remote sensing

MDPI

Article

A Two-Stage Framework for SAR Near-Shore Ship Detection via Segmentation Guidance and Enhanced Diffusion

Yuanjie Bai [1], Hangzai Luo [2], Lulu Liu [1] and Sheng Zhong [2,*]

[1] Xi'an Microelectronics Technology Institute, Xi'an 710065, China; baiyuanjie@mail.nwpu.edu.cn (Y.B.); 1102121264@stu.xidian.edu.cn (L.L.)
[2] School of Electronic Information, Northwest University, Xi'an 710069, China; hzluo@nwu.edu.cn
* Correspondence: szhong@nwu.edu.cn; Tel.: +86-153-1946-9503

Abstract

Detecting ships near the coast in Synthetic Aperture Radar (SAR) images is an important task. However, achieving accurate detection in these scenarios remains a significant challenge. The complex coastal topologies and multiple scattering effects frequently induce severe shore–sea feature aliasing, which conventional end-to-end detectors struggle to untangle due to their inherent architectural conflicts between background suppression and fine-grained localization. To address this issue, we propose a two-stage generative framework named Segmentation Guidance and Enhanced Diffusion (SGED). In the first stage, an Enhanced Attention U-Net (EAU-Net) is specifically tailored for robust shore–sea separation. By integrating adaptive Signal-to-Noise Ratio (SNR) masking, lightweight Transformer bottlenecks, and an Edge-Aware Composite Loss, EAU-Net isolates the maritime search space, achieving a Dice Similarity Coefficient (DSC) of 0.9047 and an 85.30% Near-Shore Coverage Accuracy (NSCA). Building upon this refined prior, the second stage introduces a SAR-Enhanced Diffusion Detector (SAR-DDet). It constructs a structural–statistical dual-verification mechanism by embedding Pixel Difference Convolution (PDC) and Constant False Alarm Rate (CFAR) soft attention into the multi-scale features. Coupled with a Normalized Wasserstein Distance (NWD) loss and a four-step DDIM iterative denoising process, SAR-DDet effectively mitigates small-target gradient vanishing and corrects bounding box coordinate quantization errors. Experiments on a near-shore subset of the HRSID benchmark demonstrate that SGED achieves competitive performance. It achieves an mAP of 62.35%, an AP_{75} of 74.67%, and a small-target AP_S of 60.14%, with consistent improvements over monolithic baseline architectures on this dataset.

Keywords: near-shore ship detection; semantic segmentation; diffusion model; feature aliasing; SAR

Check for updates

Academic Editors: Veraldo Liesenberg, Zhiyong Lv, Zenghui Zhang, Kefeng Ji, Zhongzhen Sun and Haohao Ren

Received: 3 June 2026
Revised: 14 July 2026
Accepted: 25 July 2026
Published: 4 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Synthetic Aperture Radar (SAR) is a crucial technology for all-weather maritime surveillance; however, accurate ship detection in near-shore environments remains a significant challenge. Unlike open-ocean scenarios, coastal regions are characterized by severe shore–sea feature aliasing [1]. This phenomenon is driven by complex electromagnetic scattering physics: geometric distortions (e.g., radar layover and shadow effects) cause land masses to physically intrude into maritime domains; multiple scattering from man-made shore infrastructures (e.g., wharves; breakwaters) generates high-intensity echoes that mimic ship signatures; and azimuth ambiguities frequently project ghost targets onto the

adjacent sea surface [2]. Consequently, target signals are deeply entangled with terrestrial clutter, leading to an inherently low signal-to-clutter ratio and high false alarm rates.

To combat these challenges, deep learning architectures have undergone a continuous evolution in the SAR domain. As systematically mapped in the evolutionary timeline of Figure 1, the trajectory has transitioned from foundational convolutional feature optimization towards global context modeling, and most recently, to generative paradigms.

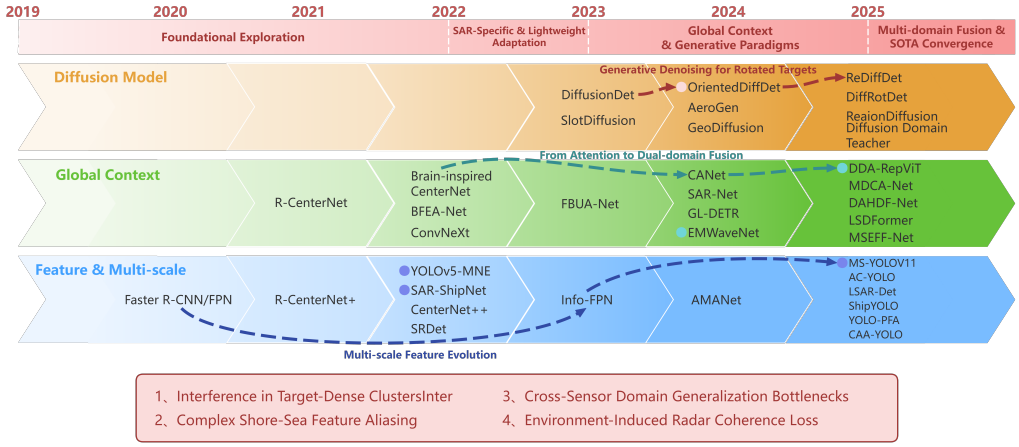


Figure 1. A Review of the evolutionary timeline and current challenges for various algorithm types.

Despite these advancements, addressing the core issue of shore–sea feature aliasing remains difficult. We can group the best current methods into three main types:

Methods Based on Feature Optimization and Multi-scale Enhancement. The core of such methods lies in strengthening target features by improving the Feature Pyramid Network (FPN) or introducing complex attention mechanisms. Faster R-CNN and its FPN variants [3] constructed a robust multi-level feature extraction framework to handle extreme scale variations; building on this, CenterNet [4] achieved stable feature refinement by explicitly modeling object centers and enhancing keypoint triplets. Regarding lightweight design, EGTB-Net [5] combined a lightweight Ghost-ECA module with Transformers, maintaining high detection accuracy while significantly reducing parameters; LSAR-Det [6] utilized the Lightweight Residual Feature Extraction (LRFE) module to further challenge the balance limit between accuracy and parameter count. Regarding SAR characteristic adaptation, SAR-ShipNet [7] added Bidirectional Coordinate Attention (BCA). It also used Multi-Resolution Feature fusion (MRF) and pioneered the elliptic Gaussian probability distribution to adapt to the slender aspect ratio of ships; Info-FPN [8] prevented high-frequency information loss in cascading operations through pixel shuffle lateral connections. Regarding the latest SOTA, AMANet [9] adopted the Adaptive Multi-Hierarchical Attention Module (AMAM) to extract fine-grained information, while MS-YOLOv11 [10] introduced 2D Haar wavelet transform to explicitly extract high-frequency edges and suppress noise from the frequency domain perspective. Although these methods perform excellently on SSDD or HRSID datasets, they perform poorly in near-shore scenarios. Due to the extreme similarity in scattering characteristics between strong shore scatterers and ships, standard attention mechanisms often fail to achieve decoupling at the feature level and may even be misled to enhance land false targets, leading to a high false alarm rate.

Methods Based on Global Context Modeling. These methods aim to utilize Transformers or large receptive field convolutions to capture global dependencies to assist in distinguishing targets from the background. Regarding mechanism fusion, CenterNet [11] simulated the hierarchical processing of the visual cortex, strengthening the fusion of deep

semantics and shallow details; FBUA-Net [12] employed the Global Context-Guided Feature Balance Pyramid (GC-FBP) for anchor-free detection. Regarding multi-modal learning, DDA-RepViT [13] and AFN-YOLO [14] introduced a framework of dual-branch parallel processing in spatial and frequency domains, utilizing Fourier transform to capture global periodic patterns to suppress speckle noise. MDCA-Net [15] maintained structural consistency of features through the Global Shift-Aware Context Aggregation (GSCA) decoder. Regarding Transformer evolution, CANet [16] introduced the Global Context Aggregation Module (GCAM) to acquire long-range dependencies, and LSDFormer [17], based on the MetaFormer architecture combined with position-sensitive attention, achieved an efficient multi-attention mechanism. In complex near-shore ship scenarios, ships, port buildings, and breakwaters are tightly adhered spatially. Excessive strong scattering interference makes the global context information itself full of ambiguity, limiting the effectiveness of these models in handling shore–ship adhesion problems.

Methods Based on Generative Diffusion Models. As an emerging framework in the detection field, diffusion models achieve precise target bounding box generation through an iterative denoising process. Regarding oriented detection, DiffusionDet [18] laid the foundation for shifting from direct regression to iterative refinement; subsequently, OrientedDiffDet [19] successfully extended this methodology to rotation target detection in remote sensing images. Regarding mathematical improvements, ReDiffDet [20] represented oriented boxes as 2D Gaussian distributions, ensuring rotation equivariance of the reverse diffusion process; DiffRotDet [21] used the Kullback–Leibler Divergence (KLD) loss function, addressing the angular periodicity problem in rotation representation. Regarding data augmentation, AeroGen [22] generated high-quality synthetic SAR data via layout-controllable diffusion, effectively alleviating the sample scarcity problem; GeoDiffusion [23] and RegionDiffusion [24] further introduced geometric constraints and text prompts, enhancing the model's zero-shot generalization capability. Directly transferring diffusion models to the SAR domain still faces huge challenges. The inherent coherent speckle noise and low signal-to-noise ratio of SAR images can severely deteriorate the conditional posterior distribution relied upon by diffusion models, leading to information loss or reconstruction shift.

Beyond detection-oriented architectures, physics-driven SAR target characterization methods have established complementary physical priors for maritime scene interpretation. In the polarimetric domain, Chen et al. built the polarimetric coherence/correlation pattern framework [25], which traces scattering matrix variation across the rotation domain to discriminate ships from port infrastructure without relying on spatial pattern matching; this line of work was extended to PolSAR ship detection by Cui et al. [26] and further generalized to a three-dimensional descriptor by Li et al. [27]. In the multi-domain characterization direction, Zhang et al. [28] proposed the polarimetric autocorrelation matrix, jointly encoding inter-polarization coherence and Doppler phase shifts to simultaneously characterize the electromagnetic and kinematic properties of maritime targets, inherently suppressing stationary land scatterers. While the present work operates on single-polarization amplitude images, the statistical anomaly-detection logic of our CFAR soft attention module is conceptually rooted in the same physics-driven discrimination principle, and extending the proposed framework to full-polarimetric input constitutes a natural direction for future investigation.

The aforementioned three methodologies mostly continue the traditional end-to-end detection architectures, forcing a single network to simultaneously solve the highly conflicting tasks of background suppression and target localization. In terms of handling physical perception and electromagnetic scattering characteristics, current architectures still require deeper fusion.

This results in the performance of models in complex near-shore scenarios being consistently constrained by the primary limitation of feature aliasing. Specifically, this overarching bottleneck can be essentially decomposed into three specific sub-challenges: (1) the inherent architectural conflict between large-scale background suppression and fine-grained target localization within a single network; (2) the severe spatial interference caused by complex shore facilities and geometric distortions; and (3) the degradation of multi-scale ship structural features, particularly for small targets, and the difficulty of precise bounding box regression against complex sea clutter. To effectively mitigate these specific challenges, the Segmentation-Guided detection approach has emerged as a highly promising direction.

Its core idea is to decompose the task: first, accurately separating sea and land areas via a semantic segmentation network to effectively suppress land interference, allowing the detector to focus solely on the clean sea surface. However, existing segmentation-guided methods often tend to directly transfer standard general-purpose segmentation networks. When dealing with the coherent speckle noise and geometric distortions intrinsic to SAR images, these models often suffer from limited segmentation accuracy due to the lack of targeted physics-informed designs. As a result, traditional sea–land segmentation algorithms struggle to be applied in practice. These early-stage segmentation errors are directly propagated to downstream detectors [29], often leading to near-shore ships being misclassified as land and permanently masked out. To bypass this severe risk of error propagation, many recent studies reluctantly stick to the end-to-end paradigm, despite its inherent limitations in handling complex feature aliasing.

In view of these intertwined challenges—specifically, (1) the inherent architectural conflict of end-to-end models in target-background separation, (2) the severe error propagation risk in conventional segmentation, and (3) the conditional feature degradation of diffusion models under SAR speckle noise—this paper proposes a two-stage diffusion model framework based on SGED, which explicitly decouples the detection task into a cascaded process to effectively mitigate the shore–sea feature aliasing problem. First, to achieve high-precision shore–sea separation, we design an Enhanced Attention U-Net (EAU-Net). This network specifically embeds a SAR noise suppression module, a Lightweight Transformer Block, and an Edge-Aware Loss function to synergistically suppress speckle noise, capture global context, and refine coastline boundaries. Subsequently, based on the precise sea mask generated by EAU-Net, we design a SAR-Enhanced Diffusion Detector (SAR-DDet). Utilizing DiffusionDet as a baseline, this model incorporates key optimizations for SAR image characteristics: we discard complex general-purpose priors and instead introduce Pixel Difference Convolution (PDC) across multiple FPN layers (P2, P3, and P5) to reinforce target structural information. Additionally, we integrate a Constant False Alarm Rate (CFAR) soft attention module at the P2 and P5 layers to adaptively suppress non-uniform sea surface clutter. This allows the denoising network to effectively guide the diffusion process, accurately reconstructing ship targets while suppressing sea clutter and distinguishing false targets.

