

ShipGen: Improving Remote Sensing Ship Detection via Layout-Conditioned Diffusion-Based Data Augmentation

Yuhan Zhang¹, Jun Liao¹, Jie Zhou¹, Yuanpei Chen¹, Xiaode Liu¹, Weihang Peng¹, Hongyu Yang¹, Qi Ming², and Yufei Guo¹, *Member, IEEE*

Abstract—Synthetic aperture radar ship detection plays a key role in the field of remote sensing. Data augmentation is frequently used to enhance detector performance due to the common challenge of limited annotated data. However, most augmented methods are limited to geometric transformations, “copy–paste,” or image inpainting, which are guided by strict constraints and suffer limited target and background diversity. To address this, we propose a novel multiphase transfer-learned framework with layout-conditioned diffusion model. It incorporates sparse target layouts as relaxed conditions into the latent diffusion model, where the conditions take the form of bounding boxes that specify target positions and sizes. This leaves other ship attributes (e.g., orientation and scattering characteristics) and backgrounds to be freely generated and enhance target and background diversity. For enhancing the model’s convergence on small datasets, we first pretrain it on large-scale datasets, and then, fine-tune it on small-scale ones for transfer learning. This enables precise layout-guided target generation even when using small datasets. For subsequent data augmentation, we adopt the widely used grounded-language-to-image generation model for practical use, and based on the transfer-learned model, we further develop a new automatic augmentation pipeline with a filtering mechanism. This pipeline integrates multiple foundation models and tools, including layout augmentation, segment-anything-model-based annotation correction, and data filtering, and can automatically adjust inappropriate annotations and screen high-quality data to serve as augmented data for downstream detection tasks. Experiments on SSDD and HRSID datasets validate that the pipeline improves multiple detectors’ performance across in-shore and off-shore scenarios, and qualitative analyses confirm high diversity and quality of generated images.

Index Terms—Data augmentation, latent diffusion model (LDM), synthetic aperture radar (SAR), ship detection.

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) is an advanced active microwave imaging technology capable of all-weather, day–night Earth observation. By synthesizing antenna motion, it

achieves high-resolution remote sensing independent of illumination and meteorological conditions, while exhibiting unique penetration capabilities through certain surface obstructions. These characteristics have positioned SAR ship detection at the forefront of remote sensing research [1], [2].

The application of deep learning (DL) in SAR ship detection has significantly advanced remote sensing capabilities, achieving state-of-the-art performance in multitarget localization and classification under complex oceanic conditions [3], [4], [5]. However, these achievements critically depend on supervised learning paradigms requiring large-scale annotated datasets—a major limitation due to the complexities of the SAR imaging process and the prohibitive labor-intensive annotation costs, collectively posing obstacles to advancement of SAR ship detection technology.

Data augmentation serves as an effective strategy to mitigate SAR data scarcity constraints. While conventional geometry-based methods (e.g., rotation, cropping, and scaling) expand dataset volume, they inherently provide limited feature-space diversity enhancement. To address this, advanced augmentation techniques, such as BoxPaste [6], [7] and Mosaic [8], have been adopted in SAR ship detection. Although these approaches enhance target diversity and yield modest performance improvements in detectors, a persistent domain gap exists between the distribution of augmented data and that of the original dataset. This distribution misalignment fundamentally constrains further detector performance gains.

To bridge this distribution gap, generative-model-based augmentation demonstrates superior capability in learning and replicating underlying data distributions. With rapid advances in generative artificial intelligence, techniques such as generative adversarial networks (GANs) [9] and diffusion models [10] have been adapted for SAR data synthesis. Yang et al. [11] employed a Pix2Pix framework to generate SAR images from edge maps, but the absence of texture details in input representations increased generation complexity, yielding outputs with suboptimal fidelity compared to authentic SAR data. Sun et al. [12] developed style-embedded ship sample augmentation that composites synthetic ship patches onto background imagery. While enhancing sample diversity and realism, this approach inherited background constraints from the original dataset, neglecting environmental diversity essential for clutter rejection learning of detectors. Ship-Go [13] addressed this limitation through denoising

Received 1 September 2025; revised 30 January 2026; accepted 11 May 2026. Date of publication 14 May 2026; date of current version 29 May 2026. (Yuhan Zhang and Jun Liao contributed equally to this work.) (Corresponding author: Yufei Guo.)

Yuhan Zhang, Jie Zhou, Yuanpei Chen, Xiaode Liu, Weihang Peng, and Yufei Guo are with the Intelligent Science and Technology Academy of CASIC, Beijing 100043, China (e-mail: yfguo@pku.edu.cn).

Jun Liao and Hongyu Yang are with the Beihang University, Beijing 100191, China.

Qi Ming is with Beijing University of Technology, Beijing 100021, China. Digital Object Identifier 10.1109/JSTARS.2026.3693312

diffusion probabilistic models (DDPM) conditioned on ship instances and trainable environmental prompts. It successfully expanded background variability for both in-shore and off-shore scenarios. However, Ship-Go [13] is an image inpainting method, and the targets in the images generated by Ship-Go are derived from the targets in existing datasets, resulting in limited diversity of its generated targets. And the lack of an automatic filtering mechanism may affect the augmentation effect. SVDDD [14] established a comprehensive diffusion-driven framework integrating a fine-tuned stable diffusion model, ControlNet [15], and multistage image filtering. This architecture significantly enhanced background diversity and demonstrated measurable improvements in downstream detector performance.

Nevertheless, the target diversity in the aforementioned diffusion-based works remains limited, leaving room for further improvement. This limitation is closely tied to the conditional inputs these methods employ. For instance, SVDDD [14] adopts segmentation masks as diffusion model conditions, while Ship-Go [13] uses visual instances. Both the conditions fall under the category of dense conditions with strict constraints. These constraints lie in the strict restriction of the target's orientation, shape, and even scattering characteristics, which in turn limits the diversity of targets generated through image inpainting guided by such strict constraints. Notably, this lack of target diversity will restrict improvement of detectors' performance, because sufficient target diversity is essential for effectively training detectors' localization modules and improving their target identification ability. In addition, sufficient background diversity is equally critical for enhancing detectors' ability to reject nontargets (e.g., large ground targets and small islands) [13]. Given the dual significance of target and background diversity, this article focuses on enhancing the diversity of target and background generation.

Therefore, we provide the latent diffusion model (LDM) with sparse conditional inputs in the form of target layouts. These target layouts are represented as a series of bounding boxes, and they specify the positions and relative sizes of targets, which are a kind of relaxed constraints for image generation. Thus, other ship attributes (such as orientation and scattering characteristics) as well as the background are left to the LDM for free generation. This helps improve diversity of target and background. However, we found that training on small datasets makes it hard for the LDM to generate diverse targets at the positions specified by the given layouts. This issue makes layout conditions unsuitable to be used as labels for target detection in subsequent augmented data. It may also require substantial resources or effort to reannotate the generated images. To avoid this problem, we propose a multiphase transfer learning framework for layout-conditioned diffusion model. We first pretrain the model on large-scale datasets, and then, fine-tune it on small-scale datasets. This enables the LDM to generate targets with precise control based on given layouts, even when using small-scale datasets. For practical implementation of this framework, we selected the widely used grounded-language-to-image generation (GLIGEN) model. In addition, building on the transfer-learned GLIGEN model, we further developed an automatic end-to-end augmentation pipeline equipped with a

filtering mechanism. This pipeline integrates multiple foundation models and tools, including layout augmentation, segment anything model (SAM)-based annotation correction, and data filtering. It can automatically adjust inappropriate annotations and screen high-quality data to serve as augmented data for downstream detection tasks.

The contributions of this article are as follows.

- 1) Layout conditions are adopted to guide the LDM in generating SAR ship images, and these conditions impose more relaxed constraints than existing augmentation methods. This lays a solid foundation for improving object and background diversity in data augmentation. Furthermore, through a novel multiphase transfer learning framework, the model can generate targets accurately according to given layouts even for small-scale datasets.
- 2) Based on the transfer-learned model, a new automatic end-to-end data augmentation pipeline with a filtering mechanism is developed. This pipeline integrates multiple foundation models and tools, including layout augmentation, SAM-based annotation correction, and data filtering. It can automatically adjust inappropriate annotations and screen high-quality data, providing usable augmented data for downstream target detection tasks.
- 3) We validate the quality and diversity of generated data using the benchmark SAR ship detection datasets including SSDD and HRSID. Through quantitative experiments, we demonstrate that our method consistently boosts the performance of diverse deep detection models by augmenting datasets across in-shore and off-shore scenarios. Complementary qualitative analyses further affirm the method's effectiveness, highlighting the rich diversity of generated images.

The rest of this article is organized as follows. Section II reviews related work, while Section III introduces the proposed methodology. Experimental results are analyzed in Section IV, and finally, Section V concludes this article.

II. RELATED WORK

A. Generative-Model-Based SAR Data Augmentation

Generative models play a pivotal role in enhancing data diversity and balancing distribution, particularly in SAR applications. In the realm of SAR, GANs [9] have demonstrated remarkable efficacy in data augmentation for tasks such as target recognition and detection. Cui et al. [16] introduced a WGAN-GP-based method with sample selection filters, generating azimuth-specific high-quality samples that significantly elevate SAR target recognition rates in small-sample scenarios, outperforming traditional linear generation techniques. Kong et al. [17] leveraged GAN-based augmentation to address dataset limitations in CNN-driven SAR target recognition, achieving promising accuracy improvements on MSTAR datasets. To tackle challenges such as sample scarcity and speckle noise, Mao et al. [18] integrated CN-GAN (blending LSGAN and Pix2Pix) with a shallow CNN, generating high signal-to-clutter-noise ratio (SCNR) data that outperform original and other enhanced datasets in recognition tasks. Du et al. [19]

enhanced SAR data augmentation with an improved WGAN-GP model, utilizing an encoder and pretrained classifier to stably produce high-quality images that bolster classification accuracy. Kwon et al. [20] proposed a cGAN-based approach for SAR ship detection, generating synthetic images to train detection networks, which proved effective in experimental evaluations. For high-resolution SAR ship detection with limited samples, Zou et al. [21] developed MW-ACGAN to generate multiscale, multiclass ship images, with Yolo v3 trained on composite datasets achieving 94% accuracy, far surpassing models trained solely on original data. Guo et al. [22] introduced RAGAN, a residual and attention-based GAN, which employs directional bounding boxes and residual-attention blocks to enhance generator performance, demonstrating efficacy in ship detection tasks.

Diffusion models [10], a newer class of generative models, have shown superiority over GANs in producing diverse and realistic images. Qosja et al. [23] pioneered the application of DDPMs to SAR, achieving better image generation quality than GANs and further improving performance through pretraining on large clutter datasets. Zhang et al. [13] proposed Ship-Go, a multicondition DDPM-based model that generates SAR images with instance annotations by constraining ship placement, supporting both in-domain and out-of-domain augmentation to enhance detector performance and dataset diversity. To address data scarcity in SAR vehicle detection, Sun et al. [24] combined diffusion models with scattering information, using DDPM for sample augmentation, extracting scattering features for similarity calculation, and integrating recognition networks to improve target recognition under limited samples. Ying et al. [25] introduced a lightweight frequency domain-aware diffusion model with a fast Fourier-transform-based frequency adjustment module, which stabilizes high-frequency component generation, resulting in realistic SAR images.

In summary, while SAR image generation has advanced data augmentation for recognition, detection-specific challenges—including the need for both target and background diversity, annotation reliability validation, and annotation automatic correction—are inadequately addressed by existing methods.

B. Diffusion Models

Diffusion models [10], [26], [27], celebrated for their stable training dynamics and superior generative performance, are progressively overtaking GANs to become the leading framework in generative tasks. Text-guided diffusion models excel at producing lifelike images, yet the inherent brevity of textual descriptions often hinders precise control over the generation process, thereby constraining their capacity for personalized content creation.

To tackle this limitation, researchers have begun integrating supplementary control conditions alongside text guidance, dramatically broadening the application landscape of diffusion models. Such extensions span a diverse range of tasks: layout-guided generation [28], [29], [30] for structured content creation, style transfer [31], [32] for aesthetic adaptation, image restoration tasks such as denoising and superresolution [33], and even video generation [34]. These applications underscore

the immense versatility of diffusion models in handling complex generative challenges.

Notably, the LDM [27] has emerged as a foundational framework in this domain. By confining the diffusion process to a low-dimensional latent space, the LDM not only retains the high fidelity of generated images but also achieves a substantial reduction in computational overhead, laying the groundwork for countless subsequent advancements in generative modeling.

Building on these advancements in diffusion models, particularly their shift toward enhanced controllability through multicondition guidance and the efficiency gains of latent space operations, our work extends this paradigm to address a critical gap in data augmentation for SAR ship detection. This approach not only aligns with the broader trend of refining diffusion-based controllability but also tailors it to the unique demands of detection-oriented data augmentation.

III. METHODOLOGY

A. Preliminaries

1) *Denoising Diffusion Probabilistic Models (DDPM)*: DDPM [35] is a class of generative models that learns data distributions through iterative noise injection and denoising. The process consists of the following two Markov chains.

Forward diffusion process: Gradually corrupts data x_0 with Gaussian noise over T steps

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where $\beta_t \in (0, 1)$ is a predefined noise schedule, and $\mathbf{I}$ denotes the identity matrix. The direct transition from x_0 to x_t is given by

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (2)$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

Reverse denoising process: Learns to reconstruct data from noise

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (3)$$

The mean is parameterized as

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (4)$$

where ϵ_θ is a neural network predicting the noise component.

The training objective minimizes the variational lower bound, simplified to

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2] \quad (5)$$

typically implemented with U-Net architectures conditioned on timestep embeddings.

2) *Latent Diffusion*: Latent Diffusion [27] extends DDPM by operating in a latent space to reduce computational cost. It employs a pretrained variational autoencoder (VAE) with encoder Enc_{VAE} and decoder Dec_{VAE} as follows:

$$\mathbf{z} = \text{Enc}_{\text{VAE}}(\mathbf{x}) \quad (6)$$

$$\hat{\mathbf{x}} = \text{Dec}_{\text{VAE}}(\mathbf{z}). \quad (7)$$

The diffusion process occurs in latent space with modified loss

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}), \epsilon, t} [\|\epsilon - \epsilon_{\theta}(\mathbf{z}_t, t, \tau(y))\|^2] \quad (8)$$

where $\tau(y)$ projects text/image conditions.

Conditional generation integrates prompts via cross-attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (9)$$

$$Q = W_Q\phi(\mathbf{z}_t), \quad K = W_K\tau(y), \quad V = W_V\tau(y). \quad (10)$$

This enables high-resolution image synthesis with $5\times$ fewer computations than pixel-space DDPM.

Stable diffusion (SD) is a text-to-image model built on latent diffusion technology. It generates high-quality images from user-provided text descriptions. To achieve this, the text descriptions are first encoded using CLIP [36] into representations that align with the latent space of images. These text representations are then incorporated into the diffusion process using cross-attention mechanisms.

3) *Grounded-Language-to-Image Generation (GLIGEN)*: GLIGEN [28] enhances pretrained text-to-image diffusion models (e.g., SD) with spatial grounding capabilities without catastrophic forgetting. Given

- 1) text prompts $C = \{c_i\}_{i=1}^m$ with token embeddings $\mathbf{c}_i$;
- 2) grounding entities $E = \{e_j\}_{j=1}^n$ with spatial coordinates $\mathbf{b}_j = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$.

