AIMS Mathematics, 11(7): 22920–22942.
DOI: 10.3934/math.2026923
Received: 29 April 2026
Revised: 01 July 2026
Accepted: 17 July 2026
Published: 29 July 2026

http://www.aimspress.com/journal/Math

*Research article*

# YOLO-Argus: An instance segmentation model based on scale-aware downsampling and parallel multi-scale edge fusion for SAR ship images

**Qiming Li[1], Hao Yin[1] and Daozheng Chen[2,*]**

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[2] Informatization Office, Shanghai Maritime University, Shanghai 201306, China

* **Correspondence:** Email: dzchen@shmtu.edu.cn; Tel: +8602138282823.

**Abstract:** With the rapid growth of global maritime transportation, accurate ship instance segmentation from SAR images has become a core technology for maritime surveillance and management. However, speckle noise, significant scale variations, and the presence of dense and structurally ambiguous targets in complex nearshore SAR scenes pose substantial challenges to feature extraction, multi-scale representation, and precise boundary delineation in instance segmentation. Therefore, we have proposed YOLO-Argus, an enhanced model for SAR ship instance segmentation. First, a spatial parallel context perception (SPCP) module is introduced to construct a dual-path feature extraction framework using multi-dilated convolutions, thereby improving multi-scale perception, particularly for small targets in complex backgrounds. Second, a multi-scale edge fusion (MSEF) module integrates edge enhancement with contextual semantic information to improve discrimination of blurred boundaries and contour structures. Furthermore, an adaptive tanh unit (ATU) embeds adaptive normalization mechanisms into attention blocks, effectively suppressing speckle noise while preserving fine-grained details. Finally, a multi-scale channel decoupling (MSCD) module leverages grouped and multi-scale convolutions to jointly optimize feature representations, enhancing robustness under noise interference. Experimental results on the public HRSID and PSeg-SSDD datasets quantitatively confirmed the effectiveness of YOLO-Argus. Compared with the baseline model, YOLO-Argus improves $AP_{50}$ and $AP_{75}$ by 4.3 and 4.7 percentage points on HRSID, and by 3.5 and 9.2 percentage points on PSeg-SSDD, respectively. The proposed method also achieved superior segmentation accuracy compared with existing state-of-the-art methods, including both general-purpose instance segmentation models and SAR ship-specific approaches. These results indicate that the proposed approach effectively mitigates the interference of noise, scale variation, and background complexity in SAR images, providing a more reliable solution for ship instance segmentation.

**Keywords:** instance segmentation; ship targets; synthetic aperture radar images; YOLOv11; multi-scale feature fusion; edge enhancement

**Mathematics Subject Classification:** 68T07, 68U10

## 1. Introduction

With the rapid expansion of global trade, maritime transportation now accounts for more than 80% of international freight volume, making accurate and efficient ship monitoring a critical component of maritime traffic safety, national maritime sovereignty protection, and the development of smart ocean systems. Traditional ship monitoring approaches, such as optical remote sensing satellites, are highly susceptible to adverse weather conditions and illumination constraints, rendering them ineffective under cloud cover, fog, rainfall, or nighttime scenarios. By contrast, synthetic aperture radar (SAR), as an active microwave remote sensing system, offers all-weather, day-and-night imaging capabilities and strong penetration through clouds and fog, and has therefore become an indispensable technology for maritime surveillance. SAR systems can continuously acquire high-resolution marine images over large spatial extents and under complex meteorological and sea-state conditions, providing a reliable data source for large-scale and near-real-time ship detection and analysis.

Early studies on SAR-based ship analysis primarily focused on target detection, aiming to determine the presence and approximate locations of ships within an image. However, with the increasing complexity of maritime operations, merely identifying ship positions is no longer sufficient. Modern maritime applications, such as traffic monitoring, risk assessment, and fine-grained vessel analysis, require more detailed structural and spatial information. Consequently, pixel-level ship instance segmentation has emerged as a crucial task, enabling precise delineation of individual ship contours and facilitating more comprehensive maritime understanding.

Despite its practical importance, SAR ship instance segmentation remains a highly challenging problem due to the intrinsic characteristics of SAR imaging and the complexity of marine environments. As a coherent imaging modality, SAR images are inherently affected by strong multiplicative speckle noise, which introduces granular textures and significantly degrades the signal-to-noise ratio and visual quality of images. This noise blurs ship boundaries and weakens the contrast between targets and background, making reliable feature extraction difficult. In addition, ship targets in SAR images exhibit substantial scale variation, ranging from small fishing vessels tens of meters in length to large cargo ships and oil tankers extending hundreds of meters. Accurately capturing features across such a wide-scale spectrum with a single model remains challenging. Furthermore, in port and nearshore regions, ships are often densely moored, leading to frequent adhesion and mutual occlusion between vessels. Meanwhile, port infrastructures such as vehicles and cranes share similar scattering characteristics with ships, increasing the likelihood of false detections. Therefore, an effective instance segmentation model must not only accurately identify ship targets but also correctly separate adjacent instances and infer complete contours for partially occluded vessels.

To address the above challenges, we propose YOLO-Argus, an enhanced model for SAR ship instance segmentation. By jointly integrating multi-scale feature modeling, structural information enhancement, and feature modulation mechanisms, the proposed method improves segmentation performance and robustness in complex marine environments. The main contributions of this work are summarized as follows:

(1) To handle significant scale variations in SAR images, we design a spatial parallel context perception (SPCP) module. By constructing a multi-path feature modeling mechanism, it effectively enhances the representation of multi-scale targets, particularly small ships.

(2) To address blurred boundaries and noise-sensitive structural information, we propose a multi-scale edge fusion (MSEF) module. By integrating edge cues with contextual semantic features, the module improves the modeling of object contours and structural continuity.

(3) To mitigate unstable feature responses caused by speckle noise, we introduce an adaptive tanh unit (ATU). By dynamically modulating feature amplitudes within attention mechanisms, it effectively suppresses noise interference while preserving fine-grained details.

(4) To improve feature robustness under complex conditions, we develop a multi-scale channel decoupling (MSCD) module. Through joint modeling across multi-scale and channel dimensions, it enhances feature representation in noisy and cluttered environments.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 elaborates on the proposed method in detail. Section 4 presents the experimental setup and results analysis. Section 5 concludes the paper and discusses future research directions.

## 2. Related work

Image instance segmentation is a fundamental task in computer vision, aiming to achieve pixel-level object recognition and delineation.

Early image segmentation methods primarily relied on traditional digital image processing techniques. Researchers proposed various algorithms based on low-level features, laying the foundation for subsequent developments. Among them, thresholding methods classify pixels using global or local optimal thresholds. For example, genetic algorithms (GAs) [1] formulate threshold selection as a fitness maximization problem, while differential evolution (DE) [2] achieves rapid convergence through single-objective optimization. Edge-based segmentation [3] first detects salient gradient boundaries and then obtains complete objects through contour closure. The watershed transform (WT) [4], a technique in mathematical morphology, has been widely used due to its sensitivity to weak boundaries. However, these methods rely heavily on handcrafted features and exhibit limited generalization and robustness under the inherent characteristics of SAR images, such as speckle noise, strong sea clutter, and significant scale variations.

As traditional methods based on handcrafted features and heuristic rules struggle to cope with complex imaging conditions and significant scale variations, instance segmentation research has gradually shifted toward data-driven paradigms centered on deep learning. Accordingly, research has focused on key aspects such as multi-scale modeling, feature fusion, and model generalization. Along these directions, extensive efforts have been made from perspectives including network architecture design, generative modeling, and open-vocabulary learning. ASF-YOLO [5] significantly improves accuracy and efficiency in cell instance segmentation by introducing scale-sequence feature fusion, a triple feature encoder, and channel and spatial attention mechanisms. DiffusionInst [6] reformulates instance segmentation as a denoising process from noise vectors to object representations, eliminating the dependence on region proposal networks and achieving competitive performance on general benchmarks. GLFRNet [7] proposes a global–local feature re-fusion network that leverages complementary features extracted from convolutional neural network (CNN) and VMamba branches

through cross-dimensional and semantic fusion. CATNet [8] aggregates multi-scale contextual information using dense feature pyramids, spatial context pyramids, and hierarchical region extractors to address challenges such as large scale variation and low contrast in remote sensing images. In the domain of open-vocabulary and zero-shot learning, MosaicFusion [9] proposes a training-free data augmentation method based on diffusion models, generating multi-instance images via a single diffusion step and extracting masks through cross-attention maps, thereby improving performance on large-vocabulary long-tail and open-vocabulary instance segmentation tasks. OpenVIS [10] introduces the InstFormer framework, which encourages comprehensive instance proposals through contrastive boundary loss and enables segmentation and tracking of arbitrary categories via an adapted InstCLIP encoder. Zhu et al. [11] proposed a normalization-free transformer architecture that challenges the necessity of LayerNorm, achieving stable training and competitive performance without normalization operations. CGNet [12] integrates local, surrounding, and global context through a context-guided module, maintaining high accuracy with significantly reduced parameters. These deep-learning approaches provide a solid foundation and diverse design paradigms for SAR ship instance segmentation.

