall Average Precision (AP) on the iVision-MRSSD dataset.

 doi:10.20944/preprints202509.2458.v1

These results demonstrate progress in global detection capability; however, they do not fully reflect robustness across different target scales.

Table 6. Objective evaluation of recent detection models on the iVision-MRSSD dataset (best results shown in bold).

Model	AP (%)	AP_s (%)	AP_m (%)	AP_l (%)
FCOS	43.5	39.3	50.1	39.6
ATSS [49]	53.2	46.7	61.4	59.2
YOLOv5n	57.5	51.1	66.4	69.5
YOLOv8n	58.1	51.5	66.7	**76.3**
YOLOv10n[50]	57.2	51.6	66.0	64.5
YOLOv11n[51]	57.9	52.1	66.1	69.6
DRC^2-Net	51.0	**71.56**	**84.15**	**78.43**

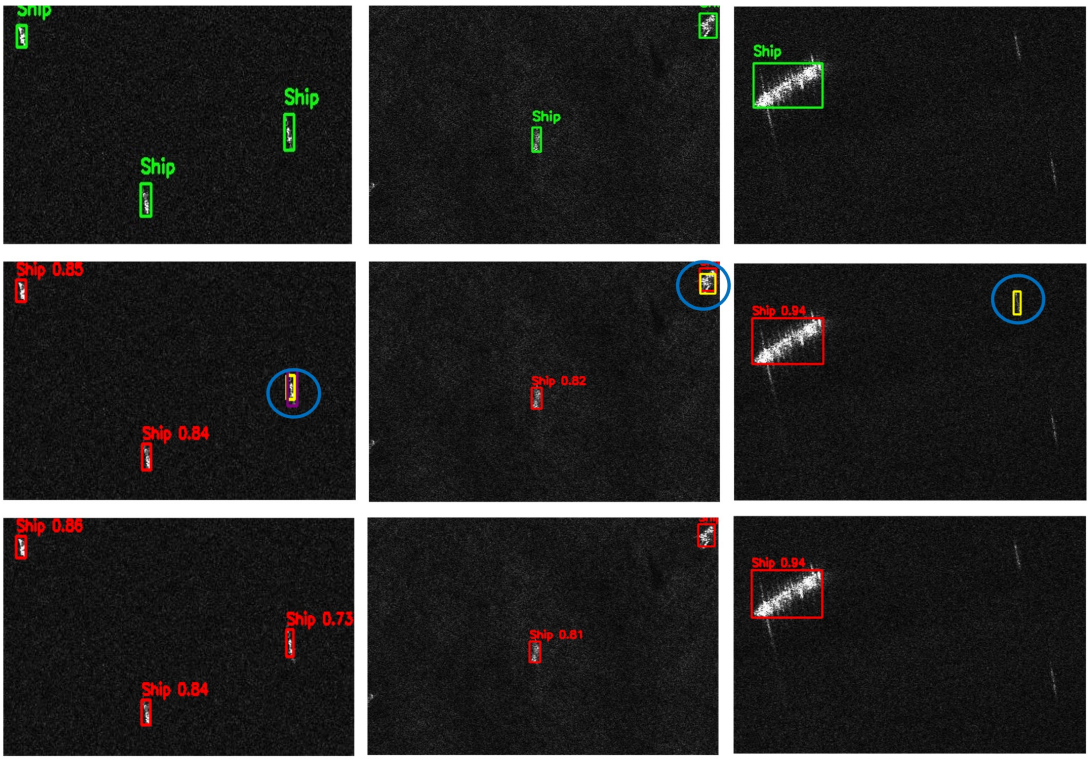
A more detailed scale-wise evaluation highlights the advantage of the proposed DRC^2-Net, which achieves **71.56%** AP_s, **84.15%** AP_m, and **78.43%** AP_l. These values significantly surpass the competing baselines, particularly in the detection of small and medium-sized ships, which are often missed by other models due to resolution loss and heavy background clutter in SAR imagery. By comparison, YOLOv8n and YOLOv11n report strong overall AP, but their AP_s values (51.5% and 52.1%, respectively) reveal notable limitations in small object detection.

