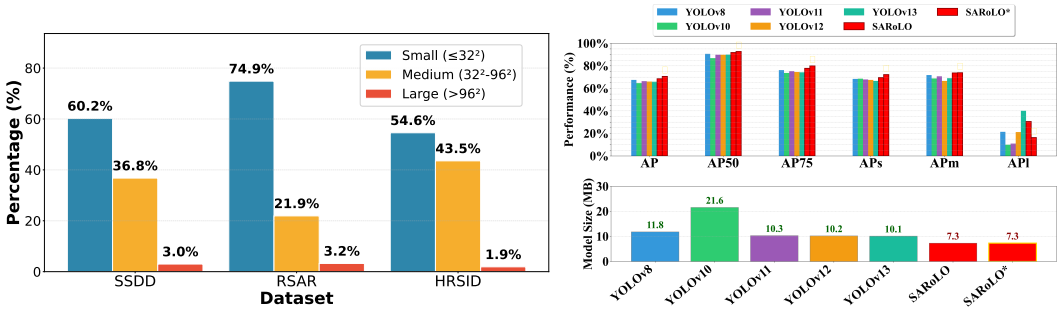
SARoLO: Structure and Scale-Aware SAR Object Detection

Anonymous authors
Paper under double-blind review

Abstract

Synthetic Aperture Radar (SAR) object detection faces persistent challenges caused by small-target dominance, strong speckle noise, and complex background clutter that obscure fine-grained structural cues. Although conventional YOLO detectors excel in general-purpose visual tasks with balanced accuracy and efficiency, they remain limited in representing scale- and structure-sensitive features due to their reliance on aggressive downsampling, information-lossy bottleneck compression, and redundant multi-scale detection heads. To address these limitations, we propose a novel structure and scale-aware SAR object detection method, called SARoLO, including a Heterogeneous Triple-Conv Downsampling (HetTriDown) module, a Heterogeneous Convolution Modulation Block (HCM Block) with dynamic non-linear modulation (DyN), a cross-stage feature aggregation module (HCMC3K), a Cross-Scale Aggregation Network (CSAN), and a dual-scale detection head, to optimize bounding box performance for SAR imagery characterized by small/medium targets and significant structural variations. Extensive experiments on the HRSID and SSDD benchmarks demonstrate SARoLO's state-of-the-art performance. On HRSID, SARoLO achieves 68.46% AP, a 2.52% gain over the latest YOLOv13 baseline. On SSDD, SARoLO attains 63.21% AP, a 0.9% improvement over YOLOv13. These results validate its effectiveness in structure- and scale-aware feature learning and efficient SAR object detection.



(a) Statistical analysis of data distribution.

(b) Experiment comparison results.

Figure 1: Comparison of detection accuracy across different scales, overall AP, and model size among different methods (left). Distribution statistics of SAR image targets across different scale intervals on various datasets (right). Our proposed SARoLO and SARoLO* achieve state-of-the-art detection performance.

1 Introduction

Synthetic Aperture Radar (SAR) imagery, due to its ability to image in all-weather and at night, plays an significant role in ship monitoring Zheng et al. (2024), maritime traffic management and ocean environmental monitoring Stoica and Li (2006); Zhou et al. (2021). However, unlike optical images, SAR images are formed through a coherent imaging mechanism, which inherently introduces speckle noise, complex scattering effects, and severe background clutter. These characteristics result in a low

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

signal-to-noise ratio and significantly increase the difficulty of reliable target perception. Furthermore, from the perspective of data distribution, SAR ship detection is dominated by small and medium-sized targets. As illustrated in Figure 1, small targets account for approximately 50%-60% in HRSID, SSDD and RSAR datasets, while medium-sized targets account for 40%-50%, and large targets are rarely observed. This highly imbalanced scale distribution makes detection performance particularly sensitive to small objects. Due to their limited spatial extent and weak scattering responses, small and medium-sized targets are highly susceptible to noise interference and background clutter, resulting in unstable feature representations and frequent false positives or missed detections. This issue is further exacerbated in complex environments such as nearshore regions, where heterogeneous backgrounds significantly reduce target discriminability.

Data-Scale Constraints Although YOLO Ultralytics (2024); Tian et al. (2025), ResNet He et al. (2016), and CSPDarknet Bochkovskiy et al. (2020) have achieved success on optical benchmarks such as COCO Lin et al. (2023)[1] and DOTA Xia et al. (2018)[2], the object scale distribution in these benchmarks is relatively balanced. For example, in the COCO dataset, small, medium, and large objects account for 41%, 34%, and 24%, respectively. In contrast, SAR imagery is predominantly composed of small and medium-sized targets. These targets typically occupy only a few to tens of pixels and rely on subtle structural cues for discrimination. This distribution characteristic makes accurate detection highly dependent on preserving fine-grained spatial resolution. However, standard downsampling operations in YOLO, including stride-2 convolutions and pooling (e.g., 3×3 kernels), inevitably reduce spatial resolution and discard discriminative structural information. As a result, small and medium-sized targets (e.g., $< 32 \times 32$ and $< 96 \times 96$) are rapidly compressed into low-resolution feature maps (e.g., $< 8 \times 8$ and $< 24 \times 24$) at early stages, where only coarse responses remain and critical structural details are severely degraded or lost. This leads to suppressed target responses and increased missed detections.

Feature Representation Constraints Modern bottleneck architectures (e.g., C3/CSP Bottleneck/C2f/C3Ghost/C2fCIB blocks in YOLO5 Jocher et al. (2020) / YOLO7 Wang et al. (2023) / YOLO8 Sohan et al. (2024) / YOLO8 Venkatesan et al. (2025) / YOLO10 Wang et al. (2024a)) enhance feature reuse and computational efficiency through compression and cross-stage stacking, while discarding task-irrelevant information and maintaining detection accuracy. However, SAR imagery exhibits unique target distribution characteristics and contains extensive speckle noise, land–sea clutter, and occlusion. During compression or cross-stage stacking, these interference factors may cause critical information to be compressed or discarded together, leading to highly uncertain detection outcomes. In addition, cross-scale feature fusion mechanisms, such as FPN, PAN, and BiFPN Lei et al. (2025), aim to enhance multi-level representations by aggregating features from different layers. However, these methods primarily rely on feature aggregation rather than feature reconstruction. Once fine-grained spatial details are lost during early or intermediate downsampling stages, they cannot be recovered through subsequent fusion. This limitation severely restricts the model's ability to maintain discriminative representations in cluttered SAR environments. Finally, most existing frameworks adopt three-scale detection heads by default. For SAR ship detection tasks dominated by small and medium-sized targets, the inclusion of large-scale detection heads introduces redundancy, leading to inefficient feature utilization and limiting overall detection performance. These limitations indicate that existing detection frameworks struggle to stabilize noisy features, enhance weak target responses, and preserve structure- and scale-aware representations in SAR imagery, ultimately constraining detection performance in nearshore dense small-object scenarios.

To address the above challenges, this paper proposes an Structure and Scale-Aware SAR Object Detection framework, called SARoLO. The contributions can be summarized as follows:

- We design and implement SARoLO, an Structure and Scale-Aware SAR Object Detection framework to address key challenges in SAR feature extraction, including feature instability, loss of fine-grained structural information, and insufficient cross-scale feature interaction. The proposed framework enhances local feature representation, preserves fine-grained

[1]COCO (Common Objects in Context): a general-purpose vision dataset built by Microsoft, used for object detection, instance segmentation, keypoint detection, etc. Contains 330K images with 80 categories. Website: `https://cocodataset.org`.

[2]DOTA (Dataset for Object deTection in Aerial images): a large-scale aerial image dataset released by Wuhan University and others for object detection. Contains 2,806 images with 15 categories, annotated with oriented bounding boxes. Website: `https://captain-whu.github.io/DOTA`.

structures, and strengthens cross-scale feature interaction, thereby improving detection accuracy for small and medium hard targets in complex scenarios.

- We introduce several key components in SARoLO, including a Heterogeneous Triple-Conv Downsampling (HetTriDown) module maintaining fine-grained spatial structures, a Heterogeneous Convolution Modulation Block (HCM Block) combining dynamic nonlinear modulation (DyN) with symmetric and asymmetric convolutions to preserve local feature fidelity and geometric structures, a Cross-Stage Feature Aggregation (HCMC3K) module that integrates multi-level pointwise convolution with HCM-enhanced modeling for efficient cross-stage feature aggregation, a Cross-Scale Aggregation Network (CSAN) promoting cross-scale semantic alignment and feature interaction, a dual-scale detection head that leverages high-resolution feature maps to enhance fine-grained representation of small and medium targets.
- We conduct extensive experiments on several benchmark datasets (e.g., HRSID, SSDD). The results demonstrate that our method achieves state-of-the-art performance. On SSDD, SARoLO achieves 63.21% AP, a 0.9% improvement over YOLO13. On HRSID, SARoLO achieves 68.46% AP, outperforming YOLO10 by 3.86% and surpassing all other YOLO variants across almost all metrics.

2 Related Work

2.1 SAR Object Detection

Research on Synthetic Aperture Radar (SAR) object detection has evolved rapidly in recent years, driven by advances in deep learning architectures and the increasing availability of high-resolution SAR datasets. Early convolutional neural network (CNN) Ketkar and Moolayil (2021) models demonstrated strong potential in SAR target recognition and detection tasks by leveraging their ability to automatically extract hierarchical features Huang et al. (2022). Representative detection frameworks such as Faster R-CNN Ren et al. (2015), RetinaNet Lin et al. (2017) and YOLO variants Ultralytics (2024); Tian et al. (2025); Lei et al. (2025) have been successfully applied to maritime SAR imagery, providing efficient end-to-end pipelines and achieving competitive accuracy in various benchmarks.

