2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SonarGAN: A Progressive GAN Framework for Sonar Image Denoising under Multi-Type Noises

Zhangrui Hu†, Yunxuan Feng†, Binyu Nie, Lei Yan, Wenjie Lu and Liang Hu*

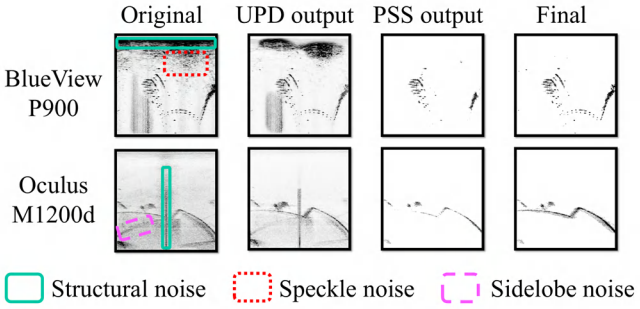
***Abstract*— Forward-looking sonar is essential for underwater perception especially in turbid waters, yet its images are often strongly degraded by various noises, including speckle, sidelobe, and structural noises, which severely hinder downstream tasks such as underwater reconstruction, positioning, and navigation. Most conventional sonar denoising methods reduce the noise at the expense of loss of fine image features or blurred image, while modern supervised learning methods demand large paired datasets that are impractical to obtain in real underwater conditions. In this paper, we propose SonarGAN, a progressive Generative Adversarial Networks (GAN) based framework that denoises sonar images under multi-type noises in one go. Unlike traditional supervised methods, SonarGAN avoids the need for costly paired datasets by combining unpaired real and simulated images, synthetic noisy–clean pairs, and joint refinement for comprehensive denoising. Extensive experiments across multiple types of sonar and underwater environments demonstrate the effectiveness of SonarGAN and its generalization in real-world conditions. We further show that SonarGAN provides high-quality inputs for downstream 3D reconstruction, significantly improving both the completeness and geometric accuracy of the reconstructed models. Our code and dataset are available at `https://github.com/Amarantos12/SonarGAN`.**

I. Introduction

Forward-looking sonar (FLS) has become an indispensable sensing modality in underwater exploration and robotic perception due to its ability to provide effective information in environments where optical cameras fail because of turbidity, low visibility, or insufficient lighting [1]. However, sonar images are often corrupted by diverse noise sources, including speckle noise caused by coherent signal interference, reverberation from multipath reflections, sidelobe noise introduced by imperfect beamforming [2], [3], as well as structural noise introduced by inherent imaging mechanisms or hardware limitations [4]. These noises reduce image contrast, blur object boundaries, and increase the difficulty of extracting reliable features, thereby limiting the performance of downstream tasks such as detection, mapping, 3D reconstruction, and autonomous navigation.

Conventional denoising methods [5], [6] reduce some noises but often over-smooth fine structures and struggle to preserve edges in dynamic underwater environments. Deep supervised learning based methods [7], [8] achieve better performance by learning noise distributions but require paired (noisy and denoised) sonar datasets that are expensive to obtain, and generalize poorly across different sonar types and underwater conditions [9]. In recent years, GAN-based methods [10], [11] have shown promising results, with CycleGAN [12] being especially attractive for sonar images since it is able to learn from unpaired data. However, CycleGAN often introduces artifacts and fails to remove structural noise, leaving room for more robust solutions.

(a) Simulation and real-world environments for sonar images collection.



(b) Denoising effects of our methods per-stage on two types of sonar.

Fig. 1: The environments for sonar data collection (a) and denoising effects of our methods (b).

To address the challenges, we propose SonarGAN, a progressive GAN-based framework that incrementally improves noise suppression while preserving target structures. SonarGAN consists of three stages: (1) Unpaired Preliminary Denoising (UPD), which employs a CycleGAN to suppress random noise using unpaired real noisy and simulated clean images; (2) Paired Structured Suppression (PSS), which leverages synthetic noisy-clean pairs in a pix2pix framework to explicitly remove structural noise; and (3) Constrained Joint Refinement (CJR), which integrates the two models for comprehensive denoising while preserving structural details.

†Equal contribution to this work. *Corresponding author.
The authors are with the College of Artificial Intelligence, Harbin Institute of Technology, Shenzhen, China. (Email: l.hu@hit.edu.cn)
This work was supported in part by the Science Center Program of National Natural Science Foundation of China under Grant 62188101, Shenzhen Science and Technology Program under Grant SYSPG20241211173609005, and under Grant JCYJ20241202123714019.

A self-attention module is embedded into all generators to capture long-range dependencies in sonar imagery.

The contributions of this paper are summarized as follows:

1) We propose a sonar image denoising framework, SonarGAN, which avoids costly annotating and removing noises in sonar images. It effectively removes various types of noise in sonar images, and generalizes easily and effectively over different types of sonars and different working conditions.
2) We introduce the Visual Geometry Group (VGG) perceptual loss and pixel supervised loss to fine-tune the generators, effectively maintain structural integrity and prevent the loss of critical target information in denoised images.
3) We conducted extensive comparison experiments on both simulated and real datasets collected from different types of sonars, demonstrating the superior performance of the proposed SonarGAN method. The results demonstrate the superior performance of SonarGAN in both qualitative visual enhancement and quantitative improvement of downstream 3D reconstruction accuracy.

The remainder of this paper is organized as follows. Related works about sonar image denoising are summarized in Section II. Section III explains the proposed SonarGAN in detail, followed by the experiments in Section IV. The conclusion and future works are given in Section V.

II. Related works

A. Classical Sonar Denoising

Early sonar denoising focused on speckle noise, using classical filters (Lee, Kuan, Frost) based on a multiplicative speckle model, which often blurred fine structures [13]. Adaptive schemes with variation-based thresholds improved edge preservation [14]. Later, generic methods such as wavelets [15], non-local means [16], and BM3D [17] achieved stronger denoising with better detail retention. However, these conventional approaches remain inadequate for suppressing sonar-specific structural noise in FLS images, including stripe and fixed-point artifacts.

B. Learning-based Sonar Denoising

With the development of deep learning, convolutional neural networks (CNNs) have been widely used in sonar image denoising. Ji et al. proposed a self-supervised method based on Noise2Void, which can effectively remove Gaussian and speckle noise [18]; Si et al. proposed WTCRNet, which combines wavelet transform and contrastive regularization for forward-looking sonar denoising [19]. In addition, blind denoising baselines such as SCUNet, which combines CNNs and Transformers, have also been explored [20]. Recently, Vishwakarma proposed a denoising and inpainting method based on convolutional sparse representation, which effectively preserves edges and textures [7].

Generative adversarial networks (GANs) have shown great potential in sonar denoising. Typically, unpaired frameworks including UIDNet [21], UNIT [22], CUT [23], StillGAN [24], and DRGAN [25] have been proposed to enhance the effectiveness of image denoising. In sonar scenarios, Zhao et al. proposed NCD-GAN, which achieves unpaired denoising through simultaneous contrastive learning [8]; Lin et al. used conditional GANs to generate masks from FLS images to filter noise [26]; STGAN combined GAN with Transformers for speckle noise suppression [27]. These methods outperform traditional ones in complex noise environments, but still face risks of unstable training and mode collapse.

In recent years, diffusion models have been widely explored for sonar image denoising. Following the seminal work of Ho et al. on DDPM [28], subsequent studies applied diffusion models to sonar super-resolution and denoising, demonstrating notable improvements on real datasets [29], [30]. However, their high inference cost limits practical real-time deployment.

III. Method

This section details the proposed SonarGAN, which adopts a three-stage progressive training strategy for denoising (Fig. 2). We first briefly review CycleGAN [12] and pix2pix [11], used in the first two stages. SonarGAN then performs initial random-noise suppression via the unpaired UPD stage, followed by sonar-specific structural noise extraction in the paired PSS stage. Finally, the CJR stage integrates the pre-trained models within a CycleGAN framework to jointly optimize denoising while preserving geometric details. For visual consistency, sonar images are inverted to grayscale.

A. Brief Introduction of CycleGAN and pix2pix

1) CycleGAN: The goal of CycleGAN is to learn a bidirectional mapping between a source domain X and a target domain Y in the absence of paired training examples [12]. The fundamental framework consists of two generators and two discriminators: the generators, $G : X \to Y$ and $F : Y \to X$, are responsible for the bidirectional mappings; while the discriminators, D_Y and D_X, are tasked with distinguishing whether the input is real or not in their respective domains. The total loss function of CycleGAN is composed of three components: adversarial loss $\mathcal{L}_{\mathrm{ad}}$, cycle consistency loss $\mathcal{L}_{\mathrm{cyc}}$ and identity loss $\mathcal{L}_{\mathrm{id}}$.

The adversarial loss constrains the distribution of generated images to align with that of the target domain. For the mapping $G : X \to Y$ and its discriminator D_Y, the loss $\mathcal{L}_{\mathrm{GAN}}(G, D_Y, X, Y)$ is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p_Y}[\log D_Y(y)] \\ + \mathbb{E}_{x\sim p_X}[\log(1 - D_Y(G(x)))]\end{aligned}. \quad (1)$$

The loss $\mathcal{L}_{\mathrm{GAN}}(F, D_X, Y, X)$ for the mapping $F : Y \to X$ and its discriminator D_X can be defined similarly. And the total adversarial loss $\mathcal{L}_{\mathrm{ad}}$ is the sum of $\mathcal{L}_{\mathrm{GAN}}(G, D_Y, X, Y)$ and $\mathcal{L}_{\mathrm{GAN}}(F, D_X, Y, X)$.

The cycle consistency loss is introduced to regularize the mapping functions and prevent them from learning arbitrary translations. This loss enforces that a forward-and-backward

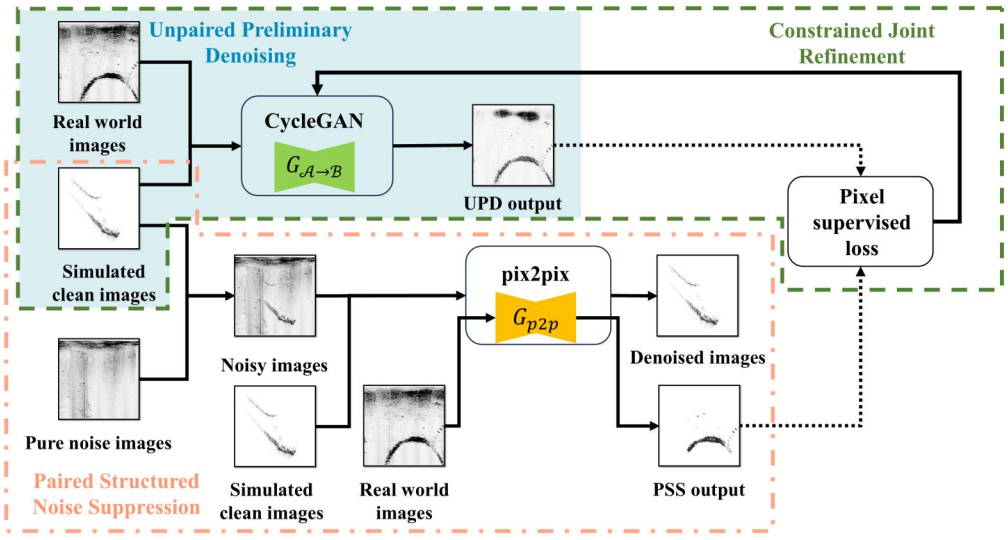


Fig. 2: The system framework of SonarGAN. The overall training is carried out in three stages. The UPD stage and the PSS stage are trained respectively to obtain two pre-trained denoising generators, which are fine-tuned in the CJR stage.

translation should be able to recover the original image, thereby approximating an identity mapping:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G,F) = \mathbb{E}_{x\sim p_X}[\|F(G(x)) - x\|_1] \\ + \mathbb{E}_{y\sim p_Y}[\|G(F(y)) - y\|_1]\end{aligned}. \tag{2}$$

The identity loss encourages the generator to preserve the input when it already belongs to the target domain. This is particularly useful for stabilizing the color and brightness distributions of the generated images:

$$\begin{aligned}\mathcal{L}_{\text{id}}(G,F) = \mathbb{E}_{y\sim p_Y}[\|G(y) - y\|_1] \\ + \mathbb{E}_{x\sim p_X}[\|F(x) - x\|_1]\end{aligned}. \tag{3}$$

The total loss function of CycleGAN is a weighted sum of these losses:

$$\mathcal{L} = \mathcal{L}_{\text{ad}} + \lambda_{\text{cyc}}\,\mathcal{L}_{\text{cyc}}(G,F) + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}}(G,F). \tag{4}$$

2) Pix2pix: Pix2pix learns a mapping from source domain X_1 to target domain Y_1 using paired data [11], consisting of a generator and a conditional discriminator. The generator translates inputs to target images, while the discriminator distinguishes real from generated pairs. Its objective combines an adversarial loss for realism and an L1 reconstruction loss to enforce pixel-level fidelity.

B. Unpaired Preliminary Denoising (UPD)

The goal of the UPD stage is to use the CycleGAN framework to train a generator $G_{\mathcal{A}\to\mathcal{B}}$ that can receive noisy sonar images and generate preliminary denoising results. Specifically, we first obtain two unpaired datasets: the noisy sonar image set $\mathcal{A}$ collected in the real world and the noise-free sonar image set $\mathcal{B}$ collected in a simulated environment. Then $G_{\mathcal{A}\to\mathcal{B}}$ is trained under the CycleGAN framework to complete the mapping from domain $\mathcal{A}$ to domain $\mathcal{B}$, as illustrated in Fig. 3.

However, using only the CycleGAN loss here cannot achieve ideal denoising because it is difficult to balance the

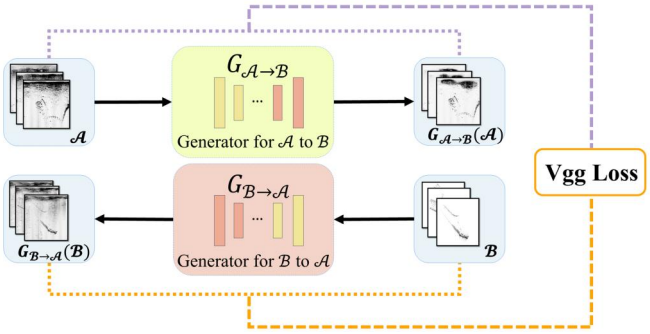


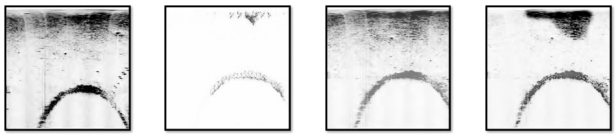
Fig. 3: The structure of UPD.

effects of adversarial loss and cycle loss. The former will cause the model to mistakenly identify some pixels with weak intensity as noise and remove them, as shown in Fig. 4(b); while the latter tends to preserve nearly all information from the input, thereby hindering the removal of complex noises, as shown in Fig. 4(c). To achieve a better balance between effective denoising and structural fidelity, we design a VGG loss $\mathcal{L}_{\text{vgg}}$ as below:

$$\begin{aligned}\mathcal{L}_{\text{vgg}} = \sum_{l\in L}\Big(\|\phi_l(G_{\mathcal{A}\to\mathcal{B}}(\mathcal{A})) - \phi_l(\mathcal{A})\|_1 \\ + \|\phi_l(G_{\mathcal{B}\to\mathcal{A}}(\mathcal{B})) - \phi_l(\mathcal{B})\|_1\Big),\end{aligned} \tag{5}$$

where $\phi_l(\cdot)$ denotes the feature map extracted by the l-th layer of VGG19 [31], and L is the number of the layers. The VGG loss measures similarity in the deep feature space, rather than in the raw pixel space. This allows the model to preserve high-level semantic information, such as object edges and textures, while having more freedom to alter pixel-level noise distributions, thereby breaking the aforementioned dilemma, as shown in Fig. 4(d).

The final loss function for the UPD stage is a weighted



(a) (b) (c) (d)

Fig. 4: (a) The original noisy image; (b) The over-denoising image; (c) The incomplete-denoising image; (d) The denoised image using VGG perceptual loss.

sum of the standard CycleGAN loss and the VGG loss:

$$\mathcal{L}_{\mathrm{UPD}} = \mathcal{L}_{\mathrm{ad}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}}. \quad (6)$$

We optimize the model with $\mathcal{L}_{\mathrm{UPD}}$ and retain the UPD outputs, yielding preliminarily denoised images $\hat{\mathcal{B}} = G_{\mathcal{A}\to\mathcal{B}}(\mathcal{A})$ that suppress random noise while preserving coarse structures. These results, together with the pretrained generators and discriminators, are used for joint optimization in the third (CJR) stage.

A key design choice is using polar rather than Cartesian sonar images. Since the PatchGAN discriminator [11] evaluates local patches, Cartesian representations contain many invalid patches outside the sector (Fig. 5), which are trivially judged as clean and can prematurely stop generator training, degrading denoising performance.

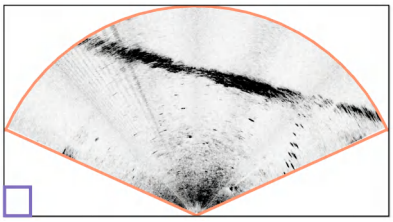


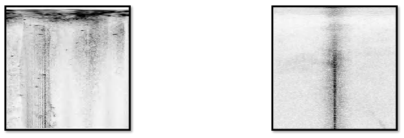
Fig. 5: The sector area denotes the sonar detection region, while the remaining white background indicates the invalid region.

C. Paired Structured Suppression (PSS)

While the UPD stage effectively suppresses random noise, it cannot handle structural noise. Such noise arises from the sonar imaging mechanism and appears as stripe artifacts, fixed-point clutter, or other stable patterns persisting across frames (Fig. 6). Because these artifacts closely resemble true targets in intensity and shape within a single image, unpaired methods face inherent ambiguity in removing them.

To address this issue, we adopt a pix2pix framework to train a generator G_{p2p} that removes structural noise from raw sonar images. We first collect pure noise images $\mathcal{N}$ in empty environments containing only background and stable structural artifacts. These are added to clean simulated images $\mathcal{B}$ to form paired samples $(\mathcal{C} = \mathcal{B} + \mathcal{N}, \mathcal{B})$ for supervised training (Fig. 7).

Since the input data are paired, the generator G_{p2p} can be trained by the pix2pix framework. The loss function



(a) (b)

Fig. 6: (a) The pure noise image of the sonar BlueView P900-130. (b) The pure noise image of the sonar Oculus M1200d.

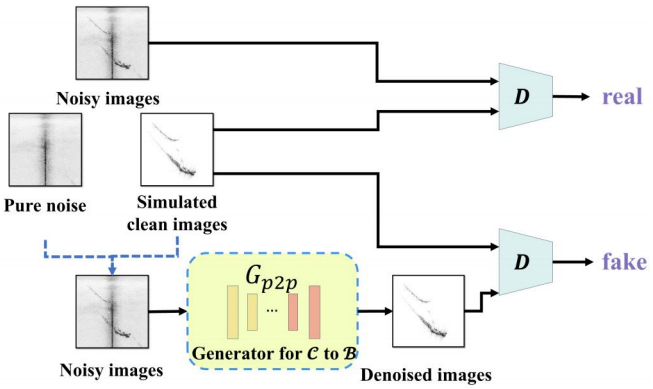


Fig. 7: Using pure noise and clean images, we synthesize noisy samples and train a conditional GAN to map from the noisy to the clean domain. The discriminator distinguishes real noisy, clean pairs from generated ones, while the generator learns to transform noisy inputs into clean-like outputs that fool the discriminator.

of the PSS stage $\mathcal{L}_{\mathrm{PSS}}$ combines the standard conditional adversarial loss with a pixel-level L1 reconstruction loss, ensuring that the generated outputs align with the target images in both distributional and content fidelity:

1) Adversarial Loss: The adversarial loss constrains the generator's output to be close to the distribution of real target images. Its form is:

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}(G_{\mathrm{p2p}}, D) = \mathbb{E}_{\mathcal{C},\mathcal{B}}[\log D(\mathcal{C},\mathcal{B})] \\ + \mathbb{E}_{\mathcal{C}}[\log(1 - D(\mathcal{C}, G_{\mathrm{p2p}}(\mathcal{C})))].\end{aligned} \quad (7)$$

Here, $D(\mathcal{C},\mathcal{B})$ determines whether the input-target pair comes from real data, while $D(\mathcal{C}, G_{\mathrm{p2p}}(\mathcal{C}))$ determines whether the input-generated pair is synthetic.

2) Reconstruction Loss: To ensure the generated image is close to the real target in pixel space, pix2pix employs the L1 loss as an additional constraint:

$$\mathcal{L}_{\mathrm{L1}}(G_{\mathrm{p2p}}) = \mathbb{E}_{\mathcal{C},\mathcal{B}}[\|\mathcal{B} - G_{\mathrm{p2p}}(\mathcal{C})\|_1]. \quad (8)$$

Combining both components, the optimization objective of pix2pix is:

$$\mathcal{L}_{\mathrm{total}}(G_{\mathrm{p2p}}, D) = \mathcal{L}_{\mathrm{GAN}}(G_{\mathrm{p2p}}, D) + \lambda\,\mathcal{L}_{\mathrm{L1}}(G_{\mathrm{p2p}}), \quad (9)$$

where λ is a weighting coefficient that balances adversarial consistency and pixel-level accuracy.

By optimizing $\mathcal{L}_{\mathrm{PSS}}$, the PSS stage yields a generator G_{p2p} that can effectively map sonar images containing structured artifacts into the noise-free domain. This generator

serves as a structural noise suppression model, which will be integrated into the third stage (CJR) to provide a critical structural prior for joint optimization.

D. Constrained Joint Refinement (CJR)

The UPD and PSS stages are designed for random speckle noise and specific structural noise, respectively. Using either one of them alone cannot guarantee both the structural integrity of targets and the comprehensive denoising effect. Therefore, to exploit the two stages synergistically in this stage, the denoising generator G_{p2p}, pretrained in the PSS stage, is used as a fixed guidance module to refine the generator $G_{\mathcal{A}\to\mathcal{B}}$ obtained from the stage of UPD.

The generator G_{p2p} is used to generate a mask that completely removes structural noise and ensures the integrity of the target structure in the original images. We input the noisy original image $\mathcal{A}$ (Fig. 4(a)) into G_{p2p} to get a denoised image $\hat{\mathcal{B}}_1 = G_{\mathrm{p2p}}(\mathcal{A})$, as shown in Fig. 8. Since the brighter area in $\hat{\mathcal{B}}_1$ comes from real echoes, which correspond to actual regions in the scene and contain most of the true targets, we apply the Otsu adaptive threshold method to generate a high-precision target mask M.

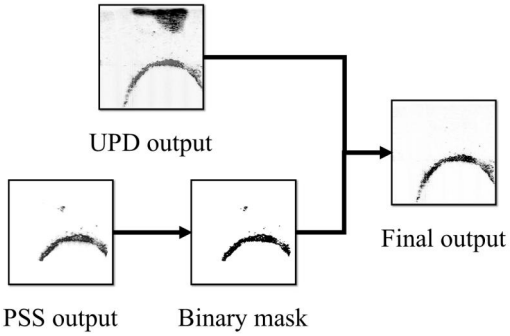


Fig. 8: We perform pixel-wise comparison between the binarized output of the PSS stage and the output of the UPD stage, and obtain the final denoising result through training.

For the target regions in the binary mask M (where $M = 1$, corresponding to the black regions in Fig. 8), the generator output should remain consistent with the noisy original image to ensure the target structure is not damaged; for the noise regions (where $M = 0$, corresponding to the white regions in Fig. 8), the generator output should approach zero as closely as possible to achieve noise suppression. To this end, we design the following pixel supervised loss function:

$$\begin{aligned}\mathcal{L}_{\mathrm{ps}} = &\|(1-M)\odot G_{\mathcal{A}\to\mathcal{B}}(\mathcal{A})\|_1 \\ &+ \|M\odot(G_{\mathcal{A}\to\mathcal{B}}(\mathcal{A})-\mathcal{A})\|_1.\end{aligned} \tag{10}$$

where $\mathcal{A}$ represents the noisy original image, $G_{\mathcal{A}\to\mathcal{B}}(\mathcal{A})$ denotes the generator output image, M is the binary target mask obtained through the Otsu method, $\odot$ indicates element-wise multiplication, and $\|\cdot\|_1$ represents the L1 norm.

By adding constraints on structural noise to the generator G_{p2p}, the generator and discriminator in the UPD stage can be further optimized. The total loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{CJR}} = &\lambda_{\mathrm{gan}}\mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}} \\ &+ \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}} + \beta\mathcal{L}_{\mathrm{ps}}.\end{aligned} \tag{11}$$

Through this design, the structural discriminative knowledge embedded in the PSS model is effectively injected into the UPD generator via the mask constraint, thereby enhancing its generalization capacity for denoising. Additionally, due to the VGG perceptual loss, the denoised output by the generator $G_{\mathcal{A}\to\mathcal{B}}$ after joint training preserves the complete structure of the target objects. The refined generator $G_{\mathcal{A}\to\mathcal{B}}$ achieves collaborative suppression of both random and structural noises while maintaining pixel-level structural fidelity in target regions, significantly improving denoising robustness and effectiveness over the original generator $G_{\mathcal{A}\to\mathcal{B}}$ obtained at the stage of UPD. The workflow of this stage is illustrated in Fig. 8.

IV. Experiments

In this section, we conducted extensive experiments on both synthetic datasets and real-world datasets. To comprehensively validate the proposed SonarGAN, we evaluate both the statistical image quality and the performance improvement on the downstream 3D reconstruction task. The proposed SonarGAN's denoising performance was compared with conventional methods [15], [16], [17], learning-based methods [20] and GAN-based methods [21], [22], [23], [24], [25]. All training and testing were performed on an Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz with four NVIDIA GeForce RTX 3090 GPUs, while inference was evaluated on a single RTX 3090 GPU. Our method achieves an average inference speed of 20 frames per second (FPS), which fully satisfies the real-time requirements of underwater robotic tasks.

A. Experiments on synthetic images

Since the simulated sonar noise distribution is relatively simple, we synthesized noise-free images collected in the simulation environment with pure noise images by sonars in real-world underwater scenes to provide more realistic experiment data. We used two simulation environments, HoloOcean [32] and DAVE Aquatic Virtual Environment [33]. The pure noise images were obtained from two different types of sonar, Oculus M1200d[1] and BlueView P900-130[2], respectively.

Fig. 9 and Table I summarize denoising results on synthetic datasets using PSNR, SSIM, and LPIPS. Conventional methods perform poorly across all datasets, while existing deep learning approaches show dataset-specific gains but suffer from limited generalization, over-denoising, or structural degradation. Although CycleGAN is relatively more stable, it still fails to preserve target structures. In contrast, our method consistently achieves the best performance, with

[1]The datasets were collected by ourselves in a laboratory tank and a river.

[2]The open-source dataset Aracati2017 [34] was acquired in the marina of the Yacht Club of Rio Grande, Brazil. The dataset is available at `https://github.com/matheusbg8/aracati2017`.

TABLE I: Quantitative Comparison of Experimental Results on Synthetic Datasets

Method	DAVE + M1200d			DAVE + P900			HoloOcean + M1200d			HoloOcean + P900		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BM3D [17]	19.33	0.07	0.41	15.48	0.06	0.55	20.05	0.14	0.45	15.72	0.14	0.60
NLM [16]	19.42	0.10	0.26	15.37	0.08	0.51	19.94	0.15	0.35	15.52	0.14	0.59
Wavelet [15]	18.37	0.07	0.46	14.70	0.06	0.61	19.38	0.13	0.56	15.06	0.13	0.68
DRGAN [25]	18.12	0.29	0.51	17.13	0.12	0.58	27.09	0.81	0.25	26.51	0.79	0.30
CUT [23]	27.67	0.81	0.20	18.04	0.10	0.60	14.86	0.10	0.62	9.99	0.08	0.77
UIDNET [21]	27.28	0.66	0.38	22.88	0.57	0.49	24.03	0.22	0.44	17.49	0.23	0.61
SCUNET [20]	19.44	0.10	0.29	15.28	0.08	0.59	19.95	0.16	0.37	15.41	0.14	0.64
StillGAN [24]	25.21	0.14	0.32	21.74	0.09	0.51	27.85	0.77	0.22	23.01	0.71	0.43
UNIT [22]	23.00	0.17	0.40	22.59	0.28	0.42	26.24	0.49	0.35	21.21	0.13	0.52
CycleGAN [12]	24.65	0.39	0.29	22.91	0.28	0.37	28.18	0.58	0.27	21.52	0.26	0.47
Ours	**32.66**	**0.93**	**0.09**	**30.41**	**0.92**	**0.15**	**31.52**	**0.91**	**0.15**	**33.21**	**0.92**	**0.15**

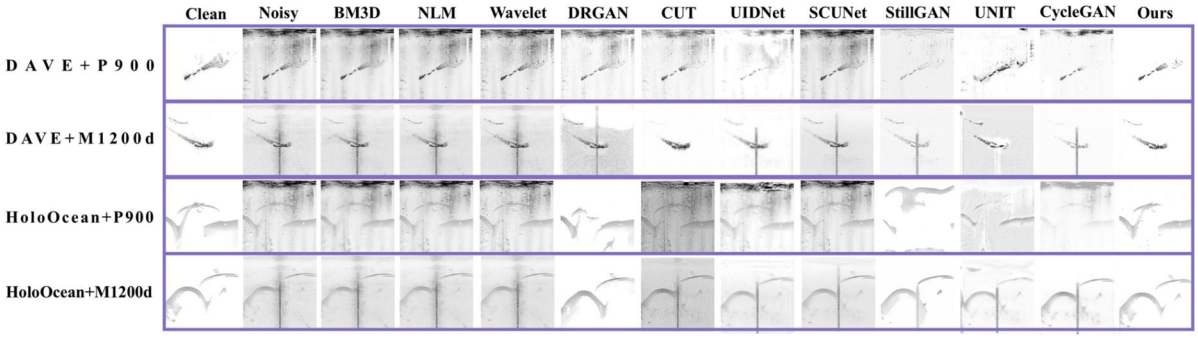


Fig. 9: Comparison of denoising results on synthetic datasets.

robust noise suppression, accurate structure preservation, and strong cross-dataset generalization.

B. Experiments on real sonar images

We used real-world data collected by the same two sonars as in the synthetic dataset experiments. Since the noise and intensity distributions of FLS do not satisfy the natural scene statistics assumption [35] on which the commonly used no-reference image quality assessment metrics are based, we adopted a new evaluation metric based on the Kullback–Leibler (KL) divergence to measure the statistical difference between denoising residuals and pure background noise here. The metric is defined as:

$$D_{\mathrm{KL}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{b=0}^{255} p_{ij,b} \ln \left(\frac{p_{ij,b}}{q_{ij,b}} \right), \tag{12}$$

where H and W are the height and width of the image, $p_{ij,b}$ represents the probability of the b-th histogram bin at pixel (i, j) in the residual images (the absolute value of the difference between the image before and after denoising), and $q_{ij,b}$ denotes the probability of the b-th bin at (i, j) in the pure noise images.

In addition, we proposed an evaluation metric based on mean images by averaging pixel values across the residual and pure noise datasets to obtain their respective mean images. PSNR and SSIM between these mean images were then calculated to evaluate spatial fidelity and noise suppression performance.

As shown in Table II and Fig. 10, DRGAN performs well on the Oculus M1200d dataset but degrades on the BlueView P900 due to highly variable noise. Traditional methods provide limited denoising; CUT distorts structures, UNIT removes structural noise at the cost of details, CycleGAN yields unbalanced results, and StillGAN is only partially effective. In contrast, our method consistently achieves strong denoising and structure preservation across both datasets, outperforming others on multiple metrics.

Furthermore, we conducted an experiment to evaluate the generalization ability of our proposed method. Specifically, we directly applied the model trained on Oculus M1200d data collected in a river to denoise the laboratory tank dataset. Surprisingly, it can still achieve a perfect denoising effect, as evident by comparative results in Fig. 11.

C. Evaluation on Downstream 3D Reconstruction

To assess the impact of SonarGAN on downstream perception, we evaluate 3D reconstruction performance of Differentiable Space Carving (DSC) [36] using images denoised by different methods. Due to space limitations, detailed quantitative and visualization results on synthetic data are provided in the supplementary video. This section reports reconstruction results from a real laboratory tank only. Following [36], an H-shaped target ($0.5 \times 0.1 \times 0.6$ m) was scanned top-down using an Oculus M1200d sonar, as illustrated in Fig. 12. Images denoised by different methods were directly fed into the DSC framework for 3D reconstruction.

As shown in Fig. 13, ambient noise degrades reconstruc-

TABLE II: Quantitative Comparison of Experimental Results on Real-world Datasets

Method	Oculus M1200d			Aracati2017 (BlueView P900-130)		
	KL Divergence ↓	PSNR ↑	SSIM ↑	KL Divergence ↓	PSNR ↑	SSIM ↑
BM3D	0.62	8.40	0.12	2.37	8.80	0.02
NLM	0.64	9.15	0.27	2.51	10.19	0.15
Wavelet	0.63	8.85	0.22	1.88	9.63	0.11
DRGAN	**0.43**	23.33	0.94	0.84	18.13	0.78
CUT	0.63	15.33	0.52	2.06	18.84	0.71
UIDNET	0.65	22.06	0.90	2.57	16.39	0.68
SCUNET	0.65	15.08	0.35	2.57	16.51	0.35
StillGAN	0.66	13.98	0.70	1.22	18.74	0.72
UNIT	0.56	14.66	0.55	0.91	10.75	0.29
CycleGAN	0.55	22.24	0.89	1.05	18.46	0.77
Ours	0.47	**31.83**	**0.96**	**0.65**	**19.87**	**0.85**

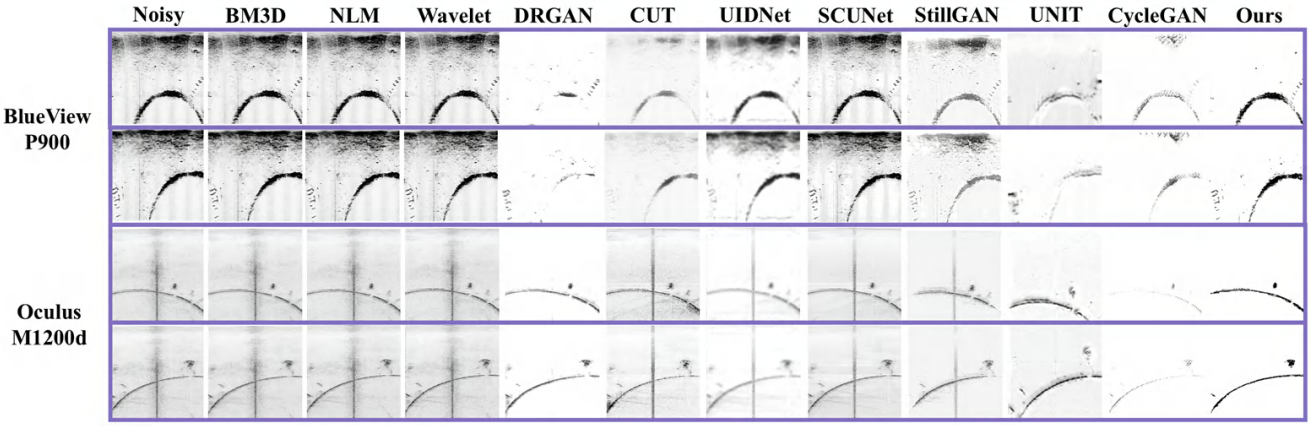


Fig. 10: Comparison of denoising results on real-world datasets.

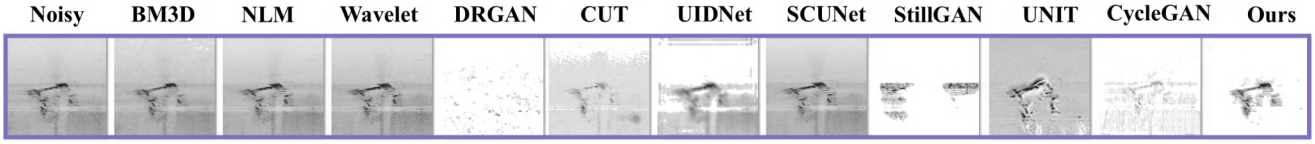


Fig. 11: Comparison of denoising results on the laboratory tank datasets.

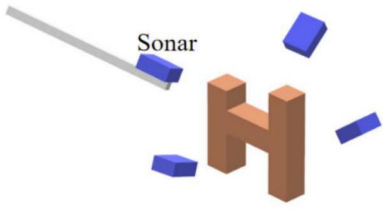


Fig. 12: Sonar perspectives for imaging in the laboratory tank [36].

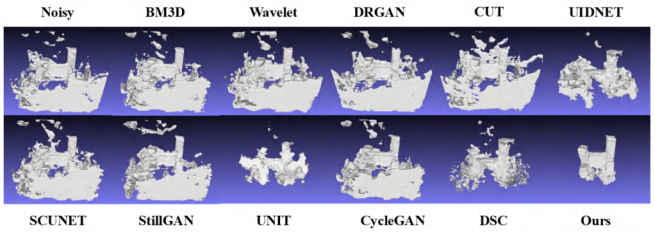


Fig. 13: 3D reconstruction of H-shape in the tank using sonar data denoised by different methods.

tion quality in real underwater conditions. The reconstruction using raw data exhibit floating artifacts, while the ones using denoised data from some baselines such as UNIT recover contours but cause geometric distortions. In contrast, SonarGAN suppresses acoustic noise and produces reconstructions with preserved geometry and structural integrity, demonstrating its effectiveness as a pre-processing step for real-world 3D reconstruction.

V. Conclusion and Future Work

This paper proposes SonarGAN, a GAN-based denoising framework for FLS images. Through progressive training, two generators are learned to sequentially remove speckle noise and sonar-specific structural noise, and are integrated

to achieve enhanced denoising. The method requires neither noise-free real data nor manual annotations, ensuring strong practicality and generalization. Extensive comparative experiments demonstrated that the denoising performance of SonarGAN is superior to existing methods. Furthermore, the denoised images are shown to effectively improve the geometric accuracy of downstream 3D reconstruction tasks. Nevertheless, SonarGAN shows a certain dependency on the noiseless sonar domain in unpaired training, as different forms of sonar images directly affect the quality of the denoised results. In future work, we plan to explore more robust approaches that remove the dependency on noiseless sonar images.

REFERENCES

[1] H. Li, G. Wang, X. Li, and Y. Lian, "A review of underwater slam technologies," in *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*. IEEE, 2023, pp. 215–225.

[2] B. Nie, W. Lu, Y. Feng, H. Gao, and K. Lin, "Removing multi-path echoes in underwater 3d reconstruction via multi-view consistency," *Pattern Recognition Letters*, