

Article

Shape-Prior Dynamic Refinement Network with Diffusion Wavelets for Fine-Grained Ship Detection in Remote Sensing Images

Zhanchao Huang ^{1,2,3,*} , Weiwang Guan ^{1,2,3}, Jiajun Zhou ^{1,2,3}, Wenjun Hong ^{1,2,3} and Hua Su ^{1,2,3} 

¹ The Academy of Digital China, Fuzhou University, Fuzhou 350108, China; 245520022@fzu.edu.cn (W.G.); 245527031@fzu.edu.cn (J.Z.); suhua@fzu.edu.cn (H.S.)

² The Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou 350108, China

³ The National & Local Joint Engineering Research Center of Satellite Geospatial Information Technology, Fuzhou University, Fuzhou 350108, China

* Correspondence: huangzc@fzu.edu.cn

Highlights

What are the main findings?

- This work proposes a shape-prior dynamic refinement network (SPDR-Net) for fine-grained ship detection. It integrates Gaussian Mixture Shape Representation, Re-Parameterized Selective Dynamic Kernels, and Diffusion-Supervised Wavelet Refinement to enhance ship morphology perception in complex remote sensing scenes.
- The proposed Gaussian Mixture Shape Representation (GMSR) explicitly models the orientation, shape, and texture distribution of ships, while diffusion-guided wavelet refinement (DSWR) recovers high-frequency boundaries and texture details through diffusion-guided wavelet reconstruction. These designs enable the model to achieve advanced fine-grained detection performance.

What are the implications of the main findings?

- The integration of Gaussian mixture modeling and diffusion-guided wavelet reconstruction in SPDR-Net offers a novel perspective for strengthening fine-grained shape and texture representation in remote sensing object detectors.
- The proposed SPDR-Net provides a robust framework for accurate fine-grained object detection in complex remote sensing scenes, supporting reliable object discrimination across diverse categories and scenarios.

Abstract

Fine-grained ship detection in remote sensing images is important for maritime surveillance and port management. Although deep learning methods have achieved promising results, fine-grained ship detection remains difficult because of complex background interference and difficulty in distinguishing similar objects. To address these limitations, we propose a Shape-Prior Dynamic Refinement Network (SPDR-Net) for fine-grained ship detection. First, we design a Gaussian Mixture Shape Representation (GMSR) mechanism to guide the learning of shapes and texture details by learnable Gaussian mixture models. Second, we develop Re-Parameterized Selective Dynamic Kernels (RSDKs) based on the GMSR shape prior to dynamically adjust the receptive field of each scale to more accurately capture morphological features. Furthermore, a Diffusion-Supervised Wavelet Refinement (DSWR) strategy is developed, which brings diffusion-based noising–denoising adversarial learning mechanism into wavelet reconstruction to recover high-frequency details, strengthening boundaries and fine-grained differential information while suppressing noise. Extensive



Academic Editors: Zhenming Peng, Yimian Dai, Liangliang Li, Fei Zhou and Xingchen Zhang

Received: 23 June 2026

Revised: 19 July 2026

Accepted: 29 July 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

experiments on public datasets demonstrate that the proposed method achieves state-of-the-art performance with mAP_{50s} of 79.3% on DOSR, 47.97% on FAIR1M, 98.1% on HRSC2016, and 97.7% on SSDD, respectively. It exhibits that SPDR-Net effectively handles both offshore and complex nearshore scenarios, offering a robust solution for fine-grained ship detection.

Keywords: ship detection; fine-grained recognition; remote sensing images; complex background; information reconstruction

1. Introduction

The fine-grained detection and recognition of ships play a pivotal role in maritime surveillance, navigation safety maintenance, traffic environment improvement, and fishery management. With the synergistic integration of high-resolution remote sensing and deep learning technologies, the efficiency and precision of fine-grained ship recognition in aerial imagery have seen significant advancements [1–3]. Nevertheless, ship detection based on remote sensing images remains a formidable challenge due to the arbitrary orientations of targets, dense arrangements, and the interference from complex backgrounds [4,5]. In addition, the target in remote sensing images is small, the difference between similar classes is small, and the fine-grained detection effect is poor.

To tackle these issues, numerous deep learning-based detectors, ranging from classical Convolutional Neural Networks (CNNs) [6–8] to recent vision Transformers [9–11], have been adapted for maritime scenarios. While these methods work well in simple cases, they often struggle when used in real-world, changing sea environments. The performance bottleneck primarily stems from the following two challenges.

(a) Morphological diversity, fine-grained similarity, and background noise interference. Ships in remote sensing images exhibit significant intra-class shape variations and subtle inter-class fine-grained structural differences, making them highly susceptible to confusion with background clutter and noise, as illustrated in Figure 1a. Existing methods typically rely on fixed feature extraction strategies, lacking explicit shape priors to model the structural and textural distributions of different ship categories. Without such priors, models struggle to distinguish ships from similarly shaped noise patterns and fail to capture discriminative details between subcategories. Furthermore, fixed receptive fields cannot adaptively focus on target regions based on the shape and attitude of ships in complex scenes, often resulting in the missed detection of small objects and a large number of redundant false alarms.

(b) Information loss and non-selective enhancement during feature reconstruction. Another critical challenge lies in the irreversible information loss caused by successive down-sampling and up-sampling operations inherent in standard neural architectures. Although U-shaped structures are widely employed to capture multi-scale contextual features (as shown in Figure 1b), the continuous down-sampling process inevitably sacrifices fine-grained spatial details in favor of high-level semantic abstraction. Existing methods often restore features without enough selectivity. This means that they strengthen both useful object details and useless background information at the same time. Such an approach fails to specifically highlight the fine-grained features of objects. As a result, the reconstructed feature maps lack the necessary clarity, preventing the precise localization and fine-grained classification of ship objects.

To alleviate the aforementioned challenges, this paper proposes a Shape-Prior Dynamic Refinement Network (SPDR-Net). The thesis of this study is to demonstrate that, by

integrating Gaussian-mixture-based ship-shape priors, prior-guided selective dynamic kernels, and Diffusion-Supervised Wavelet Refinement, it is possible to learn noise-robust and discriminative structural representations of ships, which will allow for more accurate fine-grained ship localization and classification in complex remote sensing scenes. The main contributions are summarized as follows.

(1) A Gaussian Mixture Shape Representation (GMSR) mechanism is designed. This captures the ship's orientation through strip convolutions and uses a learnable Gaussian mixture model to represent the center, orientation, shape, and texture distribution of different ship categories, providing structural priors and discriminative region cues.

(2) Re-Parameterized Selective Dynamic Kernels (RSDKs) are developed. They utilize the shape priors of the GMSR and dynamically adjust the receptive field of different layers of the network, which enables the model to focus on the target ship against complex backgrounds and capture its morphological features more accurately.

(3) A Diffusion-Supervised Wavelet Refinement (DSWR) strategy is proposed. It introduces the noise-denoising adversarial learning strategy from the diffusion model into wavelet reconstruction to recover high-frequency details that are easily lost or disturbed by noise during up-sampling and down-sampling. This effectively enhances ship boundaries and fine-grained discriminative textures while suppressing noise.

(4) The experimental results on several public datasets demonstrate that our method achieves state-of-the-art performance, with detection mAPs reaching 79.30% on DOSR dataset, 97.30% on HRSC2016 dataset, 98.60% on SSDD dataset and 47.97% on FAIR1M dataset, respectively.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. Section 3 details the proposed methodology. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes this work.

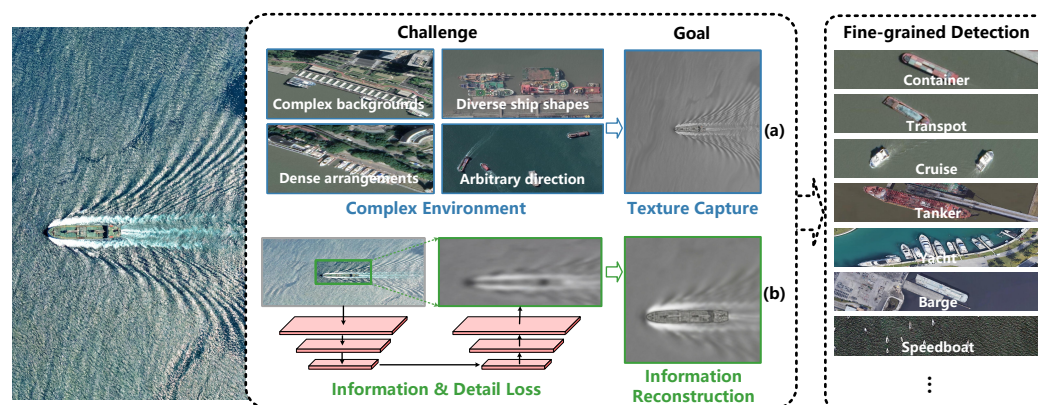


Figure 1. Ship detection faces two primary challenges: (a) complex background clutter significantly hinders feature extraction, and (b) deep sampling operations in the model inevitably lead to the loss of spatial details.

2. Related Works

2.1. Remote Sensing Ship Detection

Unlike object detection in natural scenes, remote sensing object detection is characterized by diverse scenarios and broad coverage. Furthermore, these images contain not only the targets but also numerous background objects. To address these challenges, existing methods have proposed various strategies. MLBR-YOLOX [12] and MSEIG-DETR [13] focus on removing vast sea areas and coastal backgrounds in SAR imagery while reducing sea clutter and other background distractions surrounding ships, thereby improving the detection of ships with varying sizes and orientations. CM-YOLO [14] utilizes a dual-path context enhancement neck to extract contextual information, such as location and shape,

to improve ship detection and classification in complex environments. To balance feature representation, CTF-Net [15] and CNN-SWin [16] combine Convolutional Neural Networks (CNNs) with Transformer architectures, enhancing the model to balance local feature extraction with global information modeling in SAR ship imagery. DCAPT-Net [17] introduces a densely connected attention pyramid architecture to strengthen multi-layer feature fusion, achieving effective multi-scale ship detection. SASOD [18] incorporates saliency information into the ship detection network to optimize saliency-aware features. Regarding specific scenarios, You et al. [19] proposed a DCNN-based end-to-end system designed for inshore ship detection, which excludes proposals in non-target regions to reduce false alarms on land. Finally, Yang et al. [20] developed a multi-task rotated region CNN model that predicts minimum bounding rectangles with reduced redundancy, making it particularly suitable for detecting dense ships. EFDNet [21] learns a more discriminative fine-grained feature representation for weakly supervised remote sensing object detection by decoupling localization and classification features. DRNet [22] uses dual refinement to learn more discriminative object features and better distinguish similar objects. PDDFormer [23] uses spatial gated differential attention to adaptively adjust the denoising strength at different spatial locations, so that the model suppresses noise while retaining key structural information. DiffBEV [24] utilizes a diffusion model to denoise noisy samples and progressively refine semantic features for object detection under a bird's-eye view. GDD [25] exploits the strong representation ability of diffusion models to extract robust and domain-invariant features through a progressive denoising process.

2.2. Information Restoration and Reconstruction

CNNs often use pooling and downsampling operations to reduce computation and increase the receptive field. However, these operations also have some limitations. Max pooling and average pooling both lose part of the original information. They can also affect feature consistency and introduce unwanted signals into the feature maps [26]. These problems hinder fine-grained feature extraction. To address these issues, Williams et al. [27] demonstrated that wavelet pooling retains significantly more information compared to max, average, mixed, and stochastic pooling strategies. MWCNN [28] integrates wavelet transforms into neural networks, utilizing multi-level wavelet decomposition for image restoration. WaveCNets [29] incorporate the Discrete Wavelet Transform (DWT) and Inverse Wavelet Transform (IWT) into general network layers for image classification tasks. DWSR [30] combines the information complementarity of the wavelet domain with deep convolutional networks, recovering information lost during the convolution process by processing sub-bands. Wavelet R-CNN [31] utilizes wavelet transforms to assist the detector in capturing frequency-domain information during feature extraction. FSSFFormer [32] utilizes multi-resolution wavelets to enable the model to preserve global spatial information while preserving fine details. FreDNet [33] uses the frequency domain intensity corresponding mapping to adaptively fuse the features of the encoder and decoder. Furthermore, HWANet [34] leverages low-frequency component enhancement and a frequency-domain self-attention mechanism to mitigate information loss during sampling and enhance context-aware modeling capabilities.