The model processes grounding tokens via

$$\mathbf{g}_j = \text{MLP}(\text{Enc}_{\text{text}}(e_j) \oplus \text{Fourier}(\mathbf{b}_j)) \quad (11)$$

where $\oplus$ denotes concatenation and $\text{Fourier}(\cdot)$ computes positional embeddings.

The core innovation is the *gated self-attention* mechanism

$$\text{Attn}_{\text{gate}}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

$$+ \lambda \tanh(\gamma) \cdot \text{Softmax}\left(\frac{Q_g K_g^T}{\sqrt{d_k}}\right)V_g \quad (13)$$

where Q_g and K_g are queries/keys from grounded features. The gating parameter γ (initialized at 0) controls the gradual infusion of spatial information while freezing original diffusion weights.

The training loss extends standard diffusion loss with grounding conditioning

$$\mathcal{L}_{\text{GLIGEN}} = \mathbb{E}_{t, \epsilon, C, E} [\|\epsilon - \epsilon_{\theta}(x_t, t, C, E)\|^2]. \quad (14)$$

Inference uses a two-stage strategy: activate grounding for initial k steps ($\gamma = 1$) then disable ($\gamma = 0$) to preserve output quality.

4) *Segment Anything Model (SAM)*: SAM [37] is a promptable segmentation foundation model supporting zero-shot generalization. Its architecture comprises

$$\text{Mask} = f_{\text{decoder}}(f_{\text{image}}(I), f_{\text{prompt}}(P)) \quad (15)$$

where

- f_{image} : MAE-pretrained ViT encoder generating image embeddings with input image I ;

- f_{prompt} : Encoder for sparse (points/boxes/text) or dense (masks) prompts P ;
- f_{decoder} : Transformer-based mask decoder combining embeddings.

Point prompts are encoded by

$$\mathbf{p}_i^p = \text{MLP}(\text{PE}(x_i, y_i)). \quad (16)$$

And box prompts are encoded by

$$\mathbf{p}_j^b = \text{MLP}(\text{PE}(x_{\min}, y_{\min}) \oplus \text{PE}(x_{\max}, y_{\max})) \quad (17)$$

with PE denoting positional encoding. SAM predicts multiple masks for ambiguous prompts using intersection over union (IoU)-based ranking.

SAM was trained on the SA-1B dataset, the largest segmentation dataset with 1 billion masks, enabling unprecedented generalization across domains including medical imaging, remote sensing without fine-tuning.

B. Our Method

Based on the aforementioned technique, we propose our diffusion-model-based SAR ship detection dataset enhancement pipeline, as illustrated in Fig. 1, is divided into four stages: layout augmentation, image generation, annotation correction, and data filtering.

1) *Layout Augmentation*: Layout is defined as a series of bounding box annotations of targets, which specifies targets' positions and relative sizes. It can be regarded as a kind of relaxed constraints for data augmentation when comparing it with segmentation masks in [14] and visual instances in [13], which are more strict constraints. Layout-conditioned data augmentation has already delivered greater target diversity than other diffusion-model-based methods [13], [14]. To further boost this diversity, the synthetic layout augmentation process transforms raw bounding box annotations $\mathcal{B}_{\text{raw}} = \{b_i\}_{i=1}^N$, $b_i = [x_i, y_i, w_i, h_i]$ into geometrically diversified synthetic layouts $\mathcal{B}_{\text{syn}} = \{\hat{b}_i\}_{i=1}^N$, $\hat{b}_i = [\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i]$ through stochastic spatial transformations. The parameters of each transformation are provided as follows:

$$\text{Scale: } s \sim \mathcal{U}(s_{\min}, s_{\max}) \quad (18)$$

$$\text{Horizontal flip: } f_h = \begin{cases} 1 & \text{with probability 0.5} \\ 0 & \text{otherwise} \end{cases} \quad (19)$$

$$\text{Vertical flip: } f_v = \begin{cases} 1 & \text{with probability 0.5} \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

$$\text{Translation: } \begin{cases} \Delta x = \mathcal{U}(-r, r) \cdot W \\ \Delta y = \mathcal{U}(-r, r) \cdot H. \end{cases} \quad (21)$$

Then, each transformation applies

$$\hat{w}_i = s \cdot w_i, \quad \hat{h}_i = s \cdot h_i \quad (22)$$

$$\hat{x}_i = s \cdot x_i + \Delta x, \quad \hat{y}_i = s \cdot y_i + \Delta y \quad (23)$$

$$\hat{x}_i = \begin{cases} W - (\hat{x}_i + \hat{w}_i), & f_h = 1 \\ \hat{x}_i, & f_h = 0 \end{cases} \quad (24)$$

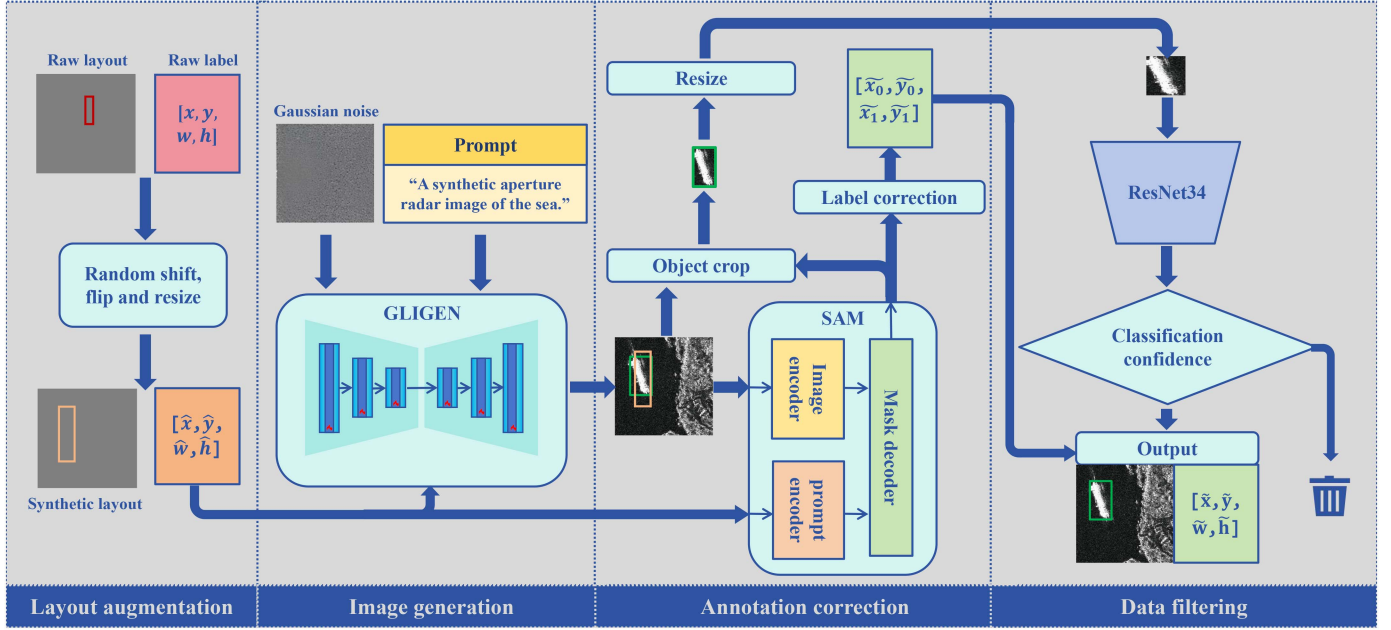


Fig. 1. Pipeline of layout-conditioned diffusion-based SAR ship detection dataset augmentation.

$$\hat{y}_i = \begin{cases} H - (\hat{y}_i + \hat{h}_i), & f_v = 1 \\ \hat{y}_i, & f_v = 0. \end{cases} \quad (25)$$

2) *Transfer Learning for Enhanced Condition Control in GLIGEN-Based Image Generation*: Synthetic SAR images are generated conditioned on synthetic layouts $\mathcal{B}_{\text{syn}}$ via spatially grounded diffusion model GLIGEN. The GLIGEN model $\mathcal{G}$ takes inputs

- 1) text prompt: $T = \text{"A synthetic aperture radar image of the sea"};$
- 2) target grounding: Entity description $T_t = \text{"ship"}$ and spatial layout $\mathcal{B}_{\text{syn}}$, which has been converted into the form of $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$.

In latent space, the denoising process integrates grounding conditions through gated self-attention

$$\mathbf{z}_t = \epsilon_{\theta}(\mathbf{z}_{t+1}, t, \phi_T(T), \text{Attn}_{\text{gate}}(\mathbf{Q}_B, \mathbf{K}_B, \mathbf{V}_B)) \quad (26)$$

where ϕ_T is text encoder of CLIP [36], and

$$\mathbf{K}_B = \mathbf{W}_k \cdot (\text{PE} + \mathbf{g}(T_t, \mathcal{B}_{\text{syn}})) \quad (27)$$

$$\mathbf{V}_B = \mathbf{W}_v \cdot (\text{PE} + \mathbf{g}(T_t, \mathcal{B}_{\text{syn}})) \quad (28)$$

where PE is a trainable position embedding, and

$$\mathbf{g}(T_t, \mathcal{B}_{\text{syn}}) = \text{MLP}(\phi_T(T_t) \oplus \text{Fourier}(\mathcal{B}_{\text{syn}})). \quad (29)$$

The generated image $\mathcal{I}_{\text{gen}}$ is decoded by the VAE decoder

$$\mathcal{I}_{\text{gen}} = \text{Dec}_{\text{VAE}}(\mathbf{z}_0). \quad (30)$$

However, we identify a critical limitation in training GLIGEN's gated attention layers on small-scale datasets

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^N \|\text{Position}(b_i^{\text{gen}}) - \text{Position}(b_i^{\text{cond}})\|_2 \quad (31)$$

TABLE I
ALIGNMENT PERFORMANCE UNDER DIFFERENT TRAINING STRATEGIES

Training Strategy	Dataset	$\mathcal{L}_{\text{align}}$ (px)	IoU
Baseline (scratch)	SSDD	28.3 ± 4.1	0.40
Baseline (scratch)	HRSID	26.7 ± 3.2	0.42
Proposed transfer	SSDD	6.3 ± 0.9	0.84
Proposed transfer	HRSID	6.5 ± 0.8	0.81
Pretrain only	SARDet-100K	7.8 ± 1.1	0.78

where $\text{Position}(b_i^{\text{gen}})$ denotes generated object coordinates and $\text{Position}(b_i^{\text{cond}})$ denotes condition coordinates. As shown in Table I, small datasets (e.g., SSDD and HRSID) yield $\mathcal{L}_{\text{align}} > 25$ pixels, indicating *significant misalignment* between generated objects and condition boxes. This is because the small variance of the conditional input leads to insufficient diversity of the gradient during the optimization process, which is not conducive to effective parameter convergence.

To address this, as shown in Fig. 2, we propose a multiphase transfer learning framework.

Large-scale pretraining: In Phase 1, we conduct warm-start training by initializing a VAE with weights from SD 1.4 and training it on the SARDet-100 K dataset [38]. This primes the model with SAR-specific features while leveraging pretrained generative capabilities.

Subsequently, in Phase 2, we perform cold-start training—optimizing GLIGEN's gated attention layers directly on SARDet-100 K from scratch

$$\theta_g^* = \underset{\theta_g}{\text{argmin}} \sum_{(x_t, c) \in \mathcal{D}_{\text{sardet}}} \mathcal{L}_{\text{GLIGEN}} \quad (32)$$

where θ_g^* is trained from scratch, with γ initialized as 0. This approach enables the model to directly learn SAR ship

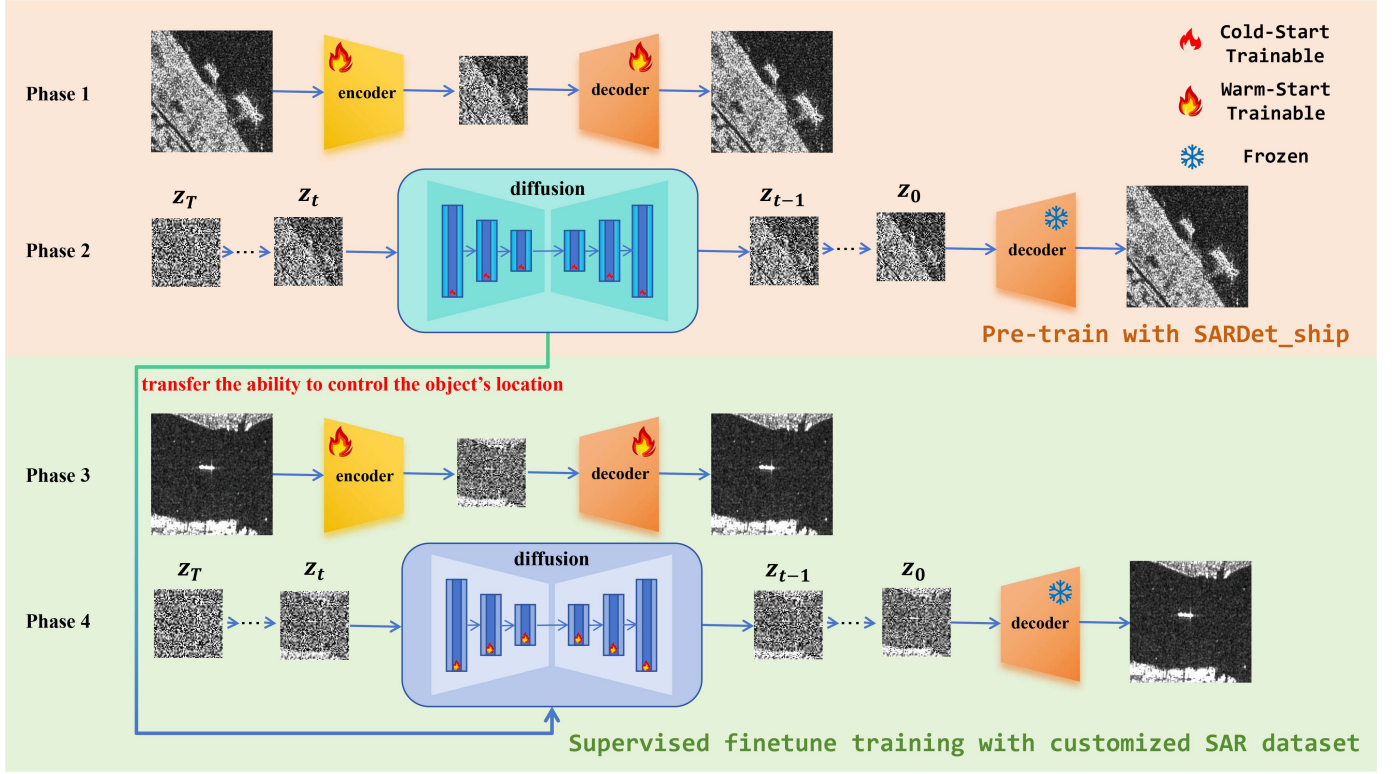


Fig. 2. Framework of transfer learning for enhanced condition control in GLIGEN-based image generation.

distributions and conditional controls from scratch. Through training on large-scale datasets, the model achieves thorough fitting of conditional control capabilities while maintaining strong generalization performance, ensuring effective adaptation to diverse SAR scenarios and detection tasks.

Domain-specific transfer: In Phase 3, we initialize the VAE with the weights of SD 1.4 and train it on a customized dataset to align the latent space with domain-specific features. Concurrently, for the target domain $\mathcal{D}_{\text{custom}}$, we initialize the GLIGEN generator parameters θ_g with pretrained weights θ_g^* and fine-tune them using the customized dataset.

$$\theta_g^{**} = \underset{\theta_g}{\operatorname{argmin}} \sum_{(x_t, c) \in \mathcal{D}_{\text{custom}}} \mathcal{L}_{\text{GLIGEN}}. \quad (33)$$

This approach facilitates the precise transfer of spatial control capabilities to novel target domains, ensuring adaptive data generation with diverse spatial conditions.