The main contributions of this paper are summarized as follows:

(1) We build a new detection system. It works in two steps to resolve the inherent architectural conflict of end-to-end models. Through the cascade of EAU-Net segmentation guidance and SAR-DDet enhanced detection, it explicitly decomposes the complex task into sea–land separation and sea–surface target refinement, providing a targeted approach to alleviating the feature aliasing problem in complex coastal environments.
(2) We designed EAU-Net, a specialized network dedicated to SAR shore–sea segmentation. It systematically integrates five components: an adaptive Signal-to-Noise Ratio (SNR)-based noise suppression module for speckle noise, a lightweight Trans-

former bottleneck for global geometric topology, a Cross-Scale Feature Fusion Module, an Adaptive Edge Refinement Module, and an Edge-Aware Composite Loss. This synergistic design delineates physical boundaries and explicitly mitigates the severe False Negative problem common in complex harbor areas. Consequently, it significantly reduces the Average Symmetric Surface Distance (ASSD) to 10.379 pixels and achieves a leading Near-Shore Coverage Accuracy (NSCA) of 85.30%, providing an optimally clean sea-surface prior for downstream detection.

(3) We propose a multi-scale SAR feature enhancement and progressive regression strategy for SAR-DDet. By extending PDC across the P2, P3, and P5 layers, and applying CFAR soft attention specifically to the high-resolution P2 and strong-semantic P5 layers, we establish a synergistic structural–statistical dual-prior verification mechanism. Incorporating small-target Copy–Paste augmentation and utilizing the Normalized Wasserstein Distance (NWD) loss, the model overcomes the gradient vanishing bottleneck inherent in traditional IoU-based small-target regression. Furthermore, the four-step Denoising Diffusion Implicit Model (DDIM) iterative inference progressively mitigates bounding box coordinate quantization errors, yielding a 3.75% surge in high-precision localization (AP_{75}), ultimately pushing the overall mAP to 62.35% and small target precision (AP_S) to 60.14%.

The remainder of this paper is organized as follows. Section 2 presents the proposed SGED framework in detail, covering EAU-Net for high-precision shore–sea segmentation (Section 2.2), SAR-DDet for enhanced diffusion-based target detection (Section 2.3), and the corresponding optimization strategies and loss functions (Section 2.4). Section 3 describes the experimental datasets, evaluation metrics, and implementation details and presents comprehensive quantitative comparisons, qualitative analyses, and ablation studies. Section 4 provides an in-depth discussion of the results from both methodological and physical perspectives. Section 5 concludes this paper and outlines directions for future work.

2. Materials and Methods

2.1. General Framework

To effectively alleviate the shore–sea feature aliasing problem in SAR near-shore ship detection, this paper proposes a novel framework named SGED, which operates in a two-stage cascaded manner. The proposed SGED framework explicitly decomposes the detection task into two specialized stages to alleviate the feature aliasing problem. The first stage utilizes EAU-Net for high-precision shore–sea separation, aimed at eliminating land background interference at the pixel level. The second stage employs SAR-DDet to perform ship detection within the refined search space of the sea area. This sequential processing ensures that the detector can focus on distinguishing targets from sea clutter without the intrusion of complex land scatterers. The framework structure is illustrated in Figure 2.

Stage 1: Segmentation Guidance. The core objective of this stage is to eliminate land background interference from the original SAR image. We designed an Enhanced Attention U-Net (EAU-Net), which is based on the Attention U-Net backbone and integrates five key specialized modules tailored to SAR image characteristics: (1) A SAR noise suppression module to adaptively handle speckle noise; (2) a Lightweight Transformer Block to capture global context and connect fractured coastlines; (3) a Cross-Scale Feature Fusion Module to combine high-level semantics with low-level details; (4) an Adaptive Edge Refinement Module to precisely correct coastline boundaries; and (5) an Edge-Aware Loss function to reinforce supervision of boundary gradients. EAU-Net receives the original SAR image as input and outputs a high-precision binary shore–sea segmentation mask. This binary mask undergoes a strategic Sea Area Expansion before guiding the detection stage. Specifically,

a morphological dilation is applied to the coastline boundary within the mask. This expansion is essential because ground-truth bounding boxes for near-shore ships—often affected by layover or shadow—inevitably contain pixels overlapping with the shore. By expanding the sea-domain labels (e.g., by a 20-pixel margin), we ensure that the complete electromagnetic scattering information of the labeled boxes is preserved. This prevents the "information truncation" that would occur with a strict pixel-level mask, thereby minimizing the risk of error propagation from segmentation to detection. The algorithm flow is shown in Figure 3.

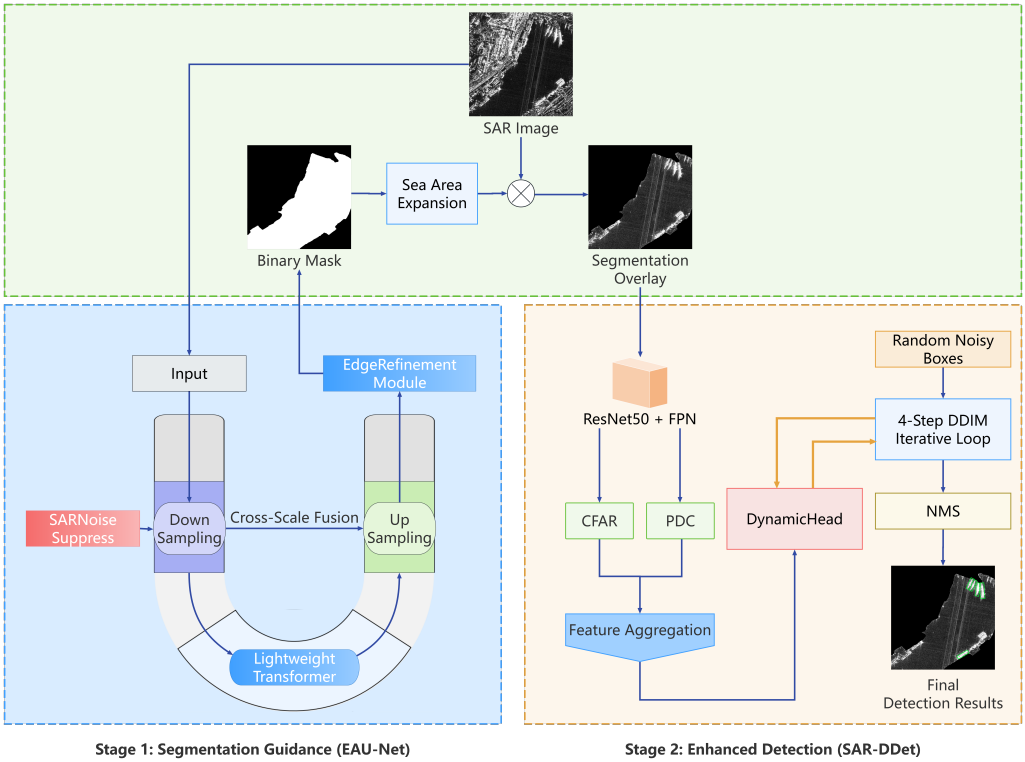


Figure 2. A two-stage cascaded framework of SGED, illustrating the inference-time architecture. Note that Copy–Paste data augmentation (detailed in Section 2.4.3) is a training-time strategy applied exclusively during training and is therefore not depicted in this architectural overview.

Stage 2: Enhanced Target Detection. This "clean" sea area image (rather than the original image) serves as the input for the second-stage detector. We designed a SAR-Enhanced Diffusion Detector (SAR-DDet). Using DiffusionDet as a baseline, this model discards complex general-purpose priors and instead adopts a lightweight enhancement strategy tailored to SAR physical characteristics: parallel embedding of PDC across multiple FPN layers (P2, P3, and P5), and a CFAR Soft Attention module specifically applied to the high-resolution P2 and strong-semantic P5 layers. This detector efficiently suppresses residual sea surface clutter and enhances ship target structural features on the segmented sea area, ultimately generating precise bounding boxes through an iterative denoising process. The algorithm flow is shown in Figure 4.

This two-stage design achieves task decoupling: the first-stage EAU-Net is specifically dedicated to handling large-scale spatial sea-land separation issues, ensuring the integrity of the target search space; meanwhile, the second-stage SAR-DDet is specifically dedicated to handling complex small targets and inter-class distinction issues against a low signal-to-noise ratio background, thereby achieving significant overall performance gains.

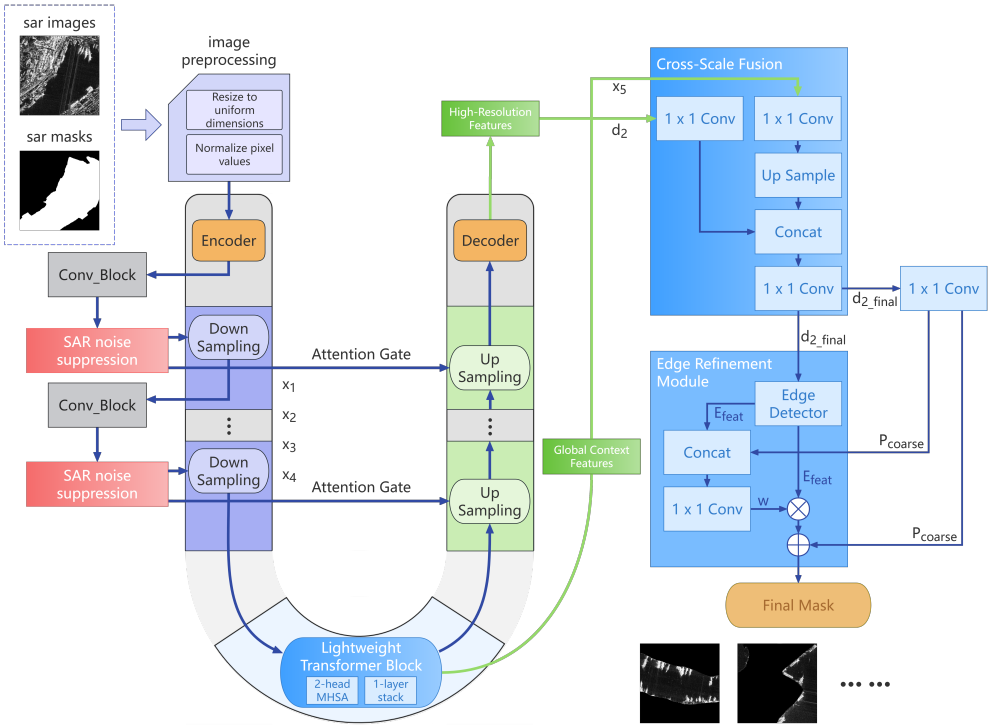


Figure 3. Detailed architecture of the proposed EAU-Net for SAR shore–sea segmentation. The framework integrates SAR Noise Suppression modules at each encoder level, a Lightweight Transformer at the bottleneck, and a Cross-Scale Fusion path linking global semantics with local details. The final boundary is optimized by the Edge Refinement Module with an adaptive weighting mechanism.

2.2. Shore–Sea Segmentation Based on EAU-Net

To achieve high-precision shore–sea separation in the first stage, we designed an Enhanced Attention U-Net (EAU-Net). This network uses Attention U-Net as its basic skeleton, but its encoder, bottleneck, decoder, and loss functions have been deeply customized for unique SAR image physical characteristics (such as coherent speckle noise, low SNR, and complex geometric structures). The detailed architecture of EAU-Net is shown in Figure 3, composed of five synergistic core modules.

2.2.1. SAR Noise Suppression

Coherent speckle noise (Speckle Noise) inherent in SAR images is a multiplicative noise with statistical properties distinct from Gaussian noise. Traditional convolution operations easily misidentify strong speckle noise as valid texture features or over-smooth real edge details during denoising.

To address this, we embedded a SAR Noise Suppression Module at each downsampling stage of the encoder. The core idea of this module is adaptive feature modulation based on local Signal-to-Noise Ratio (SNR), and its workflow is shown in Figure 5.