This performance gap emphasizes the importance of scale-aware design. The combination of RCCA for global contextual reasoning and DeCSP modules for adaptive receptive fields enables DRC^2-Net to maintain consistent performance across scales. Consequently, the proposed framework provides a robust and efficient solution for complex SAR ship detection tasks, balancing accuracy and scale sensitivity in diverse maritime environments.

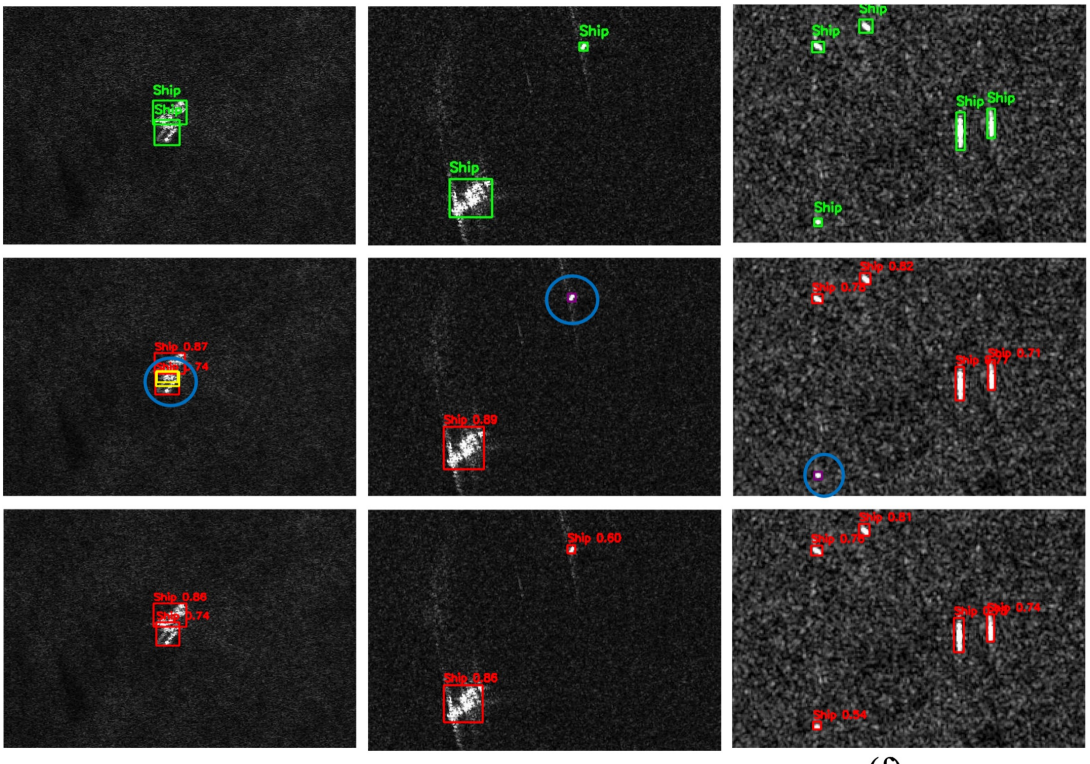
To qualitatively assess detection performance, representative scenes from the SSDD dataset are illustrated in Figure 10. Columns (a–i) cover diverse maritime conditions, including open-sea scenarios, nearshore environments, and multi-scale ship distributions within cluttered backgrounds. These examples emphasize the inherent challenges of SAR-based ship detection and provide visual evidence of the improvements achieved by the proposed DRC^2-Net.

