DenoDet V2: Phase-Amplitude Cross Denoising for SAR Object Detection

Kang Ni[1*], Minrui Zou[2*], Yuxuan Li[2], Xiang Li[2], Kehua Guo[3], Ming-Ming Cheng[2], Yimian Dai[2†]

[1]Nanjing University of Posts and Telecommunications
[2]PCA Lab, VCIP, Computer Science, Nankai University
[3]Central South University
tznikang@njupt.edu.cn, minrui.zou@mail.nankai.edu.cn, guokehua@csu.edu.cn, {xiang.li.implus, cmm, yimian.dai}@nankai.edu.cn

Abstract

One of the primary challenges in Synthetic Aperture Radar (SAR) object detection lies in the pervasive influence of coherent noise. As a common practice, most existing methods, whether handcrafted approaches or deep learning-based methods, employ the analysis or enhancement of object spatial-domain characteristics to achieve implicit denoising. In this paper, we propose DenoDet V2, which explores a completely novel and different perspective to deconstruct and modulate the features in the transform domain via a carefully designed attention architecture. Compared to DenoDet V1, DenoDet V2 is a major advancement that exploits the complementary nature of amplitude and phase information through a band-wise mutual modulation mechanism, which enables a reciprocal enhancement between phase and amplitude spectra. Extensive experiments on various SAR datasets demonstrate the state-of-the-art performance of DenoDet V2. Notably, DenoDet V2 achieves a significant 0.8% improvement on SARDet-100K dataset compared to DenoDet V1, while reducing the model complexity by half.

Code — https://github.com/GrokCV/GrokSAR
Extended Version — https://arxiv.org/pdf/2508.09392

Introduction

Synthetic Aperture Radar (SAR) has revolutionized the field of remote sensing, offering an unrivaled capability to capture high-resolution imagery regardless of lighting conditions or weather obscurity (Li et al. 2024b; Ye et al. 2025).

Despite the progress made by deep learning methods (Zhao et al. 2022; Zhang et al. 2025b; Li et al. 2025; Zhang et al. 2025a), most of them are straightforward adaptations from generic object detection in computer vision, without fully considering the unique characteristics of SAR data and its associated challenges. As SAR is a coherent imaging system, its images intrinsically contain **unavoidable speckle noise**, overlaid on the objects, significantly increasing the difficulty of object detection and identification.

To tackle the aforementioned challenges, recent works have integrated frequency-domain processing within deep learning frameworks, aiming to harness its potential in separating the speckle noise. Zhao *et al.* combined a morphological network with a feature pyramid fusion structure for speckle noise suppression (Zhao et al. 2022). Similarly, Wang *et al.* employed a dual-backbone structure integrated with Haar wavelet transform (Wang, Cai, and Yuan 2023), adept at capturing both global and intricate textural details of maritime objects. Complementing these, Li *et al.* unveiled a network synergizing a feature pyramid network with a polar Fourier transform (Li et al. 2022a), procuring rotation-invariant features and delivering enhanced performance in ship object detection.

Although these methods have demonstrated improved performance in SAR object detection by exploiting frequency-domain information, they still have some limitations.

1. **Isolated Processing of Amplitude and Phase:** These approaches tend to process amplitude and phase information in isolation. Besides, these methods predominantly focus on the amplitude components, neglecting the phase details which are crucial for preserving the original spatial relationships and structural integrity of the objects as shown in Fig. 1.

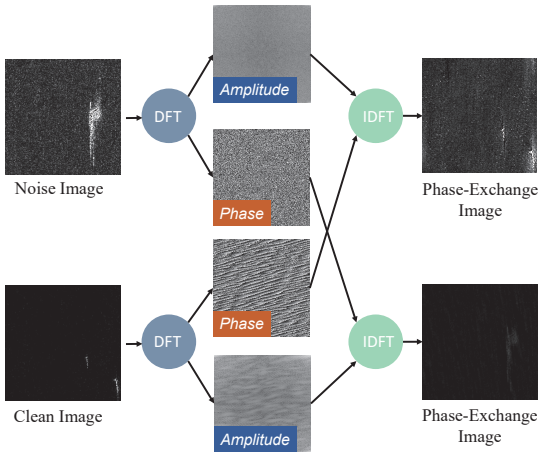

Figure 1: An example that the phase spectrum is more robust to noise. In the images after phase information exchange, the object contours from the noise images are clearly transferred while introducing only a small amount of noise interference.

*Equal Contribution.
†Corresponding Author.

2. **Increased Model Complexity and Over-Interaction:** These approaches involve interactions across the entire spectrum, which not only escalates computational complexity but may also induce an overprocessing of frequency-domain features. This overprocessing can manifest as an excessive smoothing of object features, which diminishes detection accuracy.

Recent studies (Chen et al. 2021a) establish that phase information is pivotal for maintaining structural integrity in CNNs, essential for robust recognition; amplitude, conversely, is more vulnerable to noise. This distinction is critical in SAR imagery, where coherent speckle noise severely degrades object detection. In this context, phase provides a resilient feature, remaining stable against amplitude-distorting noise. This motivates the question: *Can frequency-domain phase information be leveraged at the feature level to mitigate speckle noise effects on the amplitude spectrum, and vice versa?*

Motivated by this intriguing possibility, in this paper, we extend the idea of **attention as feature denoising** by proposing DenoDet V2, a novel method that leverages the complementary nature of amplitude and phase information through a mutual guidance mechanism for SAR image object detection. Central to our approach is a phase-guided soft-thresholding mechanism that exploits the robustness of phase information to adaptively filter noise from the amplitude spectrum. DenoDet V2 not only enhances the signal-to-noise ratio of the object features but also uses the refined amplitude spectrum to further improve phase accuracy, ensuring the preservation of essential object details. *To our knowledge, DenoDet V2 is the first to implement such a reciprocal, reference-based feature denoising strategy.*

Specifically, DenoDet V2 involves a strategic interplay within the Key and Value component of the self-attention (Vaswani et al. 2017) module, where the roles of phase and amplitude are dynamically interchanged. To further tailor this process to the intricate nature of SAR imagery, we partition the frequency spectrum into a grid of $N \times N$ local bands, within which phase and amplitude interact exclusively. This localized, band-specific approach ensures fine-grained feature denoising that is acutely attuned to each band's unique characteristics, meticulously preserving vital spatial and structural information.

