The Journal of Supercomputing (2024) 80:20793–20812
https://doi.org/10.1007/s11227-024-06237-z

SAR-ShipSwin: enhancing SAR ship detection with robustness in complex environment

Ji Tang[1] · Yonghao Han[2] · Yunting Xian[1]

Accepted: 15 May 2024 / Published online: 3 June 2024

Abstract

Contemporary synthetic aperture radar (SAR) image processing techniques face various challenges, particularly in ship detection, background noise reduction, and information preservation. To address these issues, this paper introduces a novel model we called SAR-ShipSwin, which combines the swin transformer and feature pyramid network as the backbone network structure, specifically designed for ship detection in SAR images. The backbone network optimizes computational efficiency and handles occlusion and overlap issues in SAR images successfully by introducing the improved window multi-head self-attention module. To further enhance recognition accuracy, we design the background modeling network, which efficiently identifies and eliminates complex background features. Additionally, we introduce the spatial intensity geometric pooling technique, a novel pooling strategy that preserves geometric and structural information of the original region of interest, significantly reducing information loss and distortion. Considering the diverse ship shapes in SAR images, we specially design the dynamic ship shape adaptive convolution module, which dynamically adjusts the shape of convolution kernels to better match the targets. The proposed model is validated on the SSDD and HRSID datasets, achieving state-of-the-art performance.

Keywords SAR · Ship detection · Transformer · Complex background

✉ Yunting Xian
xianyt@scut.edu.cn

Ji Tang
csjit@mail.scut.edu.cn

Yonghao Han
13606246857@163.com

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, Guangdong, People's Republic of China

[2] Taicang Maritime Safety Administration, Taicang 215400, Jiangsu, People's Republic of China

1 Introduction

Synthetic aperture radar (SAR) [1] has emerged as a pivotal technology in maritime surveillance, offering all-weather, all-day imaging capabilities with strong penetration through adverse conditions such as clouds and fog [2]. However, SAR images of ships exhibit varying scales, and the complex background in SAR images, including interference from sea surfaces, ground clutter, and speckle noise, often leads to false negatives and higher false alarms during the detection process [3]. Therefore, achieving accurate positioning and recognition of ship targets in SAR images holds promising applications.

Traditional SAR ship target detection algorithms primarily rely on contrast differences between targets and background clutter. These methods encompass techniques such as constant false alarm rate (CFAR) detection [4], template matching algorithms [5], and trace-based detection algorithms [6]. CFAR detection is one of the most commonly used techniques, and literature [7] has successfully balanced accuracy and speed in estimating CFAR parameters. Literature [8] introduced a bilateral CFAR algorithm for ship detection, reducing the impact of SAR image ambiguity and background clutter. These methods rely on handcrafted features, exhibit limited efficiency, poor generalization performance, and are unsuitable for complex detection scenarios. With the rapid development of deep learning in optical image object detection and recognition [9–14], ideas like single-stage [15], two-stage [16], anchor-free [17], and Transformers [18] have been condensed in the field of object detection, and deep learning concepts have started to be applied to SAR images, yielding significant results [19, 20]. Chen et al. [21] employed deformable convolutional neural networks to enhance feature extraction by altering the convolutional kernel's sampling points. Bai et al. [22] proposed a shallow feature enhancement network structure, employing the Inception structure along with dilated convolution to expand the feature map's visual receptive field, improving the network's adaptability to small-scale ship targets.

Features constitute the primary basis for iterative learning in object detection algorithms. Thus, optimizing the features fed into the detection network can most directly improve various algorithm aspects. Widely used modules include attention mechanisms to focus on key features and feature pyramid networks (FPN) for fusing multi-scale features. In terms of attention mechanisms, Li et al. [23] used channel attention mechanisms to convert spatial information in the image into masks, score them to extract crucial information, and provide references for the detection network. Zheng et al. [24] introduced transferable attention mechanisms, designing an attention mask that covers all positions for each attention module, highlighting the correct semantic feature regions. Regarding FPN, Zhao et al. [25] constructed a four-level scale feature pyramid network in a top-down manner. This network leveraged candidate regions and their surrounding contextual information to provide higher-quality classification confidence and final target scores, thereby enhancing semantic information extraction for small targets. Mei et al. [26] extended FPN into four parts, highly integrating features of different scales extracted by the backbone network, thereby improving the network's ability to detect small ship targets.

Although the above-mentioned methods have improved ship target detection in SAR images in various aspects, certain shortcomings remain. Firstly, these methods may lack precision or generate more false alarms when dealing with extremely small targets, such as distant ships or lifeboats, possibly due to limitations in target resolution. Secondly, when targets closely resemble the color and texture of the background, such as in the presence of sea waves, islands, or other ships, these methods may face challenges [27]. Moreover, these methods may not be robust enough, especially when dealing with noisy, missing data, or other interference in SAR images [28].

To address these issues, this paper introduces the SAR-ShipSwin (synthetic aperture radar ship detection with swin transformer integration) algorithm, building upon the faster R-CNN framework. Our main contributions include:

1. A novel backbone architecture that merges the swin transformer with a feature pyramid network (FPN), enhanced by an advanced W-MSA module. This combination is specifically designed to tackle occlusion and overlap challenges inherent in SAR imagery.
2. The design of the background modeling network (BMN), primarily for identifying and eliminating complex background features. It comprises background feature extraction layers, background attention modules, and background weakening modules, effectively reducing background-related false alarms.
3. The introduction of spatial intensity geometric pooling, a unique pooling technique that incorporates both spatial and intensity information from the region of interest (ROI). This approach is tailored to preserve the geometric and structural integrity of the original ROI, minimizing loss of information and distortion.
4. The design of the dynamic ship shape adaptive convolution (DSAC) module, which dynamically modifies the shape of the convolution kernel to more accurately conform to the observed target. This method proves more adept at capturing the true shape of ships, especially given their variable forms and potential irregularities in SAR images, compared to conventional convolution techniques.

We have conducted multiple experiments, and the results demonstrate the excellent performance of the proposed algorithm in various scenarios and conditions, effectively improving ship target detection performance and generalization capabilities.

2 Preliminary

This section elaborates on the basic mathematical concepts and theories foundational. It focuses on specialized aspects of synthetic aperture radar (SAR) image processing, advanced neural network architectures, and specific innovations in convolutional operations, providing a direct underpinning for the methodologies developed in this research.