In all qualitative figures presented in this paper, each column corresponds to a different scene, with three rows displayed vertically: the top row shows ground-truth annotations, the middle row depicts predictions from the baseline YOLOX-Tiny, and the bottom row presents predictions from DRC^2-Net. For consistency, the same color coding is applied throughout: green boxes denote ground truth, red boxes indicate correct detections, yellow boxes mark false positives, purple boxes highlight missed targets, and blue circles emphasize critical errors.

 doi:10.20944/preprints202509.2458.v1



(a) (b) (c)



(d) (e) (f)

 doi:10.20944/preprints202509.2458.v1

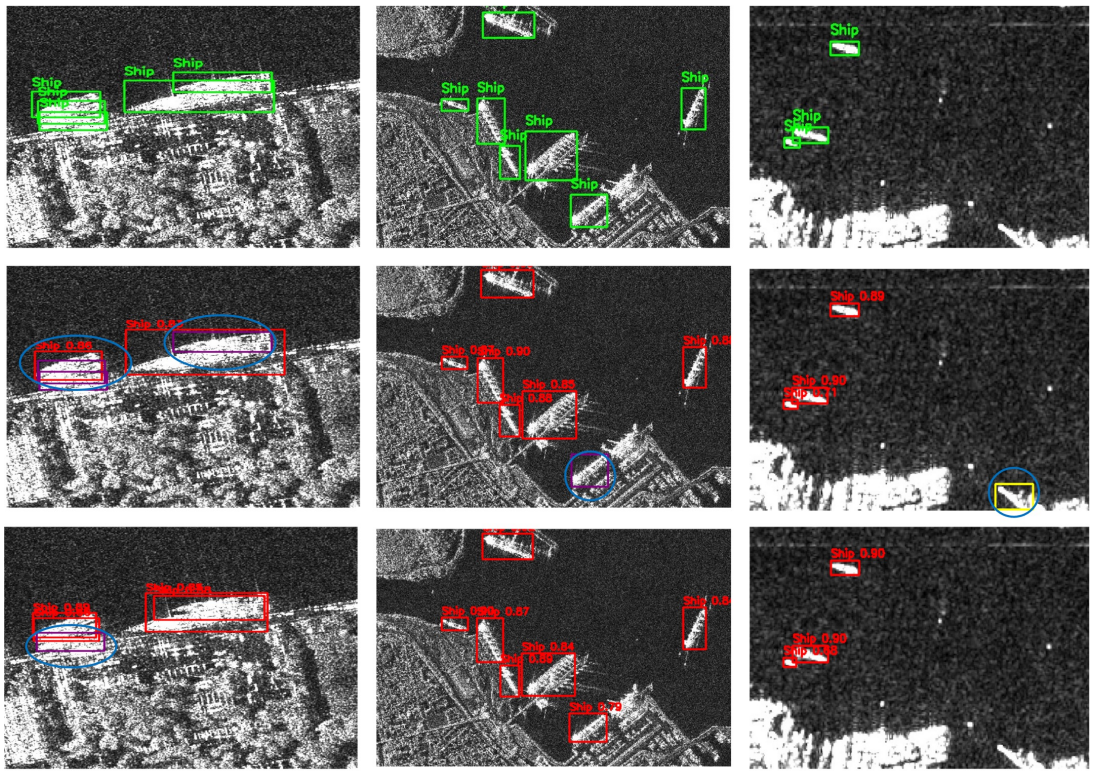


Figure 10. Visualization results across representative SSDD scenes. Groups (a–c), (d–f), and (g–i) represent diverse maritime conditions, including open sea, nearshore waters, and multi-scale cluttered backgrounds.

False alarms (yellow boxes) are most frequent in open-sea and offshore scenes (Figures 10a, 10b, 10c, and 10d), where wakes and wave patterns often resemble ships and confuse conventional detectors. DRC^2-Net alleviates these errors through deformable convolutions, which adapt the receptive fields to better distinguish targets from clutter. Missed detections (purple boxes) appear mainly in Figures 10e, 10f, and 10h, usually involving small or low-contrast ships. Notably, across all illustrated cases, DRC^2-Net missed only one target, demonstrating the advantage of RCCA in leveraging contextual cues to recover ambiguous or fragmented ships. Overall, these results confirm that DRC^2-Net achieves higher reliability by reducing false positives while improving sensitivity to challenging ship instances.