However, directly applying general segmentation techniques to SAR ship instance segmentation often fails to fully adapt to its unique imaging characteristics and complex scene conditions, thereby requiring task-specific modeling and design. To address the high cost of annotation, semi-supervised and weakly supervised methods have been widely explored. For example, GGSE-SSIS [13] introduces gradient prior guidance (GPG) and SAR image adaptive enhancement (SIAE) modules, achieving performance close to fully supervised methods using only 30% pixel-level annotations. SemiSegSAR [14] incorporates graph signal processing (GSP) to enable efficient segmentation with limited labeled data. Anchor-free methods [15], such as centroid distance loss-based designs and Gaussian heatmap-based approaches, effectively reduce computational complexity while improving performance for dense targets. Multi-scale and contextual modeling have also attracted considerable attention. MS-FPN [16] captures ship features at different scales using atrous convolution pyramids (ACP) and multi-scale attention modules (MSAMs). GCBANet [17] enhances spatial dependency and boundary accuracy through global context information modeling (GCIM) and boundary-aware box prediction (BABP). LFG-Net [18] improves small-target segmentation by combining low-level feature-guided pyramids (LFCPs) with super-resolution denoising techniques. To address the variability in ship orientations, SRNet [19] introduces rotated bounding boxes (RB-Boxes) and dual feature alignment modules (DAMs) for more accurate shape modeling. SAMSAR [20] adapts the segment anything model (SAM) by incorporating feature and positional adapters, parallel CNN branches, and cross-fusion attention (CFA), significantly improving robustness to SAR noise. For SAR and remote-sensing images, several modality-specific methods have been proposed. RSPrompter [21] employs prompt learning to guide the SAM for semantic instance segmentation in remote-sensing images. DiffSARShipInst [22] formulates SAR ship instance segmentation as a denoising and reconstruction process from noisy bounding boxes to target instances, incorporating spatial–context joint enhancement and IoU-aware loss to improve accuracy. Furthermore, FL-CSE-ROIE [23] and MAISE-Net [24] optimize feature pyramids and mask prediction through full-level context excitation and mask-aware interaction mechanisms, respectively. Overall, current research demonstrates a clear trend from general-purpose frameworks toward SAR-specific modeling, with significant progress in lightweight design, semi-supervised learning, multi-scale fusion, and noise-adaptive mechanisms.

## 3. Methodology

To improve instance segmentation performance in SAR ship images, we develop YOLO-Argus, a high-precision and robust segmentation network inspired by the design principles of YOLOv11, targeting challenges posed by noise interference, complex backgrounds, and large-scale variations. As illustrated in Figure 1, the four proposed modules progressively construct a hierarchical optimization process along the feature flow, enabling the model to effectively capture and integrate diverse feature representations across different levels, thereby improving the overall segmentation performance. Specifically, during the downsampling stage of the backbone, SPCP introduces multi-scale feature modeling to enhance the representation of targets at different scales, particularly alleviating the loss of small-target information. Subsequently, MSEF strengthens the modeling of object contours and boundary structures through multi-scale edge enhancement and feature fusion. To address unstable responses in deep semantic features, the ATU introduces an adaptive amplitude modulation mechanism that dynamically regulates feature responses through bounded nonlinear mapping in the attention-enhanced feature space, thereby improving the stability of fine-grained details. Finally, MSCD constructs a scale-decoupled representation in the deep feature space, enhancing the model's adaptability to dynamic scale variations.

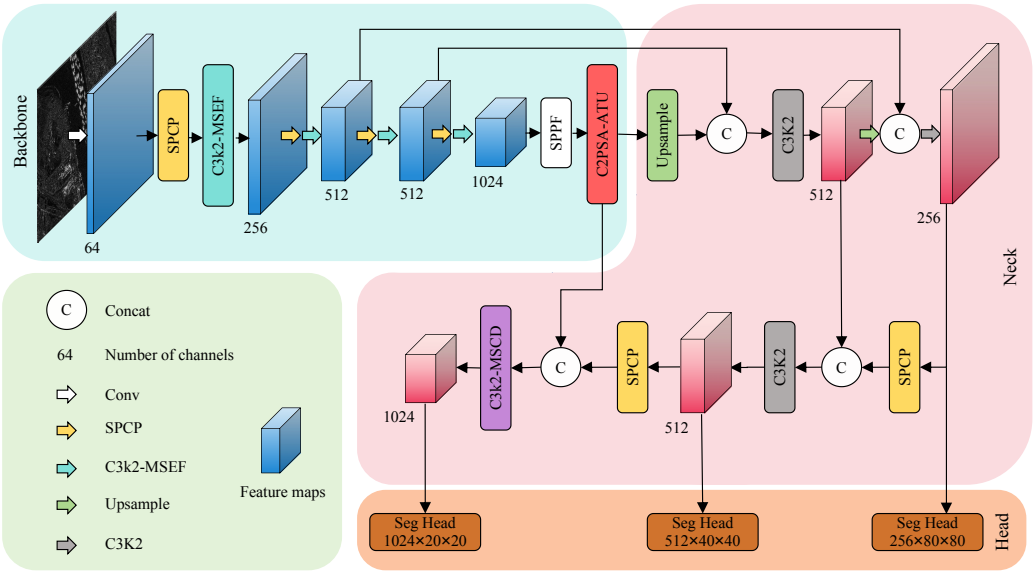


**Figure 1.** Overall architecture of YOLO-Argus. The network adopts a backbone–neck–head architecture. The highlighted yellow, blue, red, and purple modules denote the four proposed components: SPCP, C3k2-MSEF, C2PSA-ATU, and C3k2-MSCD, respectively.

### *3.1. SPCP*

During the downsampling stages of the backbone network, feature maps undergo progressive spatial resolution reduction, which requires a careful balance between semantic aggregation and preservation of spatial structure. In SAR ship images, targets exhibit pronounced scale variation and elongated shapes, and are embedded in complex noisy backgrounds. Therefore, the downsampling process should not only perform feature compression and receptive field expansion, but also provide sufficiently

discriminative multi-scale representations for subsequent segmentation branches. To this end, the SPCP module is introduced into the downsampling stages to further model intermediate features and enhance the representation capability for ship targets at different scales.

SAR ship targets exhibit a pronounced bimodal pattern in their spatial extent: offshore vessels typically occupy only a small number of pixels and rely on strictly local responses to preserve their compact boundary cues, whereas nearshore vessels are accompanied by dense coastal clutter and adjacent dock structures that demand a moderately enlarged contextual receptive field. A single fixed receptive field is therefore difficult to reconcile with both requirements. To accommodate this duality without resorting to a heavyweight multi-branch dilation pyramid, SPCP adopts a two-branch contextual aggregation strategy that explicitly couples a strictly local response with a slightly enlarged surrounding response.

As illustrated in Figure 2, given an input feature map $F_{in} \in \mathbb{R}^{H\times W\times C}$, the SPCP module first applies a $3 \times 3$ convolution with a stride of 2 to perform spatial downsampling and channel projection. This operation reduces spatial resolution while expanding the effective receptive field, thereby providing a unified feature basis for subsequent multi-scale context modeling and reducing computational overhead. The resulting downsampled feature is denoted as $F_d = \mathrm{Conv}_{3\times3,s=2}(F_{in})$. Subsequently, SPCP adopts a two-branch context aggregation structure: two parallel depthwise convolutions with different dilation rates are applied to $F_d$, and their outputs are concatenated along the channel dimension to form the multi-scale contextual representation:

$$F_{ms} = \mathrm{Concat}\left(\{\mathrm{DWConv}_{3\times3,r_i}(F_d)\}_{i=1}^{N}\right) \tag{3.1}$$

where $\mathrm{DWConv}_{3\times3,r_i}(\cdot)$ denotes a $3 \times 3$ depthwise convolution with dilation rate $r_i$, $N$ is the number of parallel branches, and $\mathrm{Concat}(\cdot)$ represents channel-wise concatenation. In our implementation, $N = 2$, with dilation rates set as $r_1 = 1$ and $r_2 = 2$, producing effective receptive fields of $3 \times 3$ and $5 \times 5$, respectively.

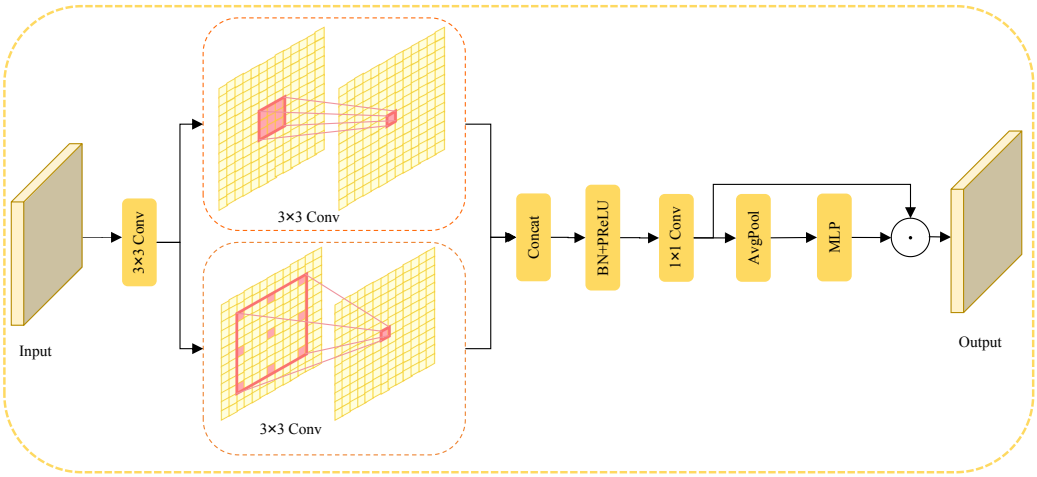


**Figure 2.** Structure of the SPCP module. The module captures local and surrounding contextual features through two complementary branches and performs channel recalibration using a lightweight MLP, enabling a synergistic balance between detail preservation and receptive field expansion.