3. Proposed Method

As illustrated in Figure 2, the overall architecture of the proposed SPDR-Net is presented. First, the model adopts a U-shaped pyramidal encoder–decoder architecture, where both the encoder and decoder are constructed using Convolutional Neural Networks (CNNs), as detailed in Figure 2a,e. Second, GMSR uses learnable Gaussian mixture to make the model learn fine-grained features of ships (Figure 2b). Then, RSDKs used the learned

ship geometry prior to extract the target ship features (Figure 2c). Finally, DSWR added noise and denoised the extracted features, predicted and filtered high-frequency noise, and realized feature refinement (Figure 2d).

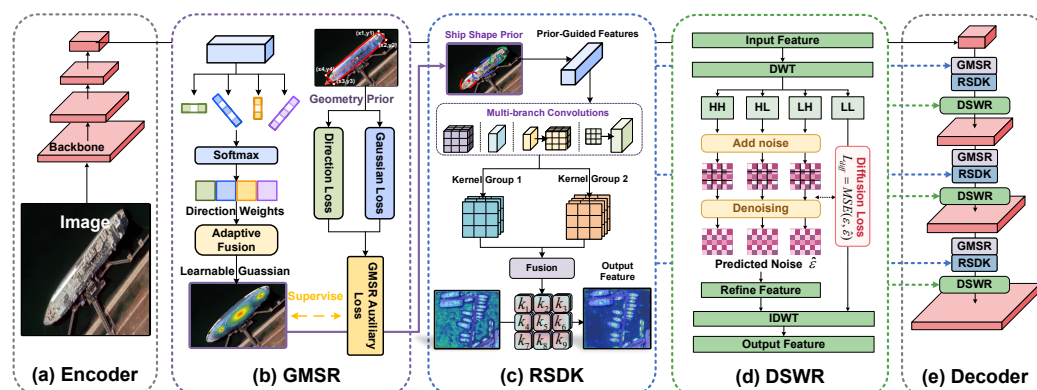


Figure 2. The SPDR-Net framework comprises (a) the CNN encoder, (b) the GMSR, (c) the RSDKs, (d) the DSWR, (e) the CNN decoder.

3.1. Gaussian Mixture Shape Representation Mechanism

Ship detection interference in remote sensing images easily occurs in complex environments. For example, the containers and objects on the port have similar structures with the buildings and shapes on the ship, and the model will have the situation of false detection. Traditional methods assign weights according to local salient regions. This will cause the model to pay too much attention to the building area on the ship, while ignoring the bow, stern and other information, and the detection box does not contain the whole ship. To alleviate these limitations, a Gaussian Mixture Shape Representation mechanism is proposed, as illustrated in Figure 3. The proposed module explicitly models ship structures from multiple orientations and introduces Gaussian mixture guidance to improve fine-grained feature extraction ability.

Given the input feature map $X \in \mathbb{R}^{C \times H \times W}$, a 1×1 convolution is first applied to obtain a compact feature representation $F = \phi(X)$, where $\phi(\cdot)$ denotes the Conv-BN-SiLU operation. To capture ship structures with arbitrary orientations, four directional structural encoders are employed to extract horizontal, vertical, and diagonal features:

$$\{F_h, F_{d1}, F_v, F_{d2}\} = \{\mathcal{B}_h(F), \mathcal{B}_{d1}(F), \mathcal{B}_v(F), \mathcal{B}_{d2}(F)\}. \quad (1)$$

An orientation estimation branch predicts pixel-wise weights for adaptive feature selection:

$$W = \text{Softmax}(\mathcal{G}(F)), \quad (2)$$

where $W \in \mathbb{R}^{4 \times H \times W}$. The directional fused feature is then used to generate a structural response map:

$$A = \sigma(\Psi(\sum_{i \in \{h, d1, v, d2\}} W_i \odot F_i)), \quad (3)$$

where $\Psi(\cdot)$ consists of a 3×3 Conv-BN-SiLU block followed by a 1×1 convolution, and $\sigma(\cdot)$ represents the sigmoid activation function. The response map highlights ship regions and suppresses background interference.

Finally, the fused feature is projected back to the original channel dimension and combined with the input through a residual connection:

$$Y = X + \gamma(\Omega(F_s) \odot A), \quad (4)$$

where $F_s \in R^{C_{mid} \times H \times W}$ denotes the fused feature map, and C_{mid} represents the channel index, $\Omega(\cdot)$ denotes a 1×1 Conv-BN-SiLU operation, and γ is a learnable scaling factor.

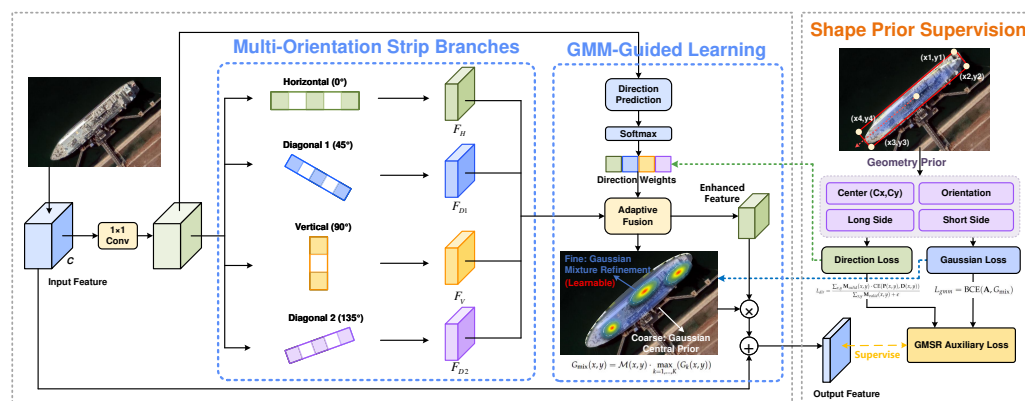


Figure 3. The architecture of the proposed GMSR. The strategy consists of: (1) multi-orientation strip branches; (2) GMM-guided learning; and (3) shape prior supervision.

To provide explicit geometric guidance, a Gaussian mixture supervision strategy is introduced. For each rotated ship, the rotated box is converted into a parametric geometric representation (c_x, c_y, l, s, θ) , where (c_x, c_y) denotes the ship center, l and s represent the lengths of the long and short axes, and $\theta \in [0, \pi)$ is the orientation angle of the long axis. For any pixel (x, y) in the feature map scale, its rotated coordinate (x_r, y_r) relative to the ship center is defined as

$$x_r = \frac{1}{s_x}(x \cdot s_x - c_x) \cos \theta + \frac{1}{s_y}(y \cdot s_y - c_y) \sin \theta, \quad (5)$$

$$y_r = -\frac{1}{s_x}(x \cdot s_x - c_x) \sin \theta + \frac{1}{s_y}(y \cdot s_y - c_y) \cos \theta, \quad (6)$$

where s_x and s_y denote the stride factors of the feature map relative to the original image in the horizontal and vertical directions, respectively.

Unlike conventional single-peak Gaussian labels, we construct a multi-component Gaussian mixture prior along the ship's long axis to capture fine-grained structural details (such as the bow, center, and stern). For a Gaussian mixture target with K components, each component $k \in \{1, \dots, K\}$ has a learnable offset o_k along the long axis and learnable standard deviation scales:

$$offsets_k = \alpha_k \cdot \frac{l}{2}, \quad (7)$$

$$\sigma_{l,k} = \sigma_{l,base} \cdot s_{l,k}, \quad \sigma_{s,k} = \sigma_{s,base} \cdot s_{s,k}, \quad (8)$$

where $\alpha_k \in [-0.85, 0.85]$ is the learnable relative displacement, and $s_{l,k}, s_{s,k} > 0.5$ are learnable scaling factors. The baseline standard deviations are computed as $\sigma_{l,base} = \max(l/(2.5 \cdot s_x), \sigma_{min})$ and $\sigma_{s,base} = \max(s/(2.5 \cdot s_y), \sigma_{min})$, where $\sigma_{min} = 1.0$ is the minimum variance boundary. The k -th anisotropic Gaussian component is formulated as

$$G_k(x, y) = \exp\left(-\frac{1}{2}\left(\frac{(x_r - offsets_k)^2}{\sigma_{l,k}^2} + \frac{y_r^2}{\sigma_{s,k}^2}\right)\right). \quad (9)$$

To restrict the supervision target strictly within the physical boundaries of the ship, the GMM components are aggregated and filtered by a hard boundary mask $\mathcal{M}(x, y)$:

$$G_{mix}(x, y) = \mathcal{M}(x, y) \cdot \max_{k=1, \dots, K} (G_k(x, y)), \quad (10)$$

where $\mathcal{M}(x, y)$ is an indicator function reflecting whether the pixel falls inside the rotated box:

$$\mathcal{M}(x, y) = \mathbf{I}\left(|x_r| \leq \frac{l}{2 \cdot s_x} \wedge |y_r| \leq \frac{s}{2 \cdot s_y}\right). \quad (11)$$

In addition, an auxiliary orientation consistency loss is introduced to improve orientation awareness. The orientation labels are divided into four categories corresponding to 0° , 45° , 90° , and 135° . Let $\mathbf{P} \in \mathbf{R}^{4 \times H \times W}$ denote the predicted orientation logits, and let $\mathbf{D} \in \{0, 1, 2, 3\}^{H \times W}$ represent the target category map. To focus optimization on the target regions, we define a binary validity mask $\mathbf{M}_{valid} \in \{0, 1\}^{H \times W}$, where $\mathbf{M}_{valid}(x, y) = 1$ if (x, y) falls inside any ground-truth ship box, and 0 otherwise.

The structural guidance loss is defined using Binary Cross-Entropy (BCE):

$$L_{gmm} = \text{BCE}(\mathbf{A}, \mathbf{G}_{\text{mix}}), \quad (12)$$

while the masked orientation consistency loss is formulated as

$$L_{dir} = \frac{\sum_{x,y} \mathbf{M}_{valid}(x, y) \cdot \text{CE}(\mathbf{P}(x, y), \mathbf{D}(x, y))}{\sum_{x,y} \mathbf{M}_{valid}(x, y) + \epsilon}, \quad (13)$$

where $\text{CE}(\cdot)$ represents the standard cross-entropy loss and $\epsilon = 10^{-6}$ is a small constant to prevent division by zero. The overall optimization objective of the GMSR module is therefore given by

$$L_{GMSR} = \lambda_g L_{gmm} + \lambda_d L_{dir}, \quad (14)$$

where λ_g and λ_d denote the balancing coefficients of the Gaussian mixture supervision and orientation consistency supervision, respectively.

Through the above processing, the model can perceive the ship area, and use the Gaussian mixture to model the target area, learn the structural fine-grained features of different types of ships, improve the fine-grained discrimination ability of the model ship and provide ship type prior for the subsequent model positioning features.

3.2. Re-Parameterized Selective Dynamic Kernels

After obtaining the prior and structural features of the ship type, the model also needs to be able to extract the ship features in the target area. Conventional convolutions use fixed kernels and cannot adapt to different ship structures. Existing attention mechanisms highlight important regions but do not change the feature extraction process. Dynamic convolution methods improve adaptability, but they often generate kernels without explicit structural guidance. To address these challenges, we propose the Re-Parameterized Selective Dynamic Kernels (RSDKs), as illustrated in Figure 4.