In summary, we advance the state-of-the-art in SAR image object detection with our contributions as follows:

1. We propose a novel concept, ***Attention as Phase-Amplitude Cross Denoising***, which leverages the inherent stability of phase information to guide the feature denoising of amplitude in SAR imagery.
2. We propose the Phase-Amplitude Token Exchange (PATE) module as a core component of our DenoDet V2. The PATE module implements a dual guidance mechanism, where phase and amplitude information mutually enhance each other.
3. Our DenoDet V2 achieves the state-of-the-art performance across various SAR datasets, especially **ranking No. 1 on the largest benchmark SARDet-100K**.

Related Work

SAR Image object Detection

Recent advancements in deep learning have significantly enhanced SAR object detection. Sun et al. (Sun et al. 2021) utilized strong scattering points to improve detection amidst interference, particularly in near-shore scenes. Ke et al. (Ke et al. 2021) integrated deformable convolution kernels into Faster R-CNN to better accommodate geometric variations. Furthermore, PVT-SAR (Zhou et al. 2022) leveraged Pyramid Vision Transformers (PVT) with self-attention mechanisms to extract multi-scale features, achieving substantial performance gains.

Despite these advancements, the inherent issue of speckle noise in SAR images remains a formidable challenge. This noise, characteristic of the radar imaging process, can obscure critical object details and lead to high false alarm rates. Recent studies have explored the application of deep learning to speckle denoising in SAR images. For instance, Shen *et al.* (Shen et al. 2021) proposed a recursive deep CNN prior model that decouples the data-fitting and regularization terms for improved denoising.

Nevertheless, these studies generally *treat denoising and object detection as separate processes*. The lack of integration makes it challenging to dynamically adjust the denoising process based on the object features, which is crucial for preserving the discriminative information for detection. Moreover, these denoising methods primarily focus on enhancing the spatial domain features, without fully exploiting the potential of frequency domain information. On the contrary, our DenoDet V2 integrates a frequency-domain denoising module into a detection framework, which differs from existing methods in the paradigm of dynamic feature denoising and phase-amplitude token exchange.

Frequency-Domain Feature Refinement

Recent studies have explored the potential of incorporating frequency-domain information into SAR object detection to enhance performance. Zhao et al. (Zhao et al. 2018) explored a visual attention mechanism that incorporates frequency-domain elements to refine ship detection in SAR images. Similarly, Li et al. (Li et al. 2022b) developed a multidimensional domain network that leverages both spatial and frequency-domain features to detect ships in SAR imagery effectively. Xu et al. (Xu et al. 2020) further emphasized the importance of frequency-domain learning, proposing a method that selectively processes frequency components to optimize network input and improve computational efficiency.

Despite these advancements, current approaches typically neglect the phase spectrum or treat phase and amplitude information in isolation, which often results in the inability to dynamically interact between phase and amplitude spectra. However, this is crucial for enhancing the detection performance in noisy SAR environments, which has been confirmed by many studies. For instance, Chen et al. (Chen et al. 2021a) found that CNNs heavily depend on the amplitude spectrum of images and are easily disturbed by noise. They proposed a data augmentation method involving phase-amplitude spectrum sample replacement to improve the gener-

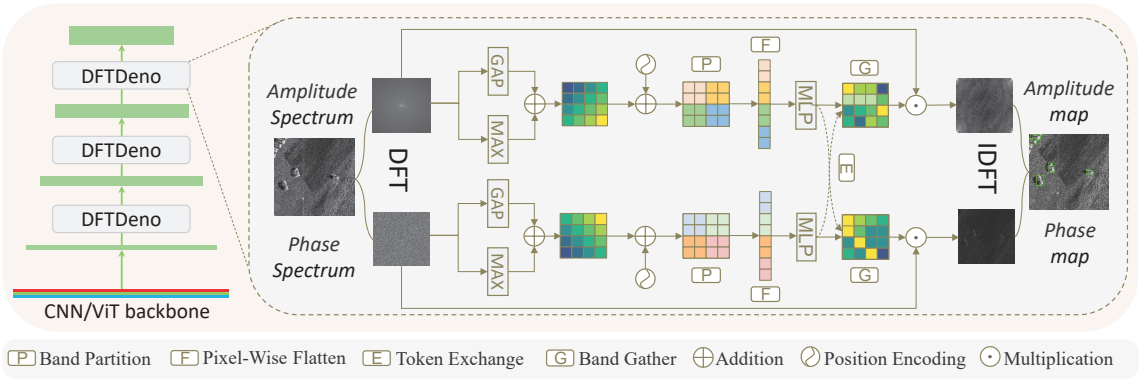

Figure 2: The overall architecture of DenoDet V2.

alization performance, further highlighting the significance of phase information. In contrast to the aforementioned works, our proposed DenoDet V2 model differs in two key aspects: Phase-Amplitude Modulation and Frequency-Domain Band Division.