For an input feature map x, we first estimate the local mean $\mu(p)$ and variance $\sigma^2(p)$ within a 3×3 neighborhood $\mathcal{N}(p)$ using average pooling. This window size represents the minimal kernel that captures all eight immediate spatial neighbors; it is well-suited to SAR speckle whose correlation length spans one to two resolution cells, providing statistically stable local moment estimates without over-smoothing the high-frequency edge information critical for ship boundary detection. Enlarging the window to 5×5 or beyond progressively blurs the SNR map and weakens ship-edge discrimination, while a 1×1 estimate is numerically unstable under multiplicative speckle. Subsequently, the SNR is formulated as follows:

$$SNR(p) = \frac{|\mu(p)|}{\sqrt{\max(\sigma^2(p), 10^{-8})}} \tag{1}$$

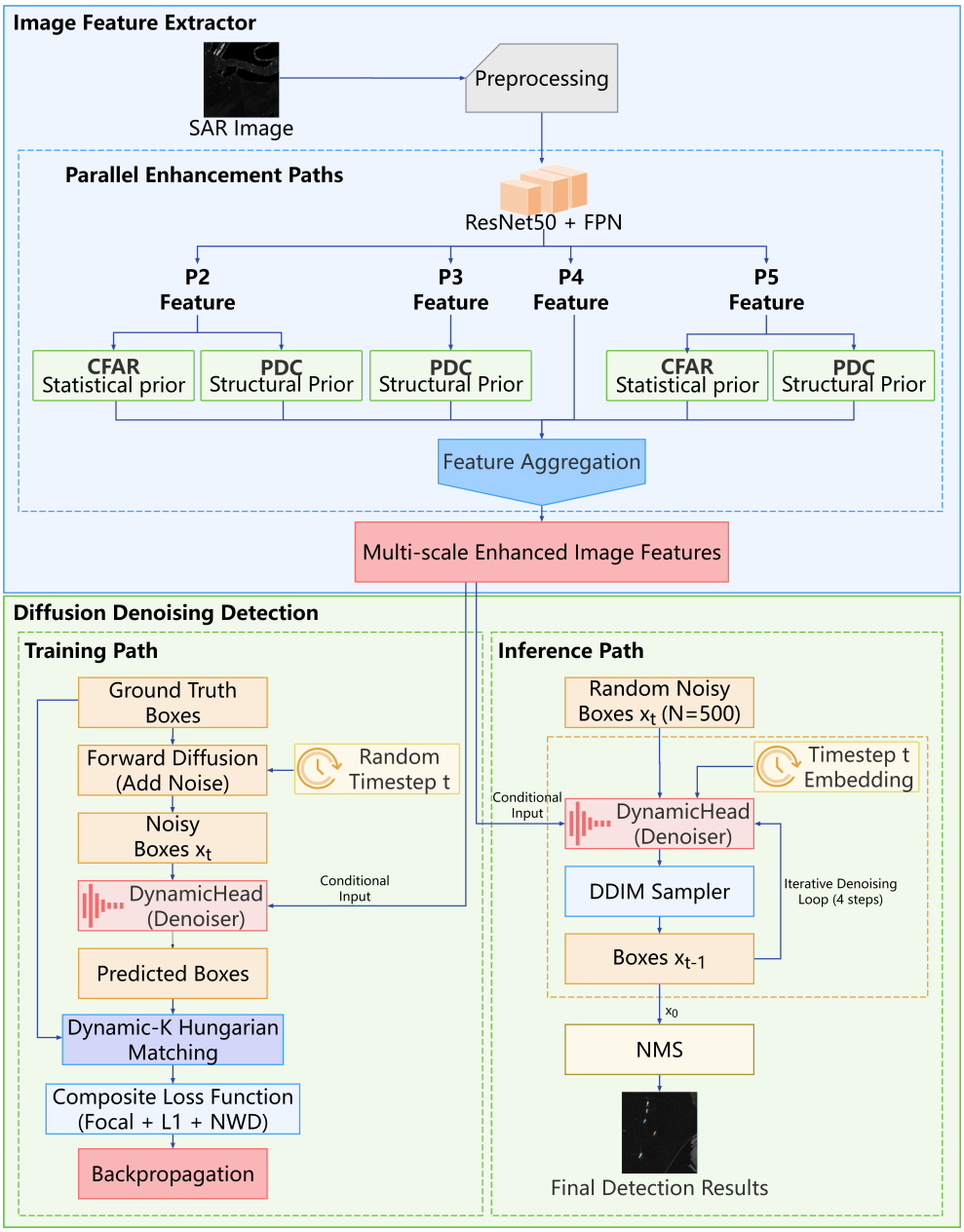


Figure 4. Illustration of the overall framework of our proposed SAR-Enhanced Diffusion Detector, taking the merged sea area image of the segmentation mask and the original image as input and outputting the predicted target image.

Based on the SNR, we construct an adaptive mask:

$$m(p) = Clamp(Sigmoid(2 \cdot (SNR(p) - r)), 0.1, 1.0) \tag{2}$$

where $Sigmoid(\cdot)$ is the standard sigmoid activation function, and the suppression ratio r is set to 0.2. Equation (2) is an original design: the $Sigmoid$ centered at r constructs a smooth, differentiable gate in which pixels with $SNR(p) > r$ (valid structural signal) receive mask values approaching 1.0, while pixels with $SNR(p) < r$ (noise-dominated regions) are suppressed toward 0.1. The scaling factor of 2 produces a moderately steep

transition to discriminate between noisy ($SNR < 0.1$) and signal-dominated ($SNR > 0.3$) regions without abrupt discontinuities. The threshold $r = 0.2$ is set based on empirical analysis of SAR homogeneous sea and land regions, where local SNR typically falls within [0.1, 0.3]. The lower clamp bound of 0.1 preserves non-zero gradient flow, preventing the suppression branch from becoming permanently inactive during training. Subsequently, this mask is multiplied element-wise with the input feature and processed by a Depthwise Convolution to obtain the enhanced representation $x_{enhanced}$. To maintain the integrity of the original backbone feature distribution, we adopt a mild refinement strategy via a residual update:

$$\hat{x} = 0.9 \cdot x + 0.1 \cdot f_{\mathrm{DW}}(x \odot m) \tag{3}$$

where the 0.9/0.1 ratio ensures that the suppression branch provides a conservative 10% correction to the backbone features. This prevents the aggressive disruption of learned representations while depthwise separable convolution (f_{DW}) is employed to minimize computational overhead. This design effectively attenuates speckle noise in homogeneous regions without compromising the structural integrity of high-frequency edges.

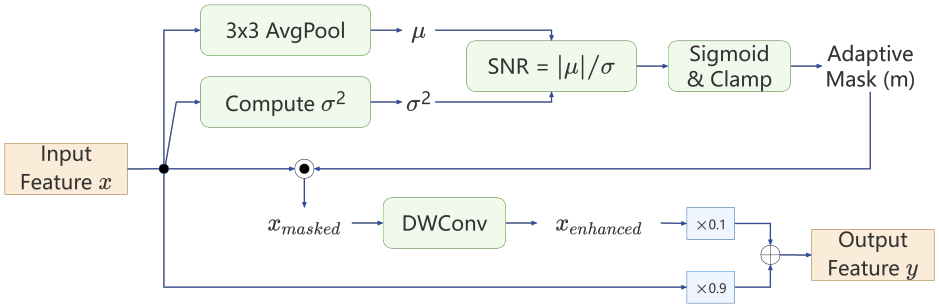


Figure 5. Schematic diagram of the SAR Noise Suppression module.

2.2.2. Lightweight Transformer Block

The receptive field of standard CNNs is strictly limited by kernel size, causing them to struggle with making correct inferences when dealing with large-scale coastline interruptions caused by geometric distortion due to a lack of global context. To address this, we introduced a Lightweight Transformer Block at the network bottleneck (lowest resolution 16×16). Specifically, given the deep feature map output by the encoder as $X_{in} \in \mathbb{R}^{C \times H \times W}$ (where $C = 1024$), we first flatten its spatial dimensions to convert it into a 1D sequence. To restrict computational overhead, if the sequence length $N = H \times W$ exceeds 1024, adaptive average pooling is applied prior to flattening.

Pixel-level long-range dependencies are then established globally through a Multi-Head Self-Attention (MHSA) mechanism utilizing a dual-head self-attention mechanism to balance representational capacity and computational efficiency on 1024-dimensional features. We adopt a Pre-LayerNorm architecture with a residual coefficient of 0.1:

$$X' = X + 0.1 \cdot \mathrm{MHA}(\mathrm{LayerNorm}(X)) \tag{4}$$

$$X'' = X' + 0.1 \cdot \mathrm{FFN}(\mathrm{LayerNorm}(X')) \tag{5}$$

The Feed-Forward Network (FFN) employs a 1× expansion ratio instead of the standard 4×, as the 1024-dimensional bottleneck features already possess sufficient representational capacity; this reduction saves approximately 75% of parameters. Crucially, a small residual coefficient of 0.1 is implemented for "soft injection" of global context, ensuring that the self-attention features complement rather than overwhelm the spatial hierarchies established by the convolutional layers.

2.2.3. Cross-Scale Feature Fusion

The cascaded upsampling process in standard U-Net decoders intrinsically dilutes high-level global context. To mitigate this semantic degradation, the proposed Cross-Scale Feature Fusion Module constructs a direct routing path. This path bridges the low-resolution global semantic features from the encoder bottleneck, denoted as $F_{global} \in \mathbb{R}^{1024\times16\times16}$, with the high-resolution spatial details from the terminal decoder stage, denoted as $F_{local} \in \mathbb{R}^{64\times256\times256}$. To strictly achieve dimensional and semantic alignment, both feature maps are first projected into a unified 64-channel latent space via 1×1 convolutions, followed by Batch Normalization (BN) and ReLU activation. The bottleneck feature is subsequently aligned spatially with the local feature via bilinear upsampling, intentionally circumventing transposed convolutions to prevent checkerboard artifacts. The aligned features are then concatenated and fused:

$$F_{fused} = BN(ReLU(W_F * [F_{local}||\hat{F}_{global}])) \tag{6}$$

where $\hat{F}_{global}$ represents the spatially upsampled bottleneck feature, $||$ denotes channel-wise concatenation, and W_F represents the 3×3 convolution weights applied for fusion.The 3×3 kernel is selected because it provides the minimal spatial receptive field necessary to capture local cross-feature correlations after concatenation; a 1×1 kernel performs only channel-wise mixing without spatial integration, whereas larger kernels introduce disproportionate overhead without meaningful accuracy gains at this fusion stage.

2.2.4. Adaptive Edge Refinement Module

The terminal decoder of U-Net directly outputs the final prediction, which easily produces irregular or blurred coastline boundaries. To achieve precise pixel-level correction, we added an Edge Refinement Module after the main decoder path. This module adopts a two-stage prediction strategy, as shown in Figure 6. First, an edge heatmap E is extracted via a lightweight edge detection network consisting of a 3×3 convolution reducing the 64 channels to 16, followed by a 1×1 convolution outputting a single-channel edge map. The 3×3 kernel matches the effective support of classical Sobel-type edge operators and captures first-order gradient responses within the 8-neighborhood—the scale at which coastline transitions manifest in the decoder feature space—while larger kernels would reduce sensitivity to the localized boundary displacements that this module is designed to correct. An adaptive weight $w \in (0, 1)$ is then generated by analyzing the concatenated coarse prediction P_{coarse} and the edge heatmap E through a 1×1 convolution and a Sigmoid activation.

The final refined prediction $P_{refined}$ is given by the adaptive weighting formula:

$$P_{refined} = P_{coarse} + \alpha \cdot w \odot E \tag{7}$$

where the refinement coefficient α is strictly set to 0.05. This small coefficient ensures the module performs conservative micro-adjustments only in edge regions deemed necessary by the network, without disrupting the structural correctness of the coarse prediction.

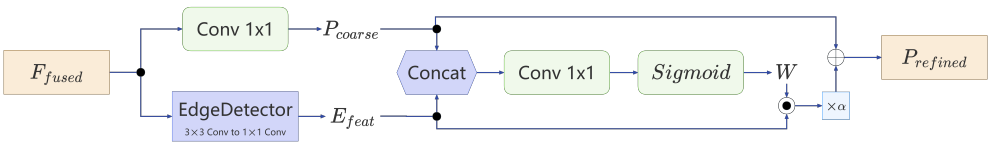


Figure 6. Schematic diagram of the Adaptive Edge Refinement Module.

2.2.5. Synergistic Mechanism for Resolving Feature Aliasing

It is important to note that the suppression of shore–sea feature aliasing is not handled by a single isolated module, but is rather achieved through the synergistic operation of the specific network components introduced above (Sections 2.2.1–2.2.4). By integrating these modules, EAU-Net systematically addresses the three primary challenges defined in the introduction:

(1) Suppressing Speckle Noise and Multiple Scattering: Operationally, the SAR Noise Suppression module (Section 2.2.1) first acts as a low-level filter, attenuating spatially isolated strong speckle noise via adaptive local SNR weighting. Subsequently, the Lightweight Transformer (Section 2.2.2) acts as a high-level spatial filter; by learning global continuity priors, it assigns low attention weights to structurally isolated "ghost" targets (caused by multiple scattering) that do not conform to physical coastline distributions, effectively preventing them from propagating to the decoder.

(2) Mitigating Geometric Distortion: To address coastline interruptions caused by radar layover or shadow effects, the Lightweight Transformer extracts a global field of view at the network bottleneck. The Cross-Scale Feature Fusion module (Section 2.2.3) then explicitly transmits this global topological prior directly to the high-resolution decoder. This routing operation guides the network to perform robust semantic interpolation across the interrupted regions based on the macro-trend, thereby restoring the spatial continuity of the coastline.

(3) Refining Adhered Boundaries: For complex harbor areas where ships and land structures are tightly adhered, the Adaptive Edge Refinement Module (Section 2.2.4) utilizes an auxiliary edge heatmap to perform explicit pixel-level correction on the coarse prediction features. Combined with the Edge-Aware Loss (detailed in Section 2.4.1), which penalizes gradient mismatches during optimization, this dual-operation ensures the final segmentation mask delineates physical boundaries, effectively mitigating land interference for the subsequent detection stage.

2.3. Enhanced Target Detection Based on Diffusion Model

After EAU-Net removes land interference, the goal of the second stage is to address the two major challenges inherent in SAR images on the "clean" sea area: sea clutter suppression and target structure identification. To this end, we designed a SAR-Enhanced Diffusion Detector (SAR-DDet). Utilizing DiffusionDet as a baseline, this model reformulates object detection as an iterative denoising process from random noise boxes to real target boxes. Our core innovation lies in avoiding the direct use of raw features from standard backbone networks; instead, we designed a SAR-specific feature enhancement layer to construct richer conditions to guide the diffusion process.

2.3.1. DiffusionDet Baseline

DiffusionDet reformulates the detection of object bounding boxes as a generative denoising Markov process. Instead of relying on heuristic anchors, it directly generates bounding boxes from random noise boxes conditioned on the extracted image features.

Forward Diffusion Process: During training, the process starts with a fixed number of ground-truth boxes, denoted as $x_0 \in \mathbb{R}^{N_{train} \times 4}$, where coordinates are normalized. This process gradually adds Gaussian noise ϵ to x_0 over T time steps. According to DDPM and reparameterization tricks, the noisy boxes x_t at any arbitrary time step t can be directly sampled:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{8}$$

where $\overline{\alpha}t = \prod i = 1^t (1 - \beta_i)$ are predefined noise scheduling coefficients.