However, due to the coherent imaging mechanism of SAR, which introduces speckle noise and anisotropic scattering, CNN-based He et al. (2016); Bochkovskiy et al. (2020) models relying on isotropic convolutional kernels struggle to capture directional structures such as elongated ship hulls or slender offshore facilities. Several studies have attempted to alleviate these issues by introducing attention mechanisms Shi et al. (2022), deformable convolutions, and multiscale feature fusion Chen et al. (2024), which improve local adaptation and scale invariance. Nevertheless, these methods still lack explicit modeling of directional dependencies and are often limited in handling elongated maritime targets characterized by high aspect ratios and strong orientation sensitivity.

2.2 Direction-Aware and Anisotropic Feature Representation

To address the challenge of capturing long-range and anisotropic contextual information, recent works have explored feature representation methods that go beyond fixed isotropic kernels. Large-kernel Ding et al. (2022) and multi-branch convolution Wang et al. (2024b; 2025) designs have been proposed to expand receptive fields and improve contextual perception. For instance, the Large Selective Kernel Network (LSKNet) Li et al. (2024) and the Poly Kernel Inception Network (PKINet) Cai et al. (2024) dynamically aggregate multi-scale contextual responses to improve feature diversity and spatial coverage. However, these methods treat all spatial orientations equally, without explicitly modeling the directional priors inherent to elongated maritime targets. Moreover, their reliance on multiple large kernels leads to increased computational cost and parameter redundancy. Another line of research focuses on introduction direction-aware or anisotropic convolutional designs. Oriented Response Networks (ORN) Zhou et al. (2017) incorporate rotation-equivariant operations to enhance orientation sensitivity, while deformable convolutional networks Ruan et al. (2021) learn spatial offsets to adapt receptive fields to object geometry. However, these models often suffer from increased computational burden or require complex geometric transformations.

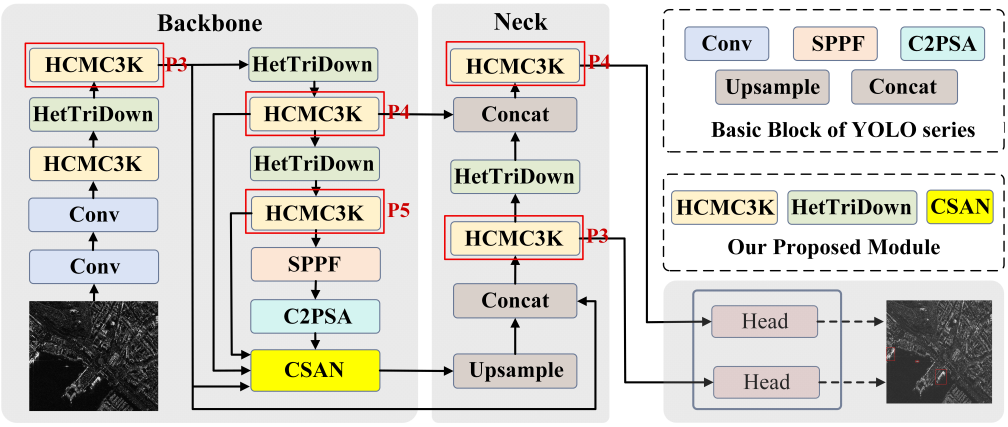


Figure 2: Overview of the proposed SARoLO architecture, consisting of HetTriDown, HCMC3K, and CSAN. HetTriDown employs heterogeneous triple-convolution for multi-scale spatial downsampling. HCMC3K integrates dynamic nonlinear modulation (DyN) with symmetric and asymmetric convolutions to form heterogeneous convolution modulation (HCM) blocks, enabling efficient cross-stage feature aggregation while preserving local fidelity and geometric structures. CSAN explicitly aligns and fuses low-level spatial details with high-level semantics via scale transformation and residual compensation, recovering degraded fine-grained structural cues.

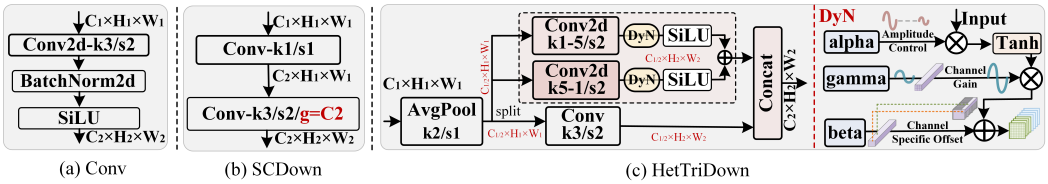


Figure 3: Architecture comparison with popular downsampling modules including (a) Conv Ultralytics (2023), (b) SCDown Wang et al. (2024a), and (c) our HetTriDown.

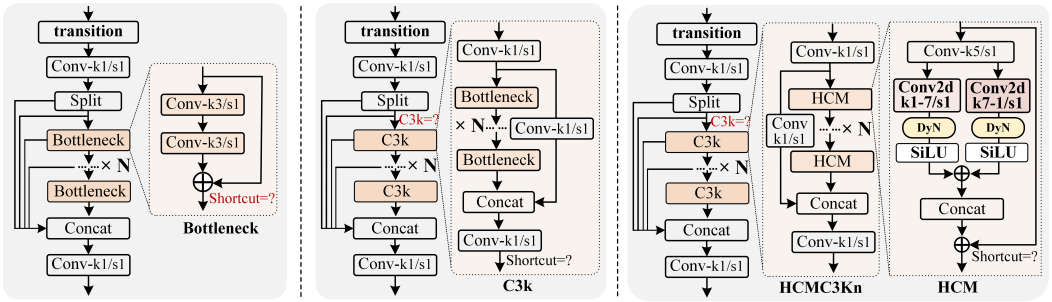
3 METHOD

The primary focus of this section is to preserve fine-grained structures, strengthen cross-scale feature interaction, and enhance local feature representation, thereby improving detection accuracy for small and medium targets in complex scenarios. In this section, we present our SARoLO architecture from the perspective of four highly efficient detection techniques. An overview of the proposed SARoLO architecture is illustrated in Figure 2. We adopt YOLO11 as our baseline architecture and employ a dual-scale detection head. By integrating the HetTriDown, HCM, HCMC3K, and CSAN modules into the backbone, neck, and head structures respectively, we achieve significantly enhanced feature representation.

3.1 HETEROGENEOUS TRIPLE-CONV DOWNSAMPLING

To address the issue that key structural details of small and weak targets in SAR imagery are easily lost or overwhelmed by noise during standard convolution downsampling or local feature extraction (as illustrated in Figure 3(a) and (b)), we propose a heterogeneous triple-convolution downsampling architecture, termed HetTriDown, as shown in Figure 3(c).

Design of Dynamic Nonlinear Modulation. To improve feature representation under noisy SAR imaging conditions, we introduce a Dynamic Nonlinear (DyN) modulation mechanism, as illustrated in Figure 3(c). Unlike conventional fixed activation functions, DyN introduces learnable modulation parameters to adaptively reshape feature responses, enabling selective enhancement of informative target-related activations while suppressing irrelevant background responses. Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, DyN is formulated as:



(a) C2f (b) C3k2 (c) HCMC3K

Figure 4: Architecture comparison with popular feature aggregation modules including (a) C2f Ultralytics (2023), (b) C3k2 Ultralytics (2024), and (c) our HCMC3K.

$$\mathbf{DyN}_{\gamma_c,\alpha,\beta_c}(\mathbf{X}) = \gamma_c \cdot \tanh(\alpha \mathbf{X}) + \beta_c, \tag{1}$$

where α, γ_c, and β_c are learnable parameters optimized during training. Here, α controls the nonlinear response range, γ_c adjusts channel-wise feature amplification, and β_c calibrates channel-specific activation shifts. Although these parameters remain fixed during inference, the nonlinear transformation produces input-dependent feature responses, allowing the network to adaptively emphasize discriminative target patterns and attenuate noisy background interference.

Design of HetTriDown. To enhance feature representation in SAR images and capture elongated structural cues, we design a DyN mechanism and a heterogeneous receptive field collaborative modeling strategy. In HetTriDown, we first apply average pooling (avg_pool2d k=2, s=1, p=0) to the input features to suppress noise, while preserving spatial structure and radiometric distribution with minimal distortion, thereby producing smoothed features. Then, a dual-path heterogeneous receptive field structure is constructed for spatial structure preservation and structure modeling: the spatial structure modeling path mimics spatial structure patterns with pronounced anisotropic distributions. It employs a pair of large-kernel asymmetric 1×5 and 5×1 convolutions (Conv2d k=5-1 / 1-5, s=2, p=2-0 / 0-2, g=1 in Figure 3(c)) to extract dominant structural responses along the horizontal and vertical directions, respectively, effectively characterizing the typical anisotropic scattering structures of SAR targets. Each response is followed by DyN and SiLU activation. The spatial structure preservation path focuses on retaining local fine-grained information and adopts a 3×3 small-receptive-field depthwise convolution (Conv2d k=3, s=2, p=1, g=1 in Figure 3(a)) to preserve high-frequency details such as edges and textures. Both paths perform downsampling with a stride of 2. Finally, the parallel fusion strategy first fuses the horizontal and vertical structural responses via element-wise summation to achieve cross-scale semantic integration. The fused result is then concatenated with the spatial structure preservation path along the channel dimension. In this design, DyN enhances feature discriminability through learnable nonlinear transformations, while average pooling suppresses noise with minimal structural distortion. The parallel fusion of horizontal and vertical asymmetric convolutions with the standard convolution jointly models anisotropic structures and isotropic details, achieving a fine-grained spatial representation of SAR targets.