2.1 SAR image processing

SAR imaging involves complex signal processing techniques to resolve features in images, particularly for ship detection.

$$I_{\mathrm{d}} = f(I_{\mathrm{s}}) \otimes K + N \tag{1}$$

where I_{d} represents the denoised image, I_{s} is the speckled SAR image, K denotes the kernel for convolution embodying the feature enhancement, $\otimes$ signifies the convolution operation, and N is the residual noise.

2.2 Swin transformer for SAR images

The swin transformer is a pivotal innovation in the field of deep learning, introducing a hierarchical structure that significantly enhances the processing of synthetic aperture radar (SAR) images. Its design is tailored to capture the inherent multi-scale nature of SAR images, making it exceptionally suited for tasks requiring fine-grained feature extraction across various scales, such as ship detection in complex maritime environments.

At its core, the swin transformer operates by partitioning the input image into non-overlapping patches, which are then treated as the basic units for the initial layer of the transformer. This patch-based processing reduces the computational complexity, enabling the model to scale to large images efficiently. The key to its hierarchical structure lies in its ability to merge patches progressively at deeper layers of the network, effectively building a pyramid of features with increasing semantic levels and decreasing spatial resolutions. The hierarchical representation can be mathematically formulated as:

$$P_{l+1} = M(P_l) \tag{2}$$

where P_l represents the set of non-overlapping patches or their feature representations at layer l, and M denotes the patch merging operation that combines adjacent patches to form P_{l+1}, the input for the next layer.

The swin transformer introduces the shifted window multi-head self-attention (SW-MSA) mechanism as a means to efficiently compute self-attention within local windows while also facilitating cross-window connection in subsequent layers. This approach significantly reduces the computational demands of traditional self-attention mechanisms, making it feasible to apply transformers to high-resolution images. The SW-MSA can be described as follows:

$$\text{SW-MSA}(Z_l) = \text{W-MSA}(\text{SHIFT}(Z_l)) \tag{3}$$

where Z_l is the input feature map to layer l, SHIFT is an operation that cyclically shifts the window partitions to enable cross-window connections, and W-MSA denotes the window-based multi-head self-attention.

The adaptive representation of features within SAR images by the swin transformer is achieved through the combination of hierarchical structuring and the

SW-MSA mechanism. This dual approach allows the model to maintain high-resolution details in early layers while aggregating more abstract semantic information in deeper layers. The process of feature extraction and representation is encapsulated in the equation:

$$Z_{l+1} = \text{SW-MSA}(\text{Norm}(Z_l)) + \text{MLP}(\text{Norm}(\text{SW-MSA}(Z_l))) + Z_l \tag{4}$$

where Norm represents layer normalization, and MLP denotes a multi-layer perceptron that is applied to the output of the SW-MSA block, followed by a residual connection that adds the input feature map Z_l to the output. This formula underscores the iterative refinement of features through self-attention and nonlinear transformations, enabling the model to capture complex dependencies and features relevant for SAR image analysis.

3 Methodology

3.1 Overall framework

In this work, we propose the SAR-ShipSwin (**S**ynthetic **A**perture **R**adar **Ship** Detection with **Swin** Transformer Integration) model, built upon the Faster R-CNN framework. The model comprises a backbone network structure and a background modeling network (BMN). The backbone network structure combines the swin transformer and feature pyramid network (FPN) to effectively address ship detection challenges in SAR images, especially when it comes to resolving the resolution requirements for small targets.

Furthermore, to tackle issues related to target occlusion and overlap in SAR images, we introduce the occlusion perceptive window multihead self-attention (OPW-MSA). In order to better capture the features of irregularly shaped ships, the model also employs the spatial intensity geometric pooling method and the dynamic ship shape adaptive convolution module. The overall architecture of the SAR-ShipSwin model is depicted in Fig. 1.

3.2 Backbone network structure

To efficiently address ship detection in SAR images, we propose a backbone network that combines the swin transformer and the feature pyramid network (FPN). This design takes into full consideration the multi-scale nature of images and the resolution requirements for small ship targets.

Traditional feature pyramid networks (FPN) [29] enhance feature representations by combining low-level positional information with high-level semantic information. By incorporating FPN, we can significantly enhance the feature map resolution for small targets, which is crucial for small ship detection in SAR images. Considering the various model options of Swin Transformer, we employ the lightweight Swin-T as the basic unit in this paper. Swin-T consists of four stages, and the features generated in each stage undergo initial feature adjustment with a 1 × 1

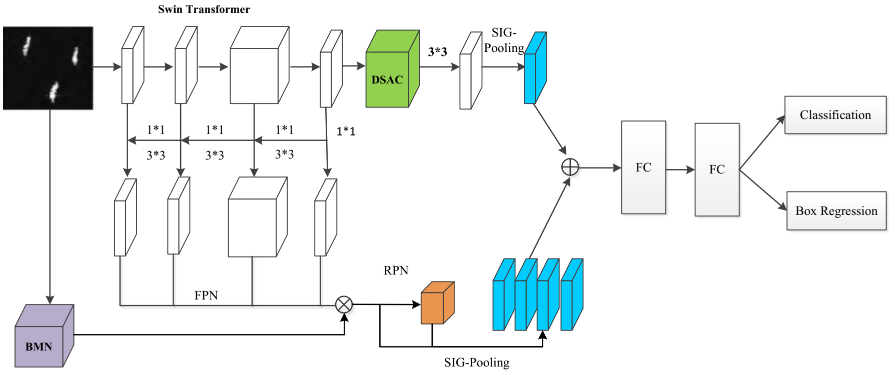

Fig. 1 Model architecture of SAR-ShipSwin

convolution. Subsequently, they are fused with features from other stages through upsampling. The fused feature maps then go through a 3 × 3 convolution for further feature extraction and output, as illustrated in Fig. 2.

Swin transformer based on the ViT architecture [30] introduces prior hierarchies, locality, and translational invariance, optimizing model computation efficiency and performance [31]. Its unique mobile window operation ensures information interaction between adjacent windows, granting the model the ability to model global information while significantly reducing computation.

The core structure of swin transformer blocks is depicted in Fig. 3. To enhance the model's information exchange capability without increasing computational complexity, we make improvements on the original W-MSA module. To address the issue of target occlusion and overlap in SAR images, we propose the occlusion perceptive window multihead self-attention (OPW-MSA).