Reverse Denoising Process: The core of the model is a denoising neural network ϵ_θ. To bridge the mathematical process with our detection task, ϵ_θ explicitly requires three inputs:

The noisy boxes $x_t \in \mathbb{R}^{N_{eval} \times 4}$ (during inference, 500 proposal boxes are initialized purely from $\mathcal{N}(0, I)$).

The temporal embedding of the current time step t.

The conditional information C, which typically consists of standard FPN feature maps.

However, directly using standard features as condition C fails in near-shore SAR scenarios due to severe feature aliasing. Therefore, in our framework, the conditional input C is strictly formulated using our SAR-specific enhanced features ($C_{enhanced}$, detailed in Section 2.3.2). Guided by this optimized $C_{enhanced}$, ϵ_θ is trained to predict the added noise ϵ, iteratively refining the N_{eval} random boxes into accurate ship coordinates. The inference stage employs a 4-step DDIM sampling strategy utilizing 500 proposal boxes. Compared to single-step inference, this iterative multi-step approach progressively refines the bounding box coordinates, significantly improving strict localization accuracy (e.g., AP_{75}) at zero additional training cost.

2.3.2. SAR-Specific Feature Enhancement

The performance of the diffusion model is highly dependent on the quality of the conditional information, denoted as C. In conventional diffusion-based detectors, this condition C is typically constructed using raw, unoptimized multi-scale features directly extracted from a general Feature Pyramid Network (FPN). However, when directly applied to SAR images, these unoptimized FPN features exhibit severe semantic ambiguity and clutter interference. Therefore, we designed a parallel, targeted enhancement strategy, this strategy operates across multiple semantic levels (P2, P3, and P5) of the FPN to ensure comprehensive enhancement from fine-grained textures to high-level semantics. This enhancement layer consists of two parallel branches:

(1) Pixel Difference Convolution (PDC)

To address edge blurring caused by Radar Cross Section (RCS) fluctuations due to pose angle changes and speckle noise in SAR targets, we introduced PDC to replace standard convolution on the P2, P3, and P5 layers. Given that SAR small ship targets (area < 32 × 32 pixels) are predominantly distributed in the high-resolution P2 and P3 layers, extending PDC to these levels is critical for amplifying small-target boundary textures and suppressing multiplicative speckle noise through relative intensity differencing. PDC enhances perception of edges and textures by calculating the difference between a pixel and its neighbors, while remaining robust to uniform speckle noise. We employ a central configuration for PDC, with its working principle shown in Figure 7.

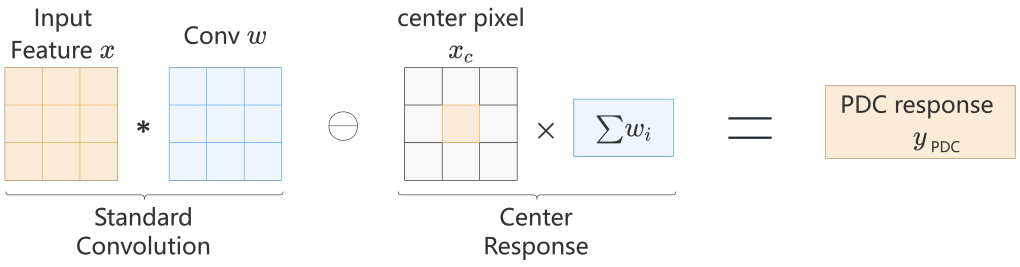


Figure 7. Schematic diagram of Pixel Difference Convolution.

The calculation formula is as follows:

$$y_{\text{PDC}} = \underbrace{\sum_{i,j}(w_{i,j} \cdot x_{i,j})}_{\text{Conv}} - \underbrace{x_{\text{center}} \cdot \sum_{i,j} w_{i,j}}_{\text{CPR}} \tag{9}$$

where $x_{i,j}$ are neighboring pixels, x_{center} is the center pixel value, and $w_{i,j}$ are convolution kernel weights. The first term (labeled Conv) computes the standard weighted sum of neighboring pixels, while the second term (labeled CPR, Central Pixel Recalibration) subtracts the center pixel's weighted contribution, thereby encoding the relative intensity difference between the center and its surroundings. The PDC module provides the denoising network with a structural prior that is more robust to SAR imaging defects while retaining original features.

(2) CFAR Soft Attention

To suppress large-scale non-uniform sea surface clutter, we apply a Cell-Averaging CFAR (CA-CFAR) soft attention module to both the high-resolution P2 layer and the low-resolution P5 layer, achieving dual-scale clutter suppression. CA-CFAR is adopted because post-land-removal sea surface clutter approximates a locally homogeneous distribution, for which the cell-averaging estimator provides a statistically stable and computationally efficient threshold without requiring additional rank-based operations. As illustrated in Figure 8, for statistical prior extraction, the 7 × 7 background window yields 40 valid estimation cells after guard exclusion, sufficient for stable local statistics at both FPN resolutions; the 3 × 3 guard corresponds to a 12-pixel spatial extent at P2 stride (×4), conservatively covering the typical small ship footprint to prevent target energy from biasing the clutter estimate. To guarantee training stability, a zero-initialization strategy is strictly adopted for the attention weights. This configuration prevents the disruption of pre-trained backbone representations during the initial training epochs, ensuring that the statistical clutter suppression priors are integrated progressively via backpropagation. The normalized response z is fed into a sigmoid function to generate soft attention weights w:

$$w = \text{sigmoid}(\alpha \cdot z) = \sigma\left(\alpha \cdot \frac{x - \mu_{\text{local}}}{\sigma_{\text{local}} + \epsilon_0}\right) \tag{10}$$

where α is a learnable soft modulation factor initialized to 0, and ϵ_0 is a precisely set small constant (10^{-8}) to ensure numerical stability during the variance normalization process. Finally, this attention weight enhances the original features via a residual connection: $x' = x \cdot w + x = x \cdot (w + 1)$. This module provides a critical statistical prior for the denoising network.

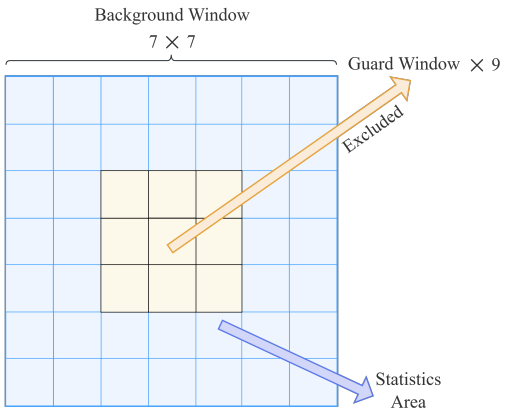


Figure 8. Schematic diagram of the CFAR Module.

2.3.3. Dynamic Head

The multi-scale features $C_{enhanced}$, enhanced by the parallel PDC and CFAR modules, are fed as the vital conditional inputs into the detection head of the diffusion model. To effectively translate these SAR-specific prior features into precise bounding boxes, this detection head employs a 6-layer cascaded Transformer decoder.

Specifically, at each denoising timestep t, the dynamic head utilizes the current noisy boxes x_t to extract Region of Interest (RoI) features directly from the multi-scale $C_{enhanced}$ maps. The cascaded decoder then processes these features through two core mechanisms:

(1) Self-Attention: It explicitly models the spatial relationships and potential overlaps between different ship proposals. In complex near-shore scenarios, this mechanism is particularly beneficial for distinguishing densely clustered ships (e.g., ships docked side-by-side at a port), effectively avoiding the suppression of valid targets.
(2) Dynamic Instance Interaction: It dynamically fuses the pooled RoI feature map information with the proposal embeddings. This allows the network to adaptively focus on the strong structural and statistical responses provided by the PDC and CFAR priors within each bounding box.

By deeply interacting with the optimized SAR features, the dynamic head guides the iterative denoising process to progressively strip away the Gaussian noise, refining the box coordinates and predicting the final ship categories.

2.3.4. Analysis of the Dual-Prior Verification Mechanism

It is important to clarify that this mechanism is not an additional network block, but rather the underlying operational logic of how the enhanced condition $C_{enhanced}$ (comprising PDC and CFAR features) physically guides the diffusion denoising process (described in Section 2.3.3).

After EAU-Net removes the land background, two residual interference categories persist on the sea surface. The first is distributed Bragg-resonance sea clutter, whose backscatter follows a locally homogeneous statistical distribution; CA-CFAR suppresses this effectively by flagging cells that deviate significantly from the surrounding mean. However, azimuth ambiguities—coherent ghost echoes arising from SAR pulse-repetition periodicity—produce isolated high-intensity returns whose statistical profile is indistinguishable from a real ship and thus evade the CFAR gate entirely. The second category involves these azimuth ambiguities and residual multiple-scattering ghosts: although statistically anomalous, they are geometrically featureless point-like returns, lacking the distributed structural gradients that metal ship hulls generate through corner-reflector and dihedral scattering at bow, stern, and superstructure boundaries; consequently, they yield weak PDC responses and are screened out by the structural prior. Wave-breaking boundaries can occasionally activate PDC via transient edge gradients, but their surrounding backscatter remains statistically homogeneous and is eliminated by the CFAR gate.

By integrating these complementary priors into the conditional input $C_{enhanced}$, we implicitly establish a "Statistical–Structural" dual-verification manifold for the dynamic head. Operationally, during the iterative reverse denoising trajectory, the diffusion network evaluates bounding box proposals by sampling this conditional space. Authentic ship targets satisfy both criteria: they exhibit distinct statistical anomalies against the sea background (triggering high CFAR responses) and possess complex geometric boundaries or corner reflectors (generating strong PDC gradients). Consequently, the denoiser assigns high confidence to these proposals, accurately reconstructing their coordinates. Conversely, azimuth ambiguities or multiple-scattering ghosts—while statistically anomalous—lack the high-frequency internal textures of real ships, yielding weak PDC responses. Similarly, strong sea clutter may present local edge gradients but fails the uniform statistical gating

of the CFAR module. Rather than relying on heuristic post-processing, this mathematically embedded dual-prior naturally steers the conditional distribution of the diffusion model, forcing it to dynamically discard proposals that fail this joint physical verification, thereby reducing false alarms in complex near-shore scenes.

2.4. Optimization and Loss Functions

The two stages of SGED are trained in a fully decoupled sequential manner, with no joint fine-tuning at any point. EAU-Net (Stage 1) is trained first to convergence on the shore–sea segmentation dataset, after which its weights are frozen. The frozen EAU-Net is then applied offline to preprocess all images, generating the masked sea-area images that serve as the fixed training inputs for SAR-DDet (Stage 2). SAR-DDet is subsequently trained from scratch on these masked inputs, with no gradient propagated back to EAU-Net. Each stage is optimized with its own task-specific loss function.

2.4.1. Edge-Aware Composite Loss for EAU-Net (Stage 1)

To address the issue of standard Binary Cross Entropy (BCE) loss lacking sensitivity to boundary pixels, we constructed a tripartite composite loss function. First, a standard Sobel operator is applied to the ground truth mask Y to extract the geometric edge map E_{gt}. A pixel-wise weighted Binary Cross Entropy ($\mathcal{L}_{\text{wBCE}}$) is employed, where the gradient contribution of pixels situated on the coastline boundary is amplified by a factor of $1 + \lambda_e$ relative to internal pixels, enforcing larger gradients at the coastline:

$$\mathcal{L}_{wBCE} = \frac{1}{HW} \sum_{i,j} (1 + \lambda_e \cdot E_{gt}(i,j)) \cdot \mathcal{L}_{BCE}(p(i,j), y(i,j)) \tag{11}$$

where $p(i,j)$ and $y(i,j)$ denote the predicted probability and the ground truth at pixel (i,j), respectively.

Second, a Dice Loss $\mathcal{L}dice$ is incorporated to directly optimize the region overlap and mitigate class imbalance. Finally, an edge consistency L1 loss $\mathcal{L}edge$ is computed by applying the Sobel operator to the predicted probability map to constrain the morphological alignment with the true edges. The total segmentation loss is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{wBCE}} + \lambda_d \cdot \mathcal{L}_{\text{dice}} + \lambda_s \cdot \mathcal{L}_{\text{edge}} \tag{12}$$

Following grid search optimization, the hyperparameters are fixed at $\lambda_e = 1.0$, $\lambda_d = 0.5$, and $\lambda_s = 0.1$, ensuring balanced gradient flow between region-based overlap and boundary-based morphological consistency.