Method

Overall Architecture

The DenoDet V2 architecture as shown in the Fig. 2 integrates our plug-and-play DFTDeno module into a generic object detector's backbone. This module refines feature extraction by performing dynamic, attention-based soft threshold denoising in the feature map transform domain.

DFTDeno module consists of a forward 2D Discrete Fourier Transform (DFT), dynamic threshold denoising, followed by an inverse 2D DFT. For a feature map $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$, the 2D DFT forward function can be defined as:

$$\mathbf{m}_{c,u,v} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{M}_{c,h,w} e^{-2\pi i(\frac{uh}{H}+\frac{vw}{W})}, \tag{1}$$

where H and W denotes the height and width of features, a pair of $h, u \in [0, H-1]$, and $w, v \in [0, W-1]$ represents the coordinate position of statistics, and $\mathbf{m}_{c,u,v}$ means the value located at coordinates (u, v) in channel c of the tensor after the forward DFT. And according to Euler's formula, we can decompose the signal component after DFT into real and imaginary part:

$$\mathcal{R}_{c,u,v} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{M}_{c,h,w} cos2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right), \tag{2}$$

$$\mathcal{I}_{c,u,v} = -\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{M}_{c,h,w} isin2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right). \tag{3}$$

Next, the real part $\mathcal{R} \in \mathbb{R}^{C\times H\times W}$ and imaginary part $\mathcal{I} \in \mathbb{R}^{C\times H\times W}$ undergo two non-linear transformations to extract the amplitude spectrum $\mathcal{A}$ and phase spectrum $\mathcal{P} \in [-\pi, \pi)$ of the frequency-domain signal as follows:

$$\mathcal{A}_{c,u,v} = \sqrt{\mathcal{R}^2_{c,u,v} + \mathcal{I}^2_{c,u,v}}, \tag{4}$$

$$\mathcal{P}_{c,u,v} = arctan2\left(\frac{\mathcal{I}_{c,u,v}}{\mathcal{R}_{c,u,v}}\right). \tag{5}$$

Eq. (1) reveals that each signal component is derived from pixel-level values in the original feature map, inherently encoding partial global information. To process such highly aggregated information, we employ an efficient attention mechanism for signal modulation. This operation suppresses noise while enhancing informative features through the following transformation:

$$\hat{\mathcal{A}} = \mathbf{G}(\mathcal{A}, \mathcal{P}) \odot \mathcal{A}, \quad \hat{\mathcal{P}} = \mathbf{G}(\mathcal{A}, \mathcal{P}) \odot \mathcal{P}, \tag{6}$$

where $\hat{\mathcal{A}}$ and $\hat{\mathcal{P}}$ represent the signal's amplitude and phase after modulation, $\odot$ means element-wise multiplication and $\mathbf{G}(\cdot) \in \mathbb{R}^{C\times H\times W}$ is the output attention map of designed module $\mathbf{G}$, which will be discussed in section in detail.

Finally, we recombine the modulated amplitude and phase information and restore frequency signals to the spatial domain via the Inverse Discrete Fourier Transform (IDFT) as follows:

$$\hat{\mathbf{m}}_{c,u,v} = \hat{\mathcal{A}}_{c,u,v} \cdot (cos\hat{\mathcal{P}}_{c,u,v} + isin\hat{\mathcal{P}}_{c,u,v}), \tag{7}$$

$$\hat{\mathbf{M}}_{\mathbf{c},\mathbf{h},\mathbf{w}} = \sum_{u=0}^{H-1}\sum_{v=0}^{W-1} \hat{\mathbf{m}}_{c,u,v} e^{2\pi i(\frac{uh}{H}+\frac{vw}{W})}. \tag{8}$$

For brevity, the normalization coefficients in both the forward and inverse transforms are omitted. And in implementation, we employed $cos(\hat{\mathcal{P}})$ and $sin(\hat{\mathcal{P}})$ to orthogonally decouple phase information, mitigating angular boundary discontinuities. A subsequent trigonometric identity harmonized these components, ensuring mathematical equivalence and parity in the resultant phase representation.

Band-wise Partition Self-Attention

In the frequency domain, features lack local correlation, necessitating the long-context modeling capability of self-attention (SA). Therefore, we adopted SA as the foundational architecture for band-wise signal modulation. Let

$\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denote the frequency domain tensor after Discrete Fourier Transform (DFT); the output of the self-attention block for signal modulation is calculated via:

$$\hat{\mathbf{X}} = \mathbf{P}_{\text{Max}}(\mathbf{X}) + \mathbf{P}_{\text{Avg}}(\mathbf{X}), \tag{9}$$

$$\widetilde{\mathbf{S}} = BPSA(\hat{\mathbf{X}}), \quad \widetilde{\mathbf{X}} = \widetilde{\mathbf{S}} \odot \mathbf{X}, \tag{10}$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{H*W}$ is the frequency spectrum attention map, $\mathbf{P}_{\text{Max}}$ and $\mathbf{P}_{\text{Avg}}$ means max and average pooling operation among channel dimension, and BPSA denotes the band-wise partition self-attention, $\widetilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$ means the output signal after modulation. In basic band-wise self-attention (BSA), the input feature $\hat{\mathbf{X}}$ is split by band frequency as $\{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, ..., \hat{\mathbf{X}}_d\}$ and $d = H * W$ for all of coordinate pixel represents an identity frequency signal in basic BSA. The result of each band from BSA can be produced as:

$$\mathbf{Q}_i = \hat{\mathbf{X}}_i\mathbf{W}_i^q, \quad \mathbf{K}_i = \hat{\mathbf{X}}_i\mathbf{W}_i^k, \quad \mathbf{V}_i = \hat{\mathbf{X}}_i\mathbf{W}_i^v, \tag{11}$$

$$\widetilde{\mathbf{S}}_i = SoftMax(\mathbf{Q}_i\mathbf{K}_i^T/\sqrt{d} + \mathbf{E})\mathbf{V}_i, \tag{12}$$

$$\widetilde{\mathbf{S}} = MLP(\sum_{i=0}^{d}\widetilde{\mathbf{S}}_i), \tag{13}$$

where $\mathbf{W}_i^q, \mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^d$ are learnable parameters producing $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{H*W}$. $\mathbf{E}$ denotes positional encoding parameters, and $\sqrt{d}$ is a scaling factor related to the frequency band count. After DFT, frequencies exhibit a non-monotonic relationship with 2D coordinates, specifically showing central symmetry about $(\frac{2}{H}, \frac{2}{W})$. We applied a central shift to this relationship to facilitate subsequent modeling.

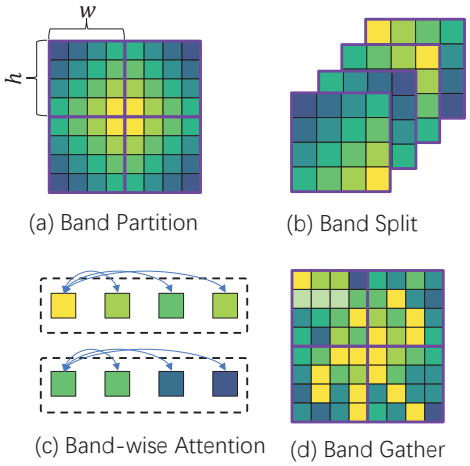

Figure 3: Process of band-wise partition attention module. All frequency of amplitude and phase signals are partitioned by (h,w) kernel size unfolding operation, then the attention is only performed on signals within the same group, after that, all of bands are gathered by folding operation into spatial dimension.

Based on this spectral property, we propose the band-wise partition self-Attention mechanism strategically decomposes the global attention operation into parallelizable sub-band computations as visualized in Fig.3, h and w denotes the Vertical and horizontal stride to partition frequency bands, and in Eq. (12), the number of frequency band group d should be rewritten as $d = \frac{H*W}{h*w}$, and $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{h*w}$ denote the query, keys and value for an identity group band. In this case, our band-wise partitioning paradigm achieves dual objectives: 1) *Preserving frequency-domain data distributions,* 2) *Enabling regional attention with reduced dimensionality.*

Phase and Amplitude Token Exchange

Conventional methods process amplitude and phase spectra independently, impeding inter-modal interaction. We propose a cross-spectral attention module as shown in Fig. 4 utilizing token exchange to facilitate adaptive, bidirectional guidance between these modalities.

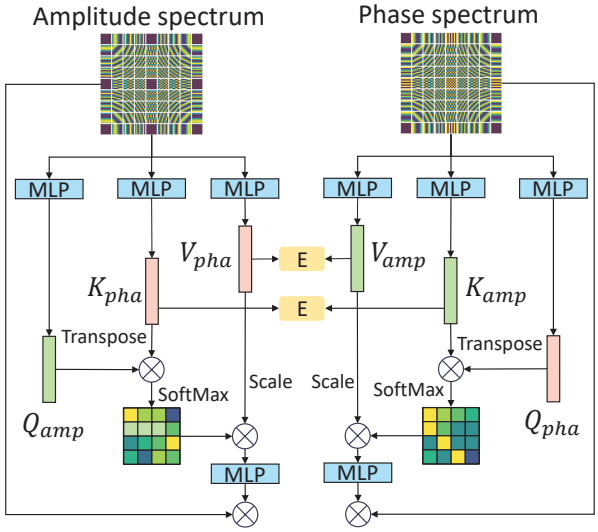

Figure 4: Illustration of DenoDet V2's token exchange scheme.

Methodologically, amplitude ($\mathcal{A}$) and phase ($\mathcal{P}$) maps ($\in \mathbb{R}^{H\times W}$) are partitioned into discrete, aligned groups ($\{\hat{\mathcal{A}}_1, ..., \hat{\mathcal{A}}_d\}$, $\{\hat{\mathcal{P}}_1, ..., \hat{\mathcal{P}}_d\}$, where $\hat{\mathcal{A}}_i, \hat{\mathcal{P}}_i \in \mathbb{R}^{h\times w}$) using a congruent stride. After separate MLP projections (generating Q, K, V representations), frequency tokens are exchanged within corresponding groups. This strategy establishes distinct intra-group mappings while maintaining strict inter-group information isolation, thereby preserving frequency characteristics. The intra-group token exchange process is defined as:

$$\mathbf{Q}_i^{\mathcal{A}} = \hat{\mathcal{A}}_i\mathbf{W}_i^{\mathcal{A}^q}, \mathbf{K}_i^{\mathcal{A}} = \hat{\mathcal{P}}_i\mathbf{W}_i^{\mathcal{P}^k}, \mathbf{V}_i^{\mathcal{A}} = \hat{\mathcal{P}}_i\mathbf{W}_i^{\mathcal{P}^v}, \tag{14}$$

$$\mathbf{Q}_i^{\mathcal{P}} = \hat{\mathcal{P}}_i\mathbf{W}_i^{\mathcal{P}^q}, \mathbf{K}_i^{\mathcal{P}} = \hat{\mathcal{A}}_i\mathbf{W}_i^{\mathcal{A}^k}, \mathbf{V}_i^{\mathcal{P}} = \hat{\mathcal{A}}_i\mathbf{W}_i^{\mathcal{A}^v}, \tag{15}$$

where $\mathbf{W}_i^{\mathcal{A}}, \mathbf{W}_i^{\mathcal{P}} \in \mathbb{R}^d$ denote the linear mapping weight for amplitude features $\mathcal{A}$ and phase features $\mathcal{P}$. $Q_i, K_i, V_i \in \mathbb{R}^{h*w\times d}$ are queries, keys and values for each group of corresponding spectrum. In this case, the swap of tokens between amplitude and phase forces the model to interact with the two modalities of information.