Before performing multi-head self-attention computation, a small neural network is used to generate an occlusion score for each pixel. This score represents the degree to which the pixel is occluded, helping us identify areas where overlap

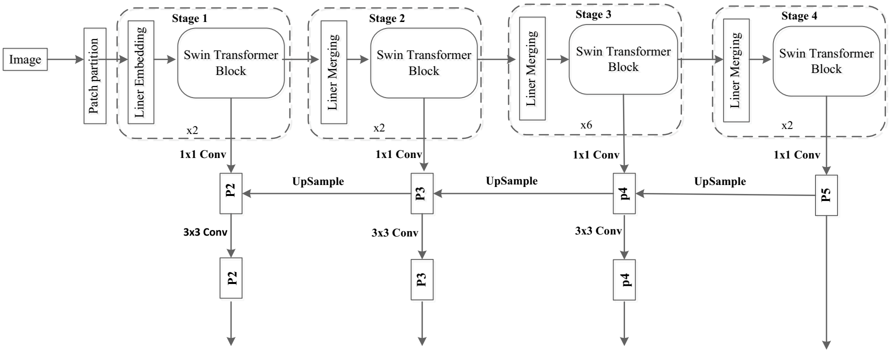

Fig. 2 Multi-scale fusion network structure of swin-T and FPN

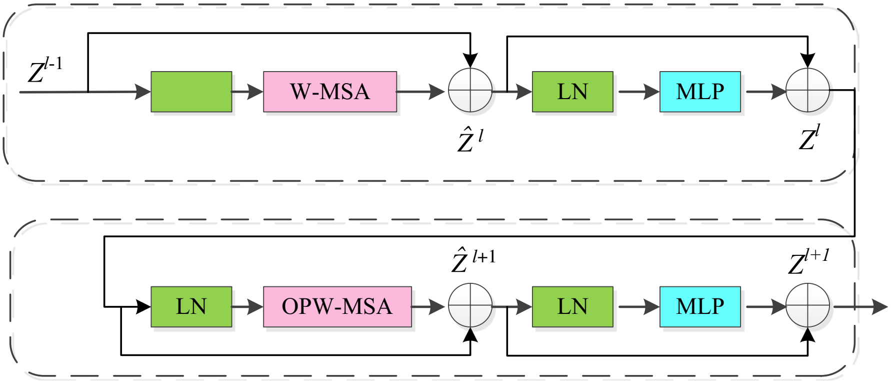

Fig. 3 Swin transformer blocks structure

or occlusion of targets may occur. We refer to this step as occlusion perceptive mapping (OPM). To assess the degree of occlusion for a region or pixel, this paper employs local gradient information from the image. High gradients may indicate the presence of boundaries, and boundaries may signify target occlusion or overlap. The computation formula for occlusion perceptive mapping (OPM) is as follows:

$$\mathrm{OPM}(x) = \nabla x \cdot w_{\mathrm{opm}} + b_{\mathrm{opm}} \tag{5}$$

where ∇x represents the gradient of pixel value x, w_{opm} and b_{opm} are learnable parameters adjusted during training to maximize occlusion recognition.

Based on the output of OPM, weights are dynamically assigned to each pixel. These weights are proportional to the occlusion scores, meaning that highly occluded areas receive higher weights:

$$\mathrm{DAW}(x) = \frac{\mathrm{OPM}(x)}{\sum_i \mathrm{OPM}(x_i)} \tag{6}$$

where $\mathrm{DAW}(x)$ represents dynamically allocated weights.

The primary stage consists of two layer normalizations (LN), a window multi-head self-attention (W-MSA) mechanism, and a multiLayer perceptron (MLP). In this stage, the W-MSA module segments the image into non-overlapping windows, effectively reducing the model's computational burden. To overcome information exchange barriers caused by non-overlapping windows, the advanced stage replaces the W-MSA module with a sliding window multihead self-attention (OPW-MSA) mechanism. The remaining parts maintain LN and MLP to construct the residual connection. The specific computation process is as follows:

$$\begin{gathered}\hat{Z}^{l} = \text{W-MSA}\left(LN(Z^{l-1})\right) + Z^{l-1},\\ Z^{l} = \text{MLP}(LN(\hat{Z}^{l})) + \hat{Z}^{l},\\ \hat{Z}^{l+1} = \text{OPW-MSA}\left(LN(Z^{l})\right) + Z^{l},\\ Z^{l+1} = \text{MLP}(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1}\end{gathered} \tag{7}$$

3.3 Background modeling network (BMN)

This sub-network focuses on identifying and eliminating complex background features, allowing the RPN to concentrate more on target extraction. BMN comprises a background feature extraction layer, a background attention module, and a background weakening module.

The background feature extraction layer consists of 2–3 convolutional layers, each followed by Batch Normalization [32] and ReLU [33] activation functions. These layers are primarily used to extract low-level background features. The computation formula for the background feature extraction layer is as follows:

$$\begin{gathered}X_1 = \text{ReLU}(\text{BN}(W_1 * X + b_1))\\ X_2 = \text{ReLU}(\text{BN}(W_2 * X_1 + b_2))\\ X_b = \text{ReLU}(\text{BN}(W_3 * X_2 + b_3))\end{gathered} \tag{8}$$

where $X \in \mathbb{R}^{H\times W\times C}$ represents the input feature map, where H and W are the height and width of the feature map, and C is the number of channels. $X_b \in \mathbb{R}^{H\times W\times C'}$ represents the output feature map, where C' is the new number of channels. $*$ denotes the convolution operation, and $W1$, $W2$, $W3$, $b1$, $b2$, $b3$ represent the convolution weights and biases, and BN denotes Batch Normalization.

The background attention module, a core component, employs spatial attention mechanisms to help the network focus on prominent background features:

$$\begin{gathered}A = \sigma(W_a * X_b + b_a)\\ X_a = X_b \odot A\end{gathered} \tag{9}$$

where $\odot$ represents element-wise multiplication, and W_a and b_a are the weights and biases of the attention module; while, σ is the sigmoid activation function.

The background weakening module employs a strategic formulation to modulate the background intensity, thereby facilitating a more focused target extraction by the region proposal network (RPN). It ingeniously incorporates a mask layer designed to generate a binary-like mask, which aligns with the spatial dimensions of the feature pyramid network (FPN) output. This mask layer is realized through a convolutional operation followed by a sigmoid activation function, σ, which maps the input feature space into a [0,1] range. In this context, values approaching 0 are indicative of background regions; whereas, values nearing 1 delineate the foreground, thus achieving a discriminative representation of the scene elements. The mathematical expression encapsulating the operation of the background weakening module can be articulated as follows:

$$\begin{aligned} M &= \sigma(W_m * X_a + b_m) \\ X_f &= X_a \odot (1 - M) + X \odot M \end{aligned} \tag{10}$$

where W_m and b_m are the weights and biases of the mask generation layer. X_f represents the feature map after weakening the background (multiplied by $1 - M$); while, the foreground regions remain unchanged (multiplied by M).