2.4.2. Composite Loss for SAR-DDet (Stage 2)

For the enhanced detection stage, we employ Hungarian matching during the training stage to find the optimal one-to-one assignment between predictions and ground truth. The total loss function $\mathcal{L}_{det}$ is calculated only on successfully matched pairs:

In the matching stage, the dynamic-K Hungarian algorithm is employed. Traditional Intersection over Union (IoU) metrics fail to provide stable gradients when predicted and ground-truth boxes for small targets have minimal overlap. To resolve this gradient vanishing issue, we replace GIoU with the Normalized Wasserstein Distance (NWD). By treating bounding boxes as 2D Gaussian distributions, NWD maintains strong gradient signals. Consequently, it performs robustly even when there is minimal overlap between predicted and ground-truth boxes. This characteristic makes the loss more tolerant to minor positional shifts and significantly more robust to the extreme scale variations of small ships, thereby effectively breaking the performance bottleneck of traditional IoU-based small-target regression. The NWD is formulated as follows:

$$NWD(b_i, b_j) = \exp\left(-\frac{W_2(b_i, b_j)^{1/2}}{c}\right) \tag{13}$$

where $c = 12.8$. The total loss function $\mathcal{L}det$ is defined as follows:

$$\mathcal{L}_{det} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{NWD}\mathcal{L}_{NWD} \tag{14}$$

where $\mathcal{L}_{cls}$ is focal loss; $\mathcal{L}_{L1}$ and $\mathcal{L}_{NWD}$ jointly supervise precise localization. The weights are set to $\lambda_{cls} = 2.0$, $\lambda_{L1} = 5.0$, and $\lambda_{NWD} = 2.5$, selected via grid search on the validation set. λ_{L1} is assigned the highest weight as precise coordinate regression is the primary bottleneck for small-target detection; $\lambda_{cls} = 2.0$ follows the standard DiffusionDet focal-loss configuration; and $\lambda_{NWD} = 2.5$ supplements $L1$ for small-target gradient stability. Moderate weight adjustments (±50%) produce mAP variations within approximately 0.5%, while extreme reweighting between λ_{L1} and λ_{NWD} causes training instability due to conflicting gradient directions. Additionally, a background weight of 0.25 is applied to penalize false positives.

2.4.3. Data Augmentation and Training Strategy

To alleviate the severe scarcity of small target samples in SAR datasets, we implement a targeted Copy–Paste small object augmentation during the training stage. Ship instances smaller than 32×32 pixels are randomly cropped from the training set and pasted into the current image with a 50% probability (maximum of 5 instances per image), significantly enriching the diversity of small-target distributions. Furthermore, the network is trained with a fixed input resolution of 800×800 pixels to eliminate spatial information loss for small targets during the downsampling process. To counter the potential increase in False Positives induced by the augmented sample sensitivity, we elevate the background penalty weight (NO_OBJECT_WEIGHT) to 0.25, ensuring an optimal balance between precision and recall.

3. Results

This section presents a comprehensive evaluation of the proposed SGED framework. We first assess the shore–sea segmentation performance of EAU-Net, then evaluate the detection performance of the full two-stage system, and finally conduct ablation studies to isolate the contribution of each component.

3.1. Datasets and Experimental Settings

3.1.1. Datasets

To facilitate the training and evaluation of our two-stage framework, this study constructed two specialized datasets for shore–sea segmentation and near-shore ship detection, respectively. It should be noted that the two stages of the SGED framework are trained and evaluated on decoupled datasets to ensure the generalization of the cascaded system.

Shore–Sea Segmentation Dataset: This dataset is utilized to train and evaluate the first-stage EAU-Net, comprising a total of 5909 single-polarization images. To ensure scene diversity while preventing data leakage, we implemented a source-isolated partitioning strategy. Specifically, 4905 images aggregated from two public datasets (sea–land segmentation V1.1 and Sea-Land-Seg-Data) were randomly partitioned into a training set (4169 images) and a validation set (736 images) at an 0.85:0.15 ratio. This split was executed using a fixed random seed (SEED = 42) to guarantee full reproducibility. To rigorously evaluate the model's generalization capabilities in completely unseen environments, 1004 images featuring typical near-shore scenes were strictly reserved as the independent test set. These test images were exclusively sourced from the self-built YGA1-SSDD dataset

with precise manual annotations, ensuring scene independence between the training and testing phases.

Near-Shore Object Detection Dataset: This dataset is dedicated to evaluating the second-stage SAR-DDet. We systematically filtered the near-shore subset of the public HRSID dataset, yielding 985 images characterized by complex coastal topologies and multi-scale ship targets (a total of 7768 bounding box instances). To strictly align with our two-stage methodology, these 985 images were first processed by the trained EAU-Net for sea–land segmentation, followed by morphological sea area expansion to generate the actual inputs for our SAR-DDet. In the primary comparison (detailed in Section 3.3), baseline detectors are evaluated on the original unmasked images to assess their inherent detection capability under full land interference. To further isolate the contribution of each stage, two supplementary controlled experiments are additionally conducted: (1) all baseline detectors are retrained and evaluated under the identical EAU-Net masked input protocol as SGED; and (2) SAR-DDet is evaluated independently on the original unmasked images, bypassing EAU-Net preprocessing. The results are presented in the subsequent detection evaluation section. The 985 images were statically partitioned into training (787 images, 80%), validation (99 images, 10%), and test (99 images, 10%) sets based on temporal and geographical non-overlap. The image IDs across the three subsets are strictly mutually exclusive, ensuring absolute scene independence and providing a highly reliable benchmark for target detection.

3.1.2. Experimental Environment

The design, training, and evaluation of the proposed framework were implemented with the PyTorch 1.12.1 deep learning framework, Python 3.9, and CUDA 11.6 on a Windows 10 operating system. To accelerate the experiments and ensure reproducibility, all processing was conducted on a unified hardware platform equipped with an Intel Core i7-14700K CPU and two NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory.

3.1.3. Evaluation Metrics

(1) Shore–Sea Segmentation (Stage 1)

The selection of evaluation metrics in this study is specifically tailored to the unique challenges of SAR near-shore maritime surveillance. We employ a multi-dimensional metric suite to ensure a comprehensive assessment:

Region Overlap and Pixel-wise Accuracy: We use Accuracy (ACC), the Dice Similarity Coefficient (DSC), and Intersection over Union (IoU) to see how well the predicted areas match the actual ones. Given the extreme class imbalance between sea and land in large-scale SAR scenes, DSC and IoU are prioritized as they are less sensitive to background pixel dominance than standard ACC.

Detection-Oriented Reliability: Recall and Precision are included to evaluate the fundamental trade-off in segmentation. In our two-stage framework, Recall is the most critical metric, as missing a ship-containing sea pixel (False Negative) in the first stage would lead to irreversible target loss in the subsequent detection stage.

Boundary and Morphological Precision: To quantify the fidelity of the extracted coastline, we employ Average Symmetric Surface Distance (ASSD) and Edge Preservation Index (EPI). ASSD measures the average displacement between predicted and actual boundaries in pixels; a lower ASSD reflects superior land-interference suppression. EPI evaluates the preservation of high-frequency structural details.

Task-Specific Metric (NSCA): To quantitatively evaluate the segmentation reliability in the critical interface where shore–sea aliasing is most severe, we utilize Near-Shore Coverage Accuracy (NSCA). Unlike global metrics, NSCA focuses exclusively on the buffer

zone where radar backscattering from land and sea is most frequently aliased. It is defined as the Recall metric specifically calculated within a morphological buffer zone extending from the ground-truth coastline. The formula is expressed as follows:

$$NSCA = \frac{|P_{sea} \cap G_{sea} \cap B|}{|G_{sea} \cap B|} \quad (15)$$

where P_{sea} and G_{sea} denote the predicted and ground-truth sea regions, respectively. B represents the near-shore buffer mask generated by dilating the ground-truth coastline with a kernel size of 41×41 (corresponding to a 20-pixel buffer radius). A higher NSCA indicates that the model more effectively preserves the potential maritime target space by minimizing the risk of sea pixels being misclassified as land in the shore–sea transition zone.

(2) Object Detection (Stage 2)

To evaluate ship detection performance, we adopt the standard COCO evaluation protocol, focusing on Mean Average Precision (mAP) and its variants across different IoU thresholds and scales. Specifically, we report mAP, AP_{50}, AP_{75}, AP_S (small targets), and AP_M (medium targets). The selection of these specific metrics is driven by the intrinsic challenges of SAR near-shore surveillance and the unique characteristics of our diffusion-based framework:

AP_{75} (High-Precision Localization): Unlike the more lenient AP_{50}, AP_{75} serves as a rigorous indicator of strict localization accuracy. It is specifically chosen to validate the unique contribution of the four-step DDIM iterative refinement process in correcting bounding box coordinate quantization errors.

AP_S (Small-target Detection): Near-shore SAR scenes are characterized by a high density of small ships, such as fishing boats and speedboats. AP_S is the decisive metric for assessing the system's reliability in minimizing the miss-detection rate for these small, low-observable targets in cluttered environments.

mAP and AP_M: These metrics provide a balanced assessment of the framework's overall detection stability and its performance on medium-scale targets where sea clutter interference is prevalent.

3.1.4. Implementation Details

To ensure the stable convergence of the two-stage framework and optimize performance across different tasks, we implemented stage-specific training and inference configurations.

(1) Stage 1: Shore–Sea Segmentation (EAU-Net)

The base parameters of EAU-Net were updated with an initial learning rate of 1×10^{-4}, while the specialized enhancement modules (SAR noise suppression, Transformer, and edge refinement) utilized a half-speed learning rate of 5×10^{-5}. We configured the Adam optimizer with a β_1 coefficient of 0.5 to better capture the high-dynamic-range characteristics of SAR data. A Cosine Annealing scheduler was employed over 20 epochs to decay the learning rate to 1×10^{-6}. To maintain structural integrity, data augmentation was limited to horizontal flipping (0.5 probability), while rotation and color jittering were omitted.

(2) Stage 2: Enhanced Target Detection (SAR-DDet)

For the detection stage, we adopted a more refined strategy based on the optimized exp016 configuration. The network was trained with a fixed input resolution of 800×800 pixels to eliminate information loss for small-targets during downsampling. We applied the AdamW optimizer with a base learning rate of 1.25×10^{-5},derived via the linear scaling rule from the DiffusionDet baseline rate of 2.5×10^{-5} (batch size 16), scaled by 0.5 to match our effective batch size of 8 constrained by GPU memory at 800×800 resolution,

and validated for stable convergence on the validation set. Crucially, the CFAR soft attention weights were strictly zero-initialized to guarantee training stability. Furthermore, we increased the background penalty weight (no-object weight) to 0.25 to effectively suppress False Positives. During the inference stage, a four-step DDIM sampling strategy with 500 proposal boxes was utilized to iteratively refine the bounding box coordinates. This specific configuration was found to be critical for achieving the notable improvement in high-precision localization observed in our results.

3.2. Shore–Sea Segmentation Results

To verify the effectiveness of the first-stage EAU-Net, we comprehensively compared it with representative state-of-the-art segmentation algorithms. It is important to clarify that to ensure a rigorous evaluation of coastline boundary adherence, all quantitative and qualitative segmentation experiments in this section are conducted strictly on the precise segmentation masks, without applying the Sea Area Expansion strategy (detailed in Section 2.1). The morphologically dilated masks are exclusively generated to secure the target search space for the second-stage detection.

3.2.1. Quantitative Analysis of Segmentation

To evaluate the shore–sea separation capability of EAU-Net, we compared it against seven representative segmentation architectures. These baselines were selected to cover a diverse range of paradigms: classic convolutional networks (U-Net [30], DeepLabv3+ [31]), attention-enhanced CNNs (Att-UNet [32]), Transformer-based architectures (TransUNet [33], SegNeXt [34]), emerging Vision Mamba models (VM-UNet [35]), and domain-adapted foundation models (SAM-Med2D [36]). The quantitative comparison results are presented in Table 1.

Table 1. Comprehensive comparison of segmentation performance across different algorithms.

Model	ACC ↑	DSC ↑	IOU ↑	R ↑	ASSD ↓	EPI ↑	NSCA ↑
U-Net	0.7910	0.7731	0.6302	0.6531	15.277	0.2382	0.5760
TransUNet	0.8121	0.7987	0.6649	0.6837	14.714	0.2455	0.6099
DeepLabv3+	0.7847	0.7606	0.6137	0.6270	16.237	0.1749	0.5264
Att-UNet	0.8203	0.7770	0.6763	0.7099	14.062	0.3341	0.6348
SegNeXt	0.8597	0.8170	0.7308	0.7724	23.907	0.2309	0.6792
SAM-Med2D	0.9057	0.8802	0.8109	0.8335	16.868	0.3052	0.8173
VM-UNet	0.8091	0.7906	0.6185	0.7129	14.785	0.2587	0.6338
SG (Ours)	**0.9089**	**0.9047**	**0.8297**	**0.8990**	**10.379**	**0.3806**	**0.8530**

↑ indicates that higher values are better, and ↓ indicates that lower values are better. Bold values highlight the best performance for each metric.

As demonstrated in Table 1, the proposed method achieves state-of-the-art performance on six out of seven metrics, including DSC (0.9047), IoU (0.8297), Recall (0.8990), ASSD (10.379), EPI (0.3806), and NSCA (0.8530). Compared to SAM-Med2D, the state-of-the-art foundation model fine-tuned for specialized domains, our method achieves superior DSC (+0.0245), IoU (+0.0188), and Recall (+0.0655). Notably, our framework is the only one to exceed the 0.90 DSC threshold, a commonly adopted threshold for high-quality segmentation in medical and remote sensing imaging tasks. We achieve a 38.5% reduction in ASSD (from 16.868 to 10.379 pixels). This indicates that while ViT-based models like SAM excel at capturing global features, they lack the fine-grained boundary localization precision of our Edge Refinement Module. Furthermore, the 3.57 percentage point absolute improvement in NSCA confirms that our design significantly reduces the risk of misclassifying near-shore ships as land, effectively preserving the maritime search space for downstream tasks.

Methods utilizing standard [0, 255] normalization (e.g., U-Net, TransUNet, and DeepLabv3+) exhibit severe Recall degradation (ranging from 0.62 to 0.68) when applied to SAR imagery. This indicates that the direct application of natural-image-trained architectures is insufficient for addressing SAR-domain near-shore segmentation. Our SAR-specific physics-informed design yields a 31.5% Recall improvement over the best-performing method in this category (TransUNet).