The branch with $r_1 = 1$ strictly preserves the local boundary cues of compact vessels and densely arranged small targets, while the branch with $r_2 = 2$ supplies a moderately enlarged neighborhood receptive field that resolves ship instances against nearby sea clutter and dock structures. We deliberately refrained from adopting larger dilation rates on the already low-resolution downsampled feature maps, because excessively dilated receptive fields integrate responses from coastlines, harbors, and broader background clutter, and this absorption of long-range scene context weakens the locality cues on which very small vessels critically depend. The two complementary branches thus strike a deliberate balance between locality preservation and contextual awareness while introducing far fewer parameters and computations than typical multi-branch dilation pyramids.

To alleviate statistical discrepancies among multi-scale features and to construct compact and stable semantic representations, SPCP introduces normalization, activation, and channel-compression operations after feature concatenation:

$$F_{sc} = \mathrm{Conv}_{1\times1}\left(\delta(\mathrm{BN}(F_{ms}))\right) \tag{3.2}$$

where $\mathrm{BN}(\cdot)$ denotes batch normalization, $\delta(\cdot)$ is a nonlinear activation function, and $\mathrm{Conv}_{1\times1}$ represents a $1 \times 1$ convolution for channel reorganization and dimensionality reduction.

This design suppresses redundant channels while enabling effective fusion of multi-scale features. Considering that SAR ship targets exhibit salient structural characteristics but are highly susceptible to noise interference, SPCP further incorporates a channel attention mechanism after channel reorganization to emphasize critical structural regions and suppress irrelevant responses. Specifically, global average pooling is first applied to extract channel-wise statistics, followed by a lightweight multilayer perceptron composed of two linear transformations to model nonlinear inter-channel dependencies:

$$z = \mathrm{GAP}(F_{sc}), \tag{3.3}$$

$$a = \sigma\left(W_2\,\phi\left(W_1 z\right)\right) \tag{3.4}$$

where $W_1$ and $W_2$ denote channel reduction and expansion mappings with dimensions $C \to \frac{C}{r}$ and $\frac{C}{r} \to C$, respectively, and $r$ is the channel reduction ratio that balances modeling capacity and computational cost. $\phi(\cdot)$ and $\sigma(\cdot)$ denote the ReLU and sigmoid activation functions, respectively.

Finally, the channel-wise attention weights are applied to the feature map via channel-wise multiplication:

$$F_{\mathrm{out}} = F_{\mathrm{sc}} \odot a \tag{3.5}$$

where $\odot$ denotes channel-wise multiplication.

This attention mechanism adaptively strengthens responses corresponding to ship bodies and boundary structures while effectively suppressing complex sea clutter and speckle noise. As a result, the output features maintain semantic consistency while exhibiting clearer and more continuous structural representations, providing more precise spatial localization cues and structural priors for subsequent instance segmentation heads.

### *3.2. MSEF*

After multi-scale feature modeling in the downsampling stage, the backbone network produces features with strong semantic representation capability. However, it remains necessary to further

enhance the modeling of spatial structural information of ship targets, particularly for regions with complex contours, elongated boundaries, and susceptibility to noise interference. Since SAR ship instance segmentation relies not only on semantic category discrimination but also heavily on accurate recovery of object shapes and boundaries, high-level semantic features alone are insufficient to support precise instance-level delineation in subsequent segmentation heads. Therefore, it is essential to introduce a feature enhancement mechanism oriented toward structural information modeling in the backbone feature extraction stage, enabling structural enhancement and boundary-aware feature representation of the multi-scale features from the previous stage.

Based on the above considerations, we design the MSEF module after the SPCP module to achieve structural information enhancement and boundary-aware feature modeling. As shown in Figure 3, to incorporate this design into the backbone feature extraction process, we further construct a C3K2-MSEF structure with MSEF as the core unit. The MSEF module takes multi-scale edge-aware modeling as its core principle. Specifically, it first performs multi-scale feature modeling to capture contextual information under different receptive fields, thereby characterizing the structural responses of ship targets across scales. On this basis, an edge-aware mechanism is further introduced to explicitly enhance high-frequency boundary information in the multi-scale features, improving the modeling of object contours and spatial continuity. While preserving the integrity of the backbone semantic representation, the module aggregates edge-enhanced features across scales, thereby strengthening the representation of boundary location, orientation, and structural consistency, and providing clearer, more stable, and discriminative structural priors for subsequent stages.

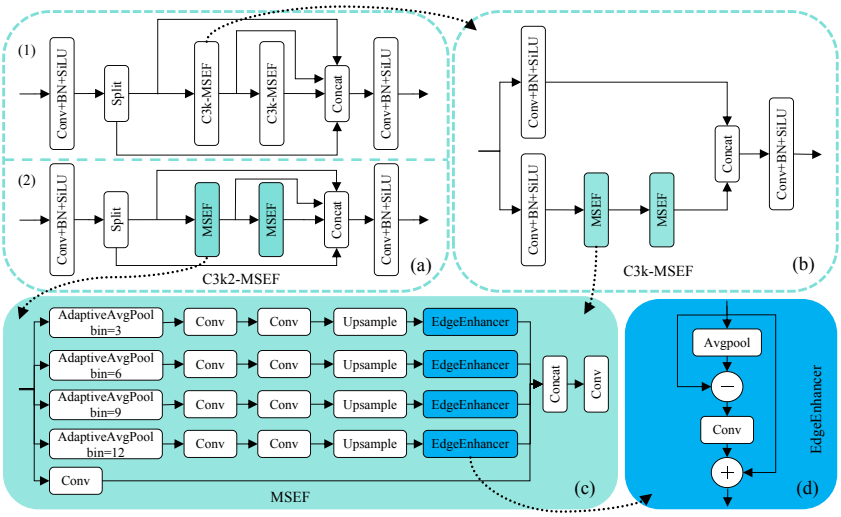


**Figure 3.** Hierarchical architecture of the C3K2-MSEF module. (a) Overall structure of the C3K2-MSEF module; (b) architecture of the C3K-MSEF module; (c) architecture of the MSEF module; (d) structure of the EdgeEnhancer unit. Specifically, Figures 3(a) and (b) illustrate the C3K2-MSEF and C3K-MSEF structures with MSEF as the core unit. In Figure 3(a), two internal structural variants are designed for different backbone stages: Structure (1) is applied at the P4/16 and P5/32 stages, corresponding to feature maps with 512 and 1024 channels, respectively; Structure (2) is applied at the P2/4 and P3/8 stages, corresponding to feature maps with 256 and 512 channels, respectively. Figures 3(c) and (d) further illustrate the proposed MSEF module and its key component, the EdgeEnhancer unit.

Following the above design, the implementation details of the MSEF module are described as follows. Let the input feature be denoted as $X \in \mathbb{R}^{H\times W\times C}$. First, to preserve local semantic information and spatial texture, a local convolutional branch is applied to encode the input feature:

$$X_0 = F_{\text{loc}}(X) \tag{3.6}$$

where $F_{\text{loc}}(\cdot)$ denotes a $3 \times 3$ convolution that provides a stable base-level semantic representation.

Next, the MSEF module introduces multi-scale adaptive average pooling to sample input features at different spatial resolutions, thereby enabling the capture of ship contour responses under varying contextual scales. In our implementation, the pooling scale set is fixed as $\mathcal{B} = \{3, 6, 9, 12\}$, corresponding to four multi-scale branches with progressively expanding contextual coverage. Together with the local convolutional branch $X_0$ in Eq (3.6), these scales construct a contour-aware pyramid that spans from compact ship-body cues at scale $b_1 = 3$ to broader hull-and-neighborhood structures at scale $b_4 = 12$, matching the schematic representation in Figure 3(c). This explicit pyramidal coverage is what enables MSEF to characterize the structural responses of SAR ship targets across heterogeneous scales. For the $k$-th scale $b_k \in \mathcal{B}$, the corresponding multi-scale feature representation is defined as:

$$X^{(k)} = \mathcal{U}\left(F_{\text{enc}}^{(k)}\left(\text{AAP}_{b_k}(X)\right)\right) \tag{3.7}$$

where $\text{AAP}_{b_k}(\cdot)$ denotes adaptive average pooling with output size $b_k \times b_k$, $F_{\text{enc}}^{(k)}(\cdot)$ consists of a $1 \times 1$ convolution followed by a depthwise separable $3 \times 3$ convolution for feature compression and scale-aware semantic encoding, and $\mathcal{U}(\cdot)$ represents bilinear upsampling to restore a unified spatial resolution.

After scale-aligned multi-scale features are obtained, an edge enhancement unit is introduced to explicitly model high-frequency structural information. Specifically, the unit first extracts low-frequency background information via local smoothing and computes a difference with the original feature to highlight high-frequency edge responses:

$$E^{(k)} = \sigma\left(F_{\text{edge}}\left(X^{(k)} - P\left(X^{(k)}\right)\right)\right) \tag{3.8}$$

where $P(\cdot)$ denotes a $3 \times 3$ average pooling operation for low-frequency smoothing, $F_{\text{edge}}(\cdot)$ represents a convolutional mapping, and $\sigma(\cdot)$ denotes the sigmoid activation function.