Instead of using standard static convolutions, RSDKs adaptively generates a selective dynamic kernel for each image. The shape and orientation of this generated kernel are guided by the ship's anisotropic Gaussian, allowing the model to select appropriate structural features and remove background noise. Furthermore, we use structural re-parameterization to keep the inference computationally efficient.

Let $\mathbf{X} \in \mathbf{R}^{C \times H \times W}$ be the input feature map, where H, W , and C define the height, width, and channel dimensions. The processing flow is formulated as follows:

First, we prioritize informative channels using a squeeze-and-excitation mechanism. The channel-enhanced feature $\mathbf{X}_{ca}$ is processed to obtain the baseline representation $\mathbf{X}_{mc}$:

$$\mathbf{X}_{mc} = \mathcal{F}_{mc}(\mathbf{X}_{ca}) = \mathcal{F}_{mc}(\mathbf{X} \odot \sigma(\text{MLP}(\text{AvgPool}(\mathbf{X})))), \quad (15)$$

where $\odot$ represents channel-wise multiplication, $\sigma(\cdot)$ is the sigmoid activation function, and $\text{MLP}(\cdot)$ captures cross-channel interactions.

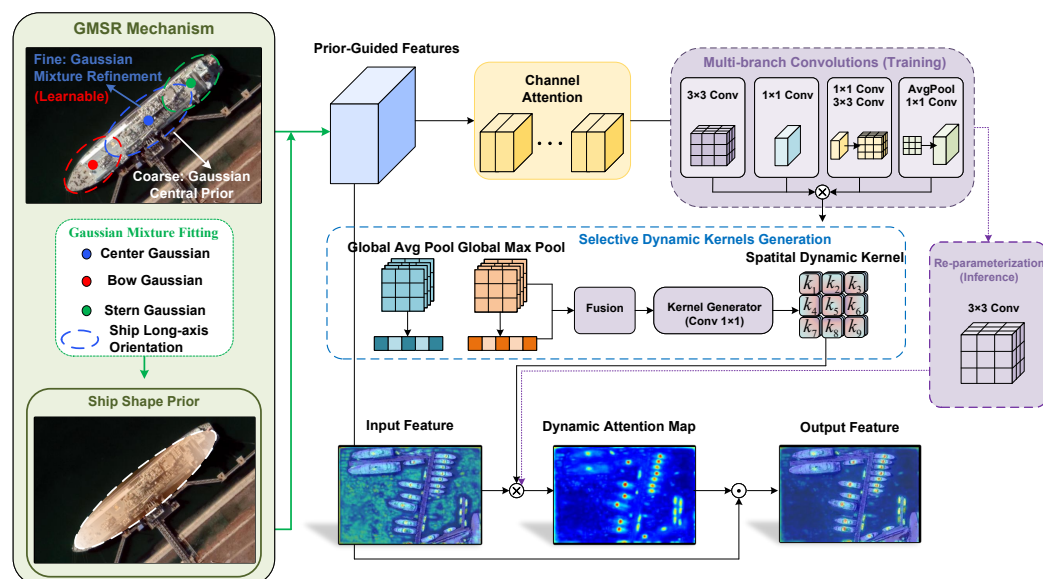


Figure 4. The architecture of the proposed RSDKs. The strategy consists of three stages: (1) channel attention for feature calibration; (2) re-parameterized convolutions which simplify to a single 3×3 convolution during inference; and (3) dynamic spatial attention, where a dynamic kernel is generated under the guidance of the anisotropic Gaussian shape to capture ship-specific structures. The residual connection preserves the original information.

During inference, these convolutions are simplified into a single 3×3 convolutional layer through structural re-parameterization to keep the model fast:

$$W_{eff} = \sum_{i=1}^n \text{pad}(W_i \otimes \gamma_i), \quad b_{eff} = \sum_{i=1}^n (\beta_i - \frac{\mu_i \gamma_i}{\sigma_i}), \quad (16)$$

where W_{eff} and b_{eff} denote the effective weight kernel and bias vector for inference, respectively. Consequently, the inference-time output is computed as

$$X_{mc} = W_{eff} * X_{ca} + b_{eff}, \quad (17)$$

where $*$ denotes the standard 3×3 convolution operator.

Unlike standard spatial attention mechanisms that use fixed, isotropic kernels, RSDKs focuses on generating a sample-specific dynamic kernel K_{dyn} that adapts to the ship's geometric shape. To achieve this, the kernel generator is optimized under the guidance of the anisotropic Gaussian mixture target (G_{mix}) from the previous module.

We first fuse the global average-pooled and max-pooled features of X_{mc} to capture the scale and orientation of the target:

$$K_{dyn} = \mathcal{G}([\text{AvgPool}(X_{mc}); \text{MaxPool}(X_{mc})]), \quad (18)$$

where $[\cdot]$ represents concatenation along the channel dimension. The generator $\mathcal{G}(\cdot)$ (implemented as a convolution layer) projects these pooled features to generate the weights of the dynamic kernel K_{dyn} .

The spatial attention map M_s is then generated by convolving the channel-averaged feature map with the dynamic kernel K_{dyn} :

$$M_s = \sigma \left(\left(\frac{1}{C} \sum_{c=1}^C X_{mc}^c \right) \circledast K_{dyn} \right), \quad (19)$$

where $M_s \in \mathbf{R}^{1 \times H \times W}$ is the spatial attention map, and the symbol $\circledast$ denotes the dynamic convolution using K_{dyn} .

Because M_s is supervised by the Gaussian mixture target G_{mix} during training, the generator $\mathcal{G}(\cdot)$ is forced to shape K_{dyn} to match the ship's physical geometry. Specifically, the generated kernel K_{dyn} stretches along the ship's long axis to aggregate structural features, while shrinking along the short axis to block background noise.

Finally, the output feature Y is obtained via a residual connection:

$$Y = X + (X_{mc} \odot M_s), \quad (20)$$

where $Y \in \mathbf{R}^{C \times H \times W}$ is the final refined feature map. This strategy enables the network to focus on key spatial regions using the Gaussian-aligned dynamic kernel while preserving original details.

3.3. Diffusion-Supervised Wavelet Refinement Strategy

After accurately extracting the features of the ship, the features include high-frequency features such as ship edges and detailed textures, as well as high-frequency noise. Traditional denoising methods weaken the edge details while filtering the noise. To separately enhance low-frequency structural features and high-frequency detailed textures, we introduce the Diffusion-Supervised Wavelet Refinement mechanism, as illustrated in Figure 5. Let $X \in \mathbf{R}^{B \times C \times H \times W}$ denote the input feature map, where B is the batch size, C represents the channel dimension, and H and W represent the spatial height and width, respectively. We first apply the Haar Discrete Wavelet Transform (DWT) to decompose X into its low-frequency subband X_{LL} and high-frequency subbands X_{HF} :

$$(X_{LL}, X_{HF}) = \text{DWT}(X) \quad (21)$$

where $X_{LL} \in \mathbf{R}^{B \times C \times \frac{H}{2} \times \frac{W}{2}}$ preserves the structural outline, and $X_{HF} \in \mathbf{R}^{B \times 3C \times \frac{H}{2} \times \frac{W}{2}}$ contains the concatenated high-frequency details.

For the low-frequency component X_{LL} , we apply a lightweight residual block using depthwise separable convolutions to refine the structural features:

$$X'_{LL} = X_{LL} + \sigma(\text{PWConv}(\text{DWConv}(X_{LL}))) \quad (22)$$

where X'_{LL} is the refined low-frequency feature, DWConv is a depthwise convolution with a 3×3 kernel, PWConv is a pointwise convolution with a 1×1 kernel, and σ denotes the SiLU activation function. The weights of PWConv are initialized to zero, which preserves the original features at the beginning of training.

Meanwhile, we employ a diffusion-based denoiser to reconstruct clean detail features from the high-frequency component X_{HF} . Unlike conventional generative diffusion models, the denoiser is inspired by diffusion models and is designed to predict noise through a noise-adding and denoising process. Let T be the total diffusion timesteps. We establish a noise schedule with cumulative variance parameters $\bar{\alpha}_t$ for $t \in [0, T - 1]$. During training,

we randomly sample a timestep $t \sim \text{Uniform}(0, T-1)$ and generate a noisy high-frequency subband X_{HF}^t by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$:

$$X_{HF}^t = \sqrt{\bar{\alpha}_t} X_{HF} + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (23)$$

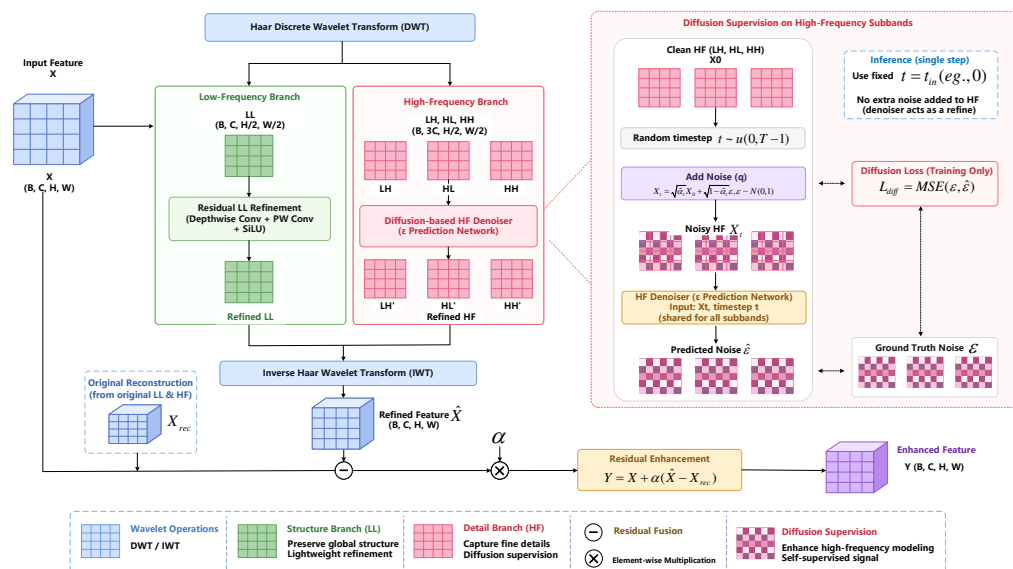


Figure 5. The architecture of the proposed DSWR. The strategy consists of: (1) 2D discrete wavelet decomposition and reconstruction; (2) diffusion-based detail restoration and enhancement.

The timestep t is mapped to a continuous embedding $e_t \in \mathbf{R}^d$ using a sinusoidal function followed by a multi-layer perceptron (MLP), such that $e_t = \text{MLP}(\text{Sinusoidal}(t))$. A high-frequency denoiser network f_θ is then trained to predict the injected noise:

$$\epsilon_\theta = f_\theta(X_{HF}^t, e_t) \quad (24)$$

where f_θ represents the denoiser network, and ϵ_θ is the predicted noise. The output layer of f_θ is initialized to zero to guarantee stable training trajectories.

During inference or forward denoising, model has already learned adaptive denoising and its parameters are fixed, additional iterative denoising provides no substantial benefit but increases inference time. Therefore, we reconstruct the refined high-frequency features X'_{HF} using a single-step denoising formulation:

$$X'_{HF} = \frac{X_{HF}^t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta}{\sqrt{\bar{\alpha}_t}} \quad (25)$$

In our implementation, we set a fixed step $t = t_{\text{infer}}$ (by default, $t_{\text{infer}} = 0$) during evaluation to directly refine the clean high-frequency input features without iterative loops.

Following the refinement of both subbands, we reconstruct the spatial feature maps using the Haar Inverse Wavelet Transform (IWT). To maintain strict feature consistency, we compute a baseline feature map X_{base} and a refined feature map X_{refined} . The final output feature map $Y \in \mathbf{R}^{B \times C \times H \times W}$ is formulated as a scaled residual modification added to the original input X :

$$Y = X + \alpha \cdot (\text{IWT}(X'_{LL}, X'_{HF}) - \text{IWT}(X_{LL}, X_{HF})) \quad (26)$$

where α is a learnable scalar parameter initialized to zero. This formulation mathematically guarantees that the module acts as an identity mapping at the start of training.