Finally, the output mask from the background weakening module is element-wise multiplied with the output of FPN to obtain a feature map with weakened background. This fused feature map serves as the input to RPN for region proposal generation.

3.4 Spatial intensity geometric pooling

The fast R-CNN architecture employs ROI (region of interest) Pooling to extract region proposal features, generating fixed-size feature maps that are then passed to fully connected layers for classification and bounding box regression to complete object detection [34]. However, since the sizes of region proposals generated by the RPN (region proposal network) can vary, performing ROI Pooling with block-wise pooling to obtain fixed 7×7-sized feature maps can disrupt the structural information of the original image, leading to imprecise object localization. Additionally, the feature maps generated by the multi-scale fusion FPN (feature pyramid network) in this paper can have inconsistent sizes, resulting in extreme aspect ratios. This can lead to significant mapping discrepancies, causing feature loss. Particularly in SAR (synthetic aperture radar) images, forcing ROIs of different sizes and shapes into uniform fixed-size feature maps through ROI Pooling may destroy or distort the structural information of small objects, such as small vessels.

To address the aforementioned issues, this paper proposes a geometrically preserving sampling method that avoids the use of traditional max or average pooling. Instead, it introduces a new operation called spatial intensity geometric pooling (SIG-pooling), which takes into account both the spatial distribution and intensity information within ROIs to calculate pooling values.

Consider an ROI region R with dimensions $h \times w$, which is divided into an $m \times n$ grid of sub-region cells. For each sub-region cell g_{ij}, a geometric weighting factor G_{ij} is defined, calculated based on the spatial distribution and intensity information of the ROI.

$$G_{ij} = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} I(x, y) \times D_{ij}(x, y) \tag{11}$$

where $I(x, y)$ is the intensity value of the ROI at coordinates (x, y), and is the reciprocal of the distance from the center of sub-region cell g_{ij} to coordinates (x, y). This ensures that the central portion of the ROI is given a higher weight.

For each sub-region cell g_{ij} , its geometrically preserved pooling value P_{ij} is computed as follows:

$$P_{ij} = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} I(x, y) \times G_{ij} \tag{12}$$

After the SIG-Pooling operation, the ROI region R is transformed into a new feature map of size $m \times n$, where each element represents the geometrically preserved pooling value of its corresponding sub-region cell. This geometrically preserved pooling method, by combining spatial and intensity information, better retains the geometric and structural information of the original ROI, reducing information loss and distortion, especially for small objects like small vessels in SAR images.

3.5 Dynamic ship shape adaptive convolution

Considering that vessels in SAR (synthetic aperture radar) images have variable and often irregular shapes, this paper introduces the dynamic ship shape adaptive convolution (DSAC) module to adapt to different vessel shapes and sizes in a specific manner. Unlike traditional convolutional kernels with fixed shapes, DSAC dynamically adjusts the shape of the convolutional kernel to fit the current target, allowing for more accurate capturing of irregular vessel features. The dynamic ship shape adaptive convolution module comprises three sub-modules: shape recognition, shape-adaptive convolution, and convolution operations.

(1) Shape Recognition Sub-module

For precise object detection, considering the shape information of the target can greatly benefit feature extraction. In SAR images where vessel shapes vary, recognizing their shapes can assist subsequent convolutional operations in extracting features more effectively. This module is a classification task and can be defined as:

$$f_{\text{shape}}(x) = \text{Softmax}\left(W_{\text{shape}} \times x + b_{\text{shape}}\right) \tag{13}$$

where x is the feature representation of the input ROI, and W_{shape} and b_{shape} are the weights and biases, respectively.

(2) Shape-Adaptive Convolution

To ensure that convolutional operations are more tailored, we need to take into account the specific shape of the input ROI. Therefore, we propose a dynamic kernel selection strategy that chooses the most suitable convolutional kernel based on the predicted shape of the input ROI. Based on the output of the shape recognition sub-module, we select a convolutional kernel set that best matches the predicted shape:

$$K = g(f_{\text{shape}}(x)) \tag{14}$$

where g is a function that selects the appropriate convolutional kernel based on the output of the shape recognition sub-module. If $f_{\text{shape}}(x)_{\text{long}} > f_{\text{shape}}(x)_{flat}$, K_{long} is chosen; otherwise, K_{flat} is chosen. K_{long} and K_{flat} are predefined sets of convolutional kernels suitable for "elongated" and "flat" shapes, respectively. For elongated vessels, we can define a convolutional kernel K_{flat} that is suitable for capturing vertical edges:

$$K_{\mathrm{long}} = \begin{bmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{bmatrix} \tag{15}$$

For flat-shaped vessels, we can define a convolutional kernel K_{flat} that is suitable for capturing horizontal edges:

$$K_{\mathrm{flat}} = \begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix} \tag{16}$$

(3) Convolution Operations

After determining the most appropriate convolutional kernel, convolution operations are performed to extract features:

$$y(p) = \sum_{q \in K} W_{\mathrm{conv}} \cdot x(p+q) + b_{\mathrm{conv}} \tag{17}$$

where p represents a pixel position in the feature map, q represents a position in the kernel K. W_{conv} and b_{conv} are the weights and biases of the convolution operation.

3.6 Loss function

Similar to the faster R-CNN model, the loss consists of two components: regression loss and classification loss. The classification loss employs the cross-entropy loss function as the classification loss function. Given a true label y for the target category and the model's predicted class probability distribution p, the cross-entropy loss can be defined as:

$$L_{\mathrm{cls}} = -\sum^{\substack{i=1 \\ C}} y_i \log(p_i) \tag{18}$$

where C is the number of categories, y_i is the ith element of the true label, with a value of 1 (if the target belongs to the ith class) or 0 (otherwise). pi is the probability of the model's prediction for the ith class.

For the regression loss, the Smooth $L1$ loss is adopted to reduce the impact of outliers when predicting bounding boxes. Given the true bounding box coordinates $t*$ and the model's predicted coordinates t, the Smooth $L1$ loss can be defined as:

$$L_{\mathrm{reg}} = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{19}$$

where $x = t - t^*$ represents the difference between predicted and true coordinates.