It is important to emphasize that the high-pass difference $X^{(k)} - P(X^{(k)})$ is computed on multi-scale encoded feature maps rather than on raw SAR intensities. After the adaptive average pooling, the $1 \times 1$ channel projection, and the depthwise separable $3 \times 3$ encoding in Eq (3.7), isolated pixel-level speckle responses have already been substantially smoothed by spatial statistics, and what remains is a semantically organized representation in which true boundary signals dominate over residual noise. Consequently, the subsequent learnable mapping $F_{\text{edge}}(\cdot)$ is required only to refine structured edge cues rather than to filter raw speckle, and the bounded sigmoid activation $\sigma(\cdot)$ further restricts the edge response to $[0, 1]$, preventing any unbounded amplification of stray high-frequency activations. This is precisely the reason why MSEF deliberately avoids handcrafted denoising or fixed-threshold filtering: in SAR images, the amplitudes of weak vessel boundaries and speckle responses are often statistically close, so any hard threshold risks eliminating legitimate small-vessel structures together with noise. By relying instead on the joint action of multi-scale pooling, learnable filtering, bounded activation, and the post-fusion convolution $F_{\text{fuse}}(\cdot)$, MSEF achieves data-driven noise suppression that adapts to the actual signal distribution at each spatial scale.

The amplitude modulation implicitly implemented within the convolutional mapping suppresses abnormal high-frequency activations induced by noise while preserving meaningful edge responses. The enhanced edge information is then injected into the original feature in a residual manner:

$$\hat{X}^{(k)} = X^{(k)} + E^{(k)}. \tag{3.9}$$

Finally, the local backbone feature and edge-enhanced features across all scales are concatenated along the channel dimension and fused through a convolutional operation:

$$Y = F_{\text{fuse}}\left(\text{Concat}\left(X_0, \hat{X}^{(1)}, \ldots, \hat{X}^{(K)}\right)\right) \tag{3.10}$$

where $F_{\text{fuse}}(\cdot)$ denotes a convolutional fusion operator that suppresses spurious edge responses introduced during multi-scale fusion and enhances overall structural consistency and discriminability.

Through this design, MSEF stably extracts ship contours and boundary structures across multiple spatial scales. While preserving semantic information, it significantly improves boundary clarity and continuity, providing more precise spatial localization cues and structural priors for subsequent instance segmentation heads, and effectively enhancing boundary accuracy and shape completeness in SAR ship instance segmentation.

### *3.3. ATU*

After high-level feature extraction and long-range dependency modeling in the backbone network, deep semantic features exhibit strong discriminative capability. However, under SAR imaging conditions, these features remain vulnerable to speckle noise and strong scattering points, which can induce abnormal amplitude stretching and localized activation accumulation. Such effects weaken the consistency of structural and boundary modeling in instance segmentation. To address this issue, an adaptive amplitude modulation structure, termed the ATU, is introduced and embedded between the attention output and the subsequent feed-forward mapping to explicitly reconstruct feature amplitudes and suppress anomalous responses in deep semantic representations.

Let the attention-enhanced input feature be $X \in \mathbb{R}^{H\times W\times C}$. The ATU performs nonlinear amplitude modulation using a parameterized hyperbolic tangent function, formulated as

$$Y = \gamma \odot \tanh(\alpha X) + \beta \tag{3.11}$$

where $\alpha \in \mathbb{R}$ is a learnable global scaling parameter controlling the saturation range of the nonlinear mapping, and $\gamma, \beta \in \mathbb{R}^C$ denote channel-wise scaling and bias parameters, respectively. The operator $\odot$ represents channel-wise multiplication.

Owing to the bounded activation property of the $\tanh(\cdot)$ function, the ATU preserves the continuity of low- and medium-amplitude semantic responses while effectively suppressing high-amplitude anomalous activations caused by strong scatterers and speckle noise in SAR images. This mitigates noise-dominated amplification in deep feature representations.

Furthermore, the channel-wise parameters $\gamma$ and $\beta$ enable adaptive modulation of amplitude distributions across different semantic channels. This allows channels associated with object structures, boundaries, and ship bodies to maintain stable response ranges. The reason for adopting this formulation in the SAR setting is that speckle and strong-scatterer responses appear in deep features as channel-heterogeneous high-amplitude outliers rather than as a uniform amplitude shift. The learnable

scalar $\alpha$ inside the tanh function directly controls the saturation interval of the nonlinear mapping, so the network can actively push these tail responses into the saturated region and obtain a soft amplitude ceiling without sacrificing low- and medium-amplitude structural information that encodes ship contours and hull textures. The channel-wise parameters $\gamma$ and $\beta$ then redistribute amplitude budgets on a per-channel basis, allowing ship-body channels, boundary channels, and clutter-affected channels to retain response ranges appropriate to their semantic role. Because $\alpha$, $\gamma$, and $\beta$ are all parametric, this amplitude stabilization remains consistent between training and inference, and across the highly variable statistics of nearshore and offshore SAR imagery, which is essential for SAR ship instance segmentation where scene statistics vary substantially.

Unlike conventional attention mechanisms that rely solely on weight reallocation, the ATU operates directly on the amplitude distribution, facilitating more robust semantic representation learning. Without introducing significant computational overhead, it achieves an effective balance between feature fidelity and noise suppression.

When combined with attention mechanisms, the ATU further enhances the stability, structural consistency, and generalization capability of deep features for SAR ship instance segmentation, thereby providing a more reliable foundation for precise instance region recovery. The architecture of the ATU module is illustrated in Figure 4.

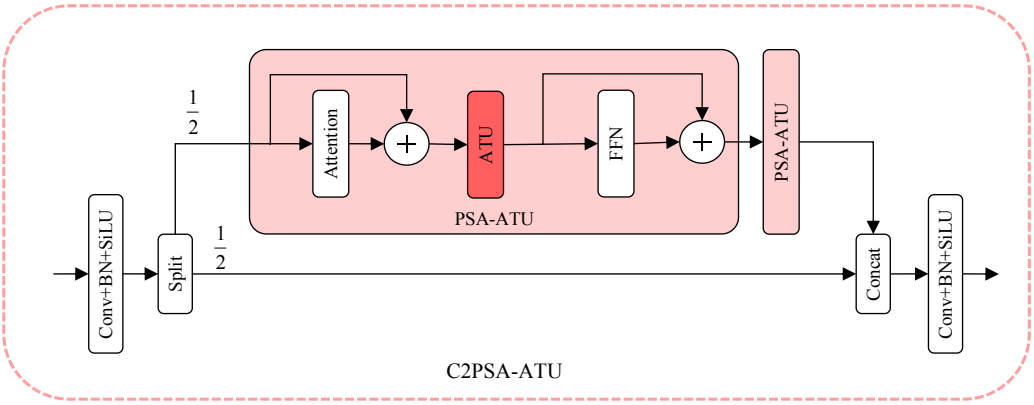


**Figure 4.** Structure of the C2PSA-ATU module. The ATU module is embedded between the attention block and the feed-forward network. Through a parameterized adaptive nonlinear transformation, it performs channel-wise recalibration of attention outputs, suppressing impulsive SAR noise while enhancing ship-related features with negligible computational overhead.

### *3.4. MSCD*

After integrating attention enhancement and feature modulation mechanisms, the feature flow enters the deep semantic modeling stage within the neck structure. This stage faces two fundamental challenges.

First, SAR ship targets exhibit highly non-uniform scale distributions and spatial layouts: large vessels in nearshore regions often show scale redundancy, whereas distant small targets suffer from low spatial resolution and weak scattering responses. Such scale heterogeneity makes it difficult for single-scale convolutions to accommodate cross-scale structural variations.

Second, although deep semantic features possess strong discriminative power, their relatively low spatial sampling density often leads to the loss of local structural details, thereby degrading boundary consistency and shape completeness in instance segmentation.

To address these challenges, the MSCD module is introduced into the neck architecture. The core idea of MSCD is to construct scale-decoupled feature subspaces along the channel dimension, enabling explicit multi-scale representation learning.

As illustrated in Figure 5, given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, the module first splits the feature along the channel dimension into two parts: an identity branch $X_{\text{cheap}}$, which preserves the original semantic information, and a set of scale-aware subspaces $X_{\text{scale}}$ dedicated to multi-scale modeling.