3.2 Feature Refinement & Aggregation

Designed for efficient feature reuse and lightweight computation, C2f (Figure 4(a)) and C3k2 (Figure 4(b)) employ symmetric $1 \times 1/3 \times 3$ convolutions with fixed activation functions in the feature pyramid. However, such isotropic operations limit the modeling of anisotropic structures and complex scattering patterns in SAR imagery, leading to insufficient structural representation. To address this issue, we propose HCMC3K (Figure 4(c)), a new feature aggregation method that integrates symmetric/asymmetric convolutions with dynamic nonlinear modulation to enhance structural representation.

Design of HCM block. Although the C2f and C3k2 modules exhibit strong feature aggregation capabilities, their bottleneck layers rely on symmetric 3×3 convolutions and fixed activation functions (e.g., SiLU). While efficient, the symmetric kernel provides no built-in directional preference, making it structurally isotropic in design. This severely limits the explicit capture of anisotropic scattering

patterns that dominate SAR imagery. Ship hulls, masts, and wake returns exhibit strongly orientation-dependent signatures; a symmetric kernel processes them identically in all directions, leaving critical structural cues unmodeled and reducing discriminability against background clutter. Moreover, fixed activations lack the ability to adapt to the wide dynamic range of SAR backscatter, often failing to amplify weak target responses in noisy regions. To address these limitations, we propose a Heterogeneous Convolution Modulation Cascade module (Figure 4(c)), which integrates dynamic nonlinear modulation (DyN) with complementary isotropic and anisotropic convolutional operators for structure-preserving representation learning. Specifically, we first employ a 5×5 convolution to capture isotropic local context with an enlarged receptive field. To further model anisotropic structural patterns, we introduce a pair of orthogonal large-kernel asymmetric convolutions (1×7 and 7×1), which decompose spatial interactions along principal axes. Their responses are aggregated via element-wise summation to encode cross-directional structural dependencies, and subsequently fused with isotropic features through channel-wise concatenation, achieving joint modeling of local details and directional structures. All convolutional outputs are adaptively recalibrated via DyN followed by SiLU, enabling input-dependent feature modulation and enhancing representation flexibility. Finally, a residual connection is incorporated to facilitate stable gradient propagation and preserve feature integrity in deep networks.

Development of HCMC3K. Building upon the HCM block, we further develop the HCMC3K module to balance representation capacity and computational efficiency. HCMC3K retains the efficient split–transform–merge paradigm of C2f/C3k2, while replacing all standard bottleneck units with the proposed heterogeneous convolutional modulation cascade to enhance structure-aware representation. Similar to C3k2, HCMC3K consists of a pointwise convolution for channel compression, N stacked HCMC3K$_n$ blocks, a concatenation operation, and a final 1×1 convolution layer (see Figure 4(c)). The input feature is first compressed by a 1×1 convolution with a reduction ratio $e = 0.25/0.5$, and then split into two branches: a shortcut branch that preserves the original gradient flow, and a transformation branch that undergoes structured feature refinement. This split design enables parallel feature propagation, improving gradient transmission while reducing redundant computation. Finally, the two branches are concatenated along the channel dimension and fused via a 1×1 convolution for feature aggregation and channel re-calibration. Unlike C2f/C3k2, which rely on cascaded isotropic convolutions with fixed nonlinearities, the proposed design explicitly introduces dynamic nonlinear modulation (DyN) together with symmetric and asymmetric convolutions to jointly preserve local feature fidelity and capture geometry-aware structural patterns.

The core of HCMC3K lies in the HCMC3K$_n$ block, where conventional bottleneck transformations are replaced by the heterogeneous convolutional modulation cascade. HCMC3K$_n$ first employs dual parallel 1×1 convolutions to construct shortcut and transformation embeddings, enabling simultaneous residual preservation and channel re-calibration. The transformation branch is then processed by n stacked HCM units, each integrating symmetric ($k \times k$) and asymmetric ($1 \times k$ / $k \times 1$) convolutions with dynamic nonlinear modulation (DyN). Such a design enables joint modeling of isotropic local patterns and anisotropic structural cues, progressively refining feature representations, enhancing directional sensitivity, and suppressing speckle-induced redundancy. Finally, the transformed features and shortcut features are concatenated and fused via a 1×1 convolution, promoting multi-level feature reuse and stabilizing gradient propagation, thereby achieving an effective trade-off between representation expressiveness and computational cost.

3.3 Cross-Scale Aggregation Network

To enhance feature representation, modules such as SPPF and C2PSA improve feature expressiveness by enlarging the receptive field or strengthening *intra-scale* feature interactions within the feature pyramid, while cross-scale architectures such as FPN and PAN focus on top-down or bottom-up feature aggregation to improve multi-level representations. However, these approaches lack explicit mechanisms for full-scale interaction and compensation between low-level spatial details and high-level semantic features (e.g., P3, P4, and P5 in Figure 2). As a result, the fine-grained structural information lost during early downsampling remains largely unrecoverable, which propagates through subsequent feature transformation stages and leads to suboptimal representations in the detection head. To address this limitation, we propose a Cross-Scale Aggregation Network (CSAN), a cross-scale feature reconstruction framework. Unlike conventional aggregation-based methods, CSAN explicitly models cross-scale semantic alignment and feature interaction, enabling the recovery of degraded fine-

grained structural cues. This design effectively mitigates the irreversible information loss introduced by downsampling, thereby enhancing feature discriminability and improving detection performance.

As illustrated in Figure 2, the proposed CSAN takes multi-level features $\{P_3, P_4, P_5\}$ extracted from different backbone stages as input, where low-level features preserve fine-grained spatial details, while high-level features encode rich semantic context. Given $\{P_3, P_4, P_5\}$, CSAN first performs channel alignment via 1×1 pointwise convolutions (Conv2d $k = 1$, $s = 1$, $p = 1$, $g = 1$), followed by explicit scale transformation to project all features into a unified resolution space. Specifically, the low-level feature P_3 is downsampled using the proposed 5×5 HetTriDown operator (HetTriDown $k = 5$, $s = 2$, $p = 2$, $g = 1$), the intermediate feature P_4 is kept unchanged, and the high-level feature P_5 is upsampled to match the target resolution. This design enables direct interaction between spatially detailed representations and semantically rich features across scales. The aligned features are then concatenated along the channel dimension and aggregated via a 1×1 convolution for joint feature transformation and channel re-calibration. Finally, a residual connection is introduced to strengthen gradient flow and alleviate the vanishing gradient problem in deep networks. Specifically, the residual branch is constructed from the input of the C2PSA module inFigure 2, which is first projected via a 1×1 pointwise convolution (Conv2d $k = 1$, $s = 1$, $p = 1$, $g = 1$) and then upsampled to form a residual compensation pathway. The overall computation can be formulated as:

$$\tilde{P}_3 = \mathrm{HetTriDown}\big(\mathrm{Conv}_{1\times1}(P_3)\big), \quad \tilde{P}_4 = P_4, \quad \tilde{P}_5 = \mathrm{Upsample}\big(\mathrm{Conv}_{1\times1}(P_5)\big), \tag{2}$$

$$F_{\text{out}} = \mathrm{Conv}_{1\times1}\big(\mathrm{Concat}[\tilde{P}_3, \tilde{P}_4, \tilde{P}_5]\big) + \mathrm{Upsample}\big(\mathrm{Conv}_{1\times1}(C2PSA)\big). \tag{3}$$

4 Experiments and Analysis

4.1 Experiment Setup

Datasets. We conduct extensive experiments on two popular SAR object detection datasets. **HRSID** Wei et al. (2020) contains 5,604 high resolution SAR images covering 16,951 SAR ship objects. The SAR dataset **SSDD** Zhang et al. (2021) contains 1160 SAR images.

Implementation details. SARoLO is trained for 300 epochs with a batch size of 24 using SGD (initial learning rate of 0.01995) and a linear decay schedule with a 3-epoch warm-up. Mosaic augmentations are employed during training. All experiments are conducted on a single RTX 3060 GPU. The model is trained from scratch on the SSDD and HRSID datasets with an input resolution of 800×800. Unless otherwise specified, all ablation studies are conducted on the HRSID dataset.

Table 1: Performance comparisons of object detection methods on the HRSID dataset.