Furthermore, a Geometric Pooling Loss is designed specifically for SIG-Pooling to ensure that the feature map after pooling effectively preserves the geometric and structural information of the original ROI. This is achieved by computing the cosine similarity between the features before and after pooling:

$$L_{\text{geometric}} = 1 - \frac{\text{feature}_{\text{pre prooling}} \cdot \text{feature}_{\text{post}-\text{prooling}}}{\parallel \text{feature}_{\text{pree proding}} \parallel \parallel \text{feature}_{\text{prost}-\text{prooling}} \parallel} \tag{20}$$

The objective of this loss function is to minimize the cosine distance between the two feature vectors, ensuring that the post-pooling feature aligns directionally with the pre-pooling feature, thus preserving the geometric and structural information of the original ROI. $\text{feature}_{\text{pre}-\text{pooling}}$ and $\text{feature}_{\text{post}-\text{pooling}}$ represent feature vectors before and after the pooling operation, respectively. $\|\cdot\|$denotes the norm operation.

The overall loss L can be defined as:

$$L = L_{\text{cls}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{geo}} L_{\text{geometric}} \tag{21}$$

where λ_{vreg} and λ_{geo} are weighting factors used to balance the various loss components.

4 Experiments

4.1 Experimental setup

Datasets To validate the effectiveness of the model proposed in this paper, experiments were conducted on the SSDD dataset [35] and the HRSID dataset [36]. The SSDD dataset comprises a total of 1160 images, with an average of 2.12 ships per image. It includes SAR images from three different sensors: Sentinel-1, Terra SAR-X, and RadarSAT-2, captured in HH, VV, VH, and HV imaging modes. The SAR image data have resolutions ranging from 1 to 15 m and cover large maritime areas as well as coastal regions with various ship targets. The HRSID dataset, released in January 2020, is a large dataset for object detection based on synthetic aperture radar (SAR). It contains 16951 instances of ships and 5604 high-resolution SAR images from Sentinel-1B, TerraSAR-X, and Tan DEM-X sensors. The dataset is designed for applications such as semantic segmentation, ship detection, and instance segmentation. Based on the COCO remote sensing image dataset, the HRSID dataset includes multi-source remote sensing images with different resolutions, polarizations, sea conditions, maritime areas, and coastal ports, with image resolutions ranging from 1 to 5 m.

Hardware and software environment The experimental hardware environment consists of an Intel Core i9-11900K CPU, 32GB of memory, and an NVIDIA GeForce RTX 3080 GPU. The operating system used is Ubuntu 20.04, and the deep learning framework employed is PyTorch.

Hyperparameters The stochastic gradient descent (SGD) algorithm is used to train our network with a batch size of 16. For our ablation experiments, the network undergoes a total of 300 epochs. A learning rate of 0.01, weight decay of 0.0005, and SGD momentum of 0.937 are set. Other unspecified hyperparameters are kept

Table 1 Comparison with other SAR ship detection methods on the SSDD dataset

Methods	*P* (%)	*R* (%)	mAP (%)
Faster R-CNN [37]	85.37	90.07	89.15
Quad-FPN [38]	95.77	89.52	95.29
Cascade R-CNN [39]	90.81	94.1	90.5
PANET [40]	91.91	86.81	91.15
DAPN [41]	91.36	85.54	90.56
HR-SDNet [42]	90.99	96.49	90.82
LS-SSDD [43]	96.1	94	97.8
Double-Head R-CNN [44]	91.91	86.96	91.1
GRid R-CNN [45]	89.71	87.77	88.92
SER Faster R-CNN [46]	92.28	86.11	91.52
SAR-ShipSwin	96.53	94.57	98.02

consistent with YOLOv5. Additionally, when comparing with other methods, we configure parameters similarly to ensure a fair comparison.

Evaluation metrics In this experiment, we utilize three metrics, namely Precision, Recall, and mean Average Precision (mAP), to analyze and verify the detection performance of the proposed method. mAP can be calculated from the Precision and Recall metrics.

4.2 Comparative experimental results

Table 1 presents a performance comparison of SAR-ShipSwin with other models on the SSDD dataset. In terms of the mAP metric, both LS-SSDD and SAR-ShipSwin exhibit outstanding performance, surpassing other methods. Notably, SAR-ShipSwin achieves the highest performance with a mAP of 98.02%. Regarding precision (P), SAR-ShipSwin leads with a value of 96.53%, followed by LS-SSDD at 96.10%, reaffirming SAR-Ship Swin's superior detection accuracy. Additionally, Quad-FPN achieves a relatively high *P* value of 95.77%, but its recall (*R*) is lower, resulting in a slightly lower overall mAP. HR-SDNet achieves the highest *R* value at 96.49%, but its *P* value is slightly lower, leading to an overall mAP of 90.82%. SAR-ShipSwin ranks third in terms of *R* value at 94.57%, but its higher *P* value places it in the lead in terms of overall mAP.

Table 2 provides a performance comparison of SAR-ShipSwin with other models on the HRSID dataset. The comparison results on the HRSID dataset demonstrate that SAR-ShipSwin exhibits outstanding performance across different scenarios, including overall scenes, nearshore scenes, and scenes far from the shore. Particularly noteworthy are its achievements in overall scenes and scenes far from the shore, where SAR-ShipSwin achieves the highest mAP values of 92.35% and 97.92%, respectively, outperforming all other models.

SAR-ShipSwin leverages a combination of Swin Transformer and FPN to handle complex SAR images, providing robust feature extraction capabilities. The inclusion of the W-MSA module, in particular, enhances the model's ability to address

Table 2 Comparison with other SAR ship detection methods on the HRSID dataset

Method	Entire scenes (%)			InShore scenes (%)			Offshore scenes (%)		
	P	*R*	mAP	*P*	*R*	mAP	*P*	*R*	mAP
Faster R-CNN	68.1	81.5	80.98	47.3	63.7	57.34	88.7	95.6	95.91
Cascade R-CNN	80.9	81.6	82.65	64.4	64	61.21	93.6	95.6	96.27
Libra R-CNN	74.3	89.3	82.58	55.5	66.1	60.41	92.2	95.5	96.21
RetinaNet	71.8	88.8	87.64	53.1	78.4	69.18	92.7	97	97.72
Swin-RetinaNet	69.6	87.1	85.94	50.2	75	65.43	91.2	96.7	97.32
FSAF	74.3	89.1	89	67.8	80.6	73.42	93.8	96.6	96.66
FreeAnchor	81.2	90.2	90	67.8	80.6	77.86	93.3	97.8	97.73
FCOS	83.4	88	88.35	70.1	77.4	73.71	95	96.3	97.17
TOOD	86	87.3	89.67	74.1	76.2	76.1	95.6	96.2	97.33
GFECI-Net	87.1	88.6	91.28	75.6	79	79.64	95.8	96.2	97.74
SAR-ShipSwin	88.5	90.4	92.35	76.9	82	80.65	96.3	96.5	97.92

Table 3 Ablation study

Module combination	Precision (*P*)	Recall (*R*)	mAP
+ Faster R-CNN Base Model	85.37	90.07	89.15
+ Backbone Network Structure	87.52	91.32	90.45
+ BMN	89.68	92.17	92.02
+ SIG-Pooling	92.23	93.04	93.65
+ DSAC (SAR-ShipSwin)	96.53	94.57	98.02

occlusion and overlapping challenges prevalent in SAR images, thereby improving detection accuracy in complex scenarios. Complex backgrounds pose a significant challenge in SAR ship detection. BMN ensures the effective removal of complex background features, thereby reducing the impact of background noise on detection. The Spatial Intensity Geometric Pooling and Dynamic Ship shape Adaptive Convolution modules guarantee that the model retains the structural information of ROIs effectively during ship detection, adapting to various ship shapes and sizes, ultimately enhancing detection accuracy.