To train the entire framework end-to-end, the auxiliary diffusion loss $\mathcal{L}_{\text{diff}}$ is calculated as the mean squared error (MSE) of the noise prediction:

$$\mathcal{L}_{\text{diff}} = \frac{1}{B \cdot 3C \cdot \frac{H}{2} \cdot \frac{W}{2}} \sum \|\epsilon - \epsilon_{\theta}\|^2 \quad (27)$$

The total loss objective $\mathcal{L}_{\text{total}}$ of the network combines the primary task loss $\mathcal{L}_{\text{task}}$ and the auxiliary diffusion loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{diff}} \quad (28)$$

where λ represents a scaling hyperparameter that balances the contribution of the diffusion loss. This module can filter the noise while recovering the high-frequency features by fitting the learning noise through a diffusion model.

Specifically, GMSR learns fine-grained structural features of ships and captures geometric information through Gaussian mixture. Furthermore, RSDKs uses the extracted geometric prior to dynamically adjust the receptive field to capture features. Finally, DSWR uses the diffusion model to further refine the extracted features from the frequency domain perspective. Thus, the fine-grained feature extraction ability of SPDR-Net is enhanced, and the accuracy of ship detection and recognition is improved.

4. Experiment and Discussion

In this section, extensive experiments are conducted on several datasets to evaluate the performance of SPDR-Net comprehensively. First, the experimental conditions, including datasets, evaluation metrics, implementation details, etc., are introduced. Then, ablation experiments are designed to validate the effectiveness of each component of the SPDR-Net model and to evaluate their performance quantitatively. We discuss in detail the key issues in ship detection. Based on this, we compare the proposed model with several representative object detection methods proposed in recent years and provide a comprehensive evaluation. Finally, we further evaluate the generalization capability of the proposed model on the public datasets.

4.1. Experimental Conditions

(1) Experimental platforms. The experiments are performed on a server with four NVIDIA GeForce RTX 4090 GPU (24 GB). We used an Auto optimizer with learning rate set to 8.6×10^{-4} and weight decay set to 0.01 via grid search. The maximum training epoch is 100 to accommodate the full convergence potential of models. To ensure fairness, identical hyperparameters were applied to all baseline methods and eliminated tuning biases.

(2) Datasets. To evaluate the performance of the SPDR-Net model more comprehensively, several public datasets are used, which cover different scenes, different shapes and categories of objects, and different data sources.

- The DOSR [35] dataset provides a widely used benchmark to evaluate the performance of ship object detection methods in remote sensing scenes. It has more than 6127 ship objects covering 20 categories in 1066 images.
- HRSC2016 [36] is a ship detection dataset consisting of 436 training images, 181 validation images, and 444 testing images from 300×300 pixels to 1500×900 pixels.
- SSDD [37] is a Synthetic Aperture Radar (SAR) remote sensing image dataset. It contains a total of 1160 SAR images of ships under various sea conditions, acquired by RadarSat-2, TerraSAR-X, and Sentinel-1 sensors, with a total of 2456 annotated ship targets.
- The DOTA v1.0 [38] dataset has more than 188,000 objects covering 15 categories in 2806 images from 800×800 pixels to 4000×4000 pixels. Due to the huge size of remote

sensing images, they are usually cropped into sub-images of 800×800 pixels with an overlap of 200 pixels on each dimension after being scaled to different sizes with the ratios of 0.5, 1.0, and 1.5.

- The FAIR1M [39] dataset provides a large-scale benchmark for fine-grained object detection and recognition in high-resolution remote sensing imagery. It contains more than 1.02 million object instances covering 37 fine-grained categories (under five coarse categories) in 15,266 images with widths ranging from 1000 to 10,000 pixels. The dataset is divided into three subsets, where 1/2 of the images are used for training, 1/6 are used for validation, and 1/3 are used for testing.

(3) Evaluation metrics. The mean Average Precision (mAP) is adopted to evaluate the detection accuracy. The mAP with an IoU threshold of 50% is represented as mAP_{50} . The $mAP_{50:95}$ means calculating the mAP when the IoU threshold is 50% to 95% at every 5% interval and then taking their mean as $mAP_{50:95}$. The speed is evaluated by frames per second (fps).

4.2. Ablation Experiments

The study uses YOLO v8 as the baseline model. To systematically evaluate the contribution of each component, a series of ablation experiments were conducted. The results of these experiments for different modules of SPDR-Net are summarized in Table 1. The proposed RSDKs, GMSR, and DSWR improved the model's mAP_{50} . The complete SPDR-Net model achieved gains of 9.9% in mAP_{50} compared to the baseline, demonstrating consistently strong performance. Furthermore, we compared the Gaussian Guided Structural Representation module (GGSr) with GMSR, Diffusion Wavelet Guided Refinement module (DWGR) with DSWR, and our own module with the single component, respectively. Unlike DSWR, DWGR is not supervised by L_{diff} loss. GMSR achieved superior performance over GGSr with mAP_{50} increased by 2.3%. Similarly, DSWR outperformed DWGR with mAP_{50} by 2.71%, which shows that L_{diff} provides effective auxiliary supervision for adaptive feature denoising, enabling the model to suppress high-frequency noise while preserving discriminative boundary and texture details. SPDR-Net improves the mAP_{50} by 6.57%, 4.5%, and 6.09% compared with RSDKs, GMSR, and DSWR, respectively. Table 1 also reports the parameters, GFLOPs, and inference speed of each configuration. The complete SPDR-Net contains 89.6 M parameters and requires 247.5 GFLOPs, while achieving an inference speed of 83.6 fps. Although the multi-orientation feature extraction and shape-guided refinement modules increase the theoretical computational complexity, RSDKs merge their multi-branch structure into a single convolutional branch through structural re-parameterization during inference. In addition, L_{diff} is used only during training, and DSWR adopts fixed single-step denoising during inference, avoiding the overhead of iterative reverse diffusion. Therefore, SPDR-Net maintains real-time inference performance while substantially improving detection accuracy, demonstrating a favorable balance between accuracy and computational efficiency.

The proposed modules significantly enhance the model's capabilities for ship detection. Specifically, the GMSR uses Gaussian mixture to model the hull and edge structural features, which effectively improves the learning ability of the model for the fine-grained structural features of the ship. Further, the RSDKs uses the Gaussian mixture geometric prior learned by the model to accurately locate the ship area and capture the ship characteristics through a selective dynamic kernel. Finally, DSWR uses the diffusion model to refine the high-frequency part of the extracted features. Finally, through the synergy of these three modules, the proposed SPDR-Net achieves superior detection performance in ship target detection in remote sensing scenes (69.4% vs. 79.3%), which improves the classification performance

of the model and eliminates false alarms. Next, we further analyze the GMSR, RSDKs and DSWR.

Table 1. Ablation study of each component on the DOSR dataset.

Methods	RSDKs	GMSR	GWGR	DSWR	L_{diff}	mAP_{50} (%)	Parameters (M)	FLOPs	Speed (fps)
Baseline						69.40	41.3	153.4	41.06
Single Component	✓					72.73	65.8	228.8	67.34
		✓				72.50	53.5	197.3	49.08
			✓			74.80	45.3	171.2	46.40
				✓		70.50	44.3	163.4	43.65
					✓	73.21	47.7	183.4	48.32
Component Combination	✓			✓		71.20	66.8	203.5	44.39
		✓		✓		73.25	76.5	214.8	43.92
			✓	✓		73.83	78.9	212.6	45.77
	✓		✓			74.80	83.2	224.5	50.34
	✓		✓	✓		75.70	85.3	235.6	55.67
Ours	✓		✓		✓	79.30	89.6	247.5	83.6

Note: Bold indicates the best result. “✓” indicates using this component.

(1) Analysis of the GMSR. The GMSR mechanism predicts the direction of the ship through the strip convolution group and uses the Gaussian mixture to model along the main axis of the ship, and supervised by Ground Truth to make the model learn the fine-grained structural features of the ship. To further verify the effectiveness of the GMSR, we present a comparative visualization of feature maps with and without its integration, as illustrated in Figure 6. “Single” stands for using only one Gaussian for fitting ship structure features, and “GMM” stands for using mixture Gaussians for fitting. We compute the single Gaussian and mixed Gaussian target ROI, target ROI represents the overlap ratio between the region of interest and the target region. It can be clearly found that using mixture Gaussian can focus on more detailed features such as the bow and stern of the ship. Furthermore, we compare the differences between “single” and “gmm”, which can reflect that GMSR can learn more detailed and complete structural features by using Gaussian mixture model than the traditional single Gaussian model. GMSR makes the model pay attention to the details such as the masts on the ship, and enhances the ability of the network to capture the characteristics of the ship and learn the characteristics of fine-grained structural differences. It not only prevents the model from misdetection caused by similar characteristic buildings on the shore, but also enhances the integrity of the representation of the ship structure, which effectively alleviates the problem of classification ambiguity of different ships.

(2) Analysis of the RSDKs. The RSDKs use the geometric prior of the hull structure learned by the GMSR module as a guide, use the multi-branch convolution architecture to capture the context features, and localize the ship target by dynamic denoising to enhance the ability of the model to extract cross-scale features. To further validate the efficacy of the RSDKs, we present a comparative visualization of feature maps with and without its integration, as illustrated in Figure 7. The results demonstrate that the RSDKs enhances the adaptive feature extraction ability of the target ship under the guidance of ship geometry prior. It not only improves the detection ability of small and medium-sized ships in complex environments, but also enhances the ability to dynamically capture features, and effectively reduces the positioning error of the model.

(3) Analysis of the DSWR. The DSWR strategy decomposes the located ship features into frequency domain for feature refinement by discrete wavelet transform. The diffusion model is used to add noise and denoise high-frequency features while retaining low-frequency features of the ship body, so that the network can further learn to predict

noise. Then, the fine frequency domain features are reconstructed. The model can filter high-frequency noise while retaining high-frequency features such as ship edges and corners, which improves the feature expression ability. To evaluate the efficacy of the DSWR, we compare the feature maps before and after its integration, as illustrated in Figure 8. We can clearly find that by adding noise and denoising, the model focuses more on the refined features of the ship, and by calculating the interpolation of the refined features and the denoised features, the red represents the enhanced area of the model, and the black represents the suppressed area of the model. Obviously, DWSR suppresses noise interference and enhances fine-grained features when reconstructing features, which improves the recognition ability of the model.

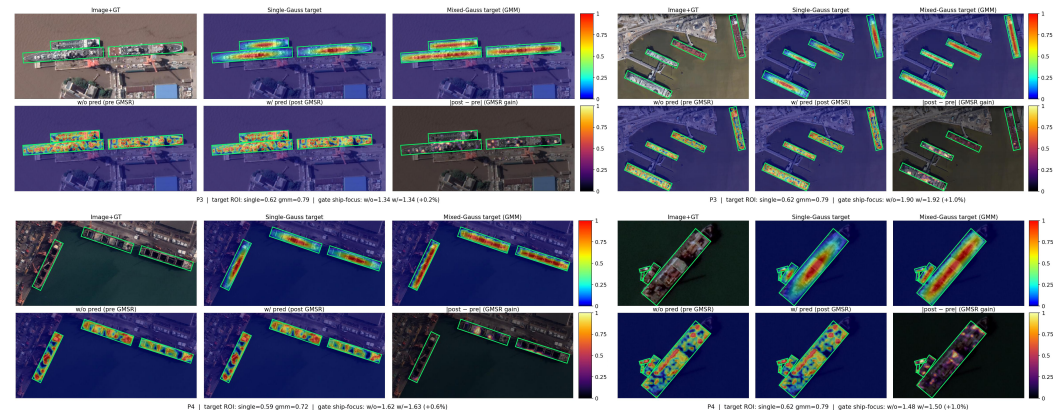


Figure 6. Visualization of the feature maps. “GMSR gain” represents the fine-grained features that the GMSR module emphasizes for model learning.