Method	**FLOPs** ↓	**#P** ↓	**mAP**↑	$\mathbf{AP_S}$↑	$\mathbf{AP_M}$↑
Two-stage					
Faster R-CNN (Ren et al. 2015)	63.2G	41.37M	39.22	32.55	47.23
Cascade R-CNN (Cai and Vasconcelos 2021)	90.99G	69.17M	53.55	49.09	62.89
Dynamic R-CNN (Zhang et al. 2020a)	63.2G	41.37M	49.75	43.12	59.72
Grid R-CNN (Lu et al. 2019)	180G	64.47M	50.05	42.43	62.01
Libra R-CNN (Pang et al. 2019)	64.02G	41.64M	52.09	45.85	63.52
ConvNeXt (Liu et al. 2022c)	63.84G	45.07M	53.15	45.67	64.55
ConvNeXtV2 (Woo et al. 2023)	120G	110.0G	53.91	47.63	64.67
LSKNet (Li et al. 2023)	53.73G	30.99M	52.39	45.15	63.59
End2end					
DETR (Carion et al. 2020)	24.94G	41.56M	45.73	37.01	58.16
Deformable DETR (Zhu et al. 2021)	51.78G	40.10M	52.00	46.99	63.58
DAB-DETR (Liu et al. 2022a)	28.94G	43.70M	43.31	34.82	56.34
Conditional DETR (Liu et al. 2022b)	28.09G	43.45M	44.04	35.25	56.47
One-stage					
FCOS (Tian et al. 2019)	51.57G	32.13M	52.52	47.01	66.13
RepPoints (Yang et al. 2019)	48.49G	36.82M	51.66	46.66	63.26
ATSS (Zhang et al. 2020b)	51.57G	32.13M	54.95	49.89	67.94
CenterNet (Zhou, Wang, and Krähenbühl 2019)	51.55G	32.12M	53.91	48.88	66.22
PAA (Kim and Lee 2020)	51.57G	32.13M	52.20	46.00	63.90
PVT-T (Wang et al. 2021)	42.19G	21.43M	46.10	38.01	59.53
RetinaNet (Lin et al. 2017)	52.77G	36.43M	46.48	40.25	59.35
TOOD (Feng et al. 2021)	50.52G	30.03M	54.65	50.20	66.72
DDOD (Chen et al. 2021c)	45.58G	32.21M	54.02	49.33	64.70
VFNet (Zhang et al. 2021)	48.38G	32.72M	53.01	47.37	65.39
AutoAssign (Zhu et al. 2020)	51.83G	36.26M	53.95	50.14	63.40
YOLOF (Chen et al. 2021b)	26.32G	42.46M	42.83	33.73	56.19
YOLOX (Ge et al. 2021)	8.53G	8.94M	34.08	28.49	43.06
GFL(w/o Deno) (Li et al. 2020)	52.36G	32.27M	55.01	49.44	67.29
DenoDet V1 (Dai et al. 2024)	52.69G	65.78M	<u>55.88</u>	<u>50.63</u>	<u>68.47</u>
⋆ DenoDet V2	52.47G	32.60M	**56.71**	**51.45**	**68.75**

Table 1: Comparison with SOTA methods on **SARDet-100K**.

Experiments

Implementation Details

Three benchmark datasets were evaluated under unified configurations: **SARDet-100K** (Li et al. 2024a), **SAR-Aircraft-1.0** (Xian et al. 2019), and **AIR-SARShip-1.0** (Zhirui et al. 2023). All experiments employed the MMDetection framework(Chen et al. 2019) with four RTX 4090 GPUs. Key configurations are detailed as follows.

SARDet-100K: This multi-class dataset contains 116,598 images (8:1:1 split) from 10 international sub-collections, covering six maritime/land objects. Images were resized to 512×512 pixels with 50% horizontal flipping. Training used DAdaptAdam optimizer for 12 epochs (batch size 16, LR 1.0, weight decay 0.05).

SAR-Aircraft-1.0: Featuring 7 aircraft categories (3,489 training/879 test images), the dataset was processed into 512×512 sub-images with 200-pixel overlaps. Identical optimizer parameters as SARDet-100K were applied for 12 epochs.

AIR-SARShip-1.0: Comprising 31 Gaofen-3 scenes (3000×3000 pixels), this ship dataset generated 512×512 chips (200-pixel overlap) containing 461 annotated vessels.

Comparison with State-of-the-Arts

Results on SARDet-100K: As shown in the Tab. 1, we quantitatively compared our DenoDet V2 with 25 state-of-the-art methods on the highly challenging SARDet-100K dataset, achieving an mAP of **56.71%** based on the COCO standard. Notably, compared to its baseline detector GFL, DenoDet V2 improved detection accuracy by **1.7%**, and DenoDet V2 achieved the highest mAP of **51.45%** and **68.75%** for small and medium objects respectively, which are particularly susceptible to noise. Furthermore, the increase in floating-point operations (FLOPs) for DenoDet V2 compared to its base model GFL is almost negligible, the balance of accuracy and efficiency demonstrates the potential of DenoDet V2 for object detection in SAR images.