Method	AP↑	AP_{50}(%)↑	AP_{75}(%)↑	AP_S(%)↑	AP_M(%)↑	AP_L(%)↑	AP_I↑	AP_O↑	Params(M)↓	FLOPs(G)↓
YOLO8 Ultralytics (2023)	67.46	90.59	76.20	68.41	71.79	21.37	51.81	79.85	3.01	8.2
YOLO10 Wang et al. (2024a)	64.60	86.78	73.71	68.58	68.78	9.96	46.25	79.34	2.71	8.4
YOLO11 Ultralytics (2024)	66.36	89.77	75.28	67.94	70.73	11.00	49.62	79.48	2.59	6.4
YOLO12 Tian et al. (2025)	65.99	89.79	74.53	67.53	66.56	21.21	49.05	78.93	2.52	6.0
YOLO13 Lei et al. (2025)	65.94	89.87	74.21	66.64	68.95	40.17	50.13	78.28	2.46	6.4
SARoLO	**68.46**	**91.83**	**77.80**	**69.51**	**73.66**	**30.62**	**53.60**	**80.51**	**1.80**	**10.0**

4.2 Experimental Evaluation

Benchmarking on Popular Datasets. We conduct comprehensive comparative analyses across different models and report detailed performance metrics, including Average Precision at IoU=50 (AP_{50}), AP_{50} for inshore (AP^I_{50}) and offshore (AP^O_{50}) scenes, AP at IoU=75 (AP_{75}), mean AP over IoU=0.5:0.95 (AP), mean AP for inshore (AP_I) and offshore (AP_O) scenes, as well as AP for small (AP_S, area $< 32^2$ pixels), medium (AP_M, $32^2 <$ area $< 96^2$ pixels), and large (AP_L, area $> 96^2$ pixels) objects on both **SSDD** and **HRSID** datasets. On the SSDD dataset (Table 2), SARoLO achieves state-of-the-art performance with 63.21% AP, 97.46% AP_{50}, and 75.01% AP_{75}, outperforming previous best methods (YOLO13) by +0.9%, +0.75%, and +3.36%, respectively.

On the HRSID dataset (Table 1), SARoLO achieves 68.46% AP, 91.83% AP_{50}, 77.80% AP_{75}, 69.51% AP_S, 73.66% AP_M, 30.62% AP_L, 53.60% AP_I, and 80.51% AP_O, demonstrating consistent improvements over state-of-the-art methods across all metrics.

Table 2: Performance comparisons of object detection methods on the SSDD dataset.

Method	AP(%)↑	AP_{50}(%)↑	AP_{75}(%)↑	AP_S(%)↑	AP_M(%)↑	AP_L(%)↑
YOLO8 Ultralytics (2023)	61.90	96.55	70.80	56.47	70.79	64.05
YOLO10 Wang et al. (2024a)	61.90	96.49	71.55	57.05	71.04	64.49
YOLO11 Ultralytics (2024)	62.25	97.20	70.98	56.56	70.62	66.45
YOLO12 Tian et al. (2025)	61.86	96.97	71.28	56.83	69.95	67.15
YOLO13 Lei et al. (2025)	62.31	96.71	71.65	56.91	70.30	67.55
SARoLO	**63.21**	**97.46**	**75.01**	**57.49**	**72.12**	**69.18**

4.3 Ablation Studies & Result Analyses

Overall Architecture and Component Analysis. We jointly analyze the effects of detection heads and core modules in Table 3. Reducing YOLO11 from three heads to two heads improves AP from 66.36% to 67.10% while reducing parameters by 28.6%, indicating that redundant detection scales may hinder efficient feature utilization. Building upon this, integrating CSAN, HetTriDown, and HCMC3K further boosts performance to 68.46% AP with only 1.80M parameters. Meanwhile, AP_I / AP_O increase from 51.27% / 79.47% to 53.60% / 80.51%, demonstrating consistent gains in both inshore and offshore scenarios. We further analyze the contribution of each component. Removing HetTriDown or HCMC3K leads to noticeable performance drops (-0.32% / -0.34% AP), while removing CSAN results in a larger degradation to 66.76% AP, highlighting its critical role in multi-scale feature aggregation.

Analyses for HCMC3K. We validate HCMC3K's effectiveness in Table 3. It improves the AP_{50} of SARoLO by 0.97% (comparing row 4 and row 7). Table 4(a) shows that compared to C2f, HCMC3K reduces parameters by 36.2% and FLOPs by 39.8%, while achieving 0.63% higher AP_{50}. It also surpasses C3k2 by 0.97% AP_{50} with fewer parameters.

Analysis of HetTriDown. We validate HetTriDown's effectiveness in Table 3 and Table 4(b). In Table 3, comparing row 4 (with HetTriDown) and row 6 (without HetTriDown), HetTriDown improves SARoLO's AP_{50} by 1.01% (91.83% vs. 90.82%) while reducing parameters by 9.5% (1.80M vs. 1.99M). Table 4(b) shows that compared to SCDown, HetTriDown achieves 0.97% higher AP_{50} (91.83% vs. 90.86%) with comparable FLOPs (10.0G vs. 9.9G); and compared to Conv, HetTriDown improves AP_{50} by 1.01% with 9.5% fewer parameters.

Influence of multiple factors on performance. Table 5 investigates the impact of various design choices. In Table 5(a), P4 channels in neck: increasing from 128 to 512 raises AP_{50} from 91.75% to 91.83% with only 0.15M additional parameters (1.80M total). Further scaling to 1024 brings negligible gain (91.68%) but adds 0.31M parameters (2.11M). Hence, we adopt 512 channels for SARoLO. In Table 5(b), P3 channels in neck: 512 channels achieves 91.83% AP_{50} with 1.80M parameters; 1024 channels yields 91.89% AP_{50} but costs 2.42M parameters. We choose 512 for efficiency. In Table 5(c), training epochs: the best AP_{50} (91.83%) is achieved at 300 epochs, while the peak AP (69.10%) occurs at 500 epochs with a slight AP_{50} drop to 91.26%. Considering the trade-off between training time and performance, we select 300 epochs for SARoLO, as the marginal AP gain from further training (+0.64%) does not justify the additional computational cost. In Table 5(d), kernel in HetTriDown: the 3×3/1×5/5×1 kernel achieves the highest AP (68.46%), AP_{50} (91.83%), and AP_{75} (77.80%) with 1.80M parameters. Its AP is 0.38% higher than the 7×7 combination. We use this kernel for SARoLO. In Table 5(e), model input image size: increasing resolution from 320 to 1024 monotonically improves AP (50.40%→70.60%) and AP_{50} (74.30%→92.70%). To balance accuracy and computational cost, we use 800×800, which achieves 68.46% AP with only 1.80M parameters. For scenarios that prioritize maximum accuracy regardless of speed, we adopt 1024×1024 to further boost performance. In Table 5(f), DyN impact on HetTriDown (D) and HCMC3K (H): jointly using both modules boosts AP_{50} from 90.67% to 91.83%, inshore AP^I_{50} from 79.92% to 81.94%, and offshore AP^O_{50} from 97.91% to 97.96%. The combination yields synergistic gains, confirming the effectiveness of both components.

Table 3: Ablation study on detection heads and modules. ✓ denotes using the component, × denotes not.

#	Model	Head	CSAN	HetTriDown	HCMC3K	Params(M)	FLOPs(G)	AP(%)	$AP_{50}(\%)$	$AP_{75}(\%)$	$AP_I(\%)$	$AP_O(\%)$
1	YOLO11	3	×	×	×	2.59	6.4	66.36	89.77	75.28	49.62	79.48
2	YOLO11	2	×	×	×	1.85	5.8	67.10	90.47	76.55	51.27	79.47
3	SARoLO	1	✓	✓	✓	1.51	9.1	68.11	91.01	77.09	52.85	80.03
4	SARoLO	2	✓	✓	✓	1.80	10.0	**68.46**	**91.83**	**77.80**	**53.60**	**80.51**
5	SARoLO	3	✓	✓	✓	2.45	10.5	68.03	91.00	76.95	52.73	80.01
6	SARoLO	2	✓	×	✓	1.99	10.6	68.14	90.82	77.70	52.83	80.37
7	SARoLO	2	✓	✓	×	2.07	9.7	68.12	90.86	77.32	53.10	80.10
8	SARoLO	2	×	✓	✓	0.97	5.2	66.76	90.38	74.88	50.64	79.47

Table 4: The impact of feature extraction and downsampling factors on performance.

(a) Feature extraction modules

Module	AP_{50}	AP^I_{50}	AP^O_{50}	Params (M)	FLOPs (G)
C3k2	90.86	80.25	97.90	2.07	9.7
C2f	91.20	81.58	97.94	2.82	16.6
HCMC3K	**91.83**	**81.94**	**97.96**	**1.80**	**10.0**

(b) Downsampling modules

Module	AP_{50}	AP^I_{50}	AP^O_{50}	Params (M)	FLOPs (G)
SCDown	90.86	80.48	97.94	**1.63**	9.9
Conv	90.82	80.37	97.92	1.99	10.6
HetTriDown	**91.83**	**81.94**	**97.96**	1.80	10.0

Table 5: The impact of multiple factors on the performance. H indicates the use of HCMC3K, and D indicates the use of HetTriDown.