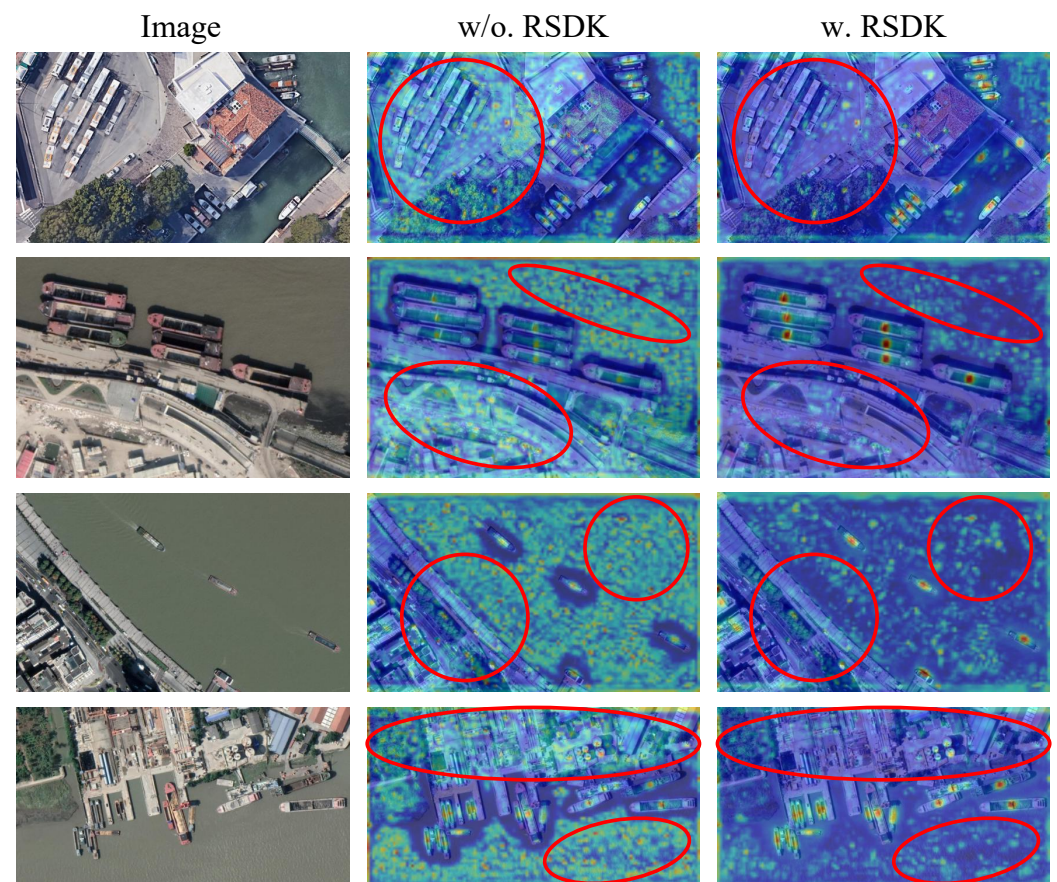


Figure 7. Visualization of the RSDKs feature maps. The red circles mark some key feature regions for comparison.

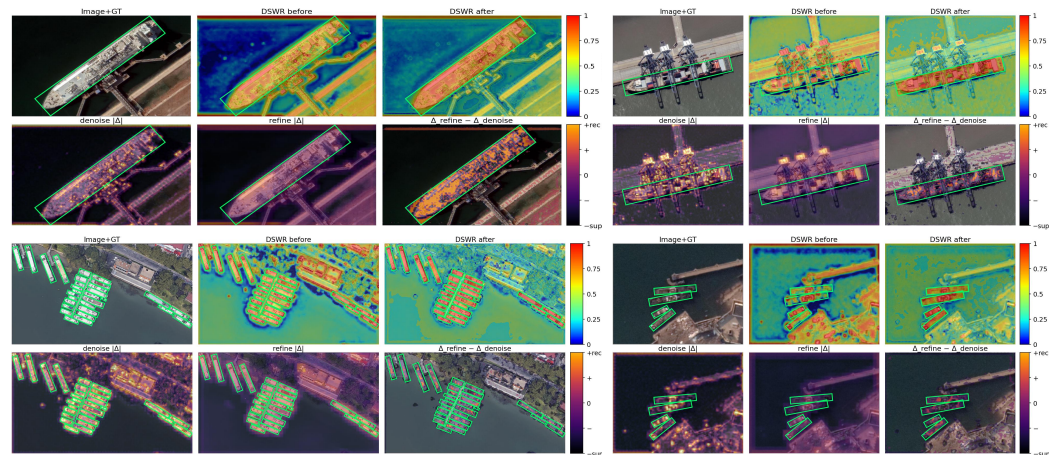


Figure 8. Visualization of the feature maps. “ $\Delta_{\text{refine}} - \Delta_{\text{denoise}}$ ” shows DWSR refined features (red region) and suppressed noise (black region). “rec” indicates that this area should be reconstructed, “sup” indicates that this area should be suppression.

4.3. Comparison Experiments

To evaluate the effectiveness of the proposed SPDR-Net, we compared it with the popular methods on several public datasets, including DOSR, HRSC2016, SSDD, DIOR, DOTA and FAIR1M.

(1) Comparison experiments of ship detection. The performance comparison of DOSR is shown in Table 2. We compare the performance of the proposed SPDR-Net with the eight existing approaches, including R2CNN (+19.3%) [40], RRPN (+14.3%) [41], SCRDet (+10.6%) [42], RetinaNet-O (+28.2%) [43], R3Det (+12.2%) [44], RSDet (+19.7%) [45], SCRDet++ (+8.6%) [46], and ReDet (+7.6%) [47]. Among them, FR-FPN-O [48], R2CNN [40], RRPN [41], and RetinaNet-O [43] are general-oriented object recognition networks, and the rest five approaches, i.e., SCRDet [42], R3Det [44], RSDet [45], SCRDet++ [46], and ReDet [47], are all designed for remote sensing object recognition. In more detail, R2CNN [40], RRPN [41], SCRDet [42], and ReDet [47] adopt the two-stage strategy, and the single-stage networks involve RetinaNet-O [43], R3Det [44], RSDet [45], and SCRDet++ [46]. It is obvious that our SPDR-Net achieves the superior performance on the DOSR dataset. For example, it demonstrates high detection accuracy across various vessel categories, including nearshore vessels (such as warships and deck ships), small-target vessels (such as fishing boats and yachts), and various cargo ships.

Furthermore, we compared the performance of the proposed method with existing advanced methods using the HRSC2016 and SSDD datasets, which are shown in Tables 3 and 4. HRSC2016 is a visible-light remote sensing image dataset, while SSDD is a Synthetic Aperture Radar (SAR) dataset. Experiments conducted on datasets of different modalities demonstrate that the proposed method outperforms latest methods. On the HRSC2016 dataset, SPDR-Net achieved mAP_{50} , mAP_{75} , and $mAP_{50:95}$ scores of 98.1%, 88.1%, and 79.0%, respectively, improving of 6.7%, 15.6%, and 17.5% over the latest method CM-YOLO [14]. Compared to existing detectors, SPDR-Net outperformed DDQ, Deformable DETR, and FCOS by 7.3%, 8.0%, and 9.6% at mAP_{50} . The significant gains in mAP_{75} and $mAP_{50:95}$ demonstrate that RSDKs and DSWR not only improved recall but also optimized localization accuracy under stricter IoU criteria (such as mAP_{75} and $mAP_{50:95}$). It is crucial to distinguish between densely moored ships and those with extreme orientations. On the SAR-based SSDD dataset, SPDR-Net achieved an mAP_{50} of 97.7%, outperforming YOLOv3 (+2.7%), CenterNet (+4.2%), DAPN (+7.6%), and CLAFANet (+6.4%). These results indicate that DSWR’s frequency-aware denoising mechanism effectively suppressed SAR speckle noise and sea clutter interference, while RSDKs’s dynamic receptive field

enhanced responsiveness to small and low-contrast ships. These progresses on optical and SAR benchmarks shown in Figure 9. It confirms that SPDR-Net generalizes well to detect objects on different modalities, which remains robust under complex backgrounds and scale variations.

Table 2. Comparison of performance (AP) on DOSR.

Category	R2CNN [40]	RRPN [41]	SCRDet [42]	Retina-O [43]	R ³ Det [44]	SCRDet++ [46]	RSDet [45]	ReDet [47]	SPDR-Net (Ours)
BCV.	56.5	62.9	65.2	49.5	65.6	61.2	55.3	64.3	85.7
Fis.	36.9	42.8	54.3	26.2	44.8	44.5	10.8	43.2	61.3
DeB.	38.8	33.2	44.0	36.9	28.5	36.1	22.8	27.9	57.2
Yac.	57.1	47.2	53.3	50.6	64.2	67.6	59.8	75.7	84.0
FTS.	26.2	43.7	49.5	27.1	49.1	62.8	52.6	65.1	88.2
Mul.	54.6	52.5	38.3	35.0	53.0	61.6	47.3	41.6	78.4
Tug	59.0	74.6	66.7	70.5	68.2	81.1	58.8	80.1	91.7
Com.	32.0	27.3	40.7	49.7	38.8	66.1	63.6	37.7	80.7
Spe.	39.3	31.2	32.2	33.7	36.0	63.5	57.2	68.0	85.8
Car.	76.7	83.3	86.7	68.7	84.2	76.3	75.3	85.1	91.3
Cru.	53.0	57.5	61.5	35.9	63.4	66.1	39.6	62.4	71.2
Flo.	5.9	33.1	22.0	11.9	34.4	13.9	16.9	37.5	50.2
Tan.	50.9	64.9	87.7	38.1	71.6	77.0	33.1	83.4	89.9
DeS.	42.4	20.6	32.8	2.0	26.7	59.0	14.2	36.6	62.8
Sub.	13.6	54.6	54.6	0.0	57.6	32.7	45.7	34.9	91.2
Con.	75.6	78.6	80.6	56.0	74.8	70.0	68.8	76.6	78.3
Bar.	46.1	53.0	49.9	12.2	42.7	18.8	26.2	44.3	89.6
Tra.	66.9	78.3	73.7	66.3	74.0	76.8	75.3	86.0	95.9
Aux.	43.5	36.1	42.8	29.7	27.2	47.3	34.0	38.6	62.8
Mil.	36.9	36.6	49.5	34.1	48.5	44.2	45.6	57.5	89.4
<i>mAP</i> ₅₀ (%)	45.6	50.6	54.3	36.7	52.7	56.3	45.2	57.3	79.3
FPS	2.84	2.71	3.84	3.71	4.90	2.83	2.77	12.04	9.83

Note: Bold indicates the best result. Ship category abbreviations follow the DOSR dataset [35]: BCV., bulk cargo vessel; Fis., fishing boat; DeB., deckbarge; Yac., yacht; FTS., flat traffic ship; Mul., multihull; Tug, tug; Com., communication ship; Spe., speedboat; Car., cargo; Cru., cruise; Flo., floating crane; Tan., tanker; DeS., deckship; Sub., submarine; Con., container; Bar., barge; Tra., transport; Aux., auxiliary ship; Mil., military ship. FPS, frames per second.

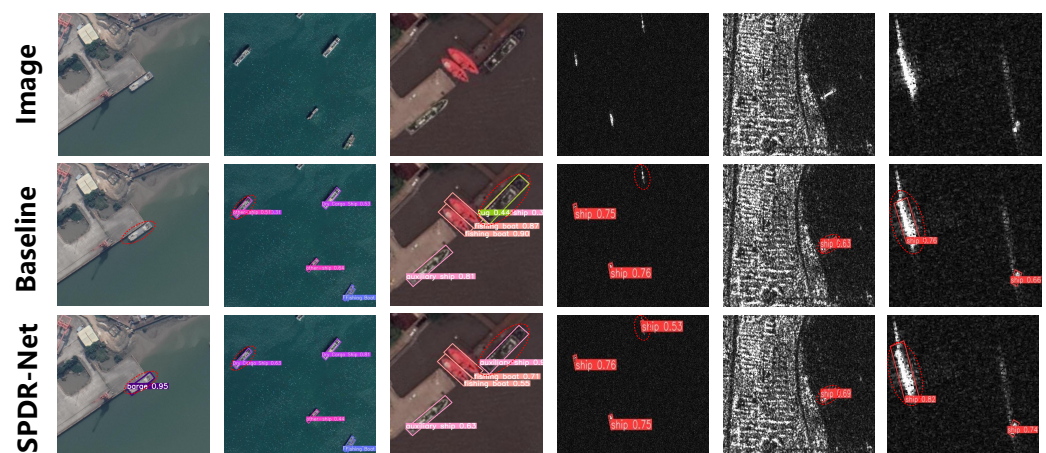


Figure 9. Detection results of fine-grained ships.

In addition, to conduct a more in-depth comparative analysis of the proposed method's fine-grained detection performance, we performed additional comparative experiments on FAIR1M, a larger remote sensing dataset with broader category coverage as shown in Figure 10, with the results presented in Table 5. Overall, SPDR-Net achieved the best mAP of 47.97% with a 3.61% improvement over the latest method PFA-C2FI (44.36%). Regarding fine-grained ship categories, SPDR-Net demonstrated superior performance

across most sub-categories, securing the best AP results for speedboats (73.37%), fishing boats (17.48%), liquid cargo ships (31.47%), and warships (44.61%). Notably, compared to the second-best methods, SPDR-Net improved detection performance for speedboats and warships by +10.06% and +9.45%, respectively. It indicates that our method has robustness in fine-grained ship recognition within complex offshore and near-shore environments. GMSR uses Gaussian mixture to learn the structural details such as the bow, mast, corner of the warship and the fuselage, craft, tail of the aircraft. Then RSDKs takes advantage of the different structural features obtained and adjusts the receptive field to better fit the target shape.

Beyond ships, SPDR-Net also achieved leading or highly competitive detection accuracy across other coarse-grained categories, including aircraft (e.g., B777 at 55.39%, B787 at 62.61%, A350 at 78.59%), vehicles (e.g., buses at 43.11%, vans at 67.39%), sports venues (tennis courts at 90.60%), and road infrastructure (intersections at 65.65%). These cross-category performance gains demonstrate that the proposed method is not limited to ships but effectively enhance discriminative feature learning and detail texture preservation for various fine-grained objects, which validates the general effectiveness of the proposed method in fine-grained detection tasks.

Table 3. Comparison of performance (AP) on HRSC2016.

Method	Year	Backbone	mAP_{50} (%)	mAP_{75} (%)	$mAP_{50:95}$ (%)
Faster R-CNN [48]	2017	R50-FPN	84.3	50.2	48.1
Libra R-CNN [49]	2019	R50-FPN	79.4	43.2	43.6
SSD [50]	2016	VGG-16	79.5	42.1	42.7
YOLOv3 [51]	2018	DarkNet-53	80.1	51.5	47.7
YOLOv5 [51]	2018	CSPDarknet	87.7	71.7	60.4
YOLOv9 [52]	2024	GELAN	73.5	46.8	43.0
YOLOv10 [53]	2024	CSPDarknet	89.3	68.8	58.4
FCOS [54]	2020	R101-FPN	88.5	61.0	55.3
YOLOX [55]	2021	CSPDarknet	85.7	63.5	55.0
DAB-DETR [56]	2022	R50-FPN	89.9	70.7	59.0
Deformable DETR [57]	2021	R50-FPN	90.1	60.5	53.1
DDQ [58]	2023	R50-FPN	90.8	60.3	54.8
DINO [59]	2022	R50-FPN	74.0	48.6	44.4
CM-YOLO [14]	2025	CSPDarknet	91.4	72.5	61.5
SPDR-Net (ours)	2026	CSPDarknet	98.1	88.1	79.0

Note: Bold indicates the best result.

Table 4. Comparison of performance (AP) on SSDD.

Methods	Year	Backbone	mAP_{50} (%)
RetinaNet-O [43]	2017	R50-FPN	89.6
Faster RCNN-O [48]	2017	R50-FPN	88.3
FCOS [54]	2020	R50-FPN	88.7
YOLOv3 [51]	2018	DarkNet-53	95.0
Cascade R-CNN [60]	2021	R50-FPN	89.5
Libra R-CNN [49]	2019	R50-FPN	89.9
SSD [50]	2016	VGG-16	94.0
CenterNet [61]	2019	HG-104	93.5
DAPN [62]	2019	R101-FPN	90.1
ARPN [63]	2020	R50-FPN	89.9
YOLOv7 [64]	2022	ELAN	93.7
OrientedFormer [65]	2024	R50-FPN	91.2
CLAFANet [66]	2025	R50-FPN	91.3
SPDR-Net (ours)	2026	CSPDarkNet	97.7

Note: Bold indicates the best result.

Table 5. Comparison of performance (AP) on FAIR1Mv1.0 dataset.

	Methods	Retina [43]	Faster [48]	RoI [67]	Casc. [60]	Glid. [68]	CHOD [69]	PCL [70]	PET [71]	SFR [72]	DR [22]	SRCNN [73]	MSC-TF [74]	PFA [75]	SPDR-Net (Ours)
Ship	Pass.	9.57	9.81	14.31	12.10	9.12	12.34	17.43	13.29	17.44	16.74	19.68	19.38	21.01	19.17
	Mot.	22.55	28.78	28.07	28.84	23.34	29.34	56.85	61.94	60.61	58.30	62.21	62.11	63.31	73.37
	Fis.	1.33	1.77	1.03	0.71	1.23	1.71	7.88	8.90	8.50	8.08	12.44	12.69	12.90	17.48
	Tug.	16.37	17.65	14.32	15.35	15.67	17.77	36.70	37.88	34.87	32.52	34.97	39.23	39.55	35.45
	Eng.	19.11	16.47	15.97	18.53	15.43	17.72	10.97	10.93	12.55	11.97	13.76	11.62	15.41	14.65
	LCS	14.26	16.19	18.04	14.63	15.32	16.78	19.91	22.05	19.71	22.07	24.68	23.76	24.96	31.47
	DCS	24.70	27.06	26.02	25.15	25.43	27.51	40.05	36.82	38.25	37.74	39.20	41.10	41.31	28.70
	War.	15.37	13.16	12.97	14.53	13.56	13.67	23.80	23.98	22.00	24.79	34.16	35.16	35.16	44.61
Airplane	B737	38.46	36.43	39.58	40.42	35.43	40.23	35.96	41.05	39.70	39.05	41.17	43.41	43.05	43.00
	B747	55.36	50.68	73.56	52.86	47.88	53.39	85.50	82.23	84.44	84.91	85.81	86.68	86.57	85.54
	B777	24.75	22.50	18.32	29.07	15.67	29.65	15.15	22.04	17.79	16.42	21.37	20.89	23.15	55.39
	B787	51.81	51.86	56.43	52.47	48.32	54.04	47.97	51.21	48.96	49.13	49.99	52.18	52.76	62.61
	C919	0.81	0.01	0.00	0.00	0.01	0.02	16.97	25.66	21.18	52.50	24.27	20.38	22.45	48.76
	A220	40.50	47.87	47.67	44.37	40.11	46.26	45.64	51.83	48.38	68.26	53.58	52.96	52.80	54.07
	A321	41.06	43.83	49.91	38.35	39.31	43.12	69.14	69.53	71.25	66.98	69.77	70.45	69.09	66.29
	A330	18.02	17.66	27.64	26.55	16.54	27.61	71.54	70.86	72.06	39.05	65.57	68.84	71.84	69.16
	A350	19.94	19.95	31.79	17.54	16.56	20.89	71.50	73.70	74.16	70.27	72.57	71.87	74.06	78.59
	ARJ21	1.70	0.13	0.00	0.00	0.01	0.00	38.68	36.90	31.47	45.45	33.61	31.72	34.34	49.41
Vehicle	SCar	65.20	68.42	68.80	68.19	66.23	70.12	56.78	68.81	57.73	58.39	60.99	62.35	61.96	67.96
	Bus	22.42	28.37	37.41	28.25	23.43	28.38	35.10	18.33	32.37	35.51	32.65	28.00	33.83	43.11
	CTrk	44.17	51.24	53.96	48.62	46.78	49.11	42.19	42.17	41.01	42.90	44.36	46.40	46.46	50.27
	DTrk	35.37	43.60	45.68	40.40	36.56	44.02	45.83	47.15	46.69	47.23	50.32	51.04	51.24	57.30
	Van	52.44	57.51	58.39	58.00	53.78	60.78	52.01	65.48	54.08	53.59	56.68	57.47	57.55	67.39
	Tra.	19.17	15.03	16.22	13.66	14.32	14.48	16.34	11.93	15.75	15.80	15.75	21.50	16.32	13.51
	Trac.	1.28	3.04	5.13	0.91	16.39	4.96	6.43	2.09	7.09	4.26	4.42	5.24	5.96	0.44
	Exc.	17.03	17.99	22.17	16.45	16.92	17.60	18.08	8.51	15.97	17.42	19.16	19.49	20.00	22.90
	TTrk	28.98	29.36	46.71	30.27	28.91	30.09	0.61	0.41	0.37	1.05	1.99	2.12	4.71	2.28
Court	Bkb.	50.58	58.26	54.84	38.81	48.41	47.91	48.45	42.64	49.43	50.43	57.21	57.81	58.25	56.91
	Ten.	81.09	82.67	80.35	80.29	80.31	82.11	78.13	78.60	79.17	82.25	83.08	82.20	81.40	90.60
	Fbl.	52.50	54.50	56.68	48.21	53.46	54.10	62.55	63.51	59.90	60.53	69.04	68.84	68.41	65.26
	Bsb.	66.76	71.71	69.07	67.90	66.93	69.97	88.23	87.40	87.90	86.96	89.97	90.13	89.42	86.21
Road	Int.	17.41	16.92	18.58	20.35	16.25	19.57	59.01	57.48	59.48	58.51	60.98	61.61	61.76	65.65
	Rnd.	60.13	59.86	58.44	55.67	59.41	59.91	20.21	20.82	22.05	24.29	25.80	24.75	26.28	31.73
	Brg.	12.58	11.87	31.81	12.62	10.39	14.38	31.67	22.57	32.76	26.58	37.42	39.75	40.07	31.87
mAP_{50} (%)		30.67	32.12	35.29	31.18	29.92	33.00	40.39	40.55	40.74	40.87	43.20	43.63	44.36	47.97

Note: Bold indicates the best result. Fine-grained abbreviations in the “Fine” column follow the FAIR1M v1.0 dataset [39]. Airplane: B737, Boeing737; B747, Boeing747; B777, Boeing777; B787, Boeing787; C919, C919; A220, A220; A321, A321; A330, A330; A350, A350; ARJ21, ARJ21. Ship: Pass., passenger ship; Mot., motorboat; Fis., fishing boat; Tug., tugboat; Eng., engineering ship; LCS, liquid cargo ship; DCS, dry cargo ship; War., warship. Vehicle: SCar, small car; Bus, bus; CTrk, cargo truck; DTrk, dump truck; Van, van; Tra., trailer; Trac., tractor; Exc., excavator; TTrk, truck tractor. Court: Bkb., basketball court; Ten., tennis court; Fbl., football field; Bsb., baseball field. Road: Int., intersection; Rnd., roundabout; Brg., bridge.

(2) Comparison experiments on other remote sensing scenarios. To further verify the generality of the proposed method, we conducted comparative experiments on large public remote sensing datasets that contain richer scenes and target categories. The performance of our method on the DIOR dataset is listed in Table 6. The proposed SPDR-Net achieves the highest mAP_{50} of 85.04% on the DIOR dataset, significantly outperforming all other advanced methods, e.g., YOLO-PKFF [76], GLEF-YOLO [77], CI-DERT [78] and EMS-Net [79]. Table 7 provides a detailed quantitative comparison on the most widely used dataset DOTA v1.0. Experimental results demonstrate that the proposed SPDR-Net surpasses the most of existing object detection methods, e.g., Oriented RepPoints [80], KFIOU [81], etc., in both accuracy and inference speed. Although our mAP is marginally lower than that of Oriented R-CNN [82], our approach surpasses all other methods in terms of inference speed and detection performance. As shown in Figure 10, SPDR-Net shows good fine-grained classification performance on Airplane category, for example, Boeing787 shows higher confidence and Boeing737 is correctly identified and identified to the intersection, which proves its effectiveness in other remote sensing scenarios.

Table 6. Comparison of performance (AP) on DIOR.

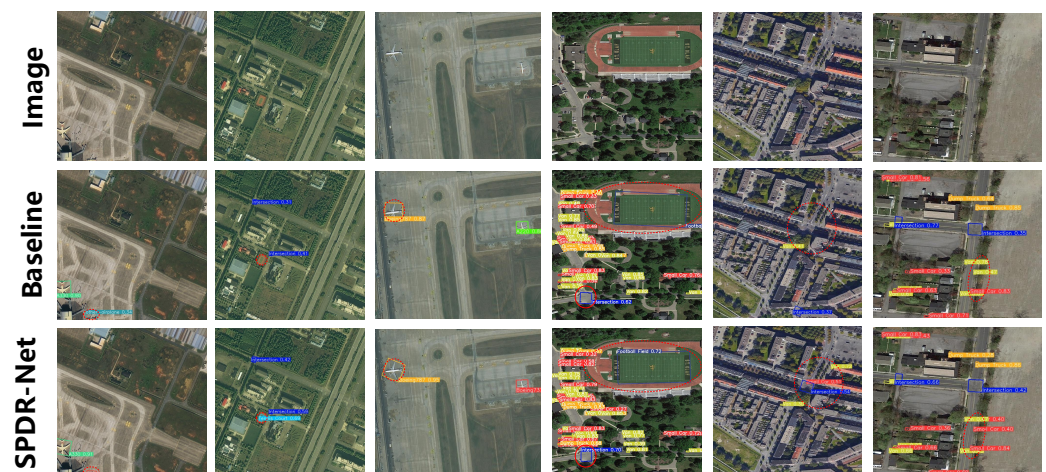
Methods	Year	Backbone	mAP_{50} (%)
RetinaNet-O [43]	2017	R50-FPN	57.55
Faster RCNN-O [48]	2017	R50-FPN	59.54
Gliding Vertex [68]	2021	R50-FPN	60.06
RoI-Transformer [67]	2018	R50-FPN	63.87
AOPG [83]	2022	R50-FPN	64.14
Oriented RepPoints [80]	2022	R50	67.71
YOLOv5 [51]	2018	CSPDarknet	70.63
YOLOv11 [84]	2024	CSPDarknet	77.71
CI-DETR [78]	2023	R50	69.52
YOLO-DA [85]	2023	CSPDarknet	71.21
GLEF-YOLOX [77]	2024	CSPDarknet	78.91
EMS-Net [79]	2025	CSPDarknet	77.23
YOLO-PKFF [76]	2025	CSPDarknet	80.15
SPDR-Net (ours)	2026	CSPDarknet	85.04

Note: Bold indicates the best result.

Table 7. Comparison of performance (AP) on DOTAv1.0.

Method	PL.	BD.	BR.	GTF.	SV.	LV.	SH.	TC.	BC.	ST.	SBF.	RA.	HA.	SP.	HC.	mAP_{50} (%)	Speed (fps)
ROI Trans. [67]	88.53	77.91	37.63	74.08	66.53	62.97	66.57	90.50	79.46	76.75	59.04	56.73	62.54	61.29	55.56	67.74	7.80
SCRDet [42]	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61	9.51
RSDet [45]	89.80	82.90	48.60	65.20	69.50	70.10	70.20	90.50	85.60	83.40	62.50	69.30	65.60	67.20	68.00	72.20	-
Gliding Vertex [68]	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02	13.10
CSL [86]	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17	8.89
Oriented R-CNN [82]	89.84	85.43	61.09	79.82	79.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11	80.87	8.10
CGCDet [87]	88.93	84.45	53.93	78.56	78.54	82.46	87.90	90.87	87.46	84.79	65.56	63.45	76.15	71.58	65.32	77.34	7.80
DODet [88]	89.34	84.31	51.39	71.04	79.04	82.86	88.15	90.90	86.88	84.91	62.69	67.63	75.47	72.22	45.54	75.49	7.98
PolarDet [89]	89.65	87.07	48.14	70.97	78.53	80.34	87.45	90.76	85.63	86.87	61.64	70.32	71.92	73.09	67.15	76.64	25.00
KFloU [81]	89.46	85.72	54.94	80.37	77.16	69.23	80.90	90.79	87.79	86.13	73.32	68.11	75.23	71.61	69.49	77.35	13.79
R ³ Det [44]	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47	12.39
Oriented RepPoints [80]	88.72	80.56	55.69	75.07	81.84	82.40	87.97	90.80	84.33	87.64	62.80	67.91	77.69	82.94	65.46	78.12	-
GWD [90]	86.96	83.88	54.36	77.53	74.41	68.48	80.34	86.62	83.41	85.55	73.47	67.77	72.57	75.76	73.40	76.30	13.86
SPDR-Net (ours)	88.56	85.63	56.34	76.91	78.94	84.26	88.98	90.82	87.84	88.37	66.78	67.67	76.38	82.56	72.67	79.51	45.24

Note: Bold indicates the best result.

**Figure 10.** Detection results of other category objects.

5. Conclusions

This paper proposes SPDR-Net for fine-grained ship detection in remote sensing images. Addressing the challenges of diverse ship shapes and subtle fine-grained differences, the proposed SPDR-Net utilizes GMSR to provide explicit shape and texture distribution priors for different ship categories. Furthermore, it develops RSDKs, in which the receptive field is dynamically adjusted with the help of GMSR shape prior, to achieve precise focusing

of the object region in the spatial domain. To address the issues of high-frequency discrimination information loss and interference from background noise during up-sampling and down-sampling, DSWR is designed to recover edge and texture details in the frequency domain through diffusion-supervised wavelet reconstruction. Specifically, GMSR models distinctive shape distributions, RSDK adaptively enhances discriminative regions, and DSWR preserves subtle boundary and texture details, jointly reducing ambiguity among visually similar ship categories. Through this collaborative design in the spatial and frequency domains, SPDR-Net effectively improves fine-grained detection accuracy and robustness under complex sea conditions. Experimental results on several public datasets show that the proposed SPDR-Net outperforms current mainstream object detection models in both detection accuracy and speed, achieving an mAP_{50} of 79.3% on the DOSR dataset, 9.9% higher than existing methods. It also achieves state-of-the-art performance on datasets with various scenes and modalities, validating its versatility.

Nevertheless, the current method offers limited improvement in fine-grained classification and requires optimization for lightweight deployment. Future work will focus on improving efficiency and generalization capabilities to achieve more refined ship category recognition.

Author Contributions: Conceptualization, Z.H. and W.G.; methodology, Z.H. and W.G.; software, W.G.; validation, W.G., W.H. and J.Z.; formal analysis, Z.H.; investigation, W.G.; resources, Z.H.; data curation, J.Z.; writing—original draft preparation, Z.H. and W.G.; writing—review and editing, Z.H. and H.S.; visualization, W.H.; supervision, Z.H. and H.S.; project administration, Z.H.; funding acquisition, Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 62501163), the Fujian Province Young and Middle-aged Teachers Education and Research Project (Grant No. JAT231007), the Fujian Provincial Natural Science Foundation of China (Grant No. 2024J01277), and the China Postdoctoral Science Foundation (2024M762192).

Data Availability Statement: The data presented in this study are derived from publicly available benchmark datasets, including DOSR, HRSC2016, SSDD, DOTA, FAIR1M, and DIOR. These datasets are available from the public sources cited in the manuscript. The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [[CrossRef](#)]
2. Lv, M.; Song, S.; Jia, Z.; Li, L.; Ma, H. Multi-Focus Image Fusion Based on Dual-Channel Rybak Neural Network and Consistency Verification in NSCT Domain. *Fractal Fract.* **2025**, *9*, 432. [[CrossRef](#)]
3. Li, L.; Song, S.; Lv, M.; Jia, Z.; Ma, H. Multi-Focus Image Fusion Based on Fractal Dimension and Parameter Adaptive Unit-Linking Dual-Channel PCNN in Curvelet Transform Domain. *Fractal Fract.* **2025**, *9*, 157. [[CrossRef](#)]
4. Xiao, X.; Zhou, Z.; Wang, B.; Li, L.; Miao, L. Ship Detection under Complex Backgrounds Based on Accurate Rotated Anchor Boxes from Paired Semantic Segmentation. *Remote Sens.* **2019**, *11*, 2506. [[CrossRef](#)]
5. Li, L.; Shi, Y.; Lv, M.; Jia, Z.; Liu, M.; Zhao, X.; Zhang, X.; Ma, H. Infrared and Visible Image Fusion via Sparse Representation and Guided Filtering in Laplacian Pyramid Domain. *Remote Sens.* **2024**, *16*, 3804. [[CrossRef](#)]
6. Yang, X.; Shao, S.; Liu, W.; Xiang, Z.; Zhang, B. TAD-YOLO: A lightweight nearshore ship detection method for small USVs in maritime trade surveillance. *Ocean. Eng.* **2025**, *341*, 122618. [[CrossRef](#)]
7. Lv, M.; Jia, Z.; Li, L.; Ma, H. Fractal Dimension-Based Multi-Focus Image Fusion via AGPCNN and Consistency Verification in NSCT Domain. *Fractal Fract.* **2026**, *10*, 1. [[CrossRef](#)]
8. Yin, J.; Zhen, T.; Xiong, G.; Yu, W. Fractal-domain Vision Graph Neural Network for Remote Sensing Ground Target Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 1–13. [[CrossRef](#)] [[PubMed](#)]