Results on SAR-Aircraft-1.0 and AIR-SARShip-1.0: As detailed in Tab. 2, DenoDet V2 sets a new state-of-the-art on both the SAR-Aircraft-1.0 and AIR-SARShip-1.0 benchmarks, outperforming 25 competing methods. It achieves a mAP of 69.93% on SAR-Aircraft-1.0 and 73.98% on AIR-SARShip-1.0, surpassing the strong RepPoints baseline by 2.8% and 4.1% respectively. These results confirm the model's exceptional capability to handle complex and noisy SAR data.

Method	FLOPs ↓	#P ↓	AIR-SARShip AP↑	SAR-Aircraft AP↑
Two-stage				
Faster R-CNN (Ren et al. 2015)	63.18G	41.35M	67.48	64.71
Cascade R-CNN (Cai and Vasconcelos 2021)	90.98G	69.15M	66.69	64.87
Dynamic R-CNN (Zhang et al. 2020a)	63.18G	41.35M	67.33	64.59
Grid R-CNN (Lu et al. 2019)	180.0G	64.47M	64.36	64.15
Libra R-CNN (Pang et al. 2019)	63.99G	41.61M	67.45	63.46
ConvNeXt (Liu et al. 2022c)	63.82G	45.05M	67.52	67.41
ConvNeXt V2 (Woo et al. 2023)	120.0G	110.0M	69.45	68.04
LSKNet (Li et al. 2023)	53.70G	30.96M	71.66	67.58
End2end				
DETR (Carion et al. 2020)	24.94G	41.56M	9.09	10.61
Deformable DETR (Zhu et al. 2021)	51.77G	40.10M	55.34	62.43
DAB-DETR (Liu et al. 2022a)	28.94G	43.70M	10.72	53.62
Conditional DETR (Liu et al. 2022b)	28.09G	43.45M	9.09	62.25
One-stage				
FCOS (Tian et al. 2019)	51.50G	32.11M	63.66	62.63
GFL (Li et al. 2020)	52.30G	32.26M	65.94	66.90
ATSS (Zhang et al. 2020b)	51.50G	32.11M	64.21	66.01
CenterNet (Zhou, Wang, and Krähenbühl 2019)	51.49G	32.11M	57.82	64.11
PAA (Kim and Lee 2020)	51.50G	32.11M	65.65	66.79
PVT-T (Wang et al. 2021)	41.62G	21.33M	65.59	61.64
RetinaNet (Lin et al. 2017)	52.20G	36.33M	65.50	66.47
TOOD (Feng et al. 2021)	50.46G	32.02M	63.92	62.66
DDOD (Chen et al. 2021c)	45.52G	32.20M	65.29	62.66
VFNet (Zhang et al. 2021)	48.32G	32.71M	64.60	66.17
AutoAssign (Zhu et al. 2020)	51.77G	36.24M	65.55	62.36
YOLOF (Chen et al. 2021b)	26.29G	42.34M	48.32	66.25
YOLOX (Ge et al. 2021)	8.52G	8.94M	59.72	63.65
RepPoints(w/o Deno) (Yang et al. 2019)	48.49G	36.82M	69.88	67.13
DenoDet V1 (Dai et al. 2024)	48.52G	70.33M	<u>72.42</u>	<u>68.60</u>
⋆ DenoDet V2	48.61G	37.15M	**73.98**	**69.93**

Table 2: Comparison with SOTA methods on **AIR-SARShip-1.0** and **SAR-Aircraft**.

Ablation Study

The necessity of phase decomposition:

As shown in Tab. 3, a quantitative ablation study validated the necessity of orthogonal phase angle decomposition. On the SARDet-100K, baseline achieved 55.6% mAP. Implementing orthogonal decomposition elevated performance to 56.1% mAP by mitigating periodic boundary discontinuities. Subsequent trigonometric rectification further improved mAP to 56.2%, demonstrating enhanced phase coherence preservation and reduced optimization landscape complexity.

The Necessity of Phase and Amplitude Refinement:

Comparative experiments validated the superior robustness of the phase spectrum over the amplitude spectrum for SAR DFT denoising in Tab. 4. On SARDet-100K, phase-only refinement surpassed amplitude-only refinement by 0.6% mAP. Critically, simultaneous refinement of both spectrums (DenoDet V2) yielded the best performance, improving mAP by 1.4% over baseline and significantly enhancing small (AP_S: +1.5%) and medium (AP_M: +1.6%) object detection. These results confirm that phase and amplitude denoising processes are complementary, non-conflicting, and robust, particularly for noise-sensitive objects.

Phase Angle Split	Phase Angle Alignment	SARDet-100K		
		mAP	AP_S	AP_M
×	×	55.6	49.5	68.5
✓	×	56.1	50.4	68.6
✓	✓	**56.2**	**51.4**	**68.6**

Table 3: Ablation study on the necessity of phase orthogonal decomposition: impact of phase angular boundary discontinuity issue.

The Necessity of Band Partition:

We investigated the necessity of band partitioning and the impact of different strides in Tab. 5. The BPSA module partitions DFT frequency data to ensure uniform intra-group frequency variation, thereby decoupling signal modulation, reducing convergence difficulty, and minimizing parameter count. Comparative experiments demonstrated that omitting partitioning (stride=1) yielded the worst performance and significantly increased parameters, confirming the necessity

DFT Amplitude Refine	DFT Phase Refine	SARDet-100K		
		mAP	AP_S	AP_M
×	×	55.0	49.4	67.3
✓	×	55.6	50.3	68.2
×	✓	56.2	**51.4**	68.6
✓	✓	**56.4**	50.9	**68.9**

Table 4: Ablation study on the necessity of phase and amplitude refine: phase spectrum vs amplitude spectrum.