(a) P4 Channels in Neck

Channels	$AP_{50}(\%)$	Params(M)	FLOPs(G)
128	91.75	1.65	9.5
256	90.93	1.69	9.6
384	91.26	1.74	9.8
512	**91.83**	**1.80**	**10.0**
640	91.01	1.86	10.2
768	91.64	1.94	10.4
1024	91.68	2.11	11.0

(b) P3 Channels in Neck

Channels	$AP_{50}(\%)$	Params(M)	FLOPs(G)
128	90.77	1.48	6.4
256	90.85	1.57	7.5
384	90.90	1.67	8.6
512	91.83	1.80	10.0
640	90.98	1.93	11.5
768	91.69	2.08	13.2
1024	91.89	2.42	17.0

(c) Training Epochs

Epoch	AP(%)	$AP_{50}(\%)$	$AP_{75}(\%)$
100	65.27	90.17	73.28
200	67.03	90.57	75.69
300	68.46	**91.83**	77.80
400	68.66	91.73	77.91
500	**69.10**	91.26	**78.37**
600	69.06	91.81	78.74
700	68.88	91.71	77.93

(d) Kernel in HetTriDown

Kernel	AP(%)	$AP_{50}(\%)$	$AP_{75}(\%)$	$AP_I(\%)$	$AP_O(\%)$	Params(M)	FLOPs(G)
3×3 / 1×3 / 3×1	68.23	90.85	77.07	52.83	80.50	1.75	9.8
3×3 / 1×5 / 5×1	**68.46**	**91.83**	**77.80**	**53.60**	**80.51**	1.80	10.0
3×3 / 1×7 / 7×1	68.08	90.86	77.24	52.67	80.45	1.84	10.1
3×3 / 1×9 / 9×1	68.22	91.15	77.01	53.41	80.01	1.89	10.3
3×3/1×11/11×1	68.03	90.89	76.86	52.38	80.46	1.93	10.4
3×3/1×13/13×1	68.11	90.99	77.00	52.68	80.33	1.98	10.6
5×5 / 1×3 / 3×1	67.93	90.95	76.77	52.60	80.29	1.93	10.4
5×5 / 1×5 / 5×1	68.22	91.26	76.97	53.30	80.37	1.98	10.6
5×5 / 1×7 / 7×1	68.42	91.11	77.18	53.49	80.25	2.02	10.7
5×5 / 1×9 / 9×1	67.95	90.94	77.39	52.65	80.01	2.07	10.9
5×5/1×11/11×1	67.90	90.93	76.12	52.42	80.13	2.11	11.0
5×5/1×13/13×1	67.52	90.73	75.92	51.47	80.48	2.16	11.1
7×7 / 1×3 / 3×1	68.08	91.08	77.15	52.87	80.11	2.20	11.3

(e) Model Input Image Size

Imgsz	AP(%)	$AP_{50}(\%)$	$AP_{75}(\%)$
320	50.40	74.30	57.10
640	65.80	89.70	73.90
800	68.46	91.83	77.80
1024	70.60	92.70	80.00

(f) DyN impact on HetTriDown and HCMC3K

H	D	$AP_{50}(\%)$	$AP^I_{50}(\%)$	$AP^O_{50}(\%)$
×	×	90.67	79.92	97.91
×	✓	90.85	80.65	97.90
✓	×	91.15	81.51	97.94
✓	✓	**91.83**	**81.94**	**97.96**

5 CONCLUSION

In this paper, we propose SARoLO, a structure- and scale-aware framework for synthetic aperture radar object detection that achieves state-of-the-art performance across multiple benchmarks. SARoLO introduces several key mechanisms, including a Heterogeneous Triple-Conv Downsampling (HetTriDown) module for preserving fine-grained structures during downsampling, a Heterogeneous Convolution Modulation (HCM) block for jointly modeling isotropic and anisotropic features, and a Cross-Stage Feature Aggregation module (HCMC3K) to balance feature reuse and representation capacity. In addition, a Cross-Scale Aggregation Network (CSAN) is designed to enable cross-scale feature interaction and structural reconstruction, while a dual-scale detection head is adopted to better align with the scale distribution of SAR targets. Comprehensive evaluations demonstrate SARoLO's generalization capabilities, introducing promising innovations to advance object detection in SAR applications.

REFERENCES

Zekang Zheng, Mingjia Wang, and Fengxiang Shang. Lightweight sar image ship target detection algorithm based on improved yolov8. In *2024 China Automation Congress (CAC)*, pages 570–575. IEEE, 2024.

Petre Stoica and Jian Li. Lecture notes-source localization from range-difference measurements. *IEEE Signal Processing Magazine*, 23(6):63–66, 2006.

Lizhang Zhou, Gang Zheng, Jingsong Yang, Xiaofeng Li, Bin Zhang, He Wang, Peng Chen, and Yan Wang. Sea surface wind speed retrieval from textures in synthetic aperture radar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

Ultralytics. Ultralytics yolov11. `https://github.com/ultralytics/ultralytics`, 2024. Accessed: 2025-07-22.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollar. pycocotools: Microsoft coco api. `https://github.com/cocodataset/cocoapi`, 2023.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

Glenn Jocher, Alex Stoken, Jirka Borovec, Liu Changyu, Adam Hogan, Laurentiu Diaconu, Jake Poznanski, Lijun Yu, Prashant Rai, Russ Ferriday, et al. ultralytics/yolov5: v3.0. *Zenodo*, 2020.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 529–545. Springer, 2024.

V Vijay Venkatesan, J Deekshitha, K Pujitha Reddy, and M Ganesan. Lightweight pcb defect detection using ghost convolutions in deep neural networks. In *2025 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)*, pages 1–6. IEEE, 2025.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024a.

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhu Zheng, Guiguang Ding, Shaoyi Du, Zongze Wu, and Yue Gao. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint arXiv:2506.17733*, 2025.

Nikhil Ketkar and Jojo Moolayil. Convolutional neural networks. In *Deep learning with Python: learn best practices of deep learning models with PyTorch*, pages 197–242. Springer, 2021.

Shiqi Huang, Xuewen Pu, Xinke Zhan, Yucheng Zhang, Ziqi Dong, and Jianshe Huang. Sar ship target detection method based on cnn structure with wavelet and attention mechanism. *Plos one*, (6):e0265599, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Hao Shi, Cheng He, Jianhao Li, Liang Chen, and Yupei Wang. An improved anchor-free sar ship detection algorithm based on brain-inspired attention mechanism. *Frontiers in neuroscience*, page 1074706, 2022.

Dongdong Chen, Rusheng Ju, Chuangye Tu, Guangwei Long, Xiaoyang Liu, and Jiyuan Liu. Gdb-yolov5s: Improved yolo-based model for ship detection in sar images. *IET Image Processing*, (11): 2869–2883, 2024.

Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *arXiv preprint arXiv:2203.06717*, 2022.

Yitong Wang, Shumin Wang, and Aixia Dou. A dual-branch fusion network based on reconstructed transformer for building extraction in remote sensing imagery. *Sensors*, 24:365, 2024b.

Jing Wang, Yuanyuan Hao, Hongxing Bai, and Lingyu Yan. Parallel attention recursive generalization transformer for image super-resolution. *Scientific Reports*, 15:8669, 2025.

Yuxuan Li, Xiang Li, Yimian Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. Lsknet: A foundation lightweight backbone for remote sensing. *arXiv preprint arXiv:2403.11735*, 2024.

Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27706–27716, 2024.

Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2017.

Shangxian Ruan, Jiating Yang, and Jianbo Chen. A deformable convolutional neural network with oriented response for fine-grained visual classification. In *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, pages 133–140, 2021.

Ultralytics. Ultralytics yolov8. `https://github.com/ultralytics/ultralytics`, 2023. Accessed: 2025-07-22.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8:120234–120254, 2020.

Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, Israr Ahmad, Dece Pan, Chang Liu, Yue Zhou, Jun Shi, and Shunjun Wei. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote. Sens.*, 13:3690, 2021. URL `https://api.semanticscholar.org/CorpusID:238244559`.

Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. Rsar: Restricted state angle resolver and rotated sar benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7416–7426, 2025.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery. *Remote Sensing*, 11(5):531, 2019.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in neural information processing systems*, 35:34899–34911, 2022.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3163–3171, 2021.

Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE transactions on geoscience and remote sensing*, 60:1–11, 2021a.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021.

Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2786–2795, 2021b.

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2849–2858, 2019.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Ying Zeng, Yushi Chen, Xue Yang, Qingyun Li, and Junchi Yan. Ars-detr: Aspect ratio-sensitive detection transformer for aerial oriented object detection. *IEEE transactions on geoscience and remote sensing*, 62:1–15, 2024.

Supplementary Materials for "SARoLO: Structure and Scale-Aware SAR Object Detection"

In this supplementary material, Section A provides a detailed introduction to the three datasets used in our experiments. Section B delineates the hyperparameter configurations and data processing pipeline, along with a comparative analysis of AP results under both default and optimized settings. Section C-F further elaborates on comprehensive performance comparison experiments and visualization analyses for SARoLO.

A Datasets details

HRSID Wei et al. (2020) is a high-resolution SAR image dataset designed for ship detection, semantic segmentation, and instance segmentation in synthetic aperture radar imagery. It contains 5,604 SAR images and 16,951 ship instances. The dataset follows the construction pipeline of the Microsoft Common Objects in Context (COCO) dataset, covering SAR images with different resolutions, polarizations, sea states, open seas, and coastal ports. The spatial resolutions of the SAR images are 0.5 m, 1 m, and 3 m. Each image exhibits varying dimensions (typically around 800×800 pixels) and contains ships of diverse scales under complex inshore and offshore backgrounds. All data processing and the partitioning of training and test sets strictly follow the official standards. Specifically, we adopt the official dataset split, where the training set contains 3,642 images with 11,047 instances, and the test set contains 1,962 images with 5,922 instances. Small instances (area $<$ 1,024 pixels) dominate the dataset, accounting for over 54% of all objects, while medium and large instances constitute approximately 43.5% and 2.0%, respectively. In terms of scene distribution, inshore images (18.4% of total) contain a higher proportion of small instances (over 73%), whereas offshore images (81.6% of total) are dominated by medium-sized instances (over 62%). Additionally, the dataset provides 400 pure-background SAR images (without any ship targets) to evaluate model robustness against land and sea clutter.