4.3 Ablation study results

In this paper, four core modules were designed within the Faster R-CNN framework: the backbone network structure, background modeling network (BMN), spatial intensity geometric pooling, and dynamic ship shape adaptive convolution (DSAC). To better evaluate the contributions of these modules to the SAR-ShipSwin model's performance, ablation experiments were conducted on the SSDD dataset, and the results are presented in Table 3.

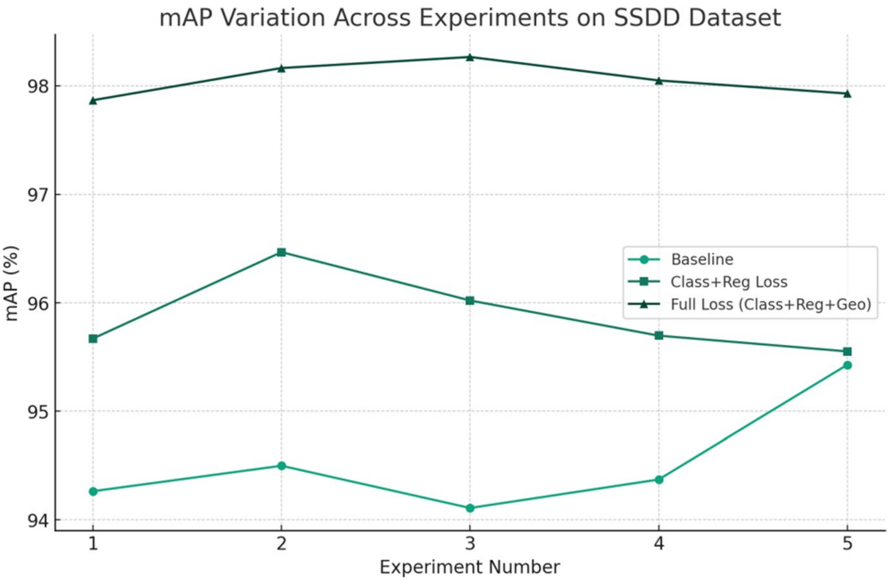

Fig. 4 Comparative experiments of different loss functions

We initially used the Faster R-CNN base model as the evaluation starting point, which already achieved a relatively high mAP of 89.15% on the SSDD dataset. Upon incorporating the backbone network structure, which combines the Swin Transformer and FPN, the model's performance improved, resulting in an mAP of 90.45%. This increase underscores the superiority of the backbone network in feature extraction. Subsequently, the Background Modeling Network (BMN) was introduced to assist the model in accurately identifying targets in SAR images with complex backgrounds, further increasing the mAP to 92.02%. This highlights BMN's contribution to enhancing detection accuracy in complex scenarios. Finally, with the application of Spatial Intensity Geometric Pooling (SIG-Pooling), we observed a further increase in mAP to 98.02%. This indicates that when the model considers spatial and intensity information within ROIs, it can better retain the geometric and structural information of the original ROIs, resulting in more accurate detections, with the most significant performance improvement observed.

To validate the loss function designed in this paper, ablation experiments were conducted to compare the performance of the SAR-ShipSwin model under different loss function configurations. Initially, the SAR-ShipSwin model was trained using the Faster R-CNN loss. Subsequently, the model was trained using two different loss function configurations: classification loss + regression loss, and classification loss + regression loss + Geometric Pooling Loss. To ensure the reliability of the results, experiments were conducted on the SSDD dataset using various configurations of loss functions, and repeated five times. The experimental results are shown in Fig. 4.

As can be seen from the figure, although the performance of all loss function configurations exhibits certain fluctuations, the complete loss function configuration (including classification, regression, and geometric pooling losses) overall

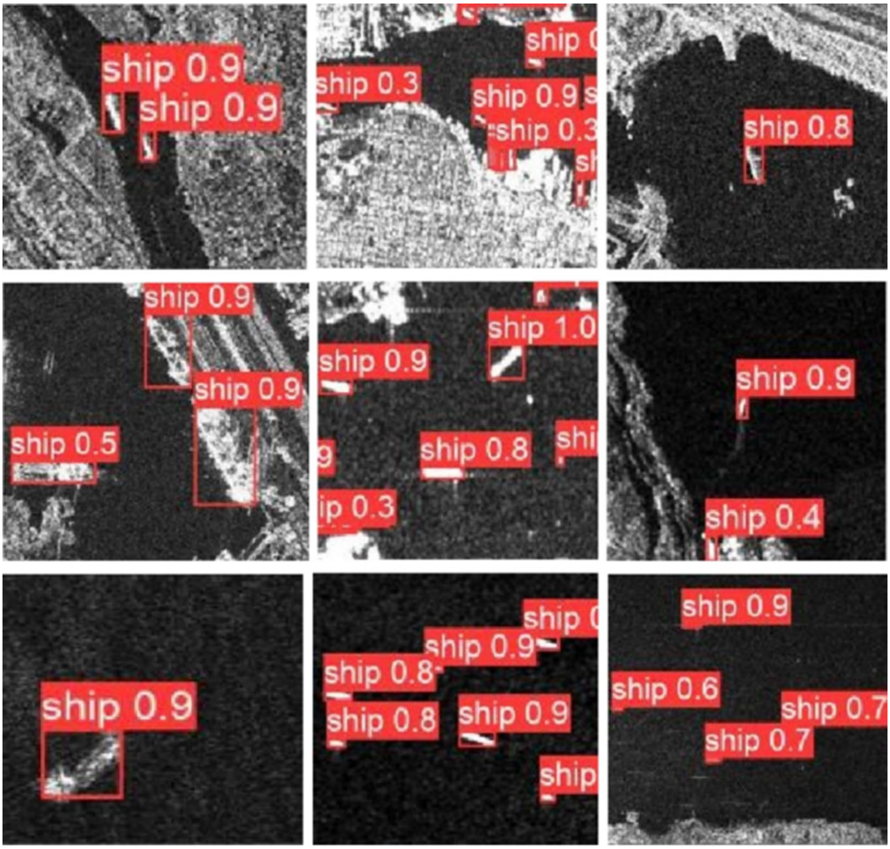

Fig. 5 Visualization of SAR-ShipSwin on the SSDD dataset

demonstrates higher and more stable performance. This validates the effectiveness of our designed loss function in enhancing the accuracy and stability of the SAR-ShipSwin model in detecting ships in complex SAR images.