3) *SAM-Based Annotation Correction:* There may be deviations between the synthetic spatial grounding $\mathcal{B}_{\text{syn}}$ and the actual annotations of the targets in the generated images. Large deviations can lead to a decline in the localization ability of the detector, so it is necessary to correct the annotations of the targets in the generated images. We feed generated images $\mathcal{I}_{\text{gen}}$ and corresponding spatial grounding $\mathcal{B}_{\text{syn}}$ to SAM for annotation refinement. The mask of object area is predicted by

$$\mathcal{M}_i = f_{\text{decoder}}(f_{\text{image}}(\mathcal{I}_{\text{gen}}), f_{\text{prompt}}(\mathcal{B}_{\text{syn}})). \quad (34)$$

Then, the annotation of objects can be derived by converting the mask

$$\mathcal{B}_{\text{sam}} = [\tilde{x}, \tilde{y}, \tilde{w}, \tilde{h}]. \quad (35)$$

4) *Detection-Oriented Data Filtering:* Given the critical importance of image quality and annotation precision in detection-oriented data augmentation, integrating augmented data that preserves object fidelity comparable to the source imagery with accurate annotations can substantially improve detection model accuracy.

To eliminate samples exhibiting suboptimal target fidelity or erroneous annotations, we implement a ResNet-34-based filtration framework.

$$\mathcal{F}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) = \text{ResNet34}(\text{Resize}(\text{Crop}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}), 224)) \quad (36)$$

where $\mathcal{I}_{\text{gen}}$ is cropped using bounding box $\mathcal{B}_{\text{sam}}$, resized to 224×224 pixels, and subsequently, processed as input to a ResNet34 architecture.

Treating image filtering as a binary classification problem, we replace the final layer of ResNet with a fully connected layer for binary classification. The training data comprises two categories: high-quality data (cropping original images using ground-truth annotations) and low-quality data (applying random shifts and resizing to ground-truth annotations, and cropping images) with quality label of $\{0 : \text{low}, 1 : \text{high}\}$.

The filtering decision uses classification result and confidence thresholding, the retention condition combines classification

quality and confidence

$$\mathbb{I}_{\text{retain}}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) = \begin{cases} 1 & \text{if } \mathcal{F}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) = 1 \text{ and } s_i \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (37)$$

where s_i is the probability (confidence) that the classifier outputs 1, and τ is its threshold for decision.

Then, filter generated data samples classified as high-quality with confidence scores exceeding a threshold, retaining them in the augmented dataset; discard data below the threshold. And, the augmented dataset is updated by

$$D_{\text{aug}}^{(t+1)} = D_{\text{aug}}^{(t)} \cup \{(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) \mid \mathbb{I}_{\text{retain}}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) = 1\}. \quad (38)$$

Discarded samples form the set

$$D_{\text{discard}} = \{(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) \mid \mathbb{I}_{\text{retain}}(\mathcal{I}_{\text{gen}}, \mathcal{B}_{\text{sam}}) = 0\}. \quad (39)$$

IV. EXPERIMENTS

A. Data Description

1) *SARDet-100K*: The SARDet-100K dataset [38] represents the first COCO-scale benchmark for large-scale SAR object detection, comprising 116 598 images and 245 653 annotated instances across six object categories: Ship, Aircraft, Bridge, Tank, Car, and Harbor. Data are collected from ten heterogeneous SAR datasets acquired by multiple satellites and airborne platforms including GF-3 (C-band), Sentinel-1 (C-band), TerraSAR-X (X-band), RadarSat-2 (C/X-band), and airborne Ka/Ku/X-band sensors. Spatial resolution spans 0.1–25 m with multipolarization modes (HH, HV, VH, VV). We used all ship images (60 792 in total) for the pretraining of the GLIGEN model.

2) *SSDD*: The SAR images included in the SSDD dataset [39] were acquired by RadarSat-2, TerraSAR-X, and Sentinel-1 sensors, with spatial resolutions spanning 1–15 m. This dataset encompasses four polarization modes: HH, VV, VH, and HV, and provides two forms of instance-level annotations, specifically masks and bounding boxes. The training and testing subsets comprise 928 and 232 patches, respectively. Observational analysis reveals that the background environments in SSDD can be categorized into two types: off-shore and in-shore. Within the training set, there are 746 - patches and 182 in-shore patches.

3) *HRSID*: The SAR images included in the HRSID dataset [40] were acquired by Sentinel-1 and TerraSAR-X sensors, featuring multiple polarization modes (HH, HV, and VV) and spatial resolutions spanning 1–5 m. This dataset provides two forms of instance-level annotations: masks and bounding boxes. The training and testing subsets comprise 3 642 and 1 962 patches, respectively. Observational analysis indicates that the background environments in HRSID can be categorized into three types: in-shore, near-island, and off-shore. The training set includes 404 in-shore patches, 201 near-island patches, and 3 037 off-shore patches.

B. Experiments Setting

The training sets of SSDD and HRSID were used for the transfer training of the GLIGEN model. In the detection dataset generation phase, the original bounding box annotations were also selected only from the training sets. To analyze the SAR ship detection performance, the detection model was trained on the training set (or the augmented training set containing generated samples) and evaluated on the test set accordingly.

To compare our method with Ship-Go [13], we adopted a similar experimental setup with it. And, we have also designed two scenarios that are augmentation with in-domain and cross-domain conditions. Suppose there are two datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, with different data distributions, the slight difference of our method lies in the following conditions.

- 1) *Augmentation with in-domain conditions*: The GLIGEN model is trained on $\mathcal{D}_1$ to generate images under condition of bounding boxes in $\mathcal{D}_1$. Subsequently, the generated images are merged into $\mathcal{D}_1$ to form the final augmented dataset, which is used for training detectors.
- 2) *Augmentation with cross-domain conditions*: The GLIGEN model undergoes training on $\mathcal{D}_1$, tasked with generating images under condition of bounding boxes in $\mathcal{D}_2$. These generated images are then merged into $\mathcal{D}_1$, thereby forming the final augmented dataset that serves for training detectors.

It is worth noting that due to the different conditions between our method and Ship-Go, our method is conditioned on a set of annotations that constitute the target layout, while Ship-Go is conditioned on visual instances. When Ship-Go performs augmentation with cross-domain conditions, the distribution of generated targets (such as frequency bands, wavelength, and polarization mode) differs from that of $\mathcal{D}_1$, introducing the diversity of cross-domain $\mathcal{D}_2$ targets. However, when our method conducts augmentation with cross-domain conditions, it focuses on changing the distribution of target layout. Since the generative model is trained on $\mathcal{D}_1$, the distribution of generated targets is similar to that of $\mathcal{D}_1$, but the diversity of layouts is increased (with the layouts derived from $\mathcal{D}_2$). More detailed analysis is illustrated in Section IV-F2.

In addition, five classical deep-learning-based object detectors are selected to quantify the quality and diversity of the generated samples.

- 1) *YOLOF*: A single-stage anchor-based detector that utilizes only single-scale feature maps for prediction. As a result, the model has a relatively small capacity and achieves faster inference speeds.
- 2) *RetinaNet*: A single-stage anchor-based detector. In contrast to YOLOF, it employs multiscale feature maps for prediction. Moreover, due to the network architecture, the classification and localization prediction tasks are decoupled.
- 3) *FCOS*: A single-stage anchor-free detector that improves upon the regression branch of RetinaNet. This model eliminates the need for presetting anchors (proposal regions) and is well-suited for detection scenarios where targets exhibit variable aspect ratios.

- 4) *Faster R-CNN*: A two-stage detector. Compared with single-stage detectors, it addresses the issue of imbalance between positive and negative samples. Its coarse-to-fine localization regression approach results in higher detection accuracy than single-stage detectors.
- 5) *Cascade R-CNN*: A multistage detector developed based on Faster R-CNN. This method raises the quality requirements for proposals during the label assignment stage, thereby enhancing the accuracy of detection results.

For the quantitative analysis of our method, the five aforementioned detectors are trained on the new dataset supplemented with generated samples. Detection performance metrics are then used to evaluate the quality of the generated samples. And we adopt average precision (AP) and average recall (AR) as metrics, which are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{AP} = \int_0^1 P(R) dR, \quad \text{AR} = \int_0^1 R(P) dP \quad (40)$$

where P and R correspond to precision and recall. TP, FP, and FN stand for true positives, false positives, and false negatives, respectively. Note that the mentioned AP specifically refers to AP50, i.e., the AP calculated with the IoU threshold set to 0.5.

C. Implementation Details

Our implementation leverages PyTorch 1.13.1 with CUDA 12.4, and the experiments were conducted on a server cluster running a 64-bit Linux operating system. The hardware configuration comprises a NVIDIA A800 GPU (80-GB memory) and an Intel(R) Xeon(R) Gold 6348 CPU at 2.60 GHz.

In the implementation of our method, for pretraining and fine-tune GLIGEN model, the AdamW optimizer [41] is employed with a base learning rate of 5×10^{-5} and 10 000 warmup iterations. Batch sizes are configured to 2 for autoencoder training and 16 for diffusion model optimization. Random horizontal/vertical flips and random crops are applied during data preprocessing for augmentation purposes.

For training ResNet34 for data filtering, a ResNet-34 model is initialized with weights pretrained on ImageNet [42]. All convolutional layers are frozen, except that the final fully connected layer is replaced with one outputting two channels for binary classification. And the final fully connected layer is optimized by AdamW optimizer [41] with a learning rate of 0.001. And the learning rate is decayed by 0.1 every five epochs during a total training process of ten epochs.

In the implementation of object detection experiments, the open-source framework MMDetection [43] served as the foundational codebase. Prior to model input, all images underwent resizing to a resolution of 512×512 pixels. For backbone network initialization, each detector employed pretrained weights derived from the ImageNet dataset [42], ensuring consistency in initial parameter states. To maintain experimental rigor across comparative trials, hyperparameters were standardized across all groups, adhering to the configuration protocols outlined in [44]. Training for each model spanned 12 epochs, with a

batch size of 4 utilized throughout the process. Although we adopted these parameter settings for training detectors publicly disclosed by Ship-Go, discrepancies in the AP and AR of our implemented baseline models may arise due to differences in versions of software packages such as PyTorch. Therefore, in our experimental analysis, we focus more on comparing the advantages and disadvantages of the performance improvement margins between our method and Ship-Go.

D. Ablation Study

Ablation studies on the SSDD and HRSID datasets (see Tables II and III) systematically validate the effectiveness of three core pipeline components (layout augmentation, annotation correction, and data filtering) across two models with distinct complexity: the lightweight YOLOF-ResNet-50 and the nonlightweight Cascade R-CNN-ResNet-101, evaluated in both off-shore and in-shore maritime scenes. The results consistently confirm that integrating all three components yields the optimal AP/AR values for both models—removing any single module leads to performance degradation, verifying the pipeline’s rational design and synergistic effect of components.

For the lightweight YOLOF-ResNet-50, the full pipeline delivers remarkable gains, especially in challenging in-shore scenes: SSDD in-shore AP surges from 32.8% (baseline) to 46.0%. HRSID in-shore AP rises from 16.5% to 25.3%, while off-shore AP also improves significantly, effectively compensating for the lightweight model’s accuracy deficit without sacrificing its speed advantages. The nonlightweight Cascade R-CNN-ResNet-101, with a higher baseline, still achieves consistent gains, confirming the components’ universal adaptability across model complexities.

Each component plays an irreplaceable, scene-specific role. When layout augmentation is removed (while retaining annotation correction and data filtering), a mild but noticeable performance drop occurs in both scenarios. Given the complex, irregular object arrangements and diverse coastal background interferences in near-shore areas, layout augmentation boosts dataset diversity for targets and backgrounds, thus improving the detector’s generalization ability. Dropping annotation correction (with layout augmentation and data filtering kept) causes a consistent performance decline across both off-shore and in-shore scenarios, emphasizing its essential role in ensuring high-quality, accurate training labels; even with some tolerance for minor annotation errors, overall precision and recall decrease, underscoring the need for precise labels in augmentation. Removing data filtering (while preserving the other two strategies) inflicts the most severe performance degradation in the in-shore scenario, as unfiltered in-shore data contains a large share of low-quality or nonship samples that disrupt model training, whereas the off-shore scenario sees a smaller performance drop, owing to simpler backgrounds, more easily generated ship samples, and the more prominent effect of annotation correction.

Notably, the components deliver far higher relative gains for the lightweight YOLOF-ResNet-50, narrowing the performance gap with the heavy Cascade R-CNN-ResNet-101. This makes the lightweight model a cost-effective choice for

TABLE II
ABLATION STUDY ON SSDD DATASET

Layout augmentation	Annotation correction	Data filtering	YOLOF-ResNet-50				Cascade R-CNN-ResNet-101			
			Off-shore		In-shore		Off-shore		In-shore	
			AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
✓	✓	✓	80.2	85.0	46.0	59.3	97.9	98.4	82.1	86.0
	✓	✓	78.5	83.2	43.1	56.7	97.9	97.8	79.8	83.7
✓		✓	79.1	84.0	44.2	57.5	97.8	98.1	80.5	84.5
✓	✓		77.3	82.5	41.8	55.2	97.9	97.8	78.3	82.3
✓			76.5	83.5	36.0	51.0	97.8	96.5	77.7	82.2
	✓		77.0	83.0	37.2	52.3	97.8	96.8	77.4	82.5
		✓	76.8	83.2	36.5	51.5	97.8	97.0	77.5	82.8
			75.5	82.9	32.8	49.6	97.8	95.7	77.3	82.1

✓ indicates that the corresponding component is used in the experiment.

TABLE III
ABLATION STUDY ON HRSID DATASET

Layout augmentation	Annotation correction	Data filtering	YOLOF-ResNet-50				Cascade R-CNN-ResNet-101			
			Off-shore		In-shore		Off-shore		In-shore	
			AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
✓	✓	✓	84.6	86.9	25.3	32.6	94.9	95.2	44.0	50.0
	✓	✓	80.1	83.3	22.1	29.8	94.5	95.1	43.8	49.7
✓		✓	82.3	85.1	23.5	30.9	94.8	95.2	43.9	49.8
✓	✓		78.6	81.9	20.8	28.4	94.7	95.2	43.7	49.7
✓			78.4	79.3	18.2	26.1	94.0	95.2	43.7	49.7
	✓		77.2	81.1	19.5	27.3	94.1	95.1	43.8	49.6
		✓	76.9	80.7	18.9	26.7	94.0	95.1	43.6	49.6
			76.8	80.2	16.5	24.3	93.9	95.1	43.5	49.5

✓ indicates that the corresponding component is used in the experiment.

resource-constrained maritime surveillance, balancing speed and accuracy, while in-shore scenes consistently benefit more from the pipeline than off-shore scenes, aligning with the components' design goal of solving complex maritime detection pain points.

E. Qualitative Analysis

In this section, we turn to qualitative assessment to unpack two key aspects: first, the role of ShipGen in enhancing data diversity through augmentation, and second, the practical efficacy of our proposed module. To substantiate these analyses, we present a series of visualizations showcasing the images generated by the respective methods, allowing for direct observation of their performance characteristics.