Partition Stride	SARDet-100K		
	mAP	AP_S	AP_M
1	52.5	45.6	63.8
2	55.7	49.5	68.7
4	56.2	51.3	68.5
8	**56.7**	**51.5**	**68.8**
16	56.1	51.0	68.2

Table 5: Ablation study on the necessity of band partition: impact of band partition stride.

Design	SARDet-100K		
	mAP	AP_S	AP_M
baseline	55.0	49.4	67.3
No-Exchange	56.4	50.9	**68.9**
Token-Exchange	**56.7**	**51.5**	68.8

Table 6: Ablation study on the necessity of token exchange: impact of token exchange model design.

of band separation. Model accuracy improved with stride, peaking at 8, but slightly decreased at 16. Consequently, DenoDet V2 adopts a partition stride of 8.

The Necessity of Token Exchange: As shown in Tab. 6, ablation studies validated the token exchange strategy. This operation increased mAP on the SARDet-100K dataset from 56.4% to 56.7%, confirming enhanced amplitude-phase interaction and improved modulation. Compared to baseline, this implementation achieved a 1.7% mAP gain overall and a 2.1% AP gain for small objects. This performance improvement demonstrates DenoDetV2's efficacy in suppressing speckle noise and enhancing object saliency via cross-spectral feature recalibration, thereby improving detection fidelity in cluttered environments.

Visual Analysis

Visual analysis validates our core hypothesis. Eigen-CAM-based visualizations shown in Fig. 5 demonstrate the DFT-Deno module's effectiveness: compared to baseline, DenoDet V2 exhibits more focused activations on objects while significantly suppressing attention dispersed in noise-affected areas. Furthermore, from Fig. 6, detection results on SAR-AIRCraft-1.0 and AIR-SARShip-1.0 confirm DenoDet V2's

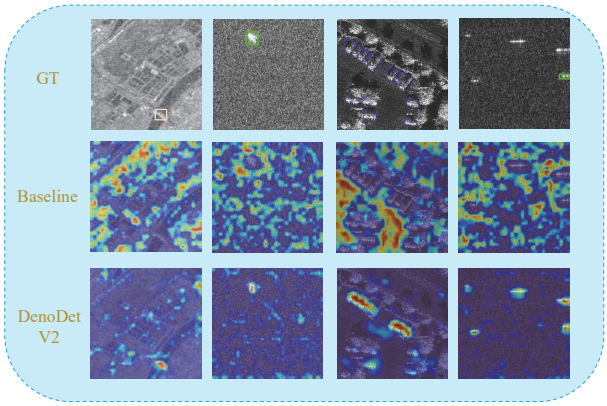

Figure 5: Feature heatmap visualization on SARDet-100K. Compared to the baseline, DenoDet V2 focuses more attention on the object areas while being less affected by the background noise.

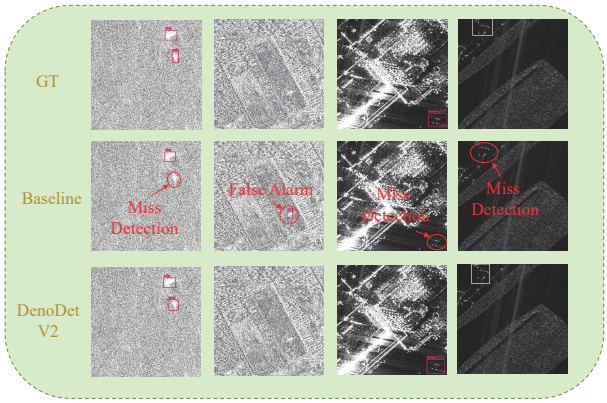

Figure 6: Detection result comparison on SAR-AIRCraft-1.0 and AIR-SARShip-1.0 dataset. DenoDet V2 achieved more accurate detection results in noisy regions and identified more ship objects.

robustness, showing accurate detection under severe noise interference without noise-induced false alarms.

Conclusion

In this paper, we presented DenoDet V2, an approach for robust SAR image object detection that leverages the complementary nature of amplitude and phase information through a mutual guidance mechanism. Our method introduces the concept of Attention as Phase-Amplitude Denoising, which exploits the inherent stability of phase to guide the feature denoising of amplitude in SAR imagery. The DFTDeno module, a key component of DenoDet V2, enables a reciprocal enhancement between phase and amplitude spectra, resulting in a synergistic feature denoising effect that boosts the accuracy of object detection. Through extensive experiments on various SAR datasets, we demonstrated the state-of-the-art performance of DenoDet V2, showcasing its potential for robust object detection in noisy SAR environments.

Acknowledgments

This work is supported by the National Science Fund of China (62101280,62301261,62206134,62472443,62225604), Shenzhen Science and Technology Program (JCYJ20250604184027034,JCYJ20240813114237048) and the Fellowship of China Postdoctoral Science Foundation (2023M731781).

References

Cai, Z.; and Vasconcelos, N. 2021. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1483–1498.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, G.; Peng, P.; Ma, L.; Li, J.; Du, L.; and Tian, Y. 2021a. Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Networks in Frequency Domain. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 448–457.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.

Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; and Sun, J. 2021b. You Only Look One-Level Feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13039–13048.

Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.-J.; and Wu, F. 2021c. Disentangle your dense object detector. In *ACM International Conference on Multimedia*, 4939–4948.