SSDD Zhang et al. (2021) is one of the earliest publicly available SAR datasets, sourced from RadarSat-2, TerraSAR, and Sentinel-1. It has since been revised to provide more detailed image annotations and define five usage standards. The official SSDD contains 1,160 SAR images of varying resolutions and sizes. We adopt the official dataset split, where the training set contains 928 images and the test set contains 232 images.

RSAR Zhang et al. (2025) dataset comprises 95,842 images, 183,534 instances, and six typical object categories, including Ship, Tank, Bridge, Aircraft, Harbor, and Car. Following the official dataset split, we use 78,837 images for training, 8,467 images for validation, and 8,538 images for testing.

B Implementation Details

Training Configuration. SARoLO is trained from scratch on the RSAR and HRSID datasets without any external pretrained weights. All experiments are conducted on a single RTX 3060 GPU. The input images are uniformly resized to 800×800 pixels. We adopt the SGD optimizer with an initial learning rate of 0.01995, a final learning rate factor of 0.01 (i.e., 1.995×10^{-4}), and a momentum of 0.96145. A weight decay of 5.7×10^{-4} is applied. The training runs for 300 epochs with a batch size of 24. A warmup phase of approximately 3.07 epochs is used, with a warmup momentum of 0.46709 and a warmup bias learning rate of 0.08463. The learning rate follows a linear decay schedule after warmup. The confidence threshold is set to 0.01 to suppress false positives.

Loss and Augmentation Settings. The loss weights are configured as follows: box loss gain = 0.2, class loss gain = 0.4374, and DFL (Distribution Focal Loss) gain = 1.48078. For data augmentation, we apply HSV hue (0.01322), HSV saturation (0.27142), translation (0.07414), scaling (0.11949), and horizontal flip (0.21951). Mosaic augmentation is enabled with a probability of 0.20645, while MixUp and copy-paste are disabled (set to 0.0). Rotation (degrees), shear, perspective, and vertical flip (flipud) are also disabled. All ablation studies are conducted on the HRSID dataset to ensure controlled comparisons, unless otherwise specified. Table 6 summarizes the complete set of hyperparameters used for training SARoLO.

Table 6: Hyperparameters for training SARoLO.

Hyperparameter	Value
Training Configuration	
Epochs	300
Batch size	24
Input image size	800×800
Optimizer	SGD
Initial learning rate (lr0)	0.01995
Final learning rate factor (lrf)	0.01
Momentum	0.96145
Weight decay	5.7×10^{-4}
Warmup epochs	3
Warmup momentum	0.46709
Warmup bias learning rate	0.08463
Confidence threshold (conf)	0.01

Hyperparameter	Value
Augmentation Parameters	
HSV hue (hsv_h)	0.01322
HSV saturation (hsv_s)	0.27142
Translation (translate)	0.07414
Scale (scale)	0.11949
Horizontal flip (fliplr)	0.21951
Mosaic	0.20645
MixUp	0.0
Copy-paste	0.0
Degrees (rotation)	0.0
Shear	0.0
Perspective	0.0
Vertical flip (flipud)	0.0

Hyperparameter	Value
Loss Parameters	
Box loss gain	0.2
Class loss gain	0.4374
DFL loss gain	1.4807

Table 7: Ship Detection Statistics Generated by Bounding Box AP on Test Set of HRSID. Model size is measured in megabytes (MB). The symbol * indicates that the SARoLO model uses an input resolution of 1024×1024; models without * use 800×800.

Model	Backbone	Model Size (MB)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN Ren et al. (2015)	ResNet-50+FPN	330.2	63.5	86.7	73.3	64.4	65.1	16.4
Faster R-CNN Ren et al. (2015)	ResNet-101+FPN	482.4	63.9	86.7	73.6	64.8	66.2	24.2
Cascade R-CNN Cai and Vasconcelos (2018)	ResNet-50+FPN	552.6	66.6	87.7	76.4	67.5	67.7	28.8
Cascade R-CNN Cai and Vasconcelos (2018)	ResNet-101+FPN	704.8	66.8	87.9	76.6	67.5	68.8	27.7
RetinaNet Wang et al. (2019)	ResNet-50+FPN	290.0	60.0	84.7	67.2	60.9	60.9	26.8
RetinaNet Wang et al. (2019)	ResNet-101+FPN	442.3	59.8	84.8	67.2	60.4	62.7	26.5
Mask R-CNN He et al. (2017)	ResNet-50+FPN	351.2	65.0	88.0	75.2	66.1	66.1	17.3
Mask R-CNN He et al. (2017)	ResNet-101+FPN	503.4	65.4	88.1	75.7	66.3	68.0	23.2
Mask Scoring R-CNN Huang et al. (2019)	ResNet-50+FPN	481.1	64.1	87.6	75.0	65.3	65.8	22.2
Mask Scoring R-CNN Huang et al. (2019)	ResNet-101+FPN	633.1	64.9	88.6	75.4	66.2	67.3	19.6
Cascade Mask R-CNN Cai and Vasconcelos (2019)	ResNet-50+FPN	615.6	67.5	88.5	77.4	68.6	67.4	22.6
Cascade Mask R-CNN Cai and Vasconcelos (2019)	ResNet-101+FPN	767.8	67.6	88.8	77.4	68.4	69.9	23.9
Hybrid Task Cascade Chen et al. (2019)	ResNet-50+FPN	639.3	68.2	87.7	78.8	69.0	71.2	38.1
Hybrid Task Cascade Chen et al. (2019)	ResNet-101+FPN	791.6	68.4	87.7	78.8	69.2	72.0	31.9
HRSDNet Wei et al. (2020)	HRFPN-W32	598.1	68.6	88.4	79.0	69.6	70.0	25.2
HRSDNet Wei et al. (2020)	HRFPN-W40	728.2	69.4	89.3	79.8	70.3	71.1	28.9
YOLO8 Ultralytics (2023)	YOLO Backbone	11.83	67.5	90.6	76.2	68.4	71.8	21.4
YOLO10 Wang et al. (2024a)	YOLO Backbone	21.59	64.6	86.8	73.7	68.6	68.8	10.0
YOLO11 Ultralytics (2024)	YOLO Backbone	10.32	66.4	89.8	75.3	67.9	70.7	11.0
YOLO12 Tian et al. (2025)	YOLO Backbone	10.23	66.0	89.8	74.5	67.5	66.6	21.2
YOLO13 Lei et al. (2025)	YOLO Backbone	10.13	65.9	89.9	74.2	66.6	69.0	40.2
SARoLO	YOLO Backbone	7.32	68.5	91.8	77.8	69.5	73.7	30.6
SARoLO*	YOLO Backbone	7.32	70.6	92.7	80.0	72.2	73.9	16.3

C Comparisons with State-of-the-Art Methods

C.1 Performance Comparison on Ship Detection.

For the HRSID dataset, based on the experimental results reported in Table 7, several data-driven observations can be made. In terms of overall detection accuracy, the proposed SARoLO* model, which uses an input resolution of 1024×1024, achieves the highest AP of 70.6%, outperforming all existing methods including HRSDNet (HRFPN-W40, 69.4%), Hybrid Task Cascade (68.4%), and Cascade Mask R-CNN (67.6%). Notably, SARoLO* attains this performance with a model size of only 7.32 MB, which is orders of magnitude smaller than two-stage detectors (minimum 290 MB) and even more compact than YOLO8 (11.83 MB) and YOLO10 (21.59 MB). The lower-resolution version SARoLO (800×800) also yields a competitive AP of 68.5%, surpassing YOLO8 (67.5%) and other lightweight models. Regarding localisation precision, SARoLO* achieves the highest AP_{50} (92.7%) and AP_{75} (80.0%) among all compared methods, indicating superior bounding box quality even at strict IoU thresholds. For reference, HRSDNet obtains 89.3% and 79.8% respectively, while Hybrid Task Cascade reaches 87.7% and 78.8%. The multi-scale analysis shows that SARoLO* excels in small and medium object detection, with AP_S of 72.2% and AP_M of 73.9%, both higher than HRSDNet (70.3% and 71.1%) and Hybrid Task Cascade (69.0% and 72.0%). Finally, comparing SARoLO (800×800) with SARoLO* (1024×1024) reveals that the higher resolution brings a +2.1% gain in AP, a +2.7% gain in AP_S, and a slight +0.2% gain in AP_M.

In the main paper, Table 2 shows that SARoLO achieves the highest AP (63.21%) among all YOLO-series detectors on SSDD. The full comparison in Table 8 (Appendix) further benchmarks SARoLO against classic two-stage detectors. The results demonstrate that SARoLO not only outperforms all two-stage methods, e.g., Hybrid Task Cascade (61.1%), HRSDNet (61.1%), and Cascade R-CNN (60.3%), but also surpasses YOLO13 (62.31%) and other YOLO variants across all metrics: AP (63.21% vs. 62.31%), AP_{50} (97.46% vs. 96.71%), AP_{75} (75.01% vs. 71.65%), AP_S (57.49% vs. 56.91%), AP_M (72.1% vs. 70.30%), and AP_L (69.18% vs. 67.55%). Notably, SARoLO achieves this state-of-the-art performance with a model size of only 7.32 MB, which is an order of magnitude smaller than YOLO13 (10.13 MB) and two orders smaller than two-stage detectors (290–790 MB). These supplementary results confirm that SARoLO establishes a new state of the art on SSDD, delivering superior accuracy across all object scales with exceptional model compactness.

Table 8: Comparison of different methods on the SSDD dataset. Model size is measured in megabytes (MB).