1) *Visual Analysis on Augmentation With In-Domain Conditions*: Figs. 3 and 4 illustrate the visual performance of augmentation with in-domain conditions via ShipGen, respectively, on the HRSID and SSDD datasets. In these figures, odd columns present images from the original dataset, while even columns display generated counterparts, with instance-level annotations marked by green boxes. To more clearly highlight the object and background diversity of generated images, ShipGen adopts the same object layout as the original dataset in image synthesis. Its primary function is to replace the objects in the original images with new in-domain objects at the same object positions, while simultaneously replacing the backgrounds of images in the original dataset with new in-domain backgrounds, thereby increasing the diversity of both objects and backgrounds.

As observed in Fig. 3, ShipGen introduces islands around ships sailing at sea. It can also transform a moving ship into one docked at the shore (such as the image in the third row, fourth column). For near-shore ships, for example, in the fourth row, second column, ShipGen preserves the mooring state of near-shore ships while appropriately adding surrounding background elements (including islands and ports), thereby providing the detection model with more diverse training samples in terms of background. In addition, ShipGen alters the ship's orientation and scattering characteristics while maintaining the object layout (e.g., the fourth column), thereby providing the detection model with more diverse training samples in terms of object.

Fig. 4 further validates ShipGen's creative generation capability in the SSDD experiment. The generated images inherit the object distribution of the original images but generate ship targets with different orientations and even different scattering characteristics at the same positions. For example, in the fourth row, the third and fourth columns show different manifestations of the cross side-lobe effect and different distributions of scattering centers. In addition, ShipGen generates contextually appropriate backgrounds around the targets. For instance, in the image of the fourth row and second column, a dock is generated between two ships berthed side by side.

2) *Visual Analysis on Augmentation With Cross-Domain Conditions*: Fig. 5 presents the visualization results of augmentation with cross-domain conditions on the HRSID dataset. In this figure, odd columns display images from the original SSDD dataset, while even columns show the generated images using HRSID-trained model. Consistently, sample generation

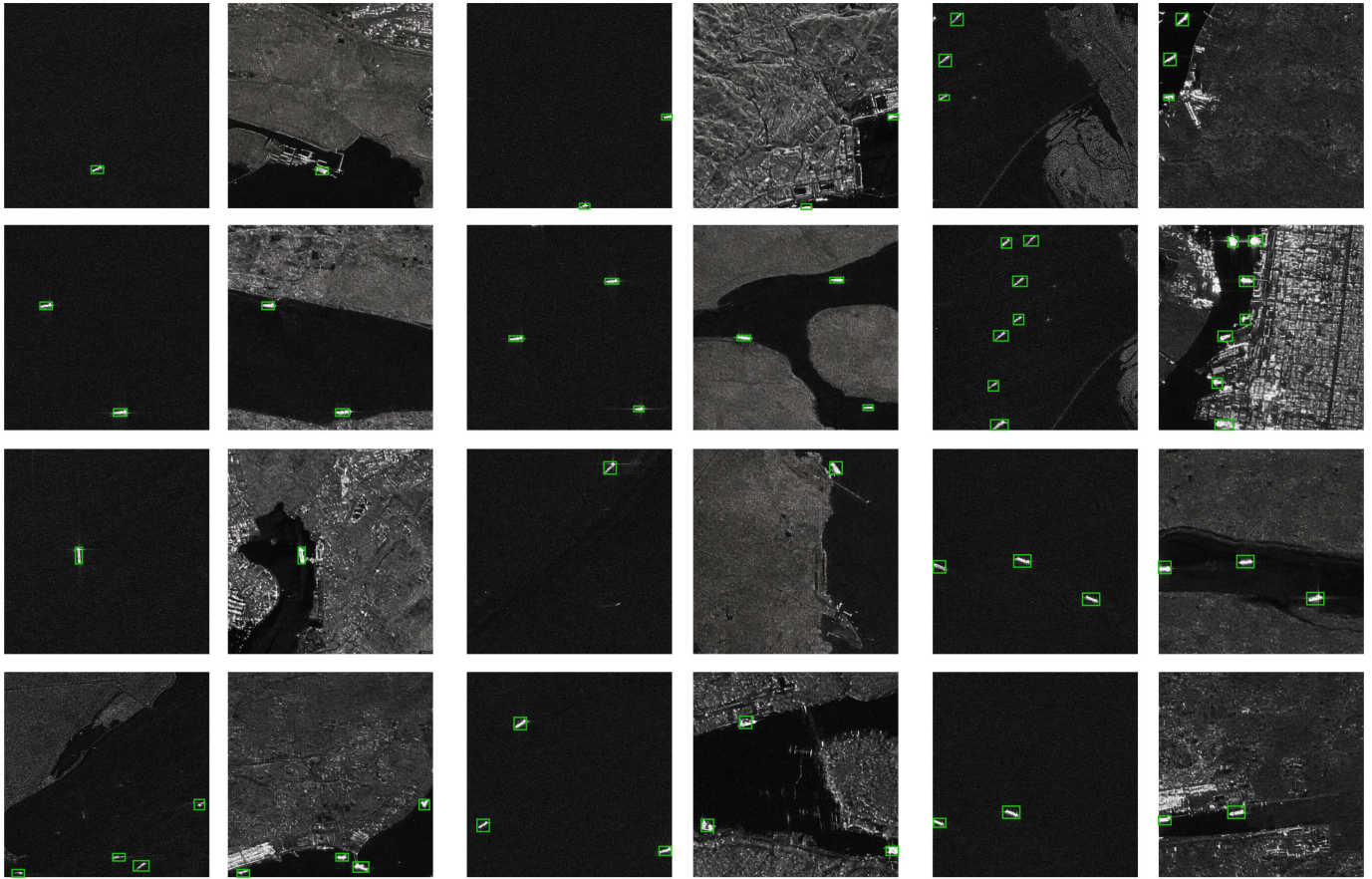


Fig. 3. Augmentation results with in-domain conditions on HRSID. Odd columns: Images from the original dataset; Even columns: Generated images. The green rectangles represent the annotations of object layout.

leverages the same object layouts as those in the cross-domain dataset images, increasing the diversity of ship sizes and positional distributions, thereby enhancing both foreground and background diversity for the original dataset.

As shown in Fig. 5, since ShipGen is trained on the HRSID dataset, it has learned the distribution patterns of objects and backgrounds based on the object layout control of this dataset. This section focuses on generating objects and backgrounds in the style of the HRSID dataset by controlling the object layout from the SSDD dataset. Even when the layout distribution and imaging size of cross-domain objects differ from those in the current domain, ShipGen can generate reasonable objects and backgrounds according to the cross-domain object layout. For example, in the third row and sixth column, based on the ship layout from the SSDD dataset, near-shore ships and surrounding coastal backgrounds are generated at the same positions. By introducing more instance layouts from out-of-domain datasets into the original dataset for training, the diversity of object layouts in the generated data is enhanced, which in turn helps improve the generalization performance of the detection model in SAR ship detection.

3) *Visual Analysis on Samples Diversity With Different Layouts*: Figs. 6 and 7 further analyze sample diversity with different layouts on the HRSID and SSDD datasets. In these figures, the odd columns present layout diagrams processed by layout

augmentation (see Section III-B1), while the even columns display the corresponding generated images. The layouts undergo operations such as translation and resizing, which enable the targets to have more diverse positions and imaging sizes, thereby introducing greater diversity to the targets in the generated images.

As shown in Fig. 6, under the constraints of different object layouts, the generated SAR images produce ship targets at corresponding positions while generating backgrounds around these targets. For instance, in the third column, the targets exhibit significant differences in positional distribution and size, and the generated images in the fourth column feature both near-shore and open-sea backgrounds, with diverse islands or coastlines in the backgrounds.

As shown in Fig. 7, under the constraints of varied object layouts, the generated SAR images accurately place ship targets at the specified positions and generate contextually coherent backgrounds around them. For example, in the fifth column, where the targets differ noticeably in their spatial arrangement and scale, the corresponding generated images in the sixth column display a range of background settings, including coastal regions with intricate port structures and off-shore areas with scattered small islands, all of which are consistent with the positional characteristics of the targets in the layout. This further demonstrates that the model can flexibly adjust background

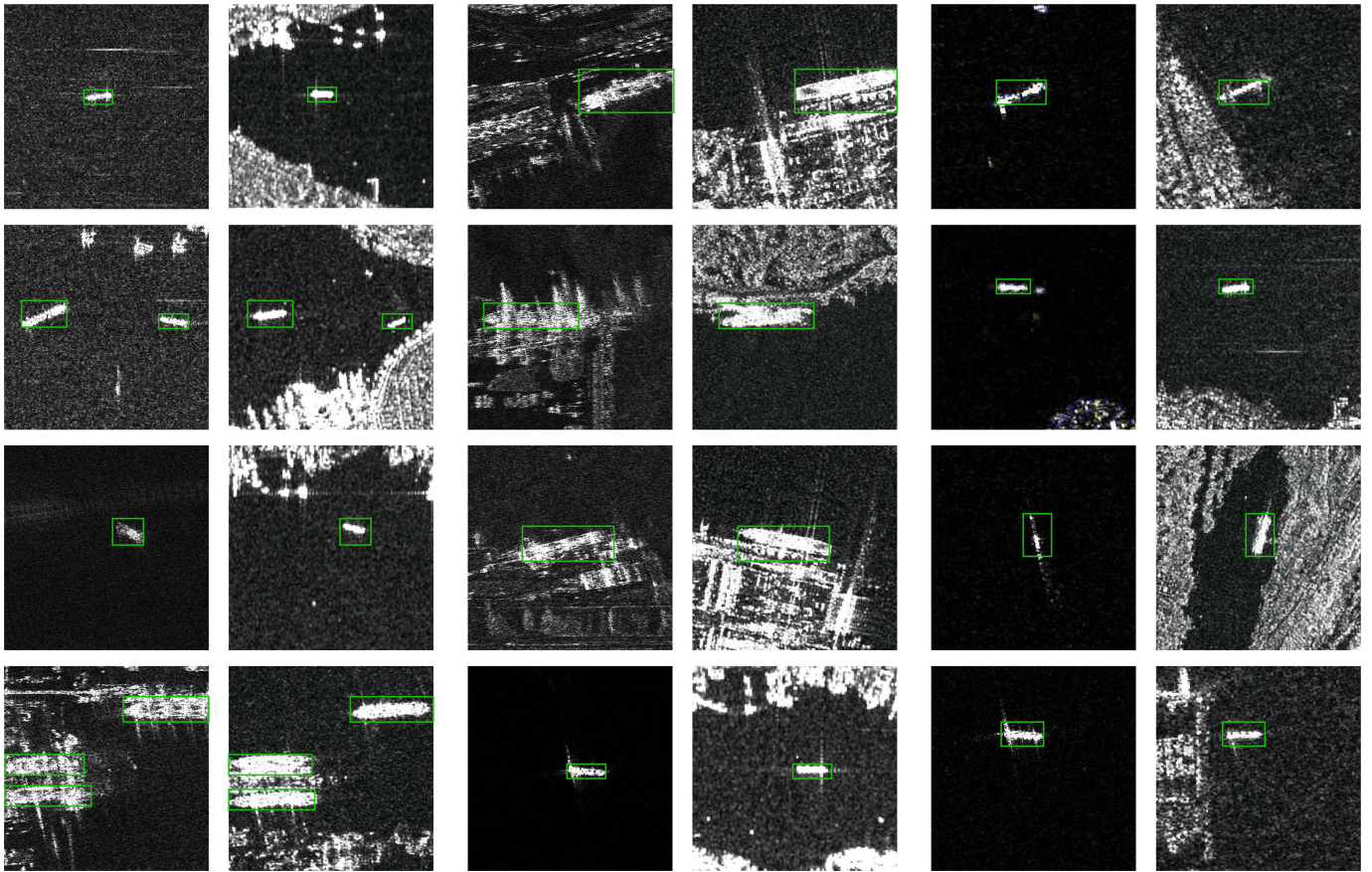


Fig. 4. Augmentation results with in-domain conditions on SSDD. Odd columns: Images from the original dataset; Even columns: Generated images. The green rectangles represent the annotations of object layout.

generation based on different object layout constraints, thereby enhancing the diversity of both targets and backgrounds in the generated samples.

4) *Visual Analysis on Samples Diversity With the Same Layout*: As shown in Figs. 8 and 9, experiments on sample diversity analysis based on the same layout were conducted on the HRSID and SSDD datasets. Images within the same column correspond to multiple SAR images generated under the constraint of the same object layout (first row).

As shown by comparing images in each column, under the constraints of the same object layout, the diversity of samples is manifested in two aspects: on one hand, it is the target-related diversity, including the diversity in the orientation of generated ships (see the second column in Fig. 8) and the diversity in the scattering characteristics of ships (refer to the first column in Fig. 8); on the other hand, it is the background-related diversity—the backgrounds around the ships can be either open-sea or near-shore (e.g., the last three columns in Fig. 8).

Similar phenomena are observed in the experiments on the HRSID dataset. For example, the targets' orientations and scattering characteristics vary significantly (see the last column in Fig. 9), and the positions of coasts and islands surrounding the ships are altered (see the seventh column in Fig. 9).

In summary, ShipGen inherently possesses randomness in image generation based on the LDM. This intrinsic randomness

ensures the diversity of both targets and backgrounds in the generated data. On one hand, it provides the ship detector with a sufficient number of false alarm training samples (such as ocean waves, islands, and partial features of other vessels), helping the detection model more accurately exclude nontarget information and reduce misjudgments. On the other hand, it offers the detector a sufficiently diverse set of positive samples (covering appearance variations under different ship types, angles, lighting conditions, and sea states), enabling the detection model to learn more comprehensive target features. As a result, the model maintains high recognition reliability for various types of ships during actual detection. Ultimately, through the optimization of these two aspects, the prediction confidence of the model is enhanced.

5) *Visual Analysis on Detection Performance*: In this section, we visualize the detection performance of detectors in complex scenes, with Faster R-CNN and FCOS employed as baseline models. As illustrated in Figs. 10 and 11, false alarms (false positives) and missed alarms (false negatives) are marked with red and green circles, respectively. It is apparent that the baseline models generate false alarms for ground targets along the shore and miss alarms for ships in near-shore areas.