To further validate the generalization capability of the proposed DRC^2-Net, instance-level visual comparisons were conducted across three representative sets of SAR scenes from the **iVision-MRSSD** dataset. These samples encompass data from six distinct satellite sensors and are grouped into three major scenarios, each highlighting specific detection challenges.

Figure 11 shows **Scenario A (a–e)**, covering shorelines, harbors, and congested maritime zones with dense vessel clusters and coastal infrastructure. These conditions often trigger false alarms and mislocalizations, particularly for small, low-resolution ships affected by scale variation and background interference. While the baseline YOLOX-Tiny frequently misses or misclassifies such targets, the proposed DRC^2-Net achieves more precise localization, especially near image edges, demonstrating greater robustness in challenging coastal environments.

 doi:10.20944/preprints202509.2458.v1

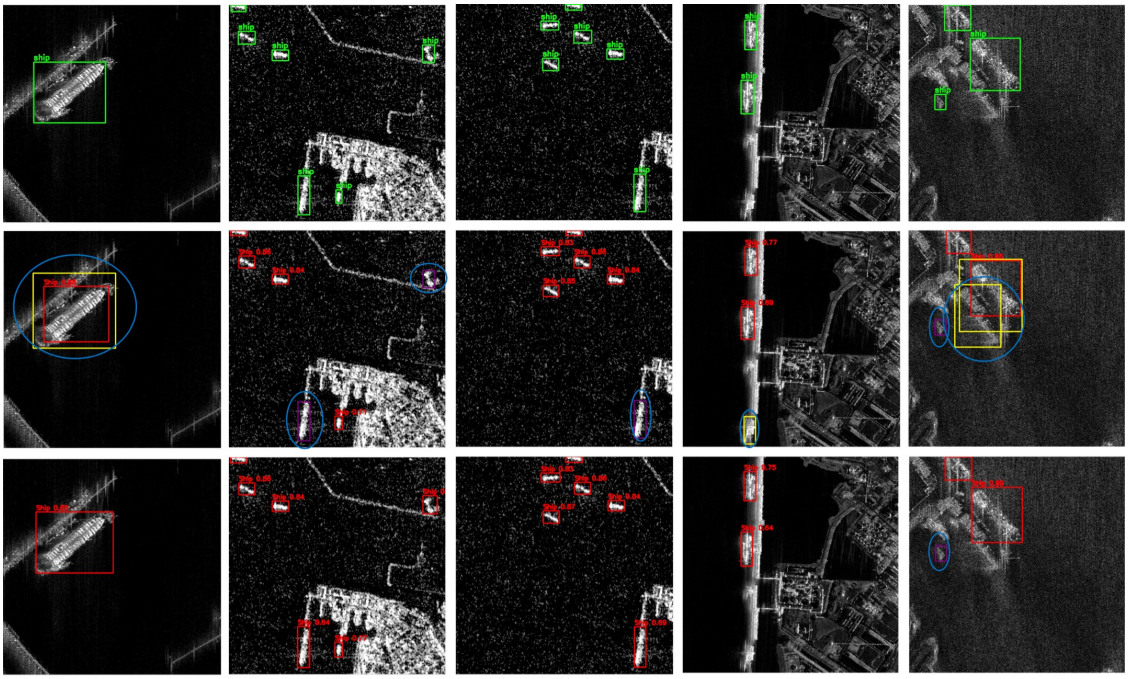


Figure 11. Visualization results for **Scenario A (a–e)** from the iVision-MRSSD dataset. Scenes represent shoreline and harbor environments with small-sized vessel clusters, occlusions, and coastal infrastructure.

Figure 12 presents **Scenario B (a–f)**, which depicts densely packed ships in far-offshore environments. These conditions are characterized by low signal-to-clutter ratios, heavy speckle noise, and ambiguous scattering patterns, all of which make target visibility and discrimination difficult. In such challenging scenes, missed detections frequently arise from faint radar returns and poorly defined object boundaries. Compared with the baseline, the proposed DRC^2-Net demonstrates stronger resilience to these issues, achieving more reliable detection under severe offshore clutter.