9. Liu, F.; Wei, H.; Zhao, W.; Li, G.; Peng, J.; Li, Z. WB-DETR: Transformer-Based Detector without Backbone. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 2959–2967. [\[CrossRef\]](#)
10. Li, L.; Ma, H.; Zhang, X.; Zhao, X.; Lv, M.; Jia, Z. Synthetic Aperture Radar Image Change Detection Based on Principal Component Analysis and Two-Level Clustering. *Remote Sens.* **2024**, *16*, 1861. [\[CrossRef\]](#)
11. Xiong, G.; Zhen, T.; Huang, W.; Min, B.; Yu, W. Fractal-domain deep learning with Transformer architecture for SAR ship classification. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 208–226. [\[CrossRef\]](#)
12. Zhang, J.; Sheng, W.; Zhu, H.; Guo, S.; Han, Y. MLBR-YOLOX: An Efficient SAR Ship Detection Network With Multilevel Background Removing Modules. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5331–5343. [\[CrossRef\]](#)
13. Gou, Z.; Wang, J.; Wang, Z.; Ma, L.; Liu, Q. SAR Ship Target Detection Algorithm Based on MSEIG-DETR. In Proceedings of the 2025 IEEE 8th International Conference on Electronic Information and Communication Technology (ICEICT), Weihai, China, 26–28 July 2025; pp. 421–426. [\[CrossRef\]](#)
14. Min, L.; Dou, F.; Zhang, Y.; Shao, D.; Li, L.; Wang, B. CM-YOLO: Context Modulated Representation Learning for Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4202414. [\[CrossRef\]](#)
15. Wu, H.; Yu, L.; Li, X.; Zhou, L.; Zhang, W.; Bai, G. CTF-Net: A Convolutional and Transformer Fusion Network for SAR Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4010005. [\[CrossRef\]](#)
16. Huang, L.; Wang, F.; Zhang, Y.; Xu, Q. Fine-Grained Ship Classification by Combining CNN and Swin Transformer. *Remote Sens.* **2022**, *14*, 3087. [\[CrossRef\]](#)
17. Song, H.; Wang, L.; Mi, L.; Yu, H.; Gan, J. DCAPT-Net: A Multi-Scale Ship Detection Algorithm for SAR Images. In Proceedings of the 2024 10th International Conference on Big Data and Information Analytics (BigDIA), Chiang Mai, Thailand, 25–28 October 2024; pp. 441–446. [\[CrossRef\]](#)
18. Ren, Z.; Tang, Y.; Yang, Y.; Zhang, W. SASOD: Saliency-Aware Ship Object Detection in High-Resolution Optical Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611115. [\[CrossRef\]](#)
19. You, Y.; Cao, J.; Zhang, Y.; Liu, F.; Zhou, W. Nearshore Ship Detection on High-Resolution Remote Sensing Image via Scene-Mask R-CNN. *IEEE Access* **2019**, *7*, 128431–128444. [\[CrossRef\]](#)
20. Yang, X.; Zhang, X.; Wang, N.; Gao, X. A Robust One-Stage Detector for Multiscale Ship Detection With Complex Background in Massive SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5217712. [\[CrossRef\]](#)
21. Yang, X.; Zhou, Z.; Yang, D. Elaborate Feature Decoupling for Weakly Supervised Fine-Grained Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5602114. [\[CrossRef\]](#)
22. Xie, X.; Cheng, G.; Li, W.; Lang, C.; Zhang, P.; Yao, Y.; Han, J. Learning Discriminative Representation for Fine-Grained Object Detection in Remote Sensing Images. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 8197–8208. [\[CrossRef\]](#)
23. Li, X.; Shi, C.; Xu, N.; Su, Y.; Kaup, A.; Liu, D.; Li, X. Position-Aware Differential Denoising Transformer for Semantic Segmentation of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 5000405. [\[CrossRef\]](#)
24. Zou, J.; Zhu, Z.; Ye, Y.; Wang, X. DiffBEV: Conditional Diffusion Model for Bird’s Eye View Perception. *arXiv* **2023**, arXiv:2303.08333. Available online: <http://arxiv.org/abs/2303.08333> (accessed on 28 July 2026).
25. He, B.; Ji, Y.; Ye, Q.; Tan, Z.; Wu, L. Generalized Diffusion Detector: Mining Robust Features from Diffusion Models for Domain-Generalized Detection. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 9921–9932. [\[CrossRef\]](#)
26. Zhang, R. Making Convolutional Networks Shift-Invariant Again. *arXiv* **2019**, arXiv:1904.11486.
27. Williams, T.; Li, R. Wavelet pooling for convolutional neural networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
28. Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; Zuo, W. Multi-level wavelet-CNN for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 773–782. [\[CrossRef\]](#)
29. Li, Q.; Shen, L.; Guo, S.; Lai, Z. Wavelet integrated CNNs for noise-robust image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtually, 14–19 June 2020; pp. 7245–7254.
30. Guo, T.; Seyed Mousavi, H.; Huu Vu, T.; Monga, V. Deep wavelet prediction for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 104–113.
31. Nigam, S.; Verma, S.; Nagabhushan, P. Wavelet RCNN: Enhancing Object detection accuracy through Spectral Information. In Proceedings of the 2023 IEEE 20th India Council International Conference (INDICON), Warangal, India, 14–17 December 2023; pp. 1323–1329.
32. Li, X.; Xu, F.; Li, J.; Su, Y.; Li, L.; Lyu, X.; Xu, Z.; Kaup, A. Frequency domain-enhanced spectral-spatial fusion transformer for semantic segmentation of remote sensing images. *Inf. Fusion* **2026**, *132*, 104248. [\[CrossRef\]](#)
33. Li, X.; Xu, F.; Zhang, J.; Zhang, H.; Lyu, X.; Liu, F.; Gao, H.; Kaup, A. Frequency-Guided Denoising Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5400217. [\[CrossRef\]](#)

34. Jin, B.; Yin, F.; Cai, W.; Li, H.; Zhu, H.; Huang, W.; Wu, Q.; Chen, H.; Sun, Z. HWANet: A Haar Wavelet-based Attention Network for remote sensing object detection. *PLoS ONE* **2025**, *20*, e0330759. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Han, Y.; Yang, X.; Pu, T.; Peng, Z. Fine-Grained Recognition for Oriented Ship Against Complex Scenes in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612318. [\[CrossRef\]](#)
36. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017.
37. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSAR DATA), Beijing, China, 13–14 November 2017; pp. 1–6. [\[CrossRef\]](#)
38. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983. [\[CrossRef\]](#)
39. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [\[CrossRef\]](#)
40. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward Arbitrary-Oriented Ship Detection With Rotated Region Proposal and Discrimination Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [\[CrossRef\]](#)
41. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [\[CrossRef\]](#)
42. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240. [\[CrossRef\]](#)
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [\[CrossRef\]](#)
44. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *arXiv* **2020**, arXiv:1908.05612. Available online: <http://arxiv.org/abs/1908.05612> (accessed on 28 July 2026).
45. Qian, W.; Yang, X.; Peng, S.; Yan, J.; Guo, Y. Learning modulated loss for rotated object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 2458–2466. [\[CrossRef\]](#)
46. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2384–2399. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. *arXiv* **2021**, arXiv:2103.07733. Available online: <http://arxiv.org/abs/2103.07733> (accessed on 28 July 2026).
48. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. *arXiv* **2019**, arXiv:1904.02701. Available online: <http://arxiv.org/abs/1904.02701> (accessed on 28 July 2026).
50. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37. [\[CrossRef\]](#)
51. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
52. Wang, C.Y.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
53. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
54. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430. Available online: <http://arxiv.org/abs/2107.08430> (accessed on 28 July 2026).
56. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. *arXiv* **2022**, arXiv:2201.12329. Available online: <http://arxiv.org/abs/2201.12329> (accessed on 28 July 2026).
57. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159. Available online: <http://arxiv.org/abs/2010.04159> (accessed on 28 July 2026).
58. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense Distinct Query for End-to-End Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7329–7338. [\[CrossRef\]](#)

59. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. Available online: <http://arxiv.org/abs/2203.03605> (accessed on 28 July 2026).
60. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [[CrossRef](#)] [[PubMed](#)]
61. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850. Available online: <http://arxiv.org/abs/1904.07850> (accessed on 28 July 2026).
62. Cui, Z.; Li, Q.; Cao, Z.; Liu, N. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [[CrossRef](#)]
63. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention Receptive Pyramid Network for Ship Detection in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2738–2756. [[CrossRef](#)]
64. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
65. Zhao, J.; Ding, Z.; Zhou, Y.; Zhu, H.; Du, W.L.; Yao, R.; El Saddik, A. OrientedFormer: An End-to-End Transformer-Based Oriented Object Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5640816. [[CrossRef](#)]
66. Qian, L.; Hu, J.; Ren, H.; Lin, J.; Luo, X.; Zou, L.; Zhou, Y. Cross-Level Adaptive Feature Aggregation Network for Arbitrary-Oriented SAR Ship Detection. *Remote Sens.* **2025**, *17*, 1770. [[CrossRef](#)]
67. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Detecting Oriented Objects in Aerial Images. *arXiv* **2018**, arXiv:1812.00155.
68. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [[CrossRef](#)] [[PubMed](#)]
69. Zhang, F.; Wang, X.; Zhou, S.; Wang, Y.; Hou, Y. Arbitrary-Oriented Ship Detection Through Center-Head Point Extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 14. [[CrossRef](#)]
70. Ouyang, L.; Guo, G.; Fang, L.; Ghamisi, P.; Yue, J. PCLDet: Prototypical Contrastive Learning for Fine-Grained Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5613911. [[CrossRef](#)]
71. Li, W.; Zhao, D.; Yuan, B.; Gao, Y.; Shi, Z. PETDet: Proposal Enhancement for Two-Stage Fine-Grained Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5602214. [[CrossRef](#)]
72. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610510. [[CrossRef](#)]
73. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M.M. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. *arXiv* **2025**, arXiv:2501.03775. Available online: <http://arxiv.org/abs/2501.03775> (accessed on 28 July 2026).
74. Qian, X.; Jian, Q.; Wang, W.; Yao, X.; Cheng, G. Incorporating Multiscale Context and Task-Consistent Focal Loss into Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5628411. [[CrossRef](#)]
75. Qian, X.; Li, X.; Wang, W.; Yao, X.; Cheng, G. Pointwise Frequency Attention and Coarse-to-Fine Learning With Image-to-Instance Level Supervision for Fine-Grained Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5619812. [[CrossRef](#)]
76. Liu, J.; Huang, B.; Lv, J. YOLO-PKFF: Remote Sensing Object Detection Enhanced With Poly Kernel Inception and Attentional Cross-Level Feature Fusion. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6010805. [[CrossRef](#)]
77. Gu, Q.; Huang, H.; Han, Z.; Fan, Q.; Li, Y. GLFE-YOLOX: Global and Local Feature Enhanced YOLOX for Remote Sensing Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2516112. [[CrossRef](#)]
78. Han, Y.; Meng, W.; Tang, W. Capsule-inferenced Object Detection for Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5260–5270. [[CrossRef](#)]
79. Chen, P.; Lyu, W.; Guo, Q.; Deng, Z.; Xu, W. EMS-Net: Efficient Multiscale Perceptual Enhancement Tiny Object Detector for Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6007805. [[CrossRef](#)]
80. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented RepPoints for Aerial Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1819–1828.
81. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; Tian, Q. The KFIOU Loss for Rotated Object Detection. *arXiv* **2023**, arXiv:2201.12558. Available online: <http://arxiv.org/abs/2201.12558> (accessed on 28 July 2026).
82. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. *arXiv* **2021**, arXiv:2108.05699. Available online: <http://arxiv.org/abs/2108.05699> (accessed on 28 July 2026).
83. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-Free Oriented Proposal Generator for Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625411. [[CrossRef](#)]
84. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. Available online: <http://arxiv.org/abs/2410.17725> (accessed on 28 July 2026).

85. Lin, J.; Zhao, Y.; Wang, S.; Tang, Y. YOLO-DA: An Efficient YOLO-Based Detector for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6008705. [[CrossRef](#)]
86. Yang, X.; Yan, J. On the Arbitrary-Oriented Object Detection: Classification based Approaches Revisited. *arXiv* **2022**, arXiv:2003.05597. Available online: <http://arxiv.org/abs/2003.05597> (accessed on 28 July 2026).
87. Wang, Y.; Zou, X.; Zhang, Z.; Xu, W.; Chen, L.; Zhong, S.; Yan, L.; Wang, G. Learning Oriented Remote Sensing Object Detection via Naive Geometric Computing. *arXiv* **2021**, arXiv:2112.00504. Available online: <http://arxiv.org/abs/2112.00504> (accessed on 28 July 2026).
88. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-Aligned Oriented Detector. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618111. [[CrossRef](#)]
89. Zhao, P.; Qu, Z.; Bu, Y.; Tan, W.; Guan, Q. PolarDet: A fast, more precise detector for rotated target in aerial images. *Int. J. Remote Sens.* **2021**, *42*, 5831–5861. [[CrossRef](#)]
90. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. *arXiv* **2022**, arXiv:2101.11952. Available online: <http://arxiv.org/abs/2101.11952> (accessed on 28 July 2026).

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