Model	**Backbone**	**Model Size (MB)**	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**
Faster R-CNN	ResNet-50+FPN	330.2	59.1	93.8	68.6	55.2	66.0	47.3
Faster R-CNN	ResNet-101+FPN	482.4	58.4	94.0	65.9	54.5	64.7	51.9
Cascade R-CNN	ResNet-50+FPN	552.6	59.7	93.1	67.6	54.8	67.1	57.8
Cascade R-CNN	ResNet-101+FPN	704.8	60.3	94.0	69.6	56.0	66.6	59.3
RetinaNet	ResNet-50+FPN	290.0	55.5	90.2	62.3	51.2	62.6	45.4
RetinaNet	ResNet-101+FPN	442.3	55.2	90.8	60.2	50.9	62.2	49.7
Mask R-CNN	ResNet-50+FPN	351.2	58.9	93.4	66.6	55.3	64.9	49.7
Mask R-CNN	ResNet-101+FPN	503.4	59.4	93.9	67.7	54.9	66.2	53.9
Mask Scoring R-CNN	ResNet-50+FPN	481.1	59.4	94.7	67.8	55.6	65.3	51.2
Mask Scoring R-CNN	ResNet-101+FPN	633.1	59.8	94.9	68.2	54.9	67.0	53.8
Cascade Mask R-CNN	ResNet-50+FPN	615.6	59.7	93.1	68.9	55.5	65.9	53.2
Cascade Mask R-CNN	ResNet-101+FPN	767.8	59.9	92.4	68.9	55.6	66.4	58.3
Hybrid Task Cascade	ResNet-50+FPN	639.3	61.1	94.5	70.3	55.9	68.4	59.2
Hybrid Task Cascade	ResNet-101+FPN	791.6	60.7	93.6	69.5	55.7	67.6	61.2
HRSDNet	HRFPN-W32	598.1	61.1	93.9	70.1	56.6	67.7	58.8
HRSDNet	HRFPN-W40	728.2	60.9	94.4	69.7	56.2	67.8	58.9
YOLO8	YOLO Backbone	11.83	61.90	96.55	70.80	56.47	70.79	64.05
YOLO10	YOLO Backbone	21.59	61.90	96.49	71.55	57.05	71.04	64.49
YOLO11	YOLO Backbone	10.32	62.25	97.20	70.98	56.56	70.62	66.45
YOLO12	YOLO Backbone	10.23	61.86	96.97	71.28	56.83	69.95	67.15
YOLO13	YOLO Backbone	10.13	62.31	96.71	71.65	56.91	70.30	67.55
SARoLO	YOLO Backbone	7.32	63.21	97.46	75.01	57.49	72.12	69.18

C.2 Performance Comparison on Multi-class Oriented Object Detection.

For the multi-class RSAR benchmark (Table 9), we follow the same evaluation protocol introduced in Section 4.2 (*Experimental Evaluation: Benchmarking on Popular Datasets*). Since RSAR contains multiple object categories, we report the category-averaged detection metrics, including mAP_{50}, mAP_{75}, and mAP. Specifically, the mean Average Precision is calculated by averaging the AP values over all object categories:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{4}$$

where N denotes the number of object categories and AP_i represents the average precision of the i-th category. Similarly, mAP_{50} and mAP_{75} are calculated under IoU thresholds of 0.5 and 0.75, respectively.

As shown in Table 9, SARoLO achieves the best overall performance among all compared OBB detection methods, obtaining 93.85% mAP_{50}, 71.19% mAP_{75}, and 63.85% mAP. Among the YOLO-series detectors, SARoLO consistently improves over recent versions, achieving gains of 1.15%, 1.17%, and 1.21% in mAP_{50}, mAP_{75}, and mAP compared with YOLO13, respectively. Compared

Table 9: Performance comparisons of OBB detection methods on the RSAR dataset.

Method	mAP50↑	mAP75↑	mAP↑
YOLO8Ultralytics (2023)	92.01	68.14	61.36
YOLO11Ultralytics (2024)	92.42	68.60	61.81
YOLO12Tian et al. (2025)	92.61	69.32	62.44
YOLO13Lei et al. (2025)	92.70	70.02	62.64
RetinaNet Tian et al. (2022)	57.67	22.72	27.65
R3Det Yang et al. (2021)	63.94	25.02	30.50
S2ANet Han et al. (2021a)	66.47	28.52	33.11
FCOS Lin et al. (2017)	66.66	31.45	34.22
Faster RCNN Ren et al. (2015)	63.18	24.88	30.46
O-RCNN Xie et al. (2021)	64.82	32.69	33.62
ReDet Han et al. (2021b)	64.71	32.84	34.30
RoI-Transformer Ding et al. (2019)	66.95	32.65	35.02
Deformable DETR Zhu et al. (2020)	46.62	13.06	19.63
ARS-DETR Zeng et al. (2024)	61.14	28.97	31.56
SARoLO	93.85	71.19	63.85

with YOLO11, the baseline architecture adopted in this work, SARoLO further improves mAP_{50}, mAP_{75}, and mAP by 1.43%, 2.59%, and 2.04%, respectively. Beyond YOLO-based detectors, SARoLO also achieves consistent improvements over representative rotated object detection methods. Compared with ReDet, RoI-Transformer, and S2ANet, SARoLO increases the mAP by 29.55%, 28.83%, and 30.74%, respectively. Compared with the transformer-based ARS-DETR, SARoLO obtains improvements of 32.71%, 42.22%, and 32.29% in mAP_{50}, mAP_{75}, and mAP, respectively. These results demonstrate that SARoLO provides more accurate multi-class oriented object detection results on RSAR.

D Ablation Studies & Result Analyses

DyN impact on HetTriDown and HCMC3K (Extended Experiment). As shown in Table 5(f), the HetTriDown and HCMC3K combination improves AP_{50} from 90.67% to 91.83% (+1.16%), inshore AP^{I}_{50} from 79.92% to 81.94% (+2.02%), and offshore AP^{O}_{50} reaches 97.96%. Extending to multi-scale metrics (Table 10),HetTriDown and HCMC3K yields improvements over the baseline (without HetTriDown or HCMC3K): AP from 67.69% to 68.46% (+0.77%), AP_{75} from 76.79% to 77.80% (+1.01%), AP_S from 68.66% to 69.51% (+0.85%), AP_M from 72.94% to 73.66% (+0.72%), and notably AP_L from 19.03% to 30.62% (+11.59%). Removing either HetTriDown or HCMC3K degrades all metrics, especially AP_L. These results confirm the effectiveness of HCMC3K (for large-scale inshore targets) and HetTriDown (for multi-scale localization accuracy).

Table 10: DyN impact on HetTriDown and HCMC3K. ✓ denotes using the DyN, × denotes not.

HCMC3K	HetTriDown	AP(%)	AP_{50}(%)	AP_{75}(%)	AP_S(%)	AP_M(%)	AP_L(%)	Params(M)	FLOPs(G)
×	×	67.69	90.67	76.79	68.66	72.94	19.03	1.80 (1,795,234)	10
×	✓	68.30	91.15	76.96	69.39	73.40	25.89	1.80 (1,795,264)	10
✓	×	67.63	90.85	75.96	68.76	73.53	14.78	1.80 (1,795,246)	10
✓	✓	**68.46**	**91.83**	**77.80**	**69.51**	**73.66**	**30.62**	1.80 (1,795,276)	10

Multi-scale analysis of detection heads and modules. Table 11 complements the results in Table 3 of the main paper (which reports AP, AP_{50}, AP_{75}, AP_I, and AP_O) by providing a detailed multi-scale evaluation across small, medium, and large objects (AP_S, AP_M, AP_L). The full SARoLO with 2 heads (row 4) achieves the highest AP_S (69.51%), AP_M (73.66%), and particularly AP_L (30.62%), outperforming the 1-head (row 3) by +7.93% on AP_L and the 3-head (row 5) by +10.04%. Removing HetTriDown (row 6) or HCMC3K (row 7) degrades AP_L to 18.42% and 23.87%, respectively, while removing both (row 8) yields the lowest AP_L (15.12%). These results confirm that the proposed components jointly enhance detection performance across all object scales, especially for large targets in complex inshore scenes, providing further evidence beyond the overall metrics reported in Table 3.

Table 11: Ablation study on detection heads and modules. ✓ denotes using the component, × denotes not. The overall metrics (AP, AP_{50}, AP_{75}, AP_I, AP_O) are reported in Table 3 of the main paper. Here we focus on multi-scale performance (AP_S, AP_M, AP_L).

#	Model	Head	CSAN	HetTriDown	HCMC3K	Params(M)	FLOPs(G)	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
1	YOLO11	3	×	×	×	2.59	6.4	67.94	70.73	11.00
2	YOLO11	2	×	×	×	1.85	5.8	68.20	69.00	20.05
3	SARoLO	1	✓	✓	✓	1.51	9.1	69.38	73.33	22.69
4	SARoLO	2	✓	✓	✓	1.80	10.0	**69.51**	**73.66**	**30.62**
5	SARoLO	3	✓	✓	✓	2.45	10.5	69.37	73.66	20.58
6	SARoLO	2	✓	×	✓	1.99	10.6	69.45	73.25	18.42
7	SARoLO	2	✓	✓	×	2.07	9.7	69.11	72.44	23.87
8	SARoLO	2	×	✓	✓	0.97	5.2	67.96	72.30	15.12

Table 12: False detection comparison on 400 background-only SAR images. "False Det. (Img.)" stands for the number of images containing at least one false detection; "FPR" denotes the false positive rate (%); "Misclass. (Obj.)" indicates the total number of misclassified target objects.