While our method may produce slightly more False Positive predictions compared to conservative predictors like DeepLabv3+, such baselines achieve this conservatism at the cost of extremely low Recall (<0.65), resulting in inferior overall DSC (<0.80). Our method adopts a balanced precision–recall profile (Recall: 0.8990) specifically optimized for the near-shore ship detection scenario, where minimizing missed targets (False Negatives) is prioritized over false alarms. The leading NSCA (0.8530) confirms that our framework effectively reduces target loss risk in the critical shore–sea buffer zone, ensuring that valid ships are well preserved for the subsequent detection stage.

3.2.2. Qualitative Analysis of Segmentation

To visually demonstrate the advantages in addressing core challenges, we selected typical scenes for visual comparison, as shown in Figure 9.

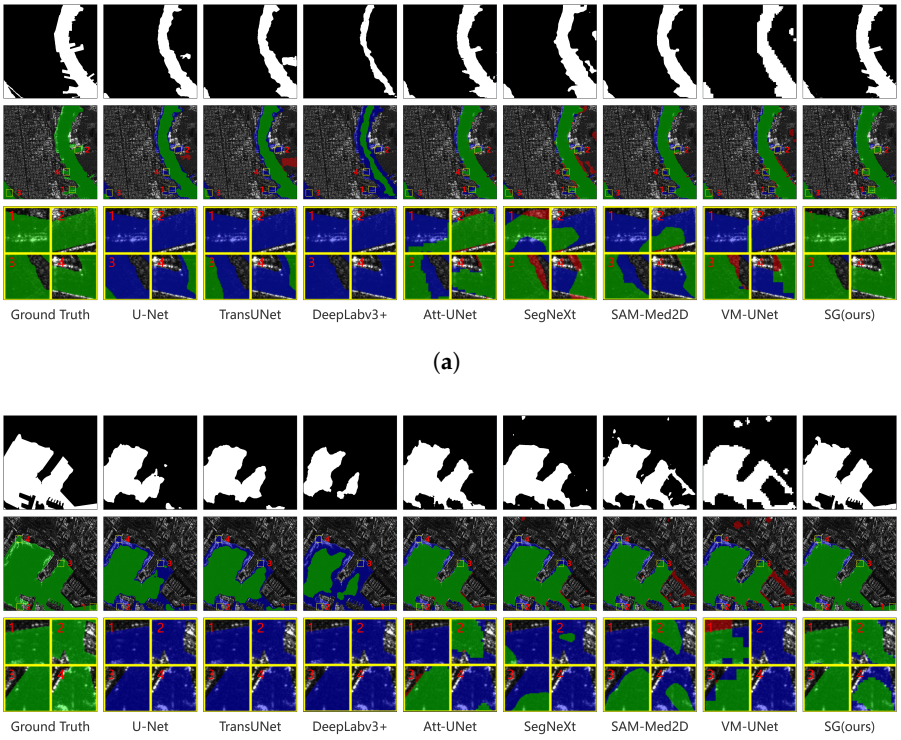


Figure 9. Qualitative evaluation of segmentation performance across (**a**) multi-class near-shore ships and (**b**) multi-type coastlines using color-coded error maps. To provide a standardized visual assessment, Green denotes True Positives (correctly classified sea/ship pixels), Red represents False Positives (land misclassified as sea), and Blue signifies False Negatives (sea/ships misclassified as land). The magnified patches in the lower rows illustrate local boundary adherence and ship-wharf separation within the yellow-boxed regions.

To qualitatively assess the robustness of EAU-Net against the inherent complexities of SAR near-shore environments, Figure 9 presents a comparison of segmentation results across diverse ship categories and coastline topologies. An analysis of the error-mapped visualizations reveals that baseline methods frequently experience the erosion of ship

features or adhesion between ships and harbor infrastructures. Specifically, in the magnified views of (Figure 9a), comparison models exhibit extensive Blue regions (False Negatives), indicating that near-shore ships are erroneously integrated into land-domain predictions. Conversely, EAU-Net preserves the morphological integrity of targets with a predominant Green response (True Positives), successfully isolating ships even under intense shore–sea feature aliasing. This high-precision sea-land separation ensures the effective suppression of land background interference, thereby providing a reliable search space for the subsequent detection stage.

Furthermore, (Figure 9b) highlights the framework's superiority in maintaining coastline integrity and smoothness across complex port facilities and natural shores. While most comparative methods produce boundaries characterized by burrs, fractures, or over-smoothing—visible as Red clusters (False Positives) where land features overflow into the sea area—EAU-Net generates contours that strictly adhere to real artificial structures. This performance underscores the synergistic impact of the Lightweight Transformer in modeling global geometric topologies to reconnect fractured segments, alongside the Edge-Aware Loss in enforcing precise pixel-level boundary supervision. These qualitative results confirm that EAU-Net achieves superior structural consistency, effectively alleviating the feature aliasing bottleneck in complex maritime scenes within the evaluated dataset.

3.2.3. Ablation Study for Shore–Sea Segmentation

The ablation study results for the segmentation stage are presented in Table 2. We analyze the contribution of each proposed component progressively.

Table 2. Ablation study results of each module.

Config	SAR	Trans	Cross	Edge	Loss	ACC ↑	DSC ↑	IOU ↑	Recall ↑	ASSD ↓	EPI ↑	NSCA ↑
E1	⊗	⊗	⊗	⊗	BCE	0.8203	0.7770	0.6763	0.7099	14.0623	0.3341	0.6348
E2	✓	⊗	⊗	⊗	BCE	0.8384	0.8112	0.7182	0.7689	12.4157	0.3576	0.6923
E3	✓	✓	⊗	⊗	BCE	0.8402	0.8274	0.7433	0.8304	10.6582	0.3752	0.7768
E4a	✓	✓	✓	⊗	BCE	0.8845	0.8843	0.7982	0.8295	12.1934	0.3938	0.7554
E4b	✓	✓	⊗	✓	BCE	0.8483	0.8286	0.7450	0.8111	10.9352	0.3639	0.7505
E4	✓	✓	✓	✓	BCE	0.8836	0.8819	0.7970	0.8355	11.6986	**0.3940**	0.7651
E5	✓	✓	✓	✓	EAL	**0.9089**	**0.9047**	**0.8297**	**0.8990**	**10.3792**	0.3806	**0.8530**

SAR: SAR Noise Suppress; Trans: Light weight Transformer; Cross: Cross Scale Fusion; Edge: Edge Refine; EAL: Edge-Aware Loss. ↑ indicates that higher values are better, and ↓ indicates that lower values are better. Bold values highlight the best performance for each metric. ✓ denotes that a module is applied, and ⊗ denotes that it is not applied.

The progressive ablation detailed in Table 2 isolates the specific functional contributions of the proposed segmentation modules. To provide intuitive physical evidence for these numerical gains, Figure 10 visualizes the feature evolution using pixel-level error maps. We analyze the mechanisms driving these performance improvements as follows:

Speckle Attenuation and Topology Restoration (E1 to E3): The feature evolution visualized in Figure 10 uncovers the sequential resolution of geometric and radiometric distortions. The baseline configuration (E1) exhibits severe fragmentation, characterized by scattered False Negatives (Blue patches) in homogeneous sea regions induced by coherent speckle noise. The integration of the SAR Noise Suppression module (E2) effectively eradicates these discrete artifacts, yielding a 3.42% DSC improvement by adaptively modulating features based on the local SNR. Building upon this cleaner feature space, the addition of the Lightweight Transformer bottleneck (E3) directly addresses structural disconnections. Large gaps along the coastline—typically induced by radar layover or shadow effects—are successfully reconnected, delivering a substantial 8.45 percentage point surge in NSCA. This confirms the efficacy of the dual-head self-attention mechanism in capturing long-range geometric dependencies to enforce macroscopic topological consistency.

Semantic Alignment and Boundary Polish (E3 to E5): While E3 restores global continuity, it suffers from semantic overflow where sea features intrude into terrestrial zones (Red patches). The introduction of Cross-Scale Feature Fusion (E4a) explicitly constrains this overflow, contracting the boundaries and significantly mitigating False Positive predictions. Compared to configuration E4a, the Edge Refinement Module (E4) further synergizes with this global constraint by targeting localized spatial displacements, sharpening the land–sea interface and reducing the ASSD by 0.49 pixels. Finally, replacing the standard BCE with the Edge-Aware Loss (E5) provides the terminal optimization. By assigning strictly heavier gradient penalties to boundary pixels, this composite loss thoroughly eliminates residual hair-like artifacts at the immediate ship-wharf interfaces, ultimately achieving a highly precise isolation of the maritime search space.

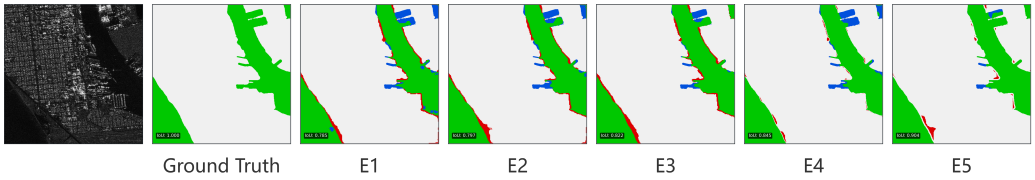


Figure 10. Qualitative evolution of the segmentation ablation study from the baseline (E1) to the proposed EAU-Net (E5). To clearly demonstrate the performance gains, error maps are employed where Green, Red, Blue, and Gray denote True Positives (TPs), False Positives (FPs), False Negatives (FNs), and True Negatives (TNs), respectively. The progression illustrates a systematic refinement trajectory characterized by noise suppression, structural connectivity, and boundary adherence.

3.3. Object Detection Results

3.3.1. Quantitative Analysis of Object Detection

This section evaluates the final performance of our two-stage framework. Having demonstrated in Section 3.2 that EAU-Net provides high-precision segmentation masks with a Near-Shore Coverage Accuracy (NSCA) of 85.30% and an Average Symmetric Surface Distance (ASSD) of 10.379 pixels, we now evaluate the detection capabilities of SAR-DDet on this optimized input.

To verify the overall superiority of this framework, we comprehensively compared the proposed two-stage approach with eight representative state-of-the-art (SOTA) object detection algorithms.It is noted that in this primary comparison (Table 3), baseline detectors are applied to the original images to evaluate their intrinsic detection capabilities under full land interference, while SGED operates on EAU-Net-masked images consistent with its two-stage design; a controlled same-input comparison is separately provided in Table 4. The baselines cover four mainstream architectures: real-time detectors (YOLOv11 [37]; YOLOv8 [38]), Transformer-based detectors (DINO [39]; RT-DETR [40]), classic two-stage detectors (Faster R-CNN [41]), and SAR-specific methods (CANet; DiffDet4SAR [42]), alongside the original DiffusionDet.

As detailed in Table 3, the complete SGED framework achieves an optimal mAP of 62.35%, showing consistent improvements over real-time detectors (e.g., YOLOv11; YOLOv8), Transformer-based architectures (e.g., DINO; RT-DETR), and classic two-stage methods. A notable improvement is observed in strict localization accuracy: our method reaches an AP_{75} of 74.67%, outperforming the latest YOLOv11 by 2.41 percentage points and the baseline DiffusionDet by 7.55 percentage points. This improvement provides a meaningful technical benefit for high-precision maritime surveillance, where accurate coordinate refinement is essential for multi-sensor data fusion, such as Automatic Identification System (AIS) integration.

Furthermore, for small ships, our framework achieves a leading AP_S of 60.14%, yielding a 30.84-point margin over Faster R-CNN (29.30%) and a 6.01-point improvement over

the baseline DiffusionDet. This confirms that the integration of the NWD loss—modeling bounding boxes as 2D Gaussian distributions—effectively overcomes the gradient vanishing bottleneck inherent in traditional IoU-based regression for sparse, small targets. Ultimately, these results validate that our task-decoupled approach and physics-informed enhancements effectively alleviate shore–sea feature aliasing within the evaluated near-shore scenario, yielding consistent performance gains over conventional monolithic end-to-end detection architectures.

Table 3. Comparison of detection performance of different algorithms on the near-shore object detection dataset. Masked: EAU-Net-preprocessed sea-area images; Original: unprocessed HRSID images. Baseline methods are evaluated on original images; SGED operates on masked images, consistent with its two-stage design.

Method	Input	mAP ↑	AP_{50} ↑	AP_{75} ↑	AP_s ↑	AP_m ↑
Faster R-CNN	Original	51.40	81.69	58.42	29.30	36.76
CANet	Original	53.10	81.10	60.10	48.70	66.70
DINO	Original	54.40	81.40	65.00	49.70	68.30
RT-DETR	Original	59.60	87.40	70.00	56.50	70.60
YOLOv8	Original	61.37	88.52	73.21	58.60	70.61
YOLOv11	Original	61.85	89.05	72.26	59.33	71.84
DiffusionDet	Original	58.95	84.14	67.12	54.13	71.91
DiffDet4SAR	Original	52.71	80.40	57.27	46.73	69.65
SGED (Ours)	Masked	**62.35**	**89.62**	**74.67**	**60.14**	**72.50**

↑ indicates that higher values are better. Bold values highlight the best performance for each metric.

Table 4. Controlled-protocol performance comparison for decoupling the contribution of the EAU-Net segmentation stage (input protocol definitions follow Table 3).