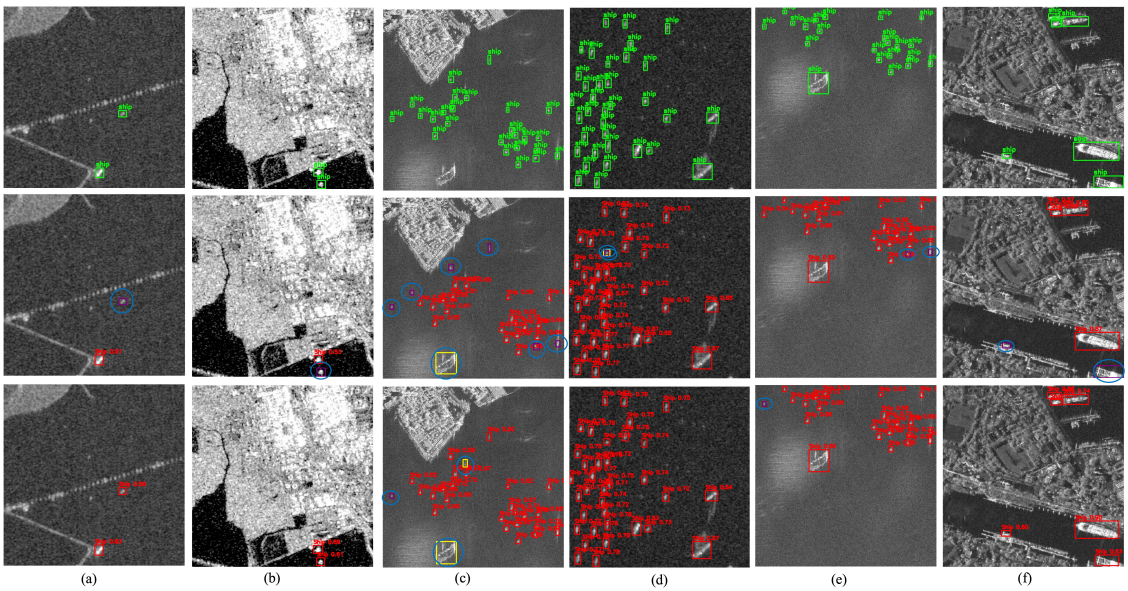


Figure 12. Visualization results of **Scenario B (a–f)** from the iVision-MRSSD dataset. The samples illustrate offshore clutter conditions, where densely distributed vessels and strong background interference increase the likelihood of false alarms and missed detections.

Figure 13 presents the final group of test scenes (**Scenario C**), featuring severe speckle noise, clutter, and ambiguous scattering patterns typical of moderate-resolution SAR imagery and rough

 doi:10.20944/preprints202509.2458.v1

sea states. These challenging conditions often lead to false positives and missed detections in baseline models. By contrast, the proposed model demonstrates stronger robustness, accurately localizing vessels despite degraded image quality and complex backgrounds.