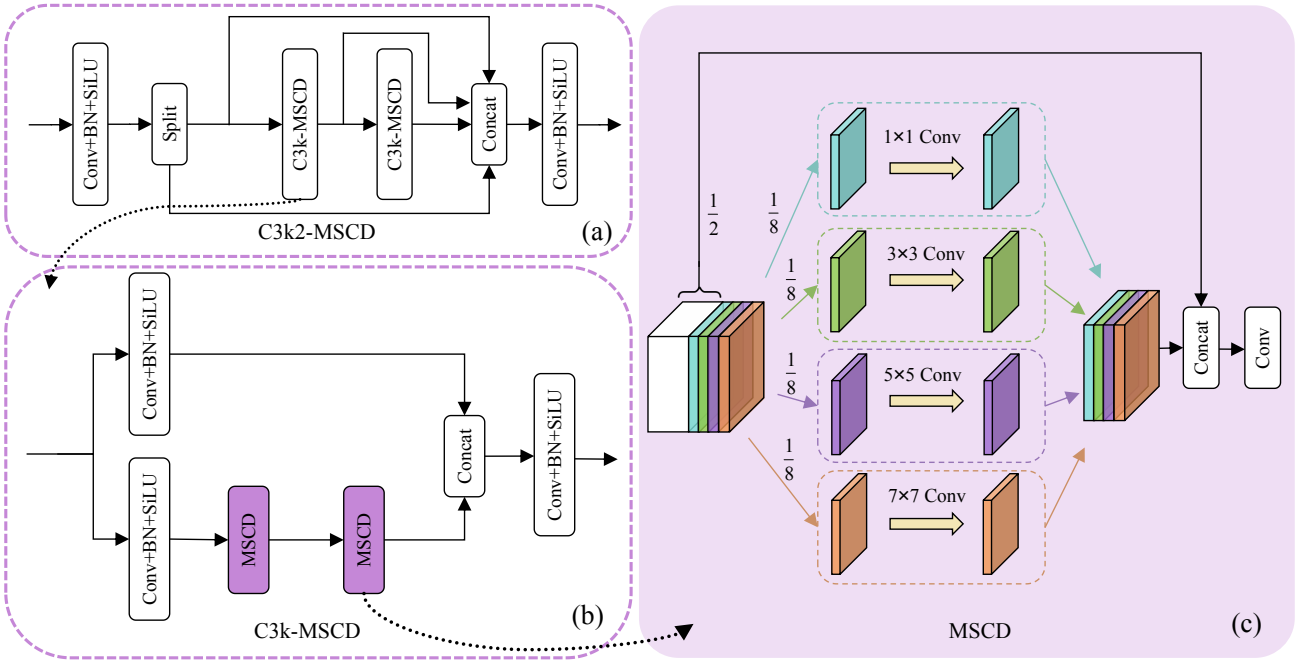


**Figure 5.** Hierarchical structure of the C3k2-MSCD module. (a) Overall structure of the C3k2-MSCD module; (b) C3k-MSCD substructure; (c) MSCD module architecture.

Specifically, the channels allocated to $X_{\text{scale}}$ are evenly divided into $G$ groups, where each group contains $C_s = C_{\text{scale}}/G$ channels. Each group is processed by a distinct convolutional branch with different receptive fields, enabling the model to capture scale-specific responses.

For each scale-specific subspace, a convolution operator with kernel size $k_g$ is applied to construct scale-aware responses. The input feature is decomposed as

$$X_{\text{scale}} = \{X_{\text{scale}}^{(g)}\}_{g=1}^{G} \tag{3.12}$$

where $X_{\text{scale}}^{(g)}$ denotes the $g$-th channel group. Each group is processed independently as

$$Y_g = F_{k_g}\left(X_{\text{scale}}^{(g)}\right), \quad g = 1, \ldots, G \tag{3.13}$$

where $F_{k_g}(\cdot)$ denotes a convolution operator with kernel size $k_g$.

The outputs of all branches are then concatenated along the channel dimension to form a unified scale-aware representation:

$$Y_{\text{scale}} = \text{Concat}(Y_1, Y_2, \ldots, Y_G). \tag{3.14}$$

Different kernel sizes correspond to different receptive fields, enabling the model to capture local contour details, structural textures, and broader morphological patterns across multiple scales.

Finally, the scale-aware features are fused with the identity branch features and passed through a $1 \times 1$ convolution for feature aggregation and channel mixing.

This design provides high flexibility in scale-aware feature representation without increasing network depth, effectively balancing cross-scale semantic consistency and local structural fidelity. Compared with conventional stacked multi-scale convolutions, MSCD achieves efficient scale-decoupled modeling via channel-wise multi-branch expansion.

As a result, it enhances structure-aware representation capability and boundary precision in SAR ship instance segmentation while maintaining low parameter complexity and computational overhead.

## 4. Experiments

### *4.1. Datasets*

To evaluate the performance of the proposed method, experiments are conducted on two publicly available SAR ship instance segmentation datasets: HRSID [25] and PSeg-SSDD [26]. The HRSID dataset consists of 5604 high-resolution SAR images with a total of 16951 annotated ship instances. The dataset is constructed by cropping 136 original SAR images with spatial resolutions finer than 5 m into image patches of uniform-size 800×800 pixels, enabling standardized training and evaluation. HRSID covers diverse maritime scenarios and ship configurations, providing a comprehensive benchmark for SAR ship instance segmentation. The PSeg-SSDD dataset is derived from the SSDD dataset originally released by Li et al. [27] in 2017 and subsequently refined and enhanced by Zhang et al.. It contains 1160 SAR images with annotations for 2456 ship instances of varying sizes. The image were collected from multiple SAR sensors, including Sentinel-1, TerraSAR-X, and RadarSat-2, encompassing a wide range of imaging conditions and acquisition geometries.

### *4.2. Experimental setup*

All experiments are conducted on a Linux-based system equipped with an NVIDIA RTX 4090 GPU and 24 GB of memory. The proposed method is implemented in Python 3.8 using the PyTorch 2.0.0 deep learning framework, with CUDA 11.8 for parallel computation. During training, the batch size is set to 32, and stochastic gradient descent (SGD) is employed as the optimizer. To mitigate overfitting, an early stopping strategy is adopted with a patience of 30 epochs. In addition, mosaic data augmentation is applied during training to enhance data diversity and improve the generalization capability of the model.

### *4.3. Evaluation metrics*

In this study, average precision (AP) is adopted as the primary evaluation metric. AP is computed based on the precision–recall (PR) curve and corresponds to the approximate area under this curve. Precision measures the correctness of model predictions, while recall reflects the model's ability to

cover ground-truth instances. They are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

where $TP$ denotes the number of true positives correctly predicted by the model, $FP$ represents the number of false positives, and $FN$ indicates the number of false negatives corresponding to missed ground-truth instances.

Following the standard common objects in context (COCO) evaluation protocol, the AP metric reported in this work is computed by averaging AP values over multiple intersection over union (IoU) thresholds ranging from 0.50 to 0.95 with a step size of 0.05:

$$AP = \frac{1}{|T|} \sum_{t \in T} AP_t, \quad T = \{0.50, 0.55, \ldots, 0.95\}. \tag{4.3}$$

To provide a more comprehensive performance analysis, $AP_{50}$ and $AP_{75}$ are also reported, corresponding to IoU thresholds of 0.50 and 0.75, respectively. In addition, scale-specific metrics, including $AP_S$, $AP_M$, and $AP_L$, are evaluated for small-, medium-, and large-scale targets.

Together, these metrics assess instance segmentation performance from multiple perspectives, including precision–recall balance, localization accuracy, and multi-scale adaptability.

### *4.4. Results and discussion*

#### 4.4.1. Ablation studies

To validate the effectiveness of the proposed components, systematic ablation experiments are conducted on the HRSID dataset. As reported in Table 1, the introduction of each module yields consistent performance improvements over the baseline. Specifically, Method A corresponds to the SPCP module, Method B to the MSEF module, Method C to the ATU module, and Method D to the MSCD module.

The standalone results in Table 1 reveal the distinct functional role of each component. When applied alone, the SPCP module raises $AP$ from 57.8% to 59.8%, and its largest gains are concentrated on small targets, where $AP_S$ improves from 55.1% to 57.6%, as well as on the strict-IoU metric $AP_{75}$, which rises from 73.7% to 75.4%. This pattern is consistent with the design intent of SPCP, which preserves local boundary cues for compact targets while supplying moderate contextual awareness for densely arranged vessels. The MSEF module yields the strongest standalone improvement, lifting $AP$ to 60.2%, and produces a clear rise in $AP_L$ from 21.2% to 24.8%, indicating that explicit boundary-aware modeling is most beneficial for large nearshore vessels whose contours are disproportionately affected by speckle interference. The ATU module produces a more moderate $AP$ gain, reaching 59.0%, but a notable improvement on $AP_L$, which rises from 21.2% to 23.0%. This is consistent with its role as an amplitude-stabilization unit that acts most visibly on high-energy responses caused by strong scatterers and large hull structures. Finally, the MSCD module delivers a balanced gain across all scales, with $AP$, $AP_S$, $AP_M$, and $AP_L$ reaching 59.2%, 56.8%, 66.9%, and 23.3%, respectively, in line with its scale-decoupled multi-branch construction. When all four modules are

jointly applied, the gains are not merely additive: the full model reaches 63.7% $AP$ and improves $AP_L$ by 8.3 percentage points over the baseline, demonstrating that the four components are functionally complementary—SPCP and MSEF strengthen structural and boundary representation in the backbone, while the ATU and MSCD stabilize and diversify the deep semantic features used by the segmentation head.