4.4 Visualization results

Figure 5 displays the visualization results for the SSDD dataset. As seen, our proposed ship detection method exhibits strong performance both near the shore and in offshore areas.

Figure 6 showcase the visualization results for the HRSID dataset. It is evident that the ship detection method proposed in this paper performs well in both nearshore and offshore areas.

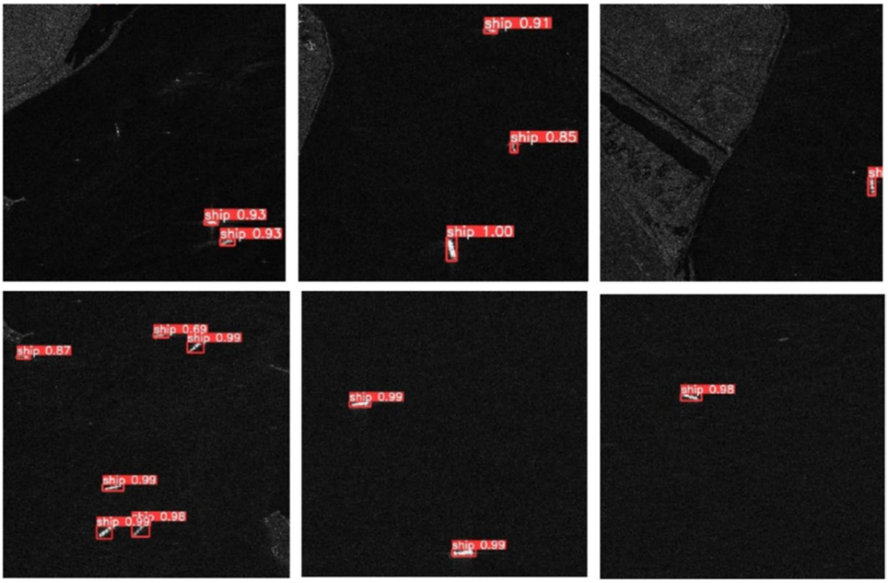

Fig. 6 Visualization of the HRSID dataset

5 Conclusion

SAR ship detection has long been a hot research topic in maritime target detection. Facing the current major challenges: first, the increased difficulty in distinguishing ship targets from complex backgrounds in SAR images, especially under various meteorological and sea conditions; second, the variability in the shape of ship targets in SAR images, along with severe occlusion and overlap [47]. In response to these issues, this paper proposes the SAR-ShipSwin model. This model, tailored to the unique characteristics of ship targets in SAR images, introduces a backbone network structure that combines Swin Transformer and FPN, effectively extracting features and optimizing model computational efficiency.

Furthermore, we propose a Background Modeling Network designed specifically to identify and eliminate complex background features, thereby improving the accuracy of target detection. Finally, considering the variability in ship shapes in SAR images, we design the Dynamic Ship shape Adaptive Convolution module, which dynamically adjusts the shape of convolutional kernels, further enhancing detection accuracy.

Through extensive comparative experiments, ablation studies, and generalization experiments, our SAR-ShipSwin demonstrates superior detection performance compared to existing baselines and some state-of-the-art algorithms. This confirms that our algorithm not only exhibits efficient detection performance but also demonstrates excellent generalization capabilities. In the future, efforts will be directed toward improving the performance of the SAR-ShipSwin model in detecting extremely small or highly occluded targets. Moreover, integrating reputation management mechanisms could further enhance our model's robustness and reliability in dynamic environments [48]. Additionally, the adoption of ConvLSTM-based approaches for improving signal processing may refine our model's ability to handle complex noise patterns and thus improve detection accuracy in challenging scenarios [49].

Author contributions Tang.J. and Han.Y. proposed requirements and concepts ; Tang.J. and Han.Y. designed and implemented methods; Tang.J. and Xian.Y. wrote the software and tested it; Tang.J., Xian.Y. and Han.Y. participated in data analysis; Tang.J. prepared the original draft; Tang.J. and Xian.Y. reviewed and edited the manuscript.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Moreira A, Prats-Iraola P, Younis M, Krieger G, Hajnsek I, Papathanassiou KP (2013) A tutorial on synthetic aperture radar. IEEE Geosci Remote Sens Mag 1(1):6–43
2. Zhang Z, Lin H, Wang M, Liu X, Chen Q, Wang C, Zhang H (2022) A review of satellite synthetic aperture radar interferometry applications in permafrost regions: Current status, challenges, and trends. IEEE Geosci Remote Sens Mag 10(3):93–114
3. Huang Y, Chen Z, Wen C, Li J, Xia X-G, Hong W (2022) An efficient radio frequency interference mitigation algorithm in real synthetic aperture radar data. IEEE Trans Geosci Remote Sens 60:1–12
4. Yang H, Zhang T, He Y, Dan Y, Yin J, Ma B, Yang J (2022) Gpu-oriented designs of constant false alarm rate detectors for fast target detection in radar images. IEEE Trans Geosci Remote Sens 60:1–14
5. Xiang D, Xie Y, Cheng J, Xu Y, Zhang H, Zheng Y (2022) Optical and SAR image registration based on feature decoupling network. IEEE Trans Geosci Remote Sens 60:1–13
6. Shen W, Wang Y, Lin Y, Li Y, Jiang W, Hong W (2023) Range-Doppler based moving target image trace analysis method in circular SAR. Remote Sens 15(8):2073
7. Li M-D, Cui X-C, Chen S-W (2021) Adaptive superpixel-level CFAR detector for SAR inshore dense ship detection. IEEE Geosci Remote Sens Lett 19:1–5
8. Wang X, Li G, Zhang X-P, He Y (2021) A fast CFAR algorithm based on density-censoring operation for ship detection in SAR images. IEEE Signal Process Lett 28:1085–1089
9. Li K, Wan G, Cheng G, Meng L, Han J (2020) Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm Remote Sens 159:296–307
10. Wu X, Sahoo D, Hoi SC (2020) Recent advances in deep learning for object detection. Neurocomputing 396:39–64
11. Xiao Y, Tian Z, Yu J, Zhang Y, Liu S, Du S, Lan X (2020) A review of object detection based on deep learning. Multimed Tools Appl 79:23729–23791
12. Liu Y, Sun P, Wergeles N, Shang Y (2021) A survey and performance evaluation of deep learning methods for small object detection. Expert Syst Appl 172:114602
13. Hou L, Wang H, Zou H, Zhou Y (2022) Robotic manipulation planning for automatic peeling of glass substrate based on online learning model predictive path integral. Sensors 22(3):1292
14. Yao J, Li C, Sun K, Cai Y, Li H, Ouyang W, Li H (2023) Ndc-scene: boost monocular 3d semantic scene completion in normalized device coordinates space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9455–9465
15. Zheng Y, Sun P, Zhou Z, Xu W, Ren Q (2021) DT-Det: adaptive dynamic refined single-stage transformer detector for arbitrary-oriented object detection in satellite optical imagery. Remote Sens 13(13):2623