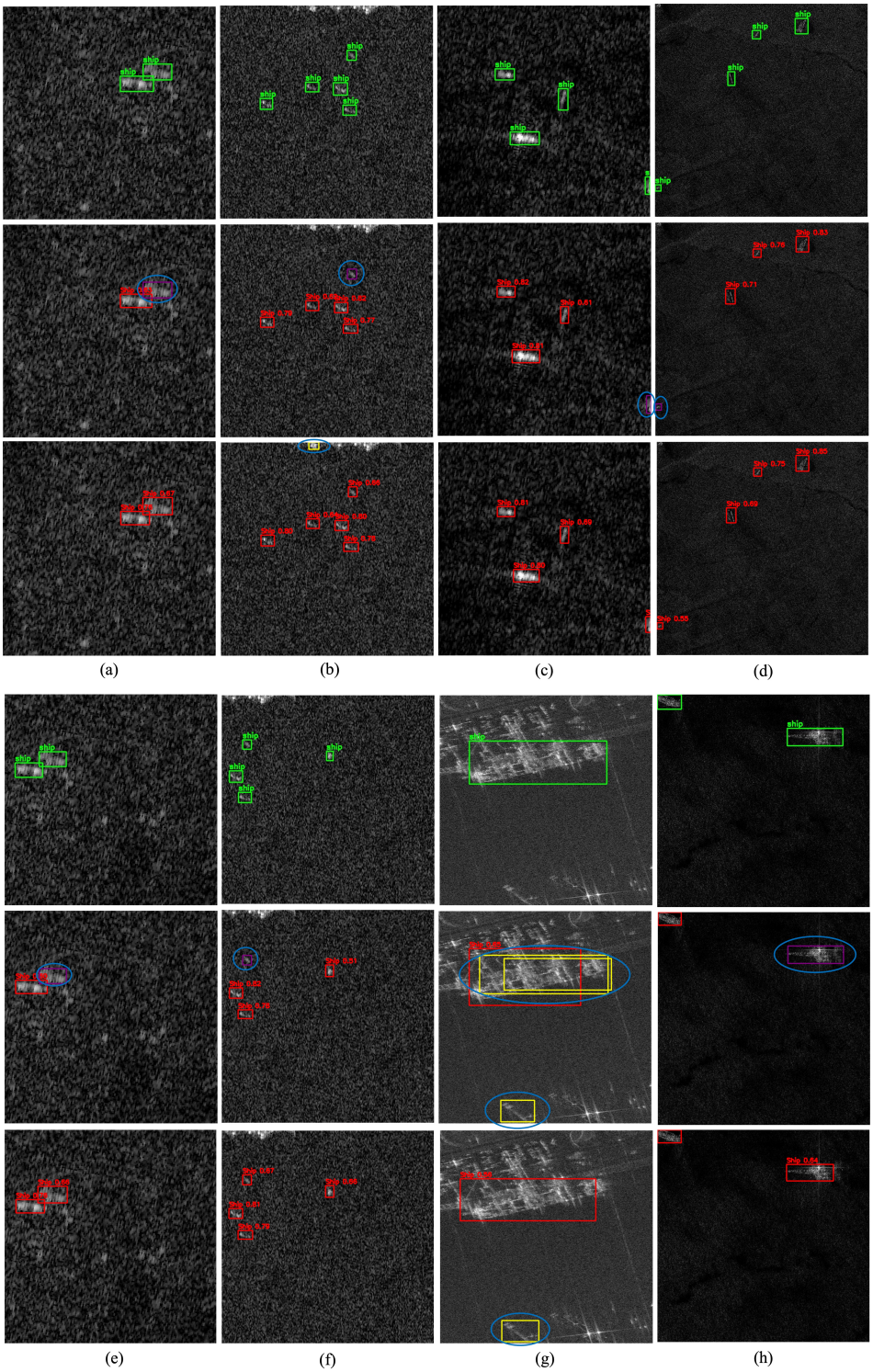


Figure 13. Visualization results of **Scenario C** (iVision-MRSSD): scenes with severe speckle noise, textured clutter, and ambiguous scattering effects.

In contrast, the proposed DRC^2-Net demonstrates enhanced robustness by combining deformable convolutions with contextual attention, effectively suppressing spurious responses and improving target discrimination. The qualitative results confirm DRC^2-Net's ability to localize small and multi-

 doi:10.20944/preprints202509.2458.v1

scale vessels even under adverse imaging conditions, highlighting its generalization capability across offshore, coastal, and noise-dominant scenarios in the iVision-MRSSD dataset.