However, after training on the augmented dataset, the enhanced background diversity equips the detector to mitigate the negative interferences introduced by complex scenes. Data

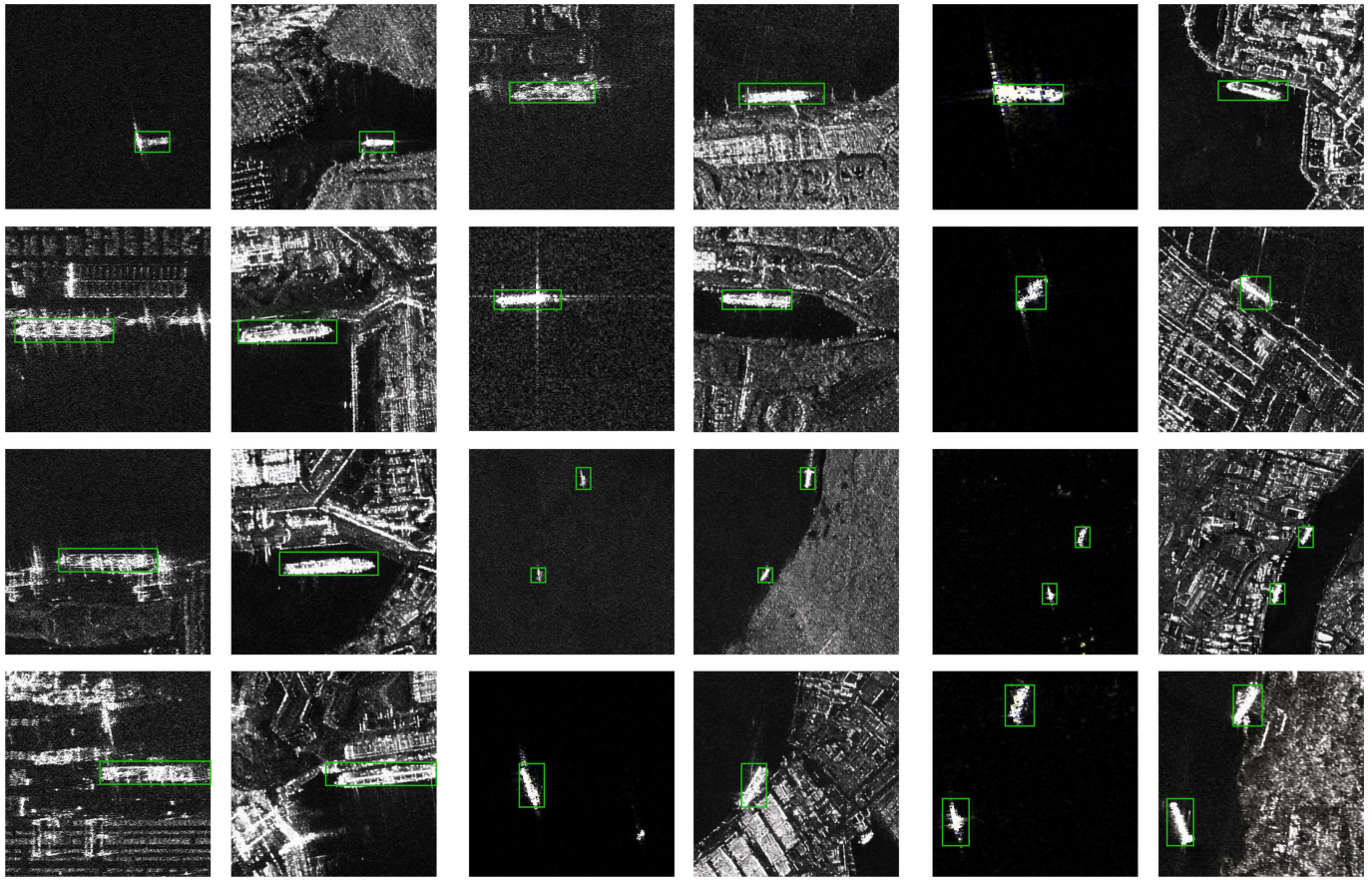


Fig. 5. Augmentation results with cross-domain conditions on HRSID. Odd columns: Images from the original SSDD dataset; Even columns: Generated images with annotations of SSDD dataset using Ship-Gen trained on the HRSID dataset. The green rectangles represent the instance-level annotations.

augmentation yields more robust representations of both backgrounds and targets, thereby enabling accurate recognition of both.

F. Quantitative Analysis

In this subsection, detection experiments are performed to evaluate the quality of generated images, covering augmentation scenarios with in-domain conditions and cross-domain conditions. For each detector type, two convolution-based backbone networks of varying scales—ResNet-50 and ResNet-101—are employed. This experimental design ensures comprehensive data quality assessment by accounting for both the diversity of detection network architectures and variations in model capacity.

1) *Main Results on Augmentation With In-Domain Conditions:* Following the augmentation protocol outlined in Ship-Go, we added 800 generated samples (200 off-shore and 600 in-shore images) to the original SSDD and HRSID datasets. Table IV evaluates the in-domain augmentation performance of our method against Ship-Go on the SSDD dataset, with bolded results indicating where our method outperforms Ship-Go. Notably, our method achieves higher AP than Ship-Go across all cases and superior AR in most scenarios. Furthermore, the AP improvement of our detection models relative to their baseline counterparts is significantly greater than that of Ship-Go. This

advantage is particularly pronounced for lightweight models such as YOLOF: in the in-shore scene, our method boosts YOLOF-ResNet-50 by 10.7% in AP, surpassing Ship-Go's 5.5% gain. For larger models such as RetinaNet-ResNet-101, our method still delivers a substantial 29.8% AP improvement in the in-shore scenario, outperforming Ship-Go's 13.5%. Even as model size increases, the performance advantage of our method remains evident. For example, Cascade R-CNN-ResNet-101 achieves 82.1% AP with our augmentation (up from its baseline of 76.7%), representing a 5.4% improvement that exceeds Ship-Go's 3.9% gain in the in-shore scenario.

The HRSID dataset serves as a supplementary validation to further assess the superiority of our method in in-domain augmentation, as shown in Table V, particularly within complex background scenarios. For off-shore scenes—where ship backgrounds are relatively simple—even as model fitting capabilities improve, the performance gains from augmentation remain modest across all detectors. This is likely due to the lower complexity of off-shore environments, which reduces the need for additional data to enhance detection accuracy. In contrast, within in-shore scenes—characterized by intricate and varied backgrounds—our augmentation method delivers substantial performance improvements across all detectors. By enriching background diversity through high-quality generated samples, our approach effectively addresses the challenges posed by

TABLE IV
EVALUATION OF AUGMENTATION WITH IN-DOMAIN CONDITIONS ON SSDD

Method	Backbone	GFLOPs	Off-shore				In-shore			
			Ours		ShipGo		Ours		ShipGo	
			AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
YOLOF	ResNet-50	25.13	71.4	79.9	71.4	80.2	25.3	47.1	28.2	44.8
YOLOF ^a	ResNet-50	25.13	80.2	85.0	78.7	85.3	46.0	59.3	33.7	50.0
YOLOF	ResNet-101	44.6	79.8	86.1	73.8	82.9	36.5	56.4	32.5	50.6
YOLOF ^a	ResNet-101	44.6	83.1	88.0	79.6	84.8	49.9	62.8	44.1	54.1
RetinaNet	ResNet-50	52.28	86.0	97.9	86.4	95.7	28.5	72.1	51.5	73.8
RetinaNet ^a	ResNet-50	52.28	95.0	97.6	89.0	98.4	63.2	80.8	56.5	81.4
RetinaNet	ResNet-101	71.76	87.4	96.3	86.7	95.5	38.1	70.9	51.4	75.0
RetinaNet ^a	ResNet-101	71.76	96.1	99.2	90.3	98.4	67.9	85.5	64.9	84.9
RetinaNet	Swin Transformer-T	54.67	92.0	97.6	88.8	98.7	58.4	84.9	59.6	85.5
RetinaNet ^a	Swin Transformer-T	54.67	96.0	98.7	90.3	99.5	73.3	90.3	71.1	92.4
FCOS	ResNet-50	50.32	96.3	97.6	90.7	99.5	67.1	79.1	67.4	84.9
FCOS ^a	ResNet-50	50.32	96.7	99.2	90.7	99.7	75.0	85.5	72.9	86.0
FCOS	ResNet-101	69.8	97.8	96.8	90.5	99.2	67.3	81.4	71.2	83.7
FCOS ^a	ResNet-101	69.8	97.8	99.2	90.6	99.7	87.6	85.5	76.6	87.2
Faster R-CNN	ResNet-50	63.25	94.8	94.9	90.8	94.9	74.9	76.2	74.7	80.8
Faster R-CNN ^a	ResNet-50	63.25	95.8	94.7	90.8	94.1	78.4	83.7	77.4	82.0
Faster R-CNN	ResNet-101	82.72	95.8	93.9	90.8	94.9	75.5	78.5	76.1	81.9
Faster R-CNN ^a	ResNet-101	82.72	96.9	98.4	90.8	96.8	79.4	84.3	77.9	86.0
Cascade R-CNN	ResNet-50	91.05	97.9	98.4	90.8	97.1	77.9	82.0	78.1	83.7
Cascade R-CNN ^a	ResNet-50	91.05	97.8	98.7	90.8	98.1	80.1	85.5	79.7	87.2
Cascade R-CNN	ResNet-101	110.52	97.9	98.4	90.7	97.3	76.7	82.0	77.3	83.1
Cascade R-CNN ^a	ResNet-101	110.52	97.9	98.4	90.8	98.3	82.1	86.0	80.2	83.7

^a indicates that the detector is trained on augmented dataset that includes generated images by different methods.
We bold the results where our augmentation method outperforms ShipGo.

TABLE V
EVALUATION OF AUGMENTATION WITH IN-DOMAIN CONDITIONS ON HRSID

Method	Backbone	GFLOPs	Off-shore				In-shore			
			Ours		ShipGo		Ours		ShipGo	
			AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
YOLOF	ResNet-50	25.13	76.6	80.0	70.6	78.2	11.7	24.0	10.0	16.8
YOLOF ^a	ResNet-50	25.13	84.6	86.9	80.1	83.3	25.3	32.6	22.2	29.0
YOLOF	ResNet-101	44.6	80.4	82.9	79.3	84.7	14.8	25.5	18.9	26.7
YOLOF ^a	ResNet-101	44.6	84.5	86.7	80.4	85.5	25.6	32.9	23.9	31.3
RetinaNet	ResNet-50	52.28	95.7	97.4	90.0	96.4	30.0	54.7	28.5	48.9
RetinaNet ^a	ResNet-50	52.28	96.2	98.0	90.2	96.6	36.3	60.3	32.5	54.4
RetinaNet	ResNet-101	71.76	95.8	97.7	90.1	96.3	30.7	54.2	29.3	49.3
RetinaNet ^a	ResNet-101	71.76	96.2	98.0	90.1	96.9	34.7	58.4	32.9	54.2
RetinaNet	Swin Transformer-T	54.67	95.6	97.4	89.7	96.9	29.6	54.2	25.8	50.5
RetinaNet ^a	Swin Transformer-T	54.67	96.0	98.0	89.9	97.2	30.5	59.9	28.6	54.2
FCOS	ResNet-50	50.32	96.5	97.2	90.5	96.6	51.1	65.4	45.3	62.1
FCOS ^a	ResNet-50	50.32	96.6	97.4	90.5	96.8	52.2	67.0	51.8	68.9
FCOS	ResNet-101	69.8	95.4	96.9	90.5	96.8	51.9	64.2	47.3	63.7
FCOS ^a	ResNet-101	69.8	95.7	97.2	90.6	96.9	52.2	65.2	52.4	69.6
Faster R-CNN	ResNet-50	63.25	91.9	93.1	90.5	91.4	39.0	45.4	40.5	44.3
Faster R-CNN ^a	ResNet-50	63.25	94.7	92.9	90.6	91.7	43.4	47.0	41.9	46.4
Faster R-CNN	ResNet-101	82.72	92.0	92.6	90.5	91.5	39.0	45.5	40.6	43.7
Faster R-CNN ^a	ResNet-101	82.72	94.7	92.7	90.5	92.4	43.8	45.7	42.0	45.4
Cascade R-CNN	ResNet-50	91.05	93.9	95.5	90.2	93.7	43.4	51.1	40.9	48.2
Cascade R-CNN ^a	ResNet-50	91.05	94.9	95.6	90.7	94.9	43.8	51.4	42.8	49.5
Cascade R-CNN	ResNet-101	110.52	93.9	95.0	90.7	94.8	43.4	49.4	40.7	47.5
Cascade R-CNN ^a	ResNet-101	110.52	94.9	95.2	90.7	95.1	44.0	50.0	42.3	48.9

^a indicates that the detector is trained on augmented dataset that includes generated images by different methods.
We bold the results where our augmentation method outperforms ShipGo.

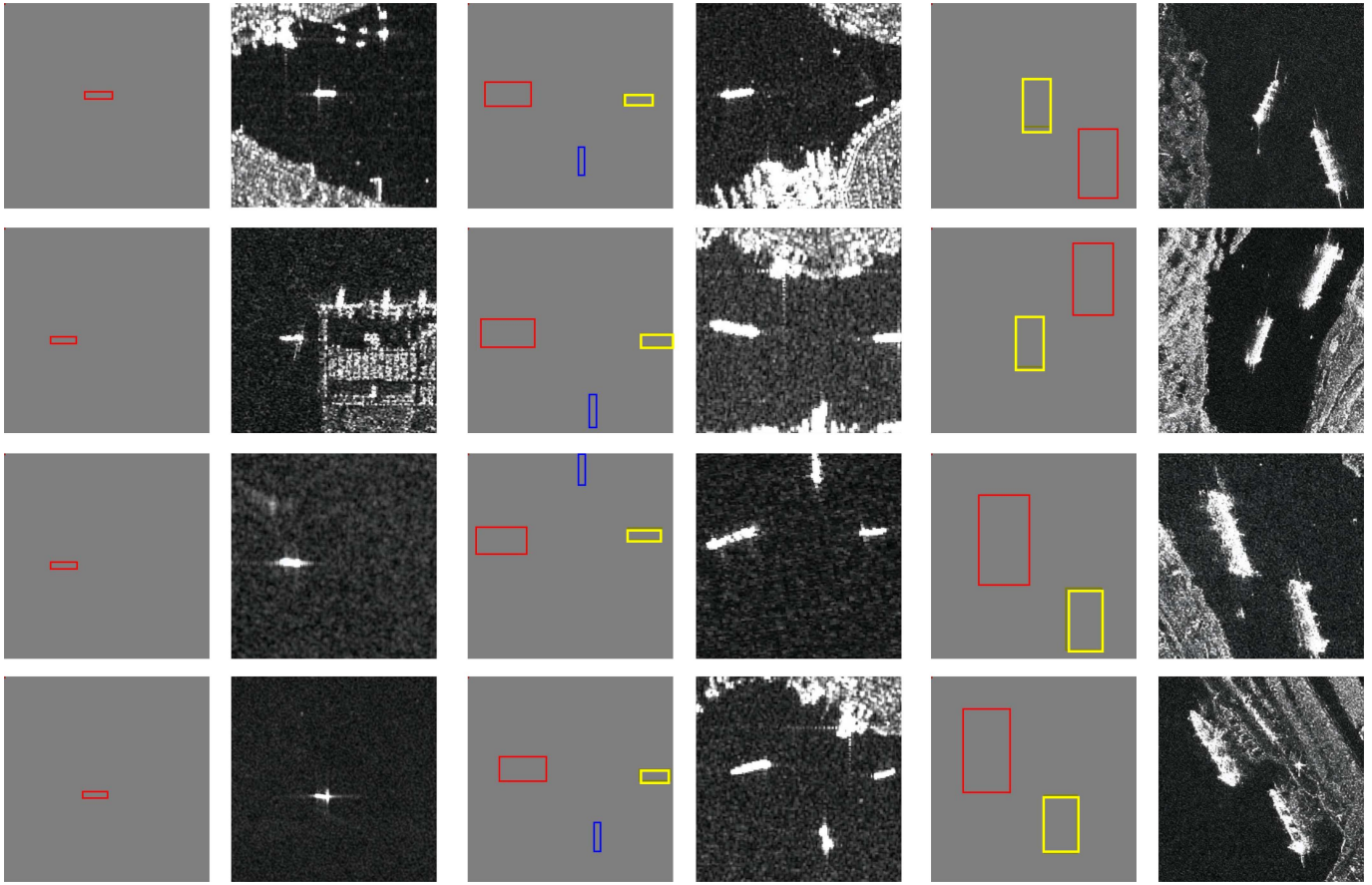


Fig. 6. Visual analysis of sample diversity caused by layout diversity on SSDD. Odd columns: The annotations of object layout; Even columns: Generated images.

complex in-shore environments. As evidenced in Table V, every detector exhibits noticeable enhancements in both AP and AR after training with our augmented dataset. In in-shore scene, our method outperforms Ship-Go across most detectors, with larger gaps than in off-shore. YOLOF benefits most: YOLOF-ResNet-50 with our method reaches 25.3% AP and 13.6% improvement, exceeding Ship-Go's 22.2% AP and 12.2% improvement. As the model size increases, since the baseline we implemented operates at a higher level compared to Ship-Go's baseline, our method has less room for improvement and faces greater difficulty in achieving gains. This results in a small number of cases where the improvement margin is smaller than that of Ship-Go. However, the AP of our augmented models remains consistently higher than that of Ship-Go. In other words, even when the baseline detection model has already reached a relatively high level on HRSID dataset, our method can still further improve the AP to some extent.