**Table 1.** Ablation study results on the HRSID dataset.

| Method | Segmentation task (%) | | | | | |
|---|---|---|---|---|---|---|
| | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | $AP$ |
| YOLOv11s | 89.1 | 73.7 | 55.1 | 65.6 | 21.2 | 57.8 |
| +A | 90.5 | 75.4 | 57.6 | 67.3 | 24.0 | 59.8 |
| +B | 91.0 | 75.8 | 58.4 | 67.9 | 24.8 | 60.2 |
| +C | 89.9 | 74.6 | 56.5 | 66.7 | 23.0 | 59.0 |
| +D | 90.1 | 74.8 | 56.8 | 66.9 | 23.3 | 59.2 |
| +A+B | 92.1 | 77.0 | 60.3 | 69.1 | 26.9 | 61.9 |
| +A+B+C | 92.8 | 77.9 | 61.5 | 70.2 | 28.5 | 62.9 |
| +A+B+C+D | 93.4 | 78.4 | 62.4 | 70.8 | 29.5 | 63.7 |

To enhance the model's capability in capturing features across different scales, the SPCP module is first introduced, where multi-dilated convolutions effectively expand the receptive field and improve multi-scale representation. To address the blurred boundaries and ambiguous contours commonly observed in SAR ship images, the MSEF module is subsequently incorporated to strengthen edge-aware structural modeling. To further suppress the intrinsic speckle noise in SAR images, the ATU module is employed to perform adaptive feature recalibration, thereby improving robustness to noise interference. Finally, to alleviate the limitation of single-scale expressiveness in high-dimensional feature representations, the MSCD module is introduced, which enriches feature diversity through the collaborative use of grouped and multi-scale convolutions.

Experimental results demonstrate that, compared with the baseline YOLOv11s model, the proposed method achieves consistent and notable improvements across multiple evaluation metrics, including $AP$, $AP_{50}$, $AP_{75}$, as well as $AP_S$, $AP_M$, and $AP_L$. These results confirm both the rationality of the proposed design and the effectiveness of each individual module in improving SAR ship instance segmentation performance.

#### 4.4.2. Comparison with the baseline model

On both the HRSID and PSeg-SSDD datasets, a visual comparison between the proposed YOLO-Argus and the baseline YOLOv11s is conducted, as illustrated in Figure 6. From the instance segmentation results, YOLO-Argus demonstrates clear advantages in nearshore scenes with complex backgrounds. It accurately identifies true ship targets while effectively suppressing false detections caused by shoreline structures and port facilities. In offshore scenarios, YOLO-Argus also maintains a high detection rate for small-scale targets, significantly reducing missed detections compared with the baseline model.

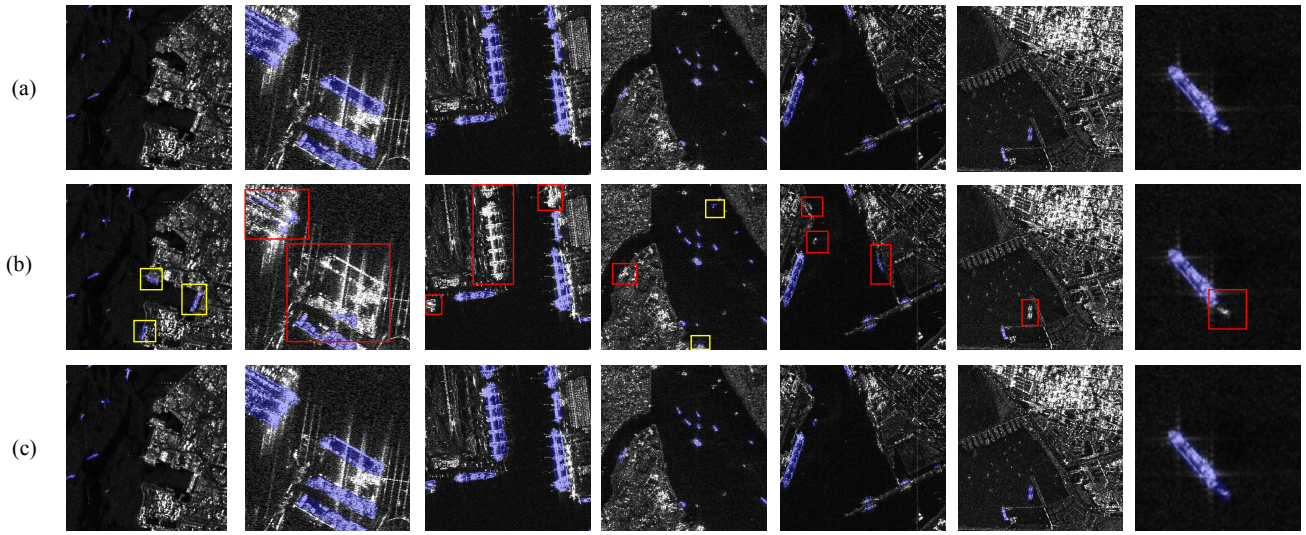


**Figure 6.** Visual comparison of ground truth, the baseline model, and YOLO-Argus. (a) Ground truth; (b) baseline; (c) YOLO-Argus. The blue masks indicate the segmentation results, while the red boxes highlight missed detections of the baseline model, and the yellow boxes denote its false positives.

To further investigate the effectiveness of the proposed structured feature modeling strategy, a visualization analysis based on Grad-CAM++ [28] is performed to compare the feature attention patterns of YOLO-Argus and the baseline model. The results are shown in Figure 7. The heatmap comparison reveals that YOLO-Argus produces attention responses that are highly concentrated on ship bodies and boundary regions, rather than diffusely spread over broader areas. This behavior is consistent with the intended effects of the SPCP multi-scale contextual modeling and the C3k2-MSEF edge fusion mechanism. In contrast, the baseline model exhibits more scattered and localized responses within target regions, with noticeable attention leakage at ship–background boundaries. These observations indicate that the proposed model effectively strengthens the representation of ship contours and boundary details through structured feature modeling, enabling the network not only to detect the presence of targets but also to focus more accurately on their true geometric structures, resulting in more complete contour recovery in the segmentation outputs.

Quantitative results reported in Table 2 further corroborate these observations. YOLO-Argus consistently outperforms the baseline model across multiple segmentation accuracy metrics, including $AP$ and $AP_{50}$, validating the effectiveness of the proposed module combination for SAR ship instance segmentation. From a cross-dataset perspective, the overall performance on the PSeg-SSDD dataset is higher than that on HRSID. This performance gap can be attributed to intrinsic differences between the two datasets. As a larger and more challenging benchmark, HRSID contains a substantial number of samples from nearshore environments with strong background interference and pronounced scale variation, posing higher demands on model robustness and generalization. In contrast, PSeg-SSDD features relatively more standardized scenes with more concentrated target distributions, making it easier for models to achieve higher segmentation accuracy. Despite these differences, YOLO-Argus maintains leading performance on both datasets, demonstrating strong adaptability across varying imaging conditions and scene complexities.

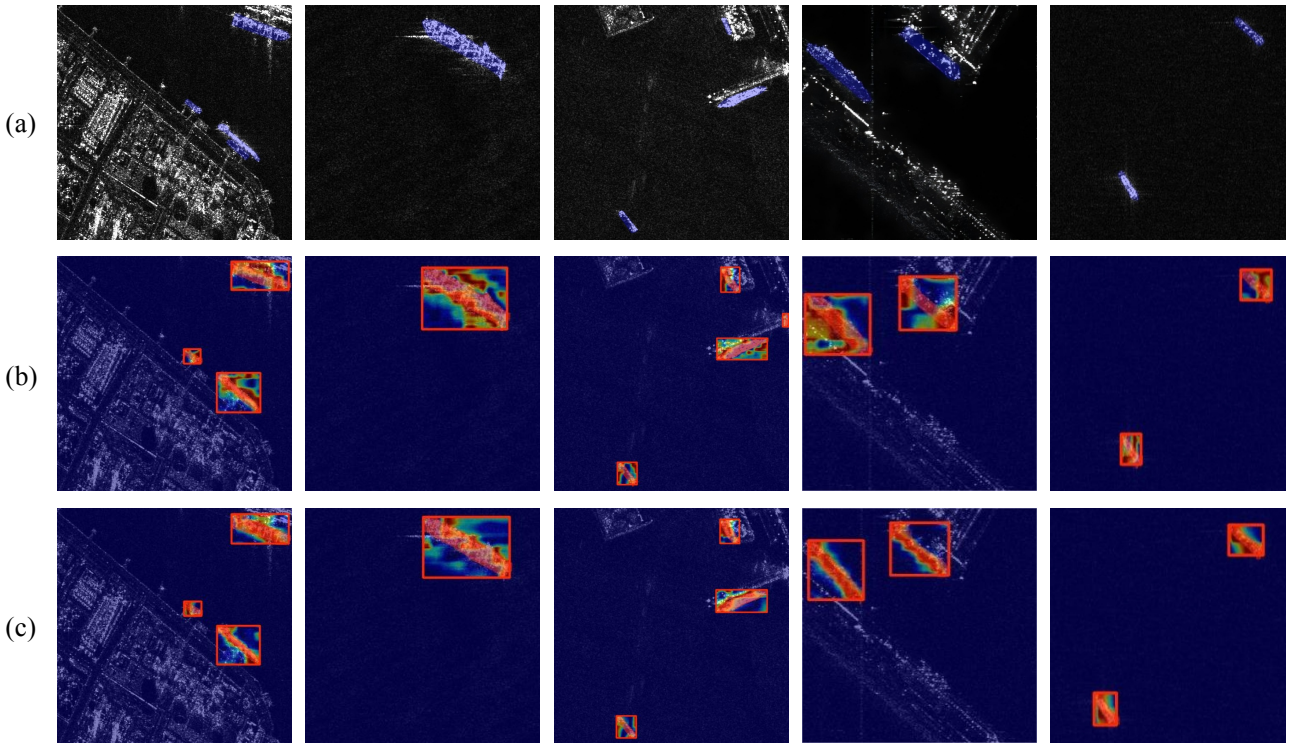


**Figure 7.** Comparison of heatmap visualizations. (a) Ground truth; (b) baseline; (c) YOLO-Argus. For clarity, only the heatmap responses corresponding to ship targets are retained, while the background regions are suppressed.