A quantitative summary of detection performance in terms of correct detections, false alarms, and missed targets is provided in Table 7. The results clearly demonstrate the superiority of the proposed model across all scenarios. In **Scenario A**, DRC^2-Net correctly detects 94.4% of ships (17/18), compared to only 61.1% for YOLOX-Tiny. In **Scenario B**, it achieves 95.7% accuracy (88/92) versus 83.7% for the baseline, reducing missed detections from 13 to 2. In the most challenging **Scenario C**, DRC^2-Net identifies 91.7% of ships (22/24), while YOLOX-Tiny captures only 58.3%. These improvements highlight the model's robustness, particularly in cluttered and noise-dominant SAR environments, achieving up to a **+33.4% gain** in detection accuracy over the baseline.

Table 7. Comparison of detection results across three SAR scenarios in terms of correct, wrong, and missed ship detections.

Scenario Type	Model	Correct	Wrong	Missed
A	GT	18	0	0
	YOLOX-Tiny	11	3	4
	DRC^2-Net	17	–	1
B	GT	92	0	0
	YOLOX-Tiny	77	2	13
	DRC^2-Net	88	2	2
C	GT	24	0	0
	YOLOX-Tiny	14	3	7
	DRC^2-Net	22	2	–

5. Conclusion

This paper presented DRC^2-Net, a lightweight extension of YOLOX-Tiny tailored for SAR ship detection. By integrating recurrent attention in the backbone and deformable convolutions in CSP-based fusion layers, the model improves semantic reasoning and geometric adaptability while preserving real-time efficiency.

On the SSDD dataset, DRC^2-Net achieves clear gains over the baseline YOLOX-Tiny: AP@50 increases by +0.9% (to 93.04%), AP_s by +1.31% (to 91.15%), AP_m by +1.22% (to 88.30%), and AP_l by +13.32% (to 89.47%). Consistent improvements were also confirmed on the more challenging iVision-MRSSD dataset, the proposed model achieves detection accuracies of 94.4%, 95.7%, and 91.7% across Scenarios A, B, and C, respectively, surpassing YOLOX-Tiny and demonstrating strong generalization in diverse maritime conditions. Qualitative results further reinforce its robustness in complex SAR environments.

Importantly, no single solution fits all SAR ship detection tasks. Effective frameworks must balance accuracy, efficiency, and adaptability to mission-specific requirements and environmental constraints. Overall, the model establishes a favorable trade-off between accuracy, efficiency, and memory footprint, providing a practical solution for real-time SAR ship detection. Future work will explore pruning, quantization, and rotated bounding boxes to further enhance deployment efficiency and precision in complex maritime environments.

References

1. Yasir, M.; Niang, A.J.; Hossain, M.S.; Islam, Q.U.; Yang, Q.; Yin, Y. Ranking Ship Detection Methods Using SAR Images Based on Machine Learning and Artificial Intelligence. *Journal of Marine Science and Engineering* **2023**, *11*, 1916.
2. Zhang, Y.; Hao, Y. A survey of SAR image target detection based on convolutional neural networks. *Remote Sensing* **2022**, *14*, 6240.