Fig. 12(a) and (b) presents an analysis of how object and background diversity from in-domain augmentation impacts in-shore ship detection performance. Augmented samples with rich, nonredundant background variations and diverse object layouts significantly enhance both precision and recall across detectors of varying sizes. Specifically, on the SSDD dataset, detectors achieve an average 14.18% improvement in AP and 7.32% in AR. For the HRSID dataset, the average gains in

AP and AR reach 3.11% and 2.30%, respectively. These results underscore two key insights: first, our method generates high-quality in-shore ship images with contextually relevant backgrounds and objects within diverse layouts; second, it is validated that augmentation with in-domain conditions, by enriching background diversity, effectively boosts in-shore detection performance. The diverse yet task-aligned backgrounds in our augmented samples provide meaningful contextual cues, enabling detectors to better distinguish ships from complex in-shore environments and generalize more robustly across scenarios.

2) *Main Results on Augmentation With Cross-Domain Conditions:* Following the augmentation protocol outlined in Ship-Go, a total of 400 generated images are added to the SSDD dataset. Table VI assesses cross-domain augmentation performance of our method versus Ship-Go on the SSDD dataset, with bolded results indicating our outperformance. Notably, our method achieves higher AP across all cases and superior AR in most scenarios, with more substantial AP improvements over baselines than Ship-Go.

This advantage is striking for lightweight models: YOLOF-ResNet-50 sees 14.3% in-shore AP gain (from 25.3% to 39.6%) with our method, dwarfing Ship-Go's 1.3% (from 28.2% to 29.5%). Even with ResNet-101, our 10.2% improvement (from 36.5% to 46.7%) far exceeds Ship-Go's 3.9% (from 32.5% to

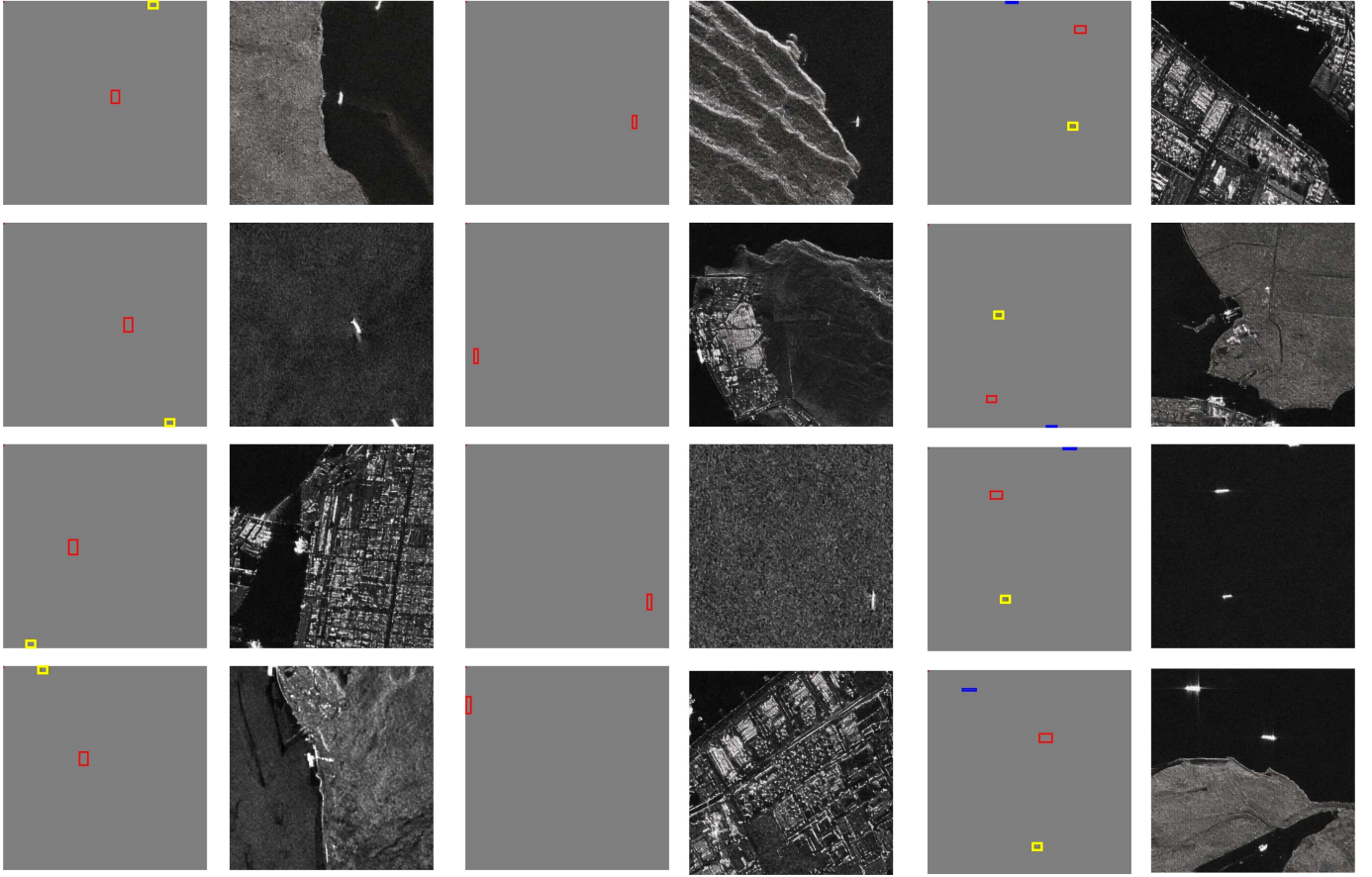


Fig. 7. Visual analysis of sample diversity caused by layout diversity on HRSID. Odd columns: The annotations of object layout; Even columns: Generated images.

36.4%), highlighting our data's ability to boost models with limited capacity.

For larger models, our superiority persists. RetinaNet-ResNet-101 gains 28.4% in-shore AP (from 38.1% to 66.5%) with us, versus Ship-Go's 5.8% (from 51.4% to 57.2%). Off-shore, RetinaNet-ResNet-50 improves 9.9% (from 86.0% to 95.9%) with our method, outperforming Ship-Go's 3.4% (from 86.4% to 89.8%), showing our data enhances precision even in simpler environments.

Even for high-capacity models, our edge holds. Cascade R-CNN-ResNet-101, with a strong 76.7% baseline, gains 4.4% AP (to 81.1%) with us, versus Ship-Go's 1.0% (from 77.3% to 78.3%). This confirms our method provides high-quality data that benefit all model scales, particularly in complex in-shore scenarios, outperforming Ship-Go consistently. Notably, in ShipGo's experimental results, the AP of cross-domain augmentation is higher than that of in-domain augmentation. In contrast, the AP of cross-domain augmentation in our method is slightly lower than that of in-domain augmentation. This is because ShipGo's cross-domain augmentation directly introduces targets from HRSID, which is a direct strong-constraint condition, while our method can only introduce the layout of HRSID targets, which is an indirect weak-constraint condition. In addition, the number of cross-domain augmented samples

(400) is less than that of in-domain augmented samples (800), which also contributes to the slightly lower AP of cross-domain augmentation in our method compared to in-domain augmentation. Even so, the AP of cross-domain augmentation in our method is still higher than that in ShipGo's cross-domain augmentation. This indicates that the data generated by our method is of higher quality and its distribution is closer to the original dataset.

3) *Quantitative Analysis on Sample Numbers*: We evaluated the impact of varying quantities of augmented data on AP for ship detection in in-shore (see Table VII) and off-shore (see Table VIII) scenarios using the SSDD dataset. The highest AP values for each detector are highlighted in bold. As shown in Table VII, within the in-shore scenario, introducing generated off-shore images to the original SSDD dataset yields a limited increase in AP (see the first three rows). This is because while off-shore images fail to enhance the diversity of in-shore backgrounds, they do contribute moderately to improving target diversity. In contrast, AP exhibits a general upward trend as more generated in-shore images are incorporated into the original dataset. This consistent improvement can be attributed to the enhanced diversity of both targets and backgrounds in the in-shore images generated by our method, which better aligns with the complexity of in-shore detection tasks.

TABLE VI
EVALUATION OF AUGMENTATION WITH CROSS-DOMAIN CONDITIONS ON SSDD

Method	Backbone	GFLOPs	Off-shore				In-shore			
			Ours		ShipGo		Ours		ShipGo	
			AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
YOLOF	ResNet-50	25.13	71.4	79.9	71.4	80.2	25.3	47.1	28.2	44.8
YOLOF ^a	ResNet-50	25.13	81.0	89.5	79.3	88.0	39.6	55.7	29.5	51.2
YOLOF	ResNet-101	44.6	79.8	86.1	73.8	82.9	36.5	56.4	32.5	50.6
YOLOF ^a	ResNet-101	44.6	82.9	88.0	77.8	85.6	46.7	60.0	36.4	52.3
RetinaNet	ResNet-50	52.28	86.0	97.9	86.4	95.7	28.5	72.1	51.5	73.8
RetinaNet ^a	ResNet-50	52.28	95.9	98.1	89.8	97.9	62.5	81.9	60.1	80.8
RetinaNet	ResNet-101	71.76	87.4	96.3	86.7	95.5	38.1	70.9	51.4	75.0
RetinaNet ^a	ResNet-101	71.76	96.7	99.5	90.1	98.7	66.5	84.9	57.2	79.1
RetinaNet	Swin Transformer-T	54.67	93.4	97.8	88.8	98.7	58.0	84.9	59.6	85.5
RetinaNet ^a	Swin Transformer-T	54.67	95.6	99.7	91.7	100	71.2	89.5	69.2	90.7
FCOS	ResNet-50	50.32	96.3	97.6	90.7	99.5	67.1	79.1	67.4	84.9
FCOS ^a	ResNet-50	50.32	96.8	99.7	94.9	100	75.5	87.2	69.0	86.6
FCOS	ResNet-101	69.8	97.8	96.8	90.5	99.2	67.3	81.4	71.2	83.7
FCOS ^a	ResNet-101	69.8	97.7	99.7	96.6	100	75.0	87.0	72.8	89.5
Faster R-CNN	ResNet-50	63.25	94.8	94.9	90.8	94.9	74.9	76.2	74.7	80.8
Faster R-CNN ^a	ResNet-50	63.25	96.8	99.7	90.8	96.3	76.6	83.0	75.2	82.6
Faster R-CNN	ResNet-101	82.72	95.8	93.9	90.8	94.9	75.5	78.5	76.1	81.9
Faster R-CNN ^a	ResNet-101	82.72	96.7	98.4	90.8	95.7	79.6	84.9	77.5	84.3
Cascade R-CNN	ResNet-50	91.05	97.9	98.4	90.8	97.1	77.9	82.0	87.1	83.7
Cascade R-CNN ^a	ResNet-50	91.05	96.8	98.7	90.9	98.7	79.5	87.2	78.8	84.9
Cascade R-CNN	ResNet-101	110.52	97.9	98.4	90.7	97.3	76.7	82.0	77.3	83.1
Cascade R-CNN ^a	ResNet-101	110.52	96.9	98.7	90.8	98.4	81.1	85.5	78.3	84.3

^a indicates that the detector is trained on augmented dataset that includes generated images by different methods.

We bold the results where our augmentation method outperforms ShipGo.

TABLE VII
AP(%) OF DIFFERENT DETECTORS AS DIFFERENT QUANTITIES OF SAMPLES WERE ADDED INTO SSDD WHEN TESTING ON *IN-SHORE* SAMPLES

$Num_{inshore}$	$Num_{offshore}$	YOLOF		RetinaNet		FCOS		Faster R-CNN		Cascade R-CNN	
		Res50	Res101	Res50	Res101	Res50	Res101	Res50	Res101	Res50	Res101
0	0	25.3	36.5	28.5	38.1	67.1	67.3	74.9	75.5	77.9	76.7
0	100	29.4	44.0	53.3	52.7	67.4	62.1	73.6	75.9	77.1	75.7
0	200	30.5	38.1	49.1	55.1	64.7	69.1	75.4	76.1	76.6	74.9
100	0	34.6	41.0	58.4	53.5	70.4	70.1	77.2	76.0	78.4	77.3
200	0	35.0	41.6	58.2	58.9	72.6	71.5	74.7	74.0	79.4	78.0
300	0	38.8	44.4	55.4	58.7	70.1	62.5	77.6	76.9	79.0	80.2
400	0	39.9	43.2	60.4	62.5	72.1	71.6	76.9	78.0	78.5	78.9
500	0	41.0	50.2	35.3	59.5	69.5	73.4	76.3	76.6	78.8	76.3
600	0	43.6	46.2	66.6	62.3	74.6	75.2	76.0	78.4	79.5	80.2

We bold the results where our augmentation method outperforms ShipGo.

TABLE VIII
AP(%) OF DIFFERENT DETECTORS AS DIFFERENT QUANTITIES OF SAMPLES WERE ADDED INTO SSDD WHEN TESTING ON *OFF-SHORE* SAMPLE

$Num_{inshore}$	$Num_{offshore}$	YOLOF		RetinaNet		FCOS		Faster R-CNN		Cascade R-CNN	
		Res50	Res101	Res50	Res101	Res50	Res101	Res50	Res101	Res50	Res101
0	0	71.4	79.8	86.0	87.4	96.3	97.8	94.8	95.8	97.9	97.9
0	100	74.5	81.7	93.8	93.3	96.8	96.9	97.6	95.9	97.9	97.9
0	200	73.9	78.3	93.8	94.5	95.2	96.9	96.8	95.0	97.9	97.8
100	0	73.3	76.3	94.7	93.0	97.0	96.3	96.7	95.9	97.8	97.9
200	0	77.0	79.6	96.1	95.3	97.6	96.9	95.8	95.9	97.8	97.9
300	0	73.8	78.2	93.5	94.6	96.2	94.8	96.6	95.9	96.9	97.9
400	0	80.5	79.7	94.2	92.7	96.7	95.9	95.8	95.9	96.8	97.9
500	0	78.2	81.4	71.4	95.1	96.8	97.8	96.7	94.9	96.9	97.9
600	0	75.1	80.5	95.8	94.6	96.8	96.6	94.8	95.9	96.8	97.9

We bold the results where our augmentation method outperforms ShipGo.