**Table 2.** Quantitative comparison with the baseline on HRSID and PSeg-SSDD.

| Dataset | Method | Segmentation task (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | $AP$ |
| HRSID | YOLOv11s | 89.1 | 73.7 | 55.1 | 65.6 | 21.2 | 57.8 |
| | YOLO-Argus | 93.4(+4.3) | 78.4(+4.7) | 62.4(+7.3) | 70.8(+5.2) | 29.5(+8.3) | 63.7(+5.9) |
| PSeg-SSDD | YOLOv11s | 92.3 | 76.9 | 60.7 | 64.1 | 64.2 | 58.3 |
| | YOLO-Argus | 95.8(+3.5) | 86.1(+9.2) | 66.3(+5.6) | 71.5(+7.4) | 76.5(+12.3) | 67.5(+9.2) |

#### 4.4.3. Comparison with other methods

To comprehensively evaluate the performance of the proposed approach, comparative experiments are conducted on both the HRSID and PSeg-SSDD datasets. To clarify the positioning of the proposed method against different categories of competitors, the compared methods are organized into two groups: general-purpose instance segmentation models and SAR ship-specific methods. As reported in Table 3, YOLO-Argus achieves the best performance across all key evaluation metrics ($AP$, $AP_{50}$, and $AP_{75}$) on both datasets.

In the comparison with general-purpose instance segmentation models, including the YOLO family, RTMDet, and recent remote-sensing-oriented frameworks such as CATNet and RSPrompter, YOLO-Argus consistently outperforms them on every metric, demonstrating its competitiveness against

mainstream state-of-the-art instance segmentation frameworks. Among general-purpose competitors, RSPrompter is the strongest baseline on both datasets but is still surpassed by the proposed method by 0.8% *AP* on HRSID and 0.2% *AP* on PSeg-SSDD.

**Table 3.** Comparison of different methods on the HRSID and PSeg-SSDD datasets (%).

| Method | HRSID | | | | | | PSeg-SSDD | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | $AP$ | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | $AP$ |
| *General-purpose instance segmentation methods* | | | | | | | | | | | | |
| YOLACT [29] | 71.1 | 41.9 | 39.5 | 46.1 | 7.3 | 39.6 | 88.0 | 52.1 | 47.3 | 53.5 | 40.2 | 48.4 |
| PANet [30] | 86.0 | 66.2 | 54.7 | 62.8 | 17.8 | 55.1 | 91.1 | 74.0 | 59.3 | 61.0 | 52.1 | 59.6 |
| NB [31] | 80.3 | 57.1 | 48.0 | 59.4 | 19.9 | 48.8 | 87.4 | 67.2 | 56.6 | 53.0 | 28.2 | 54.1 |
| RTMDet [32] | 93.1 | 75.2 | 61.6 | 70.7 | 29.1 | 62.3 | 94.7 | 66.2 | 52.5 | 69.4 | 64.0 | 56.5 |
| GR R-CNN [33] | 90.5 | 71.6 | 59.1 | 65.3 | 25.4 | 59.6 | 92.8 | 75.4 | 62.8 | 63.5 | 67.2 | 61.3 |
| YOLOv8 | 91.2 | 75.3 | 55.7 | 67.4 | 23.6 | 59.2 | 93.2 | 77.8 | 61.5 | 64.3 | 65.7 | 59.8 |
| CATNet [8] | 91.4 | 74.2 | 60.5 | 65.6 | 25.5 | 59.6 | 93.8 | 79.6 | 62.8 | 64.7 | 63.5 | 62.1 |
| RSPrompter [21] | 93.2 | 77.6 | 61.9 | 68.6 | 28.7 | 62.9 | 95.6 | 84.3 | 65.3 | 69.2 | 72.6 | 67.3 |
| *SAR ship-specific instance segmentation methods* | | | | | | | | | | | | |
| GGSE-SSIS [13] | 80.6 | 58.8 | 49.1 | 60.7 | 23.5 | 50.2 | 88.4 | 69.6 | 56.9 | 54.8 | 33.2 | 55.8 |
| AFSS-Inst [15] | 88.0 | 72.9 | 59.9 | 55.9 | 8.9 | 58.2 | 95.1 | 75.7 | 59.7 | 69.6 | 55.3 | 61.8 |
| GCBANet [17] | 88.6 | 68.9 | 57.0 | 64.3 | 25.9 | 57.3 | 93.5 | 78.8 | 63.2 | 63.0 | 55.1 | 63.1 |
| SRNet [19] | 92.1 | 77.5 | 62.1 | 70.6 | 19.6 | 62.3 | 93.6 | 75.7 | 60.0 | 68.0 | 68.0 | 61.7 |
| FL-CSE-ROIE [23] | 88.6 | 69.5 | 57.3 | 65.7 | 26.1 | 57.9 | 93.7 | 78.3 | 63.3 | 61.2 | 75.0 | 62.6 |
| SISNet [34] | 90.2 | 70.1 | 58.5 | 65.4 | 28.0 | 58.9 | 94.6 | 80.9 | 64.5 | 67.4 | 62.5 | 65.1 |
| HTC+ [35] | 90.3 | 71.0 | 58.7 | 65.7 | 26.8 | 59.1 | 94.7 | 82.3 | 64.1 | 64.9 | 65.0 | 64.3 |
| YOLO-Argus (Ours) | **93.4** | **78.4** | **62.4** | **70.8** | **29.5** | **63.7** | **95.8** | **86.1** | **66.3** | **71.5** | **76.5** | **67.5** |

In the comparison with SAR ship-specific methods, including SISNet, SRNet, HTC+, GCBANet, and FL-CSE-ROIE, YOLO-Argus also achieves the best results on both datasets. Compared with the most competitive SAR-specific baseline, YOLO-Argus improves *AP* by 1.4% on HRSID and by 2.4% on PSeg-SSDD. Notably, the proposed method shows clear advantages in $AP_S$ and $AP_L$, indicating its robustness to the extreme scale variation of SAR ship targets and its capability to preserve target boundary integrity in densely distributed nearshore environments.

These results demonstrate that, by combining context aggregation, multi-scale edge enhancement, amplitude stabilization, and scale-decoupled modeling, YOLO-Argus more effectively captures the intrinsic characteristics of ship targets in SAR images, leading to consistent improvements over both general-purpose and domain-specific competitors.

## 5. Conclusions

This work presents YOLO-Argus, a pixel-level ship instance segmentation model tailored for SAR images. To address key challenges inherent to SAR scenes—including boundary blurring caused by speckle noise, extreme scale variation of ship targets, dense mooring and instance adhesion, and

complex nearshore backgrounds—we introduce four cooperative modules: SPCP, MSEF, the ATU, and MSCD. SPCP expands the receptive field through parallel multi-dilated sampling while preserving fine-grained information; MSEF explicitly couples multi-scale contour modeling with differential edge enhancement to achieve accurate boundary localization under strong noise interference; the ATU incorporates a parameterized adaptive modulation unit that suppresses impulsive speckle responses via learnable nonlinear compression with negligible computational overhead; and MSCD employs grouped multi-scale convolutions to overcome the homogeneity limitation of serial designs, maintaining feature diversity across targets ranging from small nearshore vessels to large offshore cargo ships. Benefiting from the above design, YOLO-Argus demonstrates superior segmentation accuracy and robustness in complex marine scenarios, exhibiting strong potential for practical deployment. The proposed method can be applied to ship identification and behavior monitoring in maritime surveillance and intelligent shipping systems, and is also suitable for dense target analysis in port areas as well as maritime target monitoring under complex backgrounds. Beyond these scenarios, the method can be further promoted to industries and fields such as illegal fishing detection, maritime emergency search-and-rescue, coastal security and border monitoring, marine environmental protection, and offshore infrastructure surveillance, particularly in all-weather and day-and-night situations where optical sensors are limited by cloud cover, illumination, or harsh sea-state conditions.

Extensive experiments on the public HRSID and PSeg-SSDD datasets demonstrate that YOLO-Argus consistently achieves the best performance. Ablation studies further confirm the complementary contributions of each module. However, in a few extremely complex nearshore scenarios in HRSID, the model still shows limited recall for very small vessels. We attribute this residual limitation primarily to two coupled factors rather than to the failure of any single module. The first and dominant factor is the irreversible loss of spatial detail introduced by the multi-stage downsampling pipeline. The segmentation head of YOLO-Argus operates on $P3/8$, $P4/16$, and $P5/32$ features, and for vessels that originally span only a handful of pixels, their geometric footprint may shrink to a sub-pixel region by the time the signal reaches $P3/8$. MSEF can amplify existing boundary responses, but it cannot reconstruct contours that have been numerically erased during early downsampling. The second factor is the inherently low target-to-background separability in highly cluttered nearshore scenes: docks, coastlines, and waterfront structures themselves produce strong high-frequency responses that compete with vessel boundaries, and the ATU's amplitude stabilization, while effective against speckle and strong scatterers, cannot push a target above the noise floor when its scattering signature is intrinsically close to that of the background. We therefore view early spatial detail loss together with low target-background contrast as the dominant bottlenecks, while limited edge sensitivity at sub-pixel scales acts as a secondary contributor. This understanding directly motivates our future work on a higher-resolution $P2$ segmentation branch and small-object-specific supervision.

## Author contributions

The authors confirm contribution to the paper as follows: Qiming Li and Hao Yin: Conceptualization, Qiming Li, Hao Yin, and Daozheng Chen: methodology, Hao Yin: software, Qiming Li and Hao Yin: validation, Qiming Li and Hao Yin: formal analysis, Qiming Li, Hao Yin, and Daozheng Chen: investigation, Qiming Li and Hao Yin: resources, Hao Yin: data curation, Qiming Li and Hao Yin: writing—original draft preparation, Qiming Li, Hao Yin, and Daozheng

Chen: writing—review and editing, Hao Yin: visualization, Qiming Li: supervision, Qiming Li and Daozheng Chen: project administration, Qiming Li and Daozheng Chen: funding acquisition.

All authors reviewed the results and approved the final version of the manuscript.

## Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This research was funded by the Key Research and Development and Transformation Projects of the Xizang (Tibet) Autonomous Region Science and Technology Program (funder: the Department of Science and Technology of the Xizang (Tibet) Autonomous Region), funding (grant) number: XZ202401ZY0004.

## Conflict of interest

All authors declare no conflicts of interest in this paper.

## References

1. J. Zhang, H. Li, Z. Tang, Q. Lu, X. Zheng, J. Zhou, An improved quantum-inspired genetic algorithm for image multilevel thresholding segmentation, *Math. Probl. Eng.*, **2014** (2014), 295402. https://doi.org/10.1155/2014/295402

2. R. Storn, K. Price, Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, *J. Global Optim.*, **11** (1997), 341–359. https://doi.org/10.1023/A:1008202821328

3. W. Cao, J. Li, J. Liu, P. Zhang, Two improved segmentation algorithms for whole cardiac CT sequence images, *9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2016, 346–351. https://doi.org/10.1109/cisp-bmei.2016.7852734

4. L. Vincent, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Trans. Pattern Anal. Mach. Intell.*, **13** (1991), 583–598. https://doi.org/10.1109/34.87344

5. M. Kang, C. M. Ting, F. F. Ting, R. C. W. Phan, ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation, *Image Vis. Comput.*, **147** (2024), 105057. https://doi.org/10.1016/j.imavis.2024.105057

6. Z. Gu, H. Chen, Z. Xu, DiffusionInst: diffusion model for instance segmentation, *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, 2730–2734. https://doi.org/10.1109/icassp48485.2024.10447191

7. J. Zhao, Y. Wang, Y. Zhou, W. L. Du, R. Yao, A. El Saddik, GLFRNet: global-local feature refusion network for remote sensing image instance segmentation, *IEEE Trans. Geosci. Remote Sens.*, **63** (2025), 1–12. https://doi.org/10.1109/tgrs.2025.3530515

8. Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, C. W. Chen, Learning to aggregate multi-scale context for instance segmentation in remote sensing images, *IEEE Trans. Neural Netw. Learn. Syst.*, **36** (2024), 595–609. https://doi.org/10.1109/tnnls.2023.3336563

9. J. Xie, W. Li, X. Li, Z. Liu, Y. S. Ong, C. C. Loy, MosaicFusion: diffusion models as data augmenters for large vocabulary instance segmentation, *Int. J. Comput. Vis.*, **133** (2025), 1456–1475. https://doi.org/10.1007/s11263-024-02223-3

10. P. Guo, H. Huang, P. He, X. Liu, T. Xiao, W. Zhang, OpenVIS: open-vocabulary video instance segmentation, *Proceedings of the AAAI Conference on Artificial Intelligence*, **39** (2025), 3275–3283. https://doi.org/10.1609/aaai.v39i3.32338

11. J. Zhu, X. Chen, K. He, Y. LeCun, Z. Liu, Transformers without normalization, *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, 14901–14911. https://doi.org/10.1109/cvpr52734.2025.01388

12. T. Wu, S. Tang, R. Zhang, J. Cao, Y. Zhang, CGNet: a light-weight context guided network for semantic segmentation, *IEEE Trans. Image Process.*, **30** (2021), 1169–1179. https://doi.org/10.1109/tip.2020.3042065

13. M. Chen, T. Wang, C. Xu, J. Chen, E. Chen, Z. Pan, Gradient prior guidance and image adaptation enhancement for semi-supervised SAR ship instance segmentation, *IEEE Sensors J.*, **24** (2024), 36216–36229. https://doi.org/10.1109/jsen.2024.3467030

14. M. C. El Rai, J. H. Giraldo, M. Al-Saad, M. Darweech, T. Bouwmans, SemiSegSAR: a semi-supervised segmentation algorithm for ship SAR images, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/lgrs.2022.3185306

15. F. Gao, Y. Huo, J. Wang, A. Hussain, H. Zhou, Anchor-free SAR ship instance segmentation with centroid-distance based loss, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **14** (2021), 11352–11371. https://doi.org/10.1109/jstars.2021.3123784

16. Z. Sun, C. Meng, J. Cheng, Z. Zhang, S. Chang, A multi-scale feature pyramid network for detection and instance segmentation of marine ships in SAR images, *Remote Sens.*, **14** (2022), 6312. https://doi.org/10.3390/rs14246312

17. X. Ke, X. Zhang, T. Zhang, GCBANet: a global context boundary-aware network for SAR ship instance segmentation, *Remote Sens.*, **14** (2022), 2165. https://doi.org/10.3390/rs14092165

18. S. Wei, X. Zeng, H. Zhang, Z. Zhou, J. Shi, X. Zhang, LFG-Net: low-level feature guided network for precise ship instance segmentation in SAR images, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–17. https://doi.org/10.1109/tgrs.2022.3188677

19. X. Yang, Q. Zhang, Q. Dong, Z. Han, X. Luo, D. Wei, Ship instance segmentation based on rotated bounding boxes for SAR images, *Remote Sens.*, **15** (2023), 1324. https://doi.org/10.3390/rs15051324

20. M. Rahimi, S. Sharifian, SAMSAR: a modified SAM architecture for oceanic ship segmentation of satellite SAR images using CNN-based Cross-Fused Attention, *Expert Syst. Appl.*, **284** (2025), 127852. https://doi.org/10.1016/j.eswa.2025.127852

21. K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, et al., RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 1–17. https://doi.org/10.1109/tgrs.2024.3356074

22. X. Xu, X. Zhang, S. Wei, J. Shi, W. Zhang, T. Zhang, et al., DiffSARShipInst: diffusion model for ship instance segmentation from synthetic aperture radar imagery, *ISPRS J. Photogramm. Remote Sens.*, **223** (2025), 440–455. https://doi.org/10.1016/j.isprsjprs.2025.02.030

23. T. Zhang, X. Zhang, A full-level context squeeze-and-excitation ROI extractor for SAR ship instance segmentation, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/lgrs.2022.3166387

24. T. Zhang, X. Zhang, A mask attention interaction and scale enhancement network for SAR ship instance segmentation, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/lgrs.2022.3189961

25. S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, J. Shi, HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation, *IEEE Access*, **8** (2020), 120234–120254. https://doi.org/10.1109/access.2020.3005861

26. T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, et al., SAR ship detection dataset (SSDD): Official release and comprehensive data analysis, *Remote Sens.*, **13** (2021), 3690. https://doi.org/10.3390/rs13183690

27. J. Li, C. Qu, J. Shao, Ship detection in SAR images based on an improved faster R-CNN, *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, 2017, 1–6. https://doi.org/10.1109/bigsardata.2017.8124934

28. A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks, *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, 839–847. https://doi.org/10.1109/wacv.2018.00097

29. D. Bolya, C. Zhou, F. Xiao, Y. J. Lee, YOLACT: real-time instance segmentation, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, 9157–9165. https://doi.org/10.1109/iccv.2019.00925

30. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 8759–8768. https://doi.org/10.1109/cvpr.2018.00913

31. Z. Wang, Y. Li, S. Wang, Noisy boundaries: lemon or lemonade for semi-supervised instance segmentation, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 16805–16814. https://doi.org/10.1109/cvpr52688.2022.01632

32. C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, et al., RTMDet: an empirical study of designing real-time object detectors, *arXiv preprint*, 2022.

33. M. Yuan, H. Meng, J. Wu, S. Cai, Global recurrent Mask R-CNN: marine ship instance segmentation, *Comput. Graph.*, **126** (2025), 104112. https://doi.org/10.1016/j.cag.2024.104112

34. Z. Shao, X. Zhang, S. Wei, J. Shi, X. Ke, X. Xu, et al., Scale in scale for SAR ship instance segmentation, *Remote Sens.*, **15** (2023), 629. https://doi.org/10.3390/rs15030629

35. T. Zhang, X. Zhang, HTC+ for SAR ship instance segmentation, *Remote Sens.*, **14** (2022), 2395. https://doi.org/10.3390/rs14102395


AIMS Press