 doi:10.20944/preprints202509.2458.v1

3. Li, J.; Xu, C.; Su, H.; Gao, L.; Wang, T. Deep learning for SAR ship detection: Past, present and future. *Remote Sensing* **2022**, *14*, 2712.
4. Guan, T.; Chang, S.; Wang, C.; Jia, X. SAR Small Ship Detection Based on Enhanced YOLO Network. *Remote Sensing* **2025**, *17*, 839.
5. Zhang, L.; Zhang, Z.; Lu, S.; Xiang, D.; Su, Y. Fast superpixel-based non-window CFAR ship detector for SAR imagery. *Remote Sensing* **2022**, *14*, 2092.
6. Rihan, M.Y.; Nossair, Z.B.; Mubarak, R.I. An improved CFAR algorithm for multiple environmental conditions. *Signal, Image and Video Processing* **2024**, *18*, 3383–3393.
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. 2012, Vol. 25.
8. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015, Vol. 28.
10. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 821–830.
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn **2017**. pp. 2961–2969.
12. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction* **2023**, *5*, 1680–1716.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European conference on computer vision. Springer, 2016, pp. 21–37.
14. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection **2019**. pp. 9627–9636.
15. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems* **2016**, *29*.
16. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention receptive pyramid network for ship detection in SAR images. IEEE, 2020, Vol. 13, pp. 2738–2756.
17. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
18. Tang, X.; Zhang, J.; Xia, Y.; Xiao, H. DBW-YOLO: A high-precision SAR ship detection method for complex environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 7029–7039.
19. Zhou, K.; Zhang, M.; Wang, H.; Tan, J. Ship detection in SAR images based on multi-scale feature extraction and adaptive feature fusion. *Remote Sensing* **2022**, *14*, 755.
20. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship detection in large-scale SAR images via spatial shuffle-group enhance attention. *IEEE Transactions on Geoscience and Remote Sensing* **2020**, *59*, 379–391.
21. Sun, Z.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images. *Remote Sensing* **2021**, *13*, 4209.
22. Liu, Y.; Ma, Y.; Chen, F.; Shang, E.; Yao, W.; Zhang, S.; Yang, J. Yolov7osar: A lightweight high-precision ship detection model for Sar images based on the yolov7 algorithm. *Remote Sensing* **2024**, *16*, 913.
23. Hu, Q.; Hu, S.; Liu, S. BANet: A balance attention network for anchor-free ship detection in SAR images. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–12.
24. Zhou, H.; Chen, P.; Li, Y.; Wang, B. Enhanced detection method for small and occluded targets in large-scene synthetic aperture radar images. *Journal of Marine Science and Engineering* **2023**, *11*, 2081.
25. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognition* **2021**, *112*, 107787.
26. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* **2021**.
27. LearnOpenCV. Object detection based on lightweight YOLOX for autonomous driving, 2023.
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE, 2015, Vol. 37, pp. 1904–1916.

 doi:10.20944/preprints202509.2458.v1

30. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
32. Li, J.; Yu, Z.; Yu, L.; Cheng, P.; Chen, J.; Chi, C. A comprehensive survey on SAR ATR in deep-learning era. *Remote Sensing* **2023**, *15*, 1454.
33. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 11863–11874.
34. Dai, W.; Liu, R.; Wu, Z.; Wu, T.; Wang, M.; Zhou, J.; Yuan, Y.; Liu, J. Exploiting Scale-Variant Attention for Segmenting Small Medical Objects. arXiv 2024. *arXiv preprint arXiv:2407.07720*.
35. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation **2019**. pp. 603–612.
36. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval* **2018**, *7*, 87–93.
37. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
38. Liu, Y.; Wang, W.; Li, Q.; Min, M.; Yao, Z. DCNet: A deformable convolutional cloud detection network for remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters* **2021**, *19*, 1–5.
39. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9308–9316.
40. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sensing* **2021**, *13*, 3690.
41. Humayun, M.F.; Bhatti, F.A.; Khurshid, K. iVision MRSSD: A comprehensive multi-resolution SAR ship detection dataset for state of the art satellite based maritime surveillance applications. *Data in Brief* **2023**, *50*, 109505.
42. Khanam, R.; Hussain, M. What is YOLOv5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:2407.20892* **2024**.
43. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976* **2022**.
44. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
45. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.
46. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-time flying object detection with YOLOv8. *arXiv preprint arXiv:2305.09972* **2023**.
47. Ren, X.; Bai, Y.; Liu, G.; Zhang, P. YOLO-Lite: An efficient lightweight network for SAR ship detection. *Remote Sensing* **2023**, *15*, 3771.
48. Wang, H.; Shi, J.; Karimian, H.; Liu, F.; Wang, F. YOLOSAR-Lite: a lightweight framework for real-time ship detection in SAR imagery. *International Journal of Digital Earth* **2024**, *17*, 2405525.
49. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9759–9768.
50. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* **2024**, *37*, 107984–108011.
51. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* **2024**.