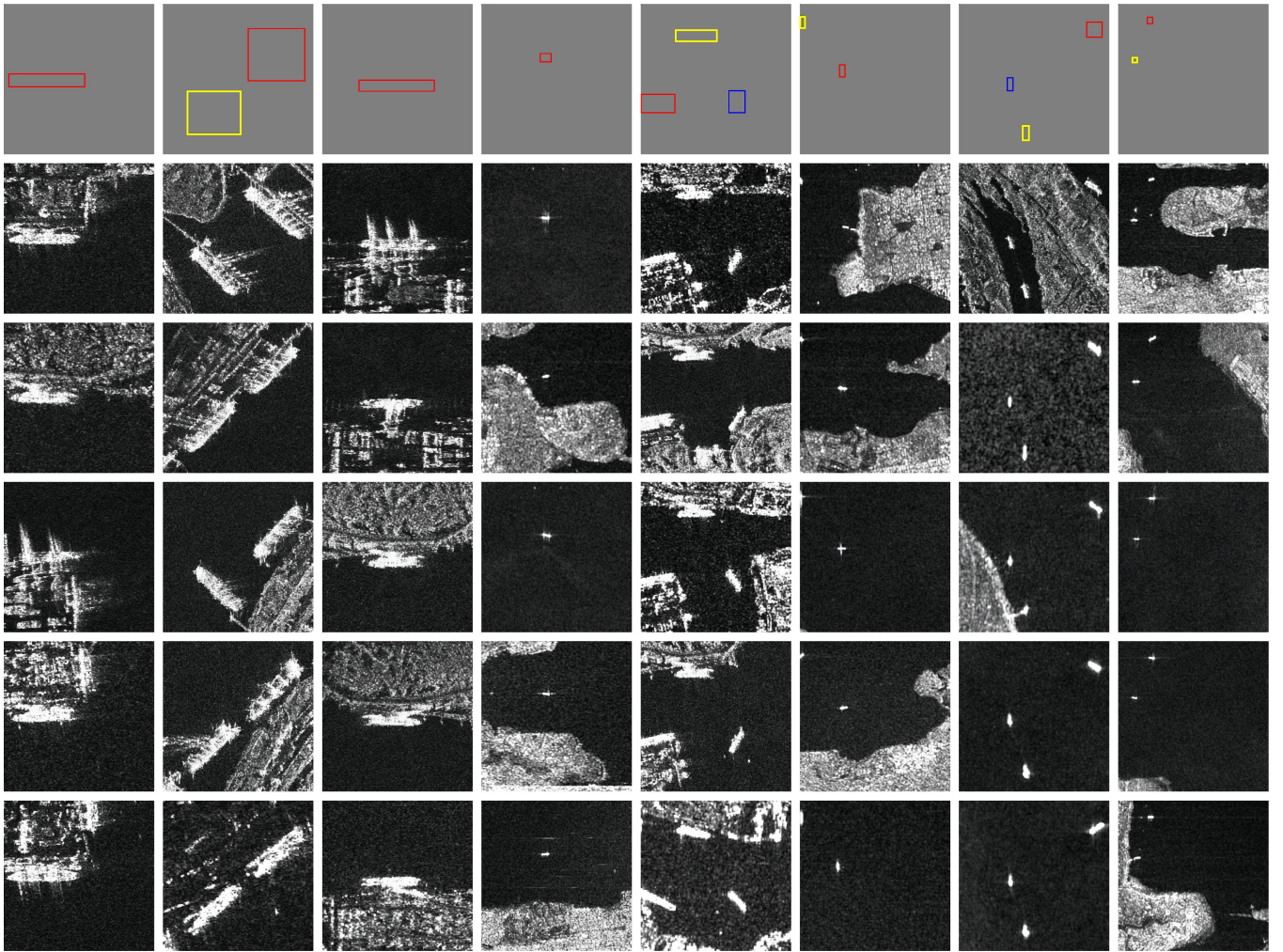


Fig. 8. Visual analysis of sample diversity given the same layout input on SSDD. The first row is the input layout, and the images in the same column represent the generated results given the same layout condition.

As observed in Table VIII, within the off-shore scenario, distinct trends emerge when augmenting the original SSDD dataset: For detectors with stronger fitting capabilities (e.g., Faster R-CNN and FCOS-ResNet-101), adding increasing numbers of generated off-shore images leads to a progressive boost in AP for off-shore detection. In contrast, detectors with fewer parameters (YOLOF, RetinaNet, and FCOS-ResNet-50) achieve larger AP gains in off-shore detection when augmented with generated in-shore images. This observation indicates that enhanced background diversity, even from in-shore contexts, can effectively improve the accuracy of off-shore ship detection, particularly for models with limited parameter capacities. Such findings underscore that background variation, when strategically incorporated, serves as a valuable supplement to target diversity, enabling smaller models to better generalize to off-shore environments despite their constrained architectural complexity.

G. Computational Overhead Analysis

We analyze the computational overhead of the proposed pipeline on an NVIDIA A800 GPU (see Table IX). Layout

TABLE IX
COMPUTATIONAL OVERHEAD ANALYSIS OF THE PROPOSED PIPELINE

Stage	GPU Memory Usage (GB)	Time Consumption (seconds)
Layout Augmentation	0	0.047
Image Generation	7.85	11.235
Annotation Correction	9.01	5.143
Data Filtering	1.16	0.054
Full Pipeline	18.02	16.479

Note: GPU memory usage is measured as the peak memory consumption on NVIDIA A800 GPU; time consumption represents the average processing time per single sample.

augmentation (0 GB, 0.047 s/sample) is a lightweight stage with negligible GPU memory and time costs, as it only involves CPU-based geometric calculations. In contrast, image generation (7.85 GB, 11.235 s/sample) and annotation correction (9.01 GB, 5.143 s/sample) dominate the resource consumption: the former relies on generative models with high memory and iterative sampling overhead, while the latter involves SAM's forward inference on high-resolution images. The full pipeline consumes 18.02-GB peak GPU memory (compatible with NVIDIA A800) and 16.479 s/sample,

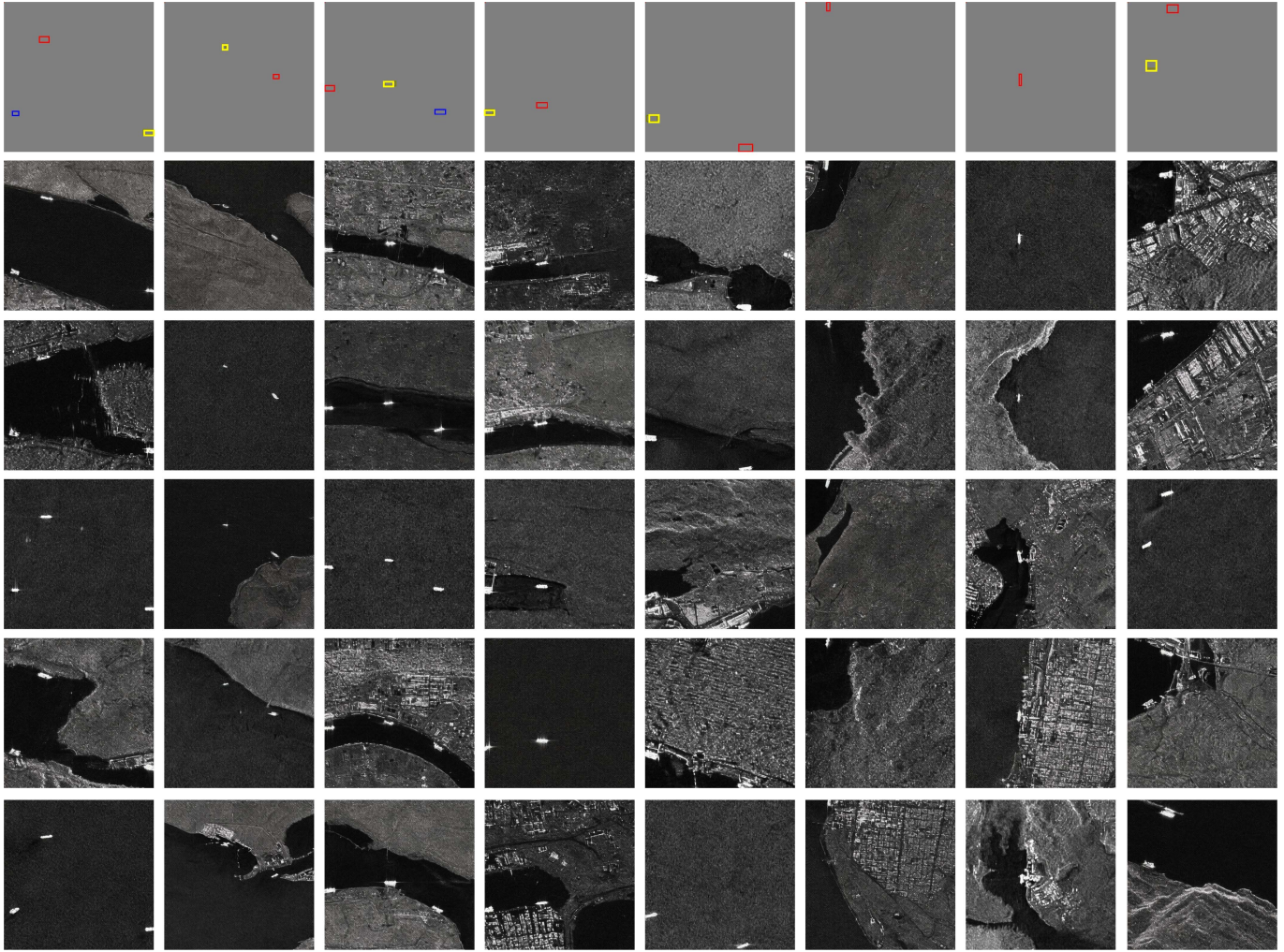


Fig. 9. Visual analysis of sample diversity given the same layout condition on HRSID. The first row is the input layout, and the images in the same column represent the generated results given the same layout condition.

with over 99% of time cost concentrated in deep-learning-based generation and correction steps—validating hardware feasibility and identifying lightweight generative models/SAM prompt optimization as key directions for efficiency improvement.

V. DISCUSSION

This study presents ShipGen, a layout-conditioned diffusion-based framework addressing limited target and background diversity in SAR ship detection data augmentation. Experimental results on SSDD and HRSID datasets validate its effectiveness, and key insights, comparisons, limitations, and future directions are discussed as follows. The core advantage of ShipGen lies in relaxed constraint generation and end-to-end high-quality pipeline design. By adopting bounding box layouts as sparse constraints, it breaks the limitations of dense constraints (e.g., segmentation masks and visual instances) in existing methods, enabling free generation of ship orientations, scattering characteristics, and backgrounds while maintaining

positional control—this is the fundamental reason for enhanced diversity. The multiphase transfer learning framework resolves the alignment issue between generated targets and layouts in small datasets, reducing alignment error from over 26 pixels to around 6 pixels and eliminating additional annotation costs. Notably, ShipGen delivers more significant relative gains for the lightweight YOLOF-ResNet-50 compared to the nonlightweight Cascade R-CNN-ResNet-101, effectively narrowing the performance gap between lightweight and heavyweight models and demonstrating practical value in resource-constrained scenarios.

Compared with related work, ShipGen outperforms in three aspects: it generates more diverse targets and backgrounds than Ship-Go [13] (which relies on instance inpainting), integrates SAM-based annotation correction and ResNet-34-based filtering to ensure data quality, and achieves consistent improvements across five detectors and multiple scenes, showing stronger generalization.

Despite promising results, ShipGen has limitations: the average generation time of 16.479 s/sample limits large-scale rapid expansion; cross-domain augmentation performance slightly

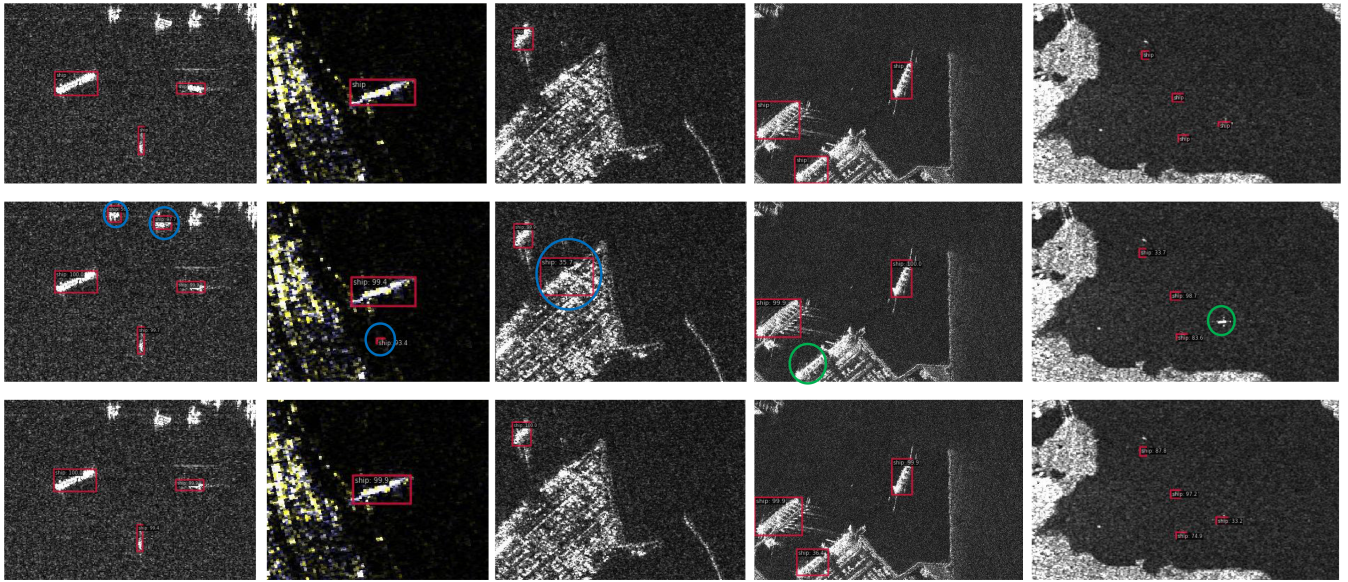


Fig. 10. Visualization of the detection results using Faster R-CNN. Top: Ground Truth; Middle: Trained on the augmented datasets; and Bottom: Baseline. The green circle represents the missing alarm. The blue circle represents the false alarm.

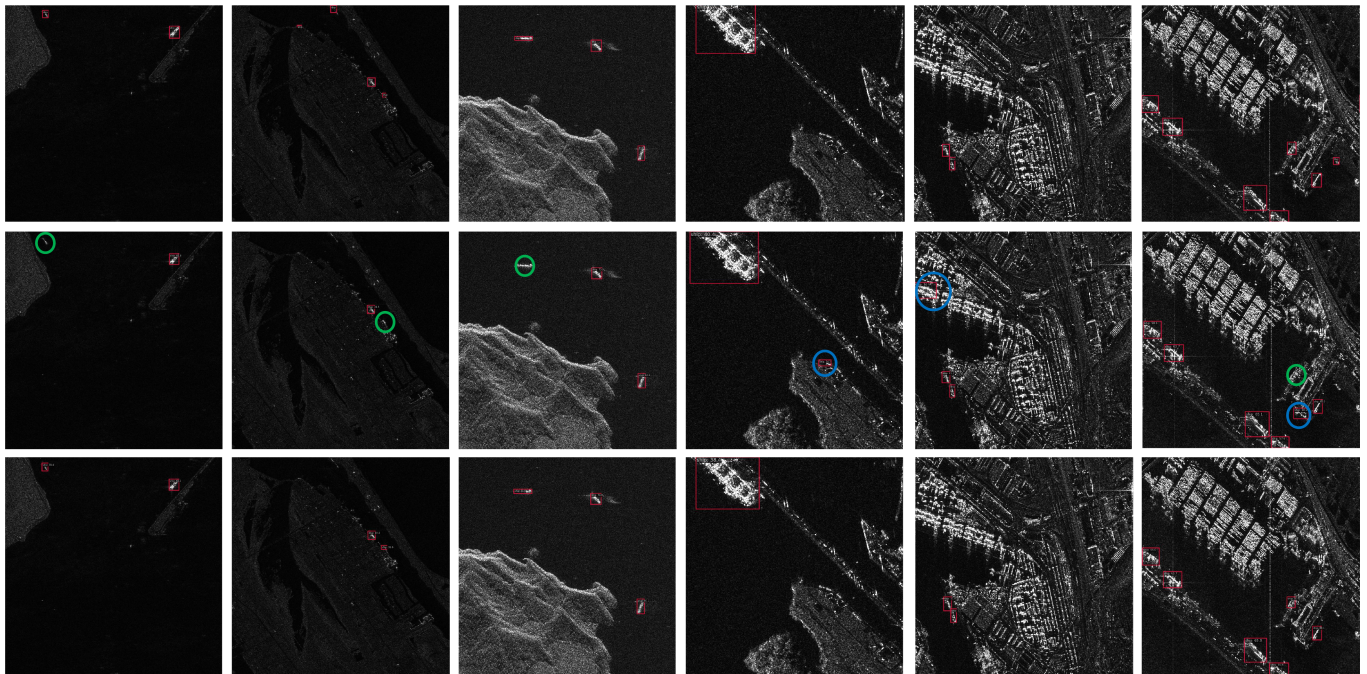


Fig. 11. Visualization of the detection results using FCOS. Top: Ground Truth; Middle: Trained on the augmented datasets; and Bottom: Baseline. The green circle represents the missing alarm. The blue circle represents the false alarm.

lags behind in-domain due to indirect layout constraints; layout augmentation only includes basic geometric transformations; and generated ship type diversity is constrained by the pretraining dataset.