Method	Input	mAP ↑	AP_{50} ↑	AP_{75} ↑	AP_s ↑	AP_m ↑
Faster R-CNN	Masked	51.60	82.76	57.55	29.63	39.95
CANet	Masked	52.70	81.40	59.90	48.10	67.50
DINO	Masked	52.90	81.30	62.20	49.20	65.70
RT-DETR	Masked	57.90	84.60	68.70	54.40	71.00
YOLOv8	Masked	60.58	87.99	72.28	58.40	69.74
YOLOv11	Masked	61.22	88.38	72.47	58.98	69.99
DiffusionDet	Masked	58.89	83.84	68.02	54.05	71.78
DiffDet4SAR	Masked	53.75	80.92	61.52	49.06	67.98
SGED (Stage 2 only, Ours)	Original	60.73	89.07	71.74	58.18	69.62
SGED (Ours)	Masked	**62.35**	**89.62**	**74.67**	**60.14**	**72.50**

↑ indicates that higher values are better. Bold values highlight the best performance for each metric.

Table 4 reveals three key findings from the controlled comparison. First, under the identical masked input protocol, SGED (62.35%) still outperforms the best-performing baseline YOLOv11 (61.22%) by 1.13 points, confirming that the performance advantage is not attributable to preprocessing alone. Second, most generic baselines show marginal or negative changes when switching from original to masked input (e.g., DINO: 54.40%→52.90%; RT-DETR: 59.60%→57.90%), indicating that land-area removal does not uniformly benefit detectors not specifically designed to exploit this prior. Third, SGED operating with Stage 2 only on original images achieves 60.73% mAP, which is 1.62 points below the full system, quantifying the specific contribution of EAU-Net and confirming the synergistic nature of the two-stage design.

3.3.2. Qualitative Analysis of Object Detection

To visually demonstrate the performance leap achieved by our framework, we selected a complex harbor scene characterized by tightly docked small ships, intense sea surface clutter, and strong azimuth ambiguities, as shown in Figure 11.

A qualitative inspection of the comparative visualizations reveals the inherent vulnerability of monolithic end-to-end detectors to shore–sea feature aliasing. Conventional architectures (e.g., Faster R-CNN, RT-DETR, and YOLOv11) exhibit high susceptibility to terrestrial background interference, frequently triggering false alarms (Yellow boxes) by erroneously regressing wharf infrastructures or radar azimuth ambiguities as valid ship targets. Furthermore, owing to the scattering degradation of small ships in dense harbor environments, these baseline models suffer from severe morphological erosion, resulting in dense clusters of missed targets (Red boxes) where ships are tightly moored against the shore.

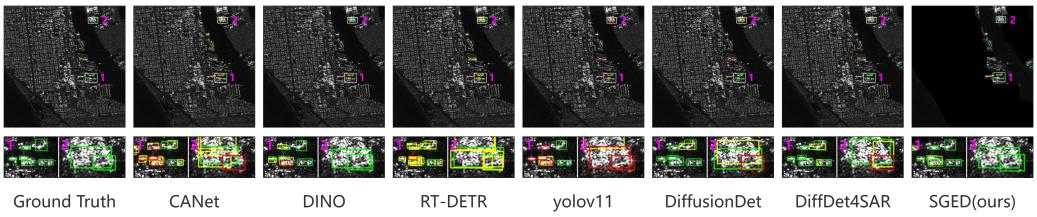


Figure 11. Qualitative comparison of object detection performance in a complex near-shore harbor scene. To provide a rigorous visual evaluation, bounding boxes are color-coded: Green represents True Positives (correctly detected ships), Yellow indicates False Positives (clutter misclassified as ships), and Red denotes False Negatives (missed targets).

In stark contrast, the proposed SGED framework demonstrates anti-interference robustness. By leveraging the prior land-clearance mask generated by EAU-Net, SAR-DDet demonstrates strong robustness against land-based strong scatterers, effectively suppressing the persistent false alarms often seen in baseline methods. Concurrently, governed by the structural–statistical dual priors and the DDIM iterative denoising process, our model effectively addresses the feature entanglement of densely docked small targets, recovering previously omitted ships (Green boxes) with notably improved boundary localization. These visual outcomes directly corroborate the significant quantitative gains in AP_{75} and AP_S, substantiating that our cascaded generative strategy effectively alleviates the inherent limitations of near-shore target detection.

3.3.3. Ablation Study

To quantify the individual and combined contributions of each component in SAR-DDet, we conduct systematic ablation experiments. To isolate the specific gains of each mechanism from the confounding effects of training dynamics, the experiments are structured into two complementary studies: a progressive accumulation study and a cross-factor study.

(1) Progressive Accumulation Study (Table 5)

We introduce the DiffusionDet evaluated on masked inputs as the starting baseline (R0, Table 5), and progressively integrate each proposed component to isolate its individual contribution. R0 adopts the original DiffusionDet architecture with a ResNet-50 backbone and FPN, initialized from COCO-pre-trained weights, trained under the identical resolution, optimizer, and schedule as SAR-DDet, with no proposed enhancements applied.

Detailed analysis of the progressive integration reveals three key mechanisms driving the performance gains:

PDC Feature Enhancement: Introducing PDC across the P2, P3, and P5 FPN layers (R1) yields an mAP of 59.66%, a gain of +0.77% over R0. The improvement is particularly pronounced in AP50 (+2.75%) and APs (+2.12%), confirming that PDC's pixel-difference mechanism effectively amplifies high-frequency structural edges suppressed by SAR speckle noise, directly improving small-target recall at coarse-to-medium precision thresholds.

Table 5. Progressive ablation study results of each module.Components are categorized as (A) architectural modules, (B) loss function, and (C) training/inference strategies.

Config	Category	Modules Added	mAP ↑	AP_{50} ↑	AP_{75} ↑	AP_s ↑	AP_m ↑
R0	-	Base	58.89	83.84	68.02	54.05	71.78
R1	A	R0 + PDC	59.66	86.59	68.07	56.17	71.85
R2	A	R1 + CFAR (P2 + P5)	59.94	86.51	68.89	55.78	73.20
R3	B	R2 + NWD Loss	60.35	86.97	69.12	56.58	72.63
R4	C	R3 + HyperOpt (STEP = 1)	60.05	86.91	68.48	57.10	71.30
R5	C	R4 + SAMPLE_STEP = 4	61.53	88.31	72.23	59.13	71.25
R6	C	R5 + Copy-Paste Aug	61.64	89.44	74.03	59.08	**72.67**
R7	C	R6 + FP Penalty	**62.35**	**89.62**	**74.67**	**60.14**	72.50

Base: DiffusionDet evaluated on EAU-Net-masked images under the same training configuration as SAR-DDet; PDC: Pixel Difference Convolution; CFAR: Constant False Alarm Rate soft attention; FP: False Positive penalty. ↑ indicates that higher values are better. Bold values highlight the best performance for each metric.

Zero-Cost Generative Refinement: A critical finding in the progressive accumulation is the impact of the generative inference process. Increasing the DDIM sampling steps from 1 to 4 (R4 to R5) yields a notable 3.75-point improvement in AP_{75} and a 1.48-point increase in overall mAP. Crucially, this represents a 'zero-training-cost' enhancement, as the performance gain is achieved purely through generative iterative refinement without altering model weights or incurring additional training overhead. This confirms that the iterative denoising formulation inherent in diffusion models provides a decisive advantage in achieving sub-pixel localization accuracy for dense near-shore ships.

Small-Target Diversity and False Positive Control: The structural diversity of small targets in SAR imagery is often insufficient. Integrating Copy–Paste augmentation (R6) significantly enriches this distribution, pushing AP_{50} up by 1.13 points and AP_{75} by 1.80 points. To counter the slight increase in False Positives induced by heightened model sensitivity, optimizing the background penalty weight (R7) yielded a final 1.06% increase in AP_S. This systematic optimization ultimately drives the complete SAR-DDet model to an optimal mAP of 62.35%, ensuring high robustness in complex shore–sea aliasing scenarios.

(2) Cross-Factor Study for Feature Enhancement (Table 6)

To strictly isolate the independent contributions of the statistical prior (CFAR) and the regression optimization (NWD), a 2 × 2 cross-factor ablation was conducted under identical training budgets (10,000 iterations) using the baseline weights.

Table 6. Cross-ablation study for CFAR and NWD.

Config	CFAR	NWD	mAP ↑	AP_{50} ↑	AP_{75} ↑	AP_s ↑	AP_m ↑
F1	⊗	⊗	59.66	86.59	68.07	56.17	71.85
F2	✓	⊗	59.94	86.51	68.89	55.78	**73.20**
F3	⊗	✓	59.85	86.68	**69.90**	**56.62**	72.01
F4	✓	✓	**60.35**	**86.97**	69.12	56.58	72.63

CFAR: Constant False Alarm Rate; NWD: Normalized Wasserstein Distance. ↑ indicates that higher values are better. Bold values highlight the best performance for each metric. ✓ denotes that a module is applied, and ⊗ denotes that it is not applied.

The isolated analysis in Table 6 clarifies the distinct and complementary roles of the statistical prior and the regression optimization. The 2 × 2 cross-factorial design, controlled under identical training budgets, explicitly quantifies their independent contributions.

The CFAR soft attention establishes a robust statistical prior that adaptively filters uniform sea clutter, yielding a targeted 1.35-point improvement for medium-sized targets (AP_m). Conversely, the NWD loss provides an essential geometric prior; by modeling bounding boxes as 2D Gaussian distributions, it neutralizes the gradient vanishing problem inherent in small-target regression, providing a 1.83-point improvement in strict localization (AP_{75}) and a 0.45-point increase in AP_S. When combined (F4, mAP 60.35%), the overall mAP surpasses both individual configurations, though CFAR integration leads to marginal reductions in AP_{75} (69.90→69.12) and AP_S (56.62→56.58) relative to NWD alone. This reflects the CFAR module's primary role in suppressing medium-scale uniform clutter (AP_m: 72.01→72.63), which slightly shifts model attention away from extreme small-target localization. Given that the AP_{75} and AP_S differences (0.78 and 0.04 points) are within acceptable ranges given the constrained 10,000-iteration training budget, these results confirm that CFAR and NWD address complementary aspects of the detection problem—clutter suppression and localization sensitivity—yielding a consistent net gain in overall mAP.

Visualizing the Decoupled Refinement Trajectory: To intuitively corroborate the quantitative gains, Figure 12 explicitly tracks the bounding box evolution across a highly challenging harbor scene. The unoptimized Baseline model struggles with severe feature degradation, generating dense clusters of False Negatives (Red boxes) among tightly moored ships. The initial integration of Pixel Difference Convolution (R1) amplifies high-frequency structural contours, visibly rescuing a portion of these missed small targets. Subsequently, the activation of the CFAR statistical prior (R2) acts as a rigorous spatial gate, decisively clearing the widespread False Positives (Yellow boxes) induced by uniform sea clutter. The application of the NWD loss (R3) specifically targets the remaining low-observable small targets, transforming residual missed regions into correct detections. Crucially, the transition to the four-step DDIM generative sampling (R4 to R5) results in a clear geometric refinement: the bounding boxes visibly contract to tightly encapsulate the ship hulls, visually explaining the exceptional 3.75% quantitative surge in AP_{75}. Finally, the terminal optimization via data augmentation and heightened background penalties (R6 to R7) further minimizes false alarms along the wharf interfaces, yielding a detection map that closely aligns with the ground truth.

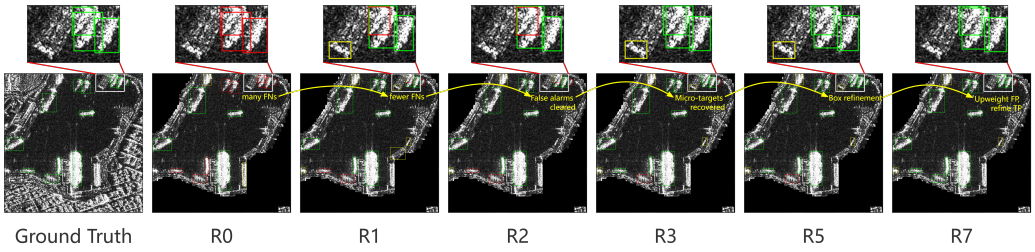


Figure 12. Qualitative visualization of the progressive ablation study for object detection. The evolution of bounding boxes from the Baseline to the final optimized model (R6 + R7) is displayed alongside a magnified local patch. To explicitly track performance gains, Green, Yellow, and Red boxes denote True Positives (TPs), False Positives (FPs), and False Negatives (FNs), respectively. The yellow arrows and annotations highlight the specific mechanisms driving target recovery and coordinate refinement.

Table 5 presents the progressive ablation tests, and Figure 12 illustrates the corresponding visual improvements. These results collectively validate the efficacy of the proposed SAR-specific modules—multi-scale PDC, CFAR soft attention, NWD loss, and DDIM

iterative sampling—which operate synergistically to maximize detection performance. Notably, empirical evaluations revealed that incorporating complex, general-purpose modules—such as multi-scale edge fusion and Bayesian geometric priors—led to performance degradation, indicating that naive application of standard deep learning components introduces redundant noise and over-processes the SAR-optimized feature representations. These findings underscore the necessity of physics-informed, domain-specific design, a principle that the subsequent Section 4 elaborates upon in terms of electromagnetic scattering mechanisms and architectural implications.

4. Discussion

The quantitative results and ablation trajectories reveal that the efficacy of near-shore maritime surveillance is strongly influenced by the precision of land-interference suppression and the robustness of bounding box regression against complex sea clutter. We synthesize our findings through the following three critical technical dimensions:

Physical Decoupling Strategy. Conventional end-to-end models frequently struggle with the competing objectives of sea–land separation and fine-grained ship localization within a single gradient flow. In the presence of strong coastal scatterers whose backscattering intensity mimics maritime ships, standard self-attention mechanisms tend to suffer from feature aliasing. Our SGED framework demonstrates the advantages of explicit task decomposition for SAR near-shore ship detection. By prioritizing the preservation of the target search space via EAU-Net (achieving an 85.30% NSCA), we establish a "clean-sea" prior that physically insulates the downstream detector from terrestrial clutter, ensuring the detection head focuses on intra-class distinction.