Model	False Det. (Img.)	FPR (%)	Misclass. (Obj.)
YOLO8	30	7.50	44
YOLO11	41	10.25	59
YOLO10	32	8.00	36
YOLO12	32	8.00	40
YOLO13	26	6.50	30
SARoLO	24	6.00	27

E Result Visualization

We have extended the analysis by incorporating Grad-CAM visualizations. Figures 5 and 6 present qualitative comparisons across different SAR scenarios. As shown in Figure 5, YOLO8–YOLO13 exhibit missed detections and inaccurate localization in dense small-object regions and nearshore scenes. In contrast, SARoLO produces more complete detections with tighter bounding boxes, especially for small and closely spaced targets. Figure 6 evaluates model behavior on background-only images (400 images containing no targets). As quantified in Table 12, YOLO8–YOLO13 generate visible false positives: for example, YOLO11 produces false detections in 41 images (FPR = 10.25%) and misclassifies 59 objects, while the best competitor (YOLO13) still yields 26 false detected images (FPR = 6.5%) and 30 misclassifications. In contrast, SARoLO produces only 24 false detected images (FPR = 6.0%) and 27 misclassified objects. Since any predicted box in these images corresponds to a false alarm, these quantitative results directly confirm that SARoLO has fewer false positives and more robust background rejection. Extending the comparison, Figure 7 presents Grad-CAM visualizations of YOLO8–YOLO13 and the proposed SARoLO. Warm colors (red/yellow) indicate high-activation regions, while cool colors (blue/green) represent low-activation areas. SARoLO exhibits stronger red intensity in its heatmaps, directly correlating with heightened spatial attention on target regions. It achieves higher attention density and displays more distinct heat signatures on texture-camouflaged and small targets. In contrast, YOLO8–YOLO13 show reduced focus accuracy and more diffuse activations under challenging conditions such as dense, small objects. The superior heatmap concentration of SARoLO further supports its precise detection capability, especially for small and closely spaced ship targets, consistent with the quantitative results above.

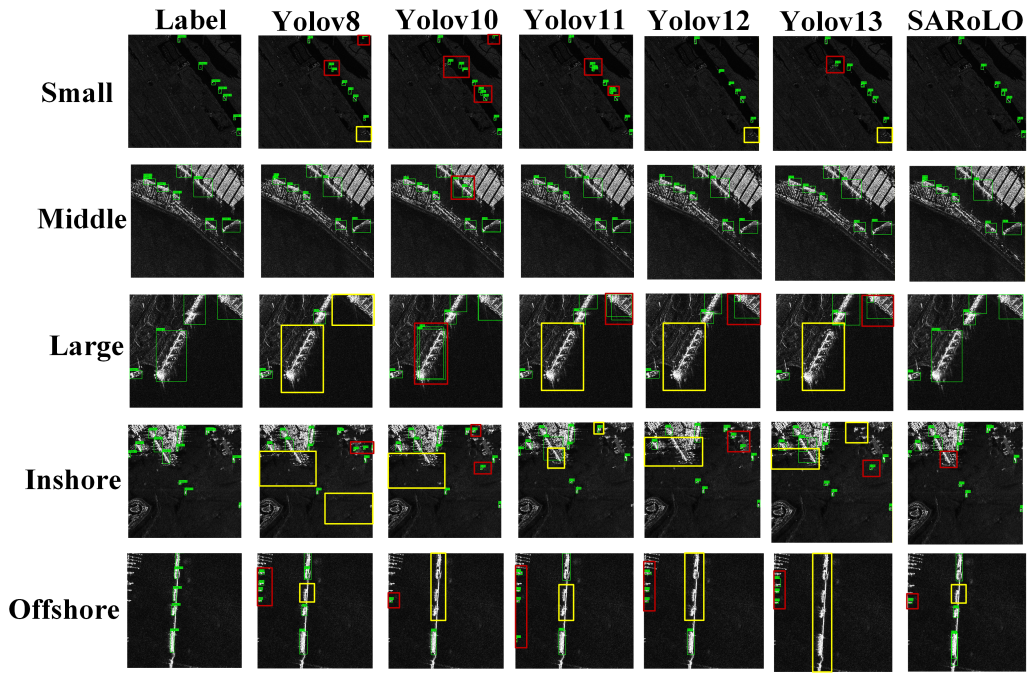


Figure 5: Qualitative comparison of detection performance across multi-scale and multi-scenario SAR images, including small, medium, large, inshore, and offshore targets. Compared with YOLO8–YOLO13, SARoLO achieves more accurate localization and fewer missed detections in dense small-object regions and complex nearshore backgrounds. Red boxes highlight missed detections, yellow boxes indicate false detections, and green denotes predicted/ground truth boxes.

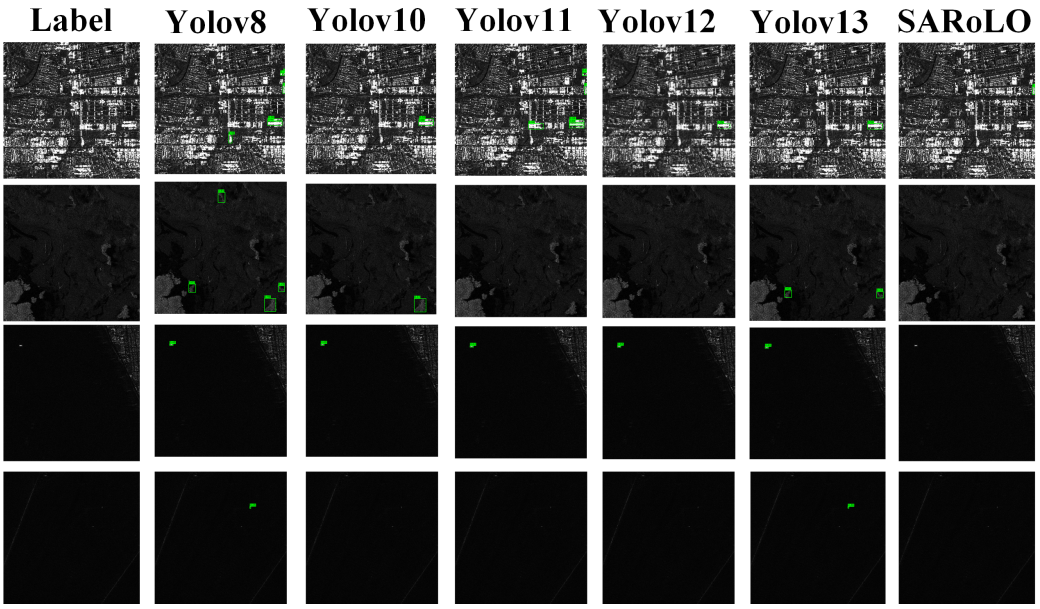


Figure 6: Qualitative comparison on background-only SAR images for evaluating false positive suppression. SARoLO effectively reduces false detections compared with other methods, demonstrating stronger robustness to background clutter. Any predicted bounding box indicates a false detection.

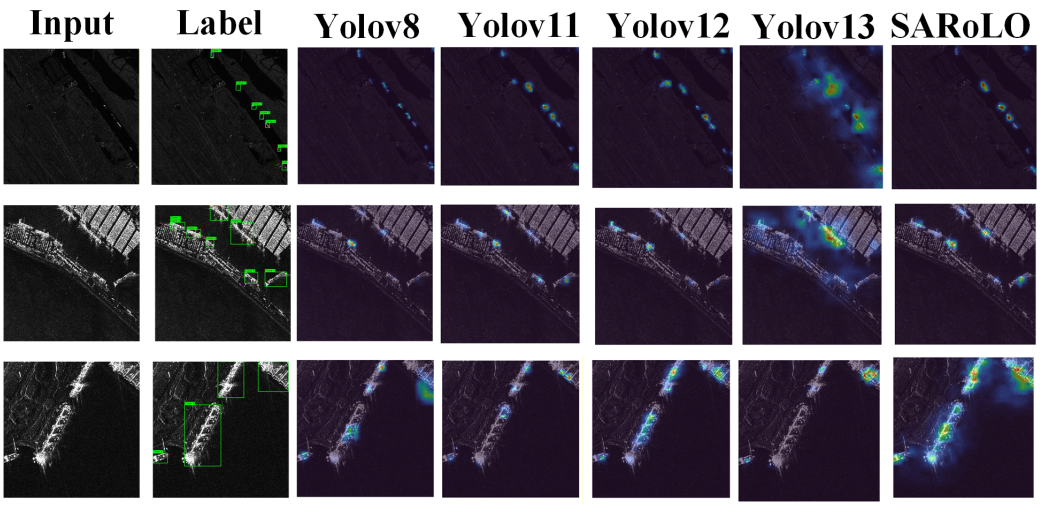


Figure 7: Comparison of GradCAM heatmaps among YOLO8, YOLO11–13, and the proposed SARoLO. Warm colors (red/yellow) indicate high-activation regions, while cool colors (blue/green) indicate low-activation areas.

AI use statement

In this work, we used generative AI tools for the following tasks: editing the research paper to improve readability, identifying relevant literature, formatting references, suggesting a structure for the research paper, and proposing a title or keywords for the research paper. Additionally, we used generative AI tools to convert manuscript sections into LaTeX code for use in Overleaf.

We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims. We have reviewed all AI-assisted work. Specifically, all AI-assisted content was manually verified by the authors for accuracy and appropriateness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.