Future work will focus on optimizing generation efficiency via lightweight diffusion models and batch processing; enhancing

cross-domain augmentation by integrating domain adaptation modules; extending layout transformations to include rotation and aspect ratio adjustment; broadening application to other SAR target detection tasks; and designing dynamic constraint adjustment mechanisms to balance diversity and alignment accuracy.

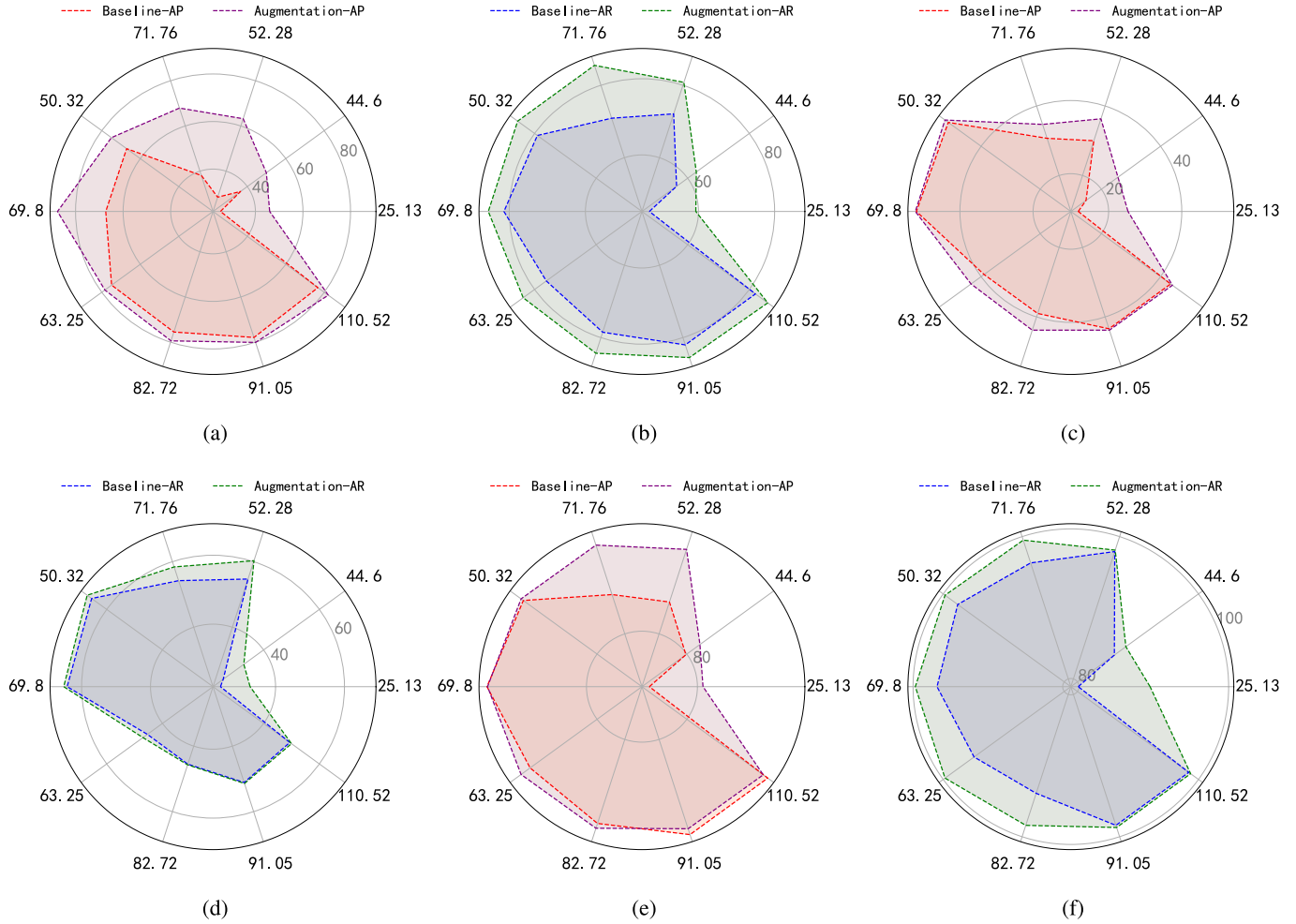


Fig. 12. Quantitative analysis of background and object diversity. (a) and (b) In-domain augmentation experiment on the SSDD dataset in the in-shore scenario, (c) and (d) in-domain augmentation experiment on the HRSID dataset in the in-shore scenario, and (e) and (f) out-of-domain augmentation experiment in the off-shore scenario. Gray numbers represent the metric axis (%) related to detection performance, and black numbers represent the model size axis (GFLOPS).

VI. CONCLUSION

This study addresses the key challenge in using data augmentation to improve SAR ship detection, which are limited target and background diversity in existing strict constraint-guided data augmentation methods. To solve these, we propose a multiphase transfer-learned framework with a layout-conditioned diffusion model and an automatic augmentation pipeline. The diffusion model uses sparse bounding box layouts as relaxed conditions, enabling free generation of ship attributes (e.g., orientation and scatter characteristics) and backgrounds to boost diversity; multiphase transfer learning ensures precise generation with limited data. The pipeline integrates layout augmentation, GLIGEN-based image generation, SAM-based annotation correction, and filtering to produce high-quality augmented data. Experiments on SSDD and HRSID datasets show the pipeline improves detector performance in in-shore and off-shore scenarios, while qualitative analyses confirm the generated images' high diversity and quality. This work provides a robust solution for enhancing SAR ship detection via data augmentation.

REFERENCES

- [1] M. Liu, B. Hou, B. Ren, L. Jiao, Z. Yang, and Z. Zhu, "Subpixel feature pyramid network for multiscale ship detection in synthetic aperture radar remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 15583–15595, 2024.
- [2] T. Zeng et al., "CFAR-DP-FW: A CFAR-guided dual-polarization fusion framework for large-scene SAR ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7242–7259, 2024.
- [3] T. Dong, T. Wang, X. Li, J. Hong, M. Jing, and T. Wei, "A large ship detection method based on component model in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 4108–4123, 2025.
- [4] Z. Wang, G. Hou, Z. Xin, G. Liao, P. Huang, and Y. Tai, "Detection of SAR image multiscale ship targets in complex inshore scenes based on improved YOLOv5," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5804–5823, 2024.
- [5] B. Hu and H. Miao, "An improved deep neural network for small-ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2596–2609, 2024.
- [6] Z. Suo, Y. Zhao, S. Chen, and Y. Hu, "BoxPaste: An effective data augmentation method for SAR ship detection," *Remote Sens.*, vol. 14, no. 22, 2022. [Online]. Available: <https://www.mdpi.com/2072-4292/14/22/5761>
- [7] Y. Tang et al., "Scene-aware data augmentation for ship detection in SAR images," *Int. J. Remote Sens.*, vol. 45, no. 10, pp. 3396–3411, 2024.

- [8] M. Zhu, G. Hu, H. Zhou, and S. Wang, "H2Det: A high-speed and high-accurate ship detector in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12455–12466, 2021.
- [9] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [10] W. G. C. Bandara, N. G. Nair, and V. Patel, "DDPM-CD: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2025, pp. 5250–5262.
- [11] L. Yang, J. Su, and X. Li, "Application of SAR ship data augmentation based on generative adversarial network in improved SSD," *Acta Armamentarii*, vol. 40, no. 12, 2019, Art. no. 2488.
- [12] K. Sun, Y. Liang, X. Ma, Y. Huai, and M. Xing, "DSDet: A lightweight densely connected sparse activated detector for ship target detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 14, 2021. [Online]. Available: <https://www.mdpi.com/2072-4292/13/14/2743>
- [13] X. Zhang et al., "Ship-Go: SAR ship images inpainting via instance-to-image generative diffusion models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 207, pp. 203–217, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271623003350>
- [14] K. Wang, Z. Pan, and Z. Wen, "SVDDD: SAR vehicle target detection dataset augmentation based on diffusion model," *Remote Sens.*, vol. 17, no. 2, 2025. [Online]. Available: <https://www.mdpi.com/2072-4292/17/2/286>
- [15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.
- [16] Z. Cui, M. Zhang, Z. Cao, and C. Cao, "Image data augmentation for SAR sensor via generative adversarial nets," *IEEE Access*, vol. 7, pp. 42255–42268, 2019.
- [17] J. Kong and F. Zhang, "SAR target recognition with generative adversarial network (GAN)-based data augmentation," in *Proc. 13th Int. Conf. Adv. Infocomm Technol.*, 2021, pp. 215–218.
- [18] C. Mao, L. Huang, Y. Xiao, F. He, and Y. Liu, "Target recognition of SAR image based on CN-GAN and CNN in complex environment," *IEEE Access*, vol. 9, pp. 39608–39617, 2021.
- [19] S. Du, J. Hong, Y. Wang, K. Xing, and T. Qiu, "Multi-category SAR images generation based on improved generative adversarial network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4260–4263.
- [20] H. Kwon, S. Jeong, S. Kim, J. Lee, and K. Sohn, "Deep-learning based SAR ship detection with generative data augmentation," *J. Korea Multimedia Soc.*, vol. 25, no. 1, pp. 1–9, 2022.
- [21] L. Zou, H. Zhang, C. Wang, F. Wu, and F. Gu, "MW-ACGAN: Generating multiscale high-resolution SAR images for ship detection," *Sensors*, vol. 20, no. 22, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/22/6673>
- [22] Y.-S. Guo, H.-C. Li, W.-S. Hu, and W.-Y. Wang, "SAR image data augmentation via residual and attention-based generative adversarial network for ship detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 439–442.
- [23] D. Qosja, S. Wagner, and D. O'Hagan, "SAR image synthesis with diffusion models," in *Proc. IEEE Radar Conf.*, 2024, pp. 1–6.
- [24] S.-K. Sun, Z. He, Z.-H. Fan, and D.-Z. Ding, "SAR image target recognition using diffusion model and scattering information," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [25] Z. Ying et al., "Diffusar: Frequency domain-aware diffusion model for SAR image generation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 11851–11866, 2025.
- [26] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.
- [27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
- [28] Y. Li et al., "GLIGEN: Open-set grounded text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22511–22521.
- [29] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra, "InstanceDiffusion: Instance-level control for image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6232–6242.
- [30] D. Zhou, Y. Li, F. Ma, X. Zhang, and Y. Yang, "MIGC: Multi-instance generation controller for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6818–6828.
- [31] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8795–8805.
- [32] Z. Wang, L. Zhao, and W. Xing, "StyleDiffusion: Controllable disentangled style transfer via diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7677–7689.
- [33] H. Chung, E. S. Lee, and J. C. Ye, "MR image denoising and super-resolution using regularized reverse diffusion," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, pp. 922–934, Nov. 2022.
- [34] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
- [35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Proc. Syst.*, vol. 33, pp. 6840–6851, 2020.
- [36] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 18–24, 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [37] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
- [38] Y. Li et al., "SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 128430–128461.
- [39] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models Methods Appl.*, 2017, pp. 1–6.
- [40] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
- [41] D. P. Kingma, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [43] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
- [44] X. Zhang et al., "Multitask learning for ship detection from synthetic aperture radar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8048–8062, 2021.

Yuhan Zhang received the Ph.D. degree in dynamical systems and control from Peking University, Beijing, China, in 2022.

She is currently an Engineer with the Intelligent Science and Technology Academy of CASIC, Beijing. Her current research interests include image generation, neuromorphic sensors, and neuromorphic computing.

Jun Liao received the bachelor's degree in mechanical design, manufacturing and automation from Sichuan University, Chengdu, China, in 2022. He is currently working toward the master's degree in electronic information engineering with the School of Artificial Intelligence, Beihang University, Beijing.

His current research interests include image generation.

Jie Zhou received the Ph.D. degree in precision instrumentation from Tsinghua University, Beijing, China, in 2011.

He is currently a Researcher with the Intelligent Science and Technology Academy of CASIC, Beijing. His current research interests include deep learning and machine vision.

Yuanpei Chen received the M.S. degree in instrumentation science and technology from Beihang University, Beijing, China, in 2018.

He is currently an Engineer with the Intelligent Science and Technology Academy of CASIC, Beijing. His current research interests focus on deep learning and machine vision.

Xiaode Liu received the Ph.D. degree in biomedical engineering from the Peking University, Beijing, China, in 2021.

He is currently a Senior Engineer with the Intelligent Science and Technology Academy of CASIC, Beijing. His interests include neuromorphic computing, diffusion model, and intelligent perception.

Weihang Peng received the B.S. degree in optical information science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2015, and the M.S. degree in instrument science and technology from the China Academy of Launch Vehicle Technology, Beijing, China, in 2018.

He is currently an Senior Engineer with the Intelligent Science and Technology Academy, CASIC, Beijing. His current research interests include computer vision and machine learning.

Hongyu Yang received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2019.

She is currently an Associate Professor with the School of Artificial Intelligence, Beihang University. Her current research interests include image generation and 3-D generation.

Qi Ming received the B.S. degree in automation from the School of Automation, Beijing Institute of Technology (BIT), Beijing, China, in 2018, and the Ph.D. degree in control science and engineering from the BIT in 2024.

He is currently a Professor with the College of Computer Science, Beijing University of Technology (BJUT), Beijing. His research interests include computer vision, multimodality, and remote sensing.

Yufei Guo (Member, IEEE) received the B.E. and Ph.D. degrees in mechanics from the Beijing Institute of Technology and Peking University, Beijing, China, in 2015 and 2020, respectively.

He is currently a Senior Engineer with the Intelligent Science and Technology Academy, China Aerospace Science and Industry Corporation (CASIC), Beijing. He has authored more than 50 papers in international top-tier journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (IEEE TPAMI), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (IEEE TNNLS), Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), International Conference on Machine Learning (ICML), Annual Conference on Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), among others. His research interests include neuromorphic computing, generative model, and LLM agent.

Dr. Guo serves as a Reviewer for IEEE TPAMI, IEEE TNNLS, CVPR, ICCV, ECCV, ICML, NeurIPS, and ICLR.