Synergistic Structural–Statistical Prior Verification. Unlike generic deep learning modules whose naive stacking often introduces redundant noise, SAR-DDet employs a dual-prior system in which PDC and CFAR target distinct electromagnetic scattering characteristics. Our evaluations confirm that these modules function in a complementary manner: the CFAR module acts as a statistical gate to suppress uniform sea clutter, while the PDC module reinforces structural integrity by amplifying edge gradients against speckle noise. Together, they constitute a joint verification mechanism in which a candidate proposal must satisfy both statistical anomaly and structural edge criteria to be confirmed as a valid target, effectively suppressing false alarms from azimuth ambiguities and multiple-scattering ghosts.

Zero-Cost Generative Refinement for the Regression Barrier. The bottleneck in achieving strict localization (AP_{75}) for small-targets lies in the insensitivity of traditional IoU-based loss functions to sub-pixel coordinate shifts. By modeling bounding boxes as 2D Gaussian distributions via the NWD loss, our network maintains consistent gradient signals even for ships with minimal spatial occupancy. Furthermore, the four-step DDIM iterative sampling serves as a "zero-training-cost" generative refinement mechanism, resulting in a 3.75% increase in AP_{75}. This confirms that the iterative denoising formulation inherent in diffusion models is more capable of overcoming coordinate quantization errors than single-step regression architectures, particularly in ship-dense, near-shore environments. It should be noted that all evaluations are conducted on a near-shore subset of the HRSID dataset, and the framework's generalizability across different SAR sensors, polarization modes, and coastline types remains to be validated in future cross-domain studies.

Multi-Sensor Perspectives. From a practical surveillance standpoint, SAR constitutes one node in a broader multi-sensor observation chain. High-Frequency Surface Wave Radars (HFSWRs) provide a complementary near-shore modality with over-the-horizon coverage; recent advances in high-resolution HFSWR target detection—covering deep-feature-fusion-based methods [43], high-resolution Range-Doppler mapping with

MUSIC-type azimuth estimation [44], and bi-frequency configurations for small-vessel tracking [45]—demonstrate strong complementarity with SAR-based detection. Integrating SGED outputs with HFSWR tracks and AIS records represents a promising direction for robust coastal surveillance.

Practical Limitations. Regarding target separability, the self-attention head can resolve adjacent ships with inter-ship gaps of approximately 5–8 pixels at the P2 FPN scale; below this threshold, overlapping scattering signatures cause bounding-box merging during NMS—a limitation that oriented or instance-segmentation representations could address. Regarding computational complexity, the combined pipeline achieves roughly 4–5 images per second on an RTX 3090 GPU, which is sufficient for batch satellite processing but not real-time video; model compression or reduced proposal counts could improve throughput for latency-sensitive applications. Regarding weather robustness, SAR is inherently immune to optical obscuration, but high sea states elevate Bragg-resonance backscatter and can narrow the CFAR margin, potentially increasing false alarm rates under storm conditions.

5. Conclusions

This paper addressed the challenge of SAR near-shore ship detection by proposing the SGED framework, a two-stage approach that explicitly decomposes the detection task into two specialized sub-tasks:

(1) High-Precision Shore–Sea Separation: We developed EAU-Net, which integrates SAR noise suppression, lightweight Transformers, and an Edge-Aware Composite Loss. It achieves a DSC of 0.9047 and an ASSD of 10.379 pixels, eliminating land background interference while maintaining target integrity (85.30% NSCA).

(2) Enhanced Target Reconstruction: We designed SAR-DDet, which utilizes a structural–statistical dual-verification mechanism and zero-training-cost iterative DDIM inference. The proposed framework achieves an mAP of 62.35%, with a strict localization accuracy (AP_{75}) of 74.67% and small-target detection (AP_S) of 60.14%. Extensive evaluations confirm that our methodology mitigates the challenge of shore–sea feature aliasing. This research substantiates that explicit task decoupling, when synergized with physics-informed SAR feature enhancements and the generative refinement of diffusion models, constitutes a promising and effective approach for addressing near-shore maritime surveillance challenges under the evaluated conditions. Future work will proceed along several directions: (1) extending the framework to cross-sensor evaluations covering diverse SAR platforms, polarization modes, and coastline morphologies; (2) improving dense-target separability through oriented bounding box or instance-segmentation representations; (3) reducing inference latency via model compression for near-real-time deployment; (4) evaluating performance across labeled sea-state categories to quantify CFAR robustness under adverse conditions; and (5) fusing SGED outputs with complementary near-shore modalities such as HFSWR tracks and AIS records toward a multi-sensor maritime surveillance system.

Author Contributions: Methodology, Y.B.; software, Y.B. and H.L.; validation, Y.B. and L.L.; investigation, Y.B. and S.Z.; data curation, Y.B.; writing—review and editing, Y.B. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets analyzed in this study are openly accessible. The custom datasets generated during the current study are available from the author upon reasonable request.

Acknowledgments: The authors wish to thank the editors and the reviewers for their valuable suggestions. During the preparation of this manuscript, the authors used Gemini 3.1 Pro for the purposes of English language polishing and LaTeX formatting assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic Aperture Radar
SGED	Segmentation Guidance and Enhanced Diffusion
SOTA	State-of-the-Art
mAP	Mean Average Precision
PDC	Pixel Difference Convolution
CFAR	Constant False Alarm Rate
EAU-Net	Enhanced Attention U-Net
FPN	Feature Pyramid Network
NSCA	Near-Shore Coverage Accuracy

References

1. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In *Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*; IEEE: New York, NY, USA, 2017; pp. 1–6.
2. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
3. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 936–944. [CrossRef]
4. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 6568–6577. [CrossRef]
5. Zhou, S.; Zhang, M.; Wu, L.; Yu, D.; Li, J.; Fan, F.; Zhang, L.; Liu, Y. Lightweight SAR Ship Detection Network Based on Transformer and Feature Enhancement. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4845–4858. [CrossRef]
6. Zhang, P.; Jiang, Y.; Zhu, X. LSAR-Det: A Lightweight YOLOv11-Based Model for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2025**,*22*,4012505. [CrossRef]
7. Deng, Y.; Guan, D.; Chen, Y.; Yuan, W.; Ji, J.; Wei, M. SAR-shipnet: SAR-ship detection neural network via bidirectional coordinate attention and multi-resolution feature fusion. In *Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2022; pp. 3973–3977.
8. Chen, S.; Zhao, J.; Zhou, Y.; Wang, H.; Yao, R.; Zhang, L.; Xue, Y. Info-FPN: An Informative Feature Pyramid Network for object detection in remote sensing images. *Expert Syst. Appl.* **2023**, *214*, 119132. [CrossRef]
9. Ma, X.; Cheng, J.; Li, A.; Zhang, Y.; Lin, Z. AMANet: Advancing SAR Ship Detection With Adaptive Multi-Hierarchical Attention Network. *IEEE Access* **2024**, *12*, 105952–105967. [CrossRef]
10. Liu, H.; Li, X.; Wang, L.; Zhang, Y.; Wang, Z.; Lu, Q. MS-YOLOv11: A Wavelet-Enhanced Multi-Scale Network for Small Object Detection in Remote Sensing Images. *Sensors* **2025**, *25*, 6008. [CrossRef] [PubMed]
11. Shi, H.; He, C.; Li, J.; Chen, L.; Wang, Y. An improved anchor-free SAR ship detection algorithm based on brain-inspired attention mechanism. *Front. Neurosci.* **2022**, *16*, 1074706. [CrossRef] [PubMed]
12. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. A novel anchor-free detector using global context-guide feature balance pyramid and united attention for SAR ship detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4003005. [CrossRef]
13. Zhang, Z.; Yang, Z.; Zhang, X.; Ding, F. Lightweight SAR Ship Detection via Dual-Domain Adaptive RepViT and Context Enhancement. In *Proceedings of the 2025 3rd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV)*; IEEE: New York, NY, USA, 2025; pp. 97–104. [CrossRef]
14. Dong, S.; Zhang, B. AFN-YOLO: Adaptive Frequency Network for SAR Ship Detection in Complex Inshore Background. In *Proceedings of the 2024 5th International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*; IEEE: New York, NY, USA, 2024; pp. 768–771. [CrossRef]
15. Chen, T.; Pan, J.; Xu, J.; Hu, J.; Cao, L.; Li, J. MDCA-Net: A multi-directional alignment and dynamic context aggregation network for optical and SAR image fusion. *Geo-Spat. Inf. Sci.* **2025**, *29*, 1281–1299. [CrossRef]

16. Li, C.; Yue, C.; Li, H.; Wang, Z. Context-aware SAR image ship detection and recognition network. *Front. Neurorobot.* **2024**, *18*, 1293992. [CrossRef] [PubMed]
17. Jiang, R.; Shi, H.; Ni, J.; Li, J.; Feng, Y.; Chen, X.; Li, Y. LSDFormer: Lightweight SAR Ship Detection Enhanced With Efficient Multiattention and Structural Reparameterization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 1359–1377.
18. Chen, S.; Sun, P.; Song, Y.; Luo, P. DiffusionDet: Diffusion Model for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 19830–19843.
19. Wang, L.; Jia, J.; Dai, H. OrientedDiffDet: Diffusion model for oriented object detection in aerial images. *Appl. Sci.* **2024**, *14*, 2000. [CrossRef]
20. Zhao, J.; Ding, Z.; Zhou, Y.; Zhu, H.; Du, W.L.; Yao, R. ReDiffDet: Rotation-equivariant Diffusion Model for Oriented Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 24429–24439.
21. Xie, L.; Zhang, X.; Xiong, X.; Liu, T.; Jiang, W. DiffRotDet: Diffusion-Based Rotated Object Detection Network. In *Proceedings of the 2025 10th International Conference on Signal and Image Processing (ICSIP)*; IEEE: New York, NY, USA, 2025; pp. 416–421.
22. Tang, D.; Cao, X.; Wu, X.; Li, J.; Yao, J.; Bai, X.; Jiang, D.; Li, Y.; Meng, D. AeroGen: Enhancing remote sensing object detection with diffusion-driven data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 3614–3624.
23. Chen, K.; Xie, E.; Chen, Z.; Wang, Y.; Hong, L.; Li, Z.; Yeung, D.Y. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv* **2023**, arXiv:2306.04607.
24. Li, K.; Luo, S.; Tseng, K.K. RegionDiffusion: Generative data augmentation for object detection with diffusion models. *Neurocomputing* **2025**, *654*, 131115. [CrossRef]
25. Chen, S.W. Polarimetric coherence pattern: A visualization and characterization tool for PolSAR data investigation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 286–297. [CrossRef]
26. Cui, X.C.; Tao, C.S.; Su, Y.; Chen, S.W. PolSAR ship detection based on polarimetric correlation pattern. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 471–475. [CrossRef]
27. Li, M.D.; Xiao, S.P.; Chen, S.W. Three-dimension polarimetric correlation pattern interpretation tool and its application. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5238716. [CrossRef]
28. Zhang, X.; Gao, G.; Chen, S.W. Polarimetric autocorrelation matrix: A new tool for joint characterizing of target polarization and Doppler scattering mechanism. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5213522. [CrossRef]
29. Kwon, S.; Lim, G.; Han, Y. SPAR-Det: Segmentation-guided and Prior-Aided Routing for Small Object Detection. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2026; pp. 2146–2155. [CrossRef]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
31. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September2018*, Springer: Cham, Switzerland, 2018; pp. 833–851.
32. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. In Proceedings of the Medical Imaging with Deep Learning (MIDL), Amsterdam, The Netherlands, 4–6 July 2018. Available online: https://openreview.net/forum?id=Skft7cijM (accessed on 12 July 2026).
33. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef] [PubMed]
34. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. *arXiv* **2022**, arXiv:2209.08575.
35. Ruan, J.; Li, J.; Xiang, S. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *ACM Trans. Multimed. Comput. Commun. Appl.* 2025, *accepted* . [CrossRef]
36. Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. SAM-Med2D. *arXiv* **2023**, arXiv:2308.16184.
37. Jocher, G.; Team, U. Ultralytics YOLOv11. GitHub Repository . 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 12 July 2026).
38. Jocher, G.; Team, U. Ultralytics YOLOv8. GitHub Repository. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 12 July 2026).

39. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=3mRwyG5one (accessed on 12 July 2026).
40. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [CrossRef]
41. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
42. Zhou, J.; Xiao, C.; Peng, B.; Liu, Z.; Liu, L.; Liu, Y.; Li, X. DiffDet4SAR: Diffusion-Based Aircraft Target Detection Network for SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4007905. [CrossRef]
43. Wu, M.; Zhang, L.; Niu, J.; Wu, Q.J. Target detection in clutter/interference regions based on deep feature fusion for HFSWR. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5581–5595. [CrossRef]
44. Golubović, D.; Erić, M.; Vukmirović, N. High-resolution Doppler and azimuth estimation and target detection in HFSWR: Experimental study. *Sensors* **2022**, *22*, 3558. [CrossRef] [PubMed]
45. Golubović, D.; Erić, M.; Vukmirović, N.; Orlić, V. High-Resolution Sea Surface Target Detection Using Bi-Frequency High-Frequency Surface Wave Radar. *Remote Sens.* **2024**, *16*, 3476. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.