16. Lu X, Li Q, Li B, Yan J (2020) Mimicdet: bridging the gap between one-stage and two-stage object detection. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, proceedings, part XIV 16. Springer, pp 541–557
17. Detector A-FO (2022) Fcos: a simple and strong anchor-free object detector. IEEE Trans Pattern Anal Mach Intell 44(4):15
18. Yao J, Wu T, Zhang X (2023) Improving depth gradient continuity in transformers: a comparative study on monocular depth estimation with CNN. arXiv preprint arXiv:2308.08333
19. Yasir M, Jianhua W, Mingming X, Hui S, Zhe Z, Shanwei L, Colak ATI, Hossain MS (2023) Ship detection based on deep learning using SAR imagery: a systematic literature review. Soft Comput 27(1):63–84
20. Li J, Yu Z, Yu L, Cheng P, Chen J, Chi C (2023) A comprehensive survey on SAR ATR in deep-learning era. Remote Sens 15(5):1454
21. Chen P, Zhou H, Li Y, Liu P, Liu B (2023) A novel deep learning network with deformable convolution and attention mechanisms for complex scenes ship detection in sar images. Remote Sens 15(10):2589
22. Bai L, Yao C, Ye Z, Xue D, Lin X, Hui M (2023) Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection. IEEE J Select Top Appl Earth Observ Remote Sensing 16:1042–1056
23. Li Y, Zhu Z, Li Y, Zhang J, Li X, Shang S, Zhu D (2023) Ctmu-net: an improved u-net for semantic segmentation of remote-sensing images based on the combined attention mechanism. IEEE J Select Top Appl Earth Observ Remote Sens 16:10148–10161
24. Zhang Y, Guo X, Leung H, Li L (2023) Cross-task and cross-domain SAR target recognition: a meta-transfer learning approach. Pattern Recogn 138:109402
25. Zhao B, Sui H, Liu J (2023) Siam-dwenet: Flood inundation detection for sar imagery using a cross-task transfer siamese network. Int J Appl Earth Obs Geoinf 116:103132
26. Mei J, Zheng Y-B, Cheng M-M (2023) D2anet: difference-aware attention network for multi-level change detection from satellite imagery. Comput Vis Media 9(3):563–579
27. Li J, Li Z, Zhang B, Wu Y (2023) A multi-channel attention network for SAR interferograms filtering applied to TomoSAR. Remote Sens 15(18):4401
28. Feng J, Liu L, Hou X, et al (2023) Qoe fairness resource allocation in digital twin-enabled wireless virtual reality systems. IEEE J Select Areas Commun
29. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2117–2125
30. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y et al (2022) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87–110
31. Yao J, Pan X, Wu T, Zhang X (2023) Building lane-level maps from aerial images. arXiv preprint arXiv:2312.13449
32. Hou L, Wang H, Zou H, Wang Q (2021) Efficient robot skills learning with weighted near-optimal experiences policy optimization. Appl Sci 11(3):1131
33. Agarap AF (2018) Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375
34. Yao J, Zhang J (2023) Depthssc: depth-spatial alignment and dynamic voxel resolution for monocular 3d semantic scene completion. arXiv preprint arXiv:2311.17084
35. Zhang T, Zhang X, Li J, Xu X, Wang B, Zhan X, Xu Y, Ke X, Zeng T, Su H et al (2021) SAR ship detection dataset (SSDD): official release and comprehensive data analysis. Remote Sens 13(18):3690
36. Wei S, Zeng X, Qu Q, Wang M, Su H, Shi J (2020) HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation. IEEE Access 8:120234–120254
37. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:15
38. Zhang T, Zhang X, Ke X (2021) Quad-FPN: A novel quad feature pyramid network for SAR ship detection. Remote Sens 13(14):2771
39. Cai Z, Vasconcelos N (2018) Cascade r-cnn: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6154–6162
40. Wang K, Liew JH, Zou Y, Zhou D, Feng J (2019) Panet: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9197–9206

41. Cui Z, Li Q, Cao Z, Liu N (2019) Dense attention pyramid networks for multi-scale ship detection in SAR images. IEEE Trans Geosci Remote Sens 57(11):8983–8997
42. Wei S, Su H, Ming J, Wang C, Yan M, Kumar D, Shi J, Zhang X (2020) Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet. Remote Sens 12(1):167
43. Zhou Y, Liu H, Ma F, Pan Z, Zhang F (2023) A sidelobe-aware small ship detection network for synthetic aperture radar imagery. IEEE Trans Geosci Remote Sens 61:1–16
44. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, part I 14. Springer, pp 21–37
45. Lu X, Li B, Yue Y, Li Q, Yan J (2019) Grid r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7363–7372
46. Lin Z, Ji K, Leng X, Kuang G (2018) Squeeze and excitation rank faster R-CNN for ship detection in SAR images. IEEE Geosci Remote Sens Lett 16(5):751–755
47. Liu L, Feng J, Mu X et al (2023) Asynchronous deep reinforcement learning for collaborative task computing and on-demand resource allocation in vehicular edge computing. IEEE Trans Intell Transp Syst
48. Liu L, Feng J, Wu C, et al (2023) Reputation management for consensus mechanism in vehicular edge metaverse. IEEE J. Select Areas Commun
49. Han H et al (2023) A ConvLSTM-based blind receiver for physical layer wireless communication. IEEE Trans Veh Technol. https://doi.org/10.1109/TVT.2023.3342169

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.