Dai, Y.; Zou, M.; Li, Y.; Li, X.; Ni, K.; and Yang, J. 2024. DenoDet: Attention as Deformable Multi-Subspace Feature Denoising for Target Detection in SAR Images. *IEEE Transactions on Aerospace and Electronic Systems (TAES)*.

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. TOOD: Task-aligned One-stage Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3490–3499. IEEE Computer Society.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint, arXiv:2107.08430*.

Ke, X.; Zhang, X.; Zhang, T.; Shi, J.; and Wei, S. 2021. SAR ship detection based on an improved faster R-CNN using deformable convolution. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3565–3568.

Kim, K.; and Lee, H. S. 2020. Probabilistic Anchor Assignment with IoU Prediction for Object Detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 355–371. Springer.

Li, D.; Liang, Q.; Liu, H.; Liu, Q.; Liu, H.; and Liao, G. 2022a. A Novel Multidimensional Domain Deep Learning Network for SAR Ship Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Li, D.; Liang, Q.; Liu, H.; Liu, Q.; Liu, H.; and Liao, G. 2022b. A novel multidimensional domain deep learning network for SAR ship detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5203213.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *Advances in Neural Information Processing Systems*.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023. Large selective kernel network for remote sensing object detection. In *IEEE International Conference on Computer Vision*, 16794–16805.

Li, Y.; Li, X.; Li, W.; Hou, Q.; Liu, L.; Cheng, M.-M.; and Yang, J. 2024a. SARDet-100K: towards open-source benchmark and toolKit for large-scale SAR object detection. In *Advances in Neural Information Processing Systems*, 128430–128461. Brookline, MA, USA: PMLR.

Li, Y.; Li, X.; Li, Y.; Yicheng, Z.; Dai, Y.; Hou, Q.; Cheng, M.-M.; and Yang, J. 2024b. SM3Det: A Unified Model for Multi-Modal Remote Sensing Object Detection. *arXiv*.

Li, Y.; Zhang, Y.; Tang, W.; Dai, Y.; Cheng, M.-M.; Li, X.; and Yang, J. 2025. Visual Instruction Pretraining for Domain-Specific Foundation Models. *arXiv*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. DAB-DETR: Dynamic anchor boxes are better queries for DETR. *arXiv preprint arXiv:2201.12329*.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022b. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3651–3660.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022c. A ConvNet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Lu, X.; Li, B.; Yue, Y.; Li, Q.; and Yan, J. 2019. Grid R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7363–7372.

Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; and Lin, D. 2019. Libra R-CNN: Towards Balanced Learning for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 821–830.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in neural information processing systems*, 28.

Shen, H.; Zhou, C.; Li, J.; and Yuan, Q. 2021. SAR image despeckling employing a recursive deep CNN prior. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1): 273–286.

Sun, Y.; Sun, X.; Wang, Z.; and Fu, K. 2021. Oriented ship detection based on strong scattering points network in large-scale SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5218018.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9627–9636.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; Cai, Z.; and Yuan, J. 2023. Automatic SAR Ship Detection Based on Multifeature Fusion Network in Spatial and Frequency Domains. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16133–16142.

Xian, S.; Zhirui, W.; Yuanrui, S.; Wenhui, D.; Yue, Z.; and Kun, F. 2019. AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *Journal of Radars*, 8(6): 852–863.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; and Ren, F. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1737–1746.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. RepPoints: Point Set Representation for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9657–9666.

Ye, Y.; Teng, X.; Yang, H.; Chen, S.; Sun, Y.; Bian, Y.; Tan, T.; Li, Z.; and Yu, Q. 2025. 3MOS: a multi-source, multi-resolution, and multi-scene optical-SAR dataset with insights for multi-modal image matching. *Visual Intelligence*, 3(1): 19.

Zhang, H.; Chang, H.; Ma, B.; Wang, N.; and Chen, X. 2020a. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, 260–275. Springer.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. VarifocalNet: An IoU-Aware Dense Object Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8514–8523.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020b. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9759–9768.

Zhang, X.; Li, D.; Dong, X.; Wu, T.; Yu, H.; Wang, J.; Li, Q.; and Li, X. 2025a. UniChange: Unifying Change Detection with Multimodal Large Language Model. *arXiv preprint arXiv:2511.02607*.

Zhang, X.; Yang, X.; Li, Y.; Yang, J.; Cheng, M.-M.; and Li, X. 2025b. RSAR: Restricted State Angle Resolver and Rotated SAR Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7416–7426.

Zhao, C.; Fu, X.; Dong, J.; Qin, R.; Chang, J.; and Lang, P. 2022. SAR Ship Detection Based on End-to-End Morphological Feature Pyramid Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 4599–4611.

Zhao, J.; Zhang, Z.; Yu, W.; and Truong, T.-K. 2018. A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images. *IEEE Access*, 6: 50693–50708.

Zhirui, W.; Yuzhuo, K.; Xuan, Z.; Yuelei, W.; Ting, Z.; and Xian, S. 2023. SAR-AIRcraft-1.0: High-resolution SAR aircraft detection and recognition dataset. *Journal of Radars*, 12(4): 906–922.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as Points. *arXiv preprint arXiv:1904.07850*.

Zhou, Y.; Jiang, X.; Xu, G.; Yang, X.; Liu, X.; and Li, Z. 2022. PVT-SAR: An Arbitrarily Oriented SAR Ship Detector With Pyramid Vision Transformer. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16: 291–305.

Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; and Sun, J. 2020. AutoAssign: Differentiable Label Assignment for Dense Object Detection. *arXiv preprint arXiv:2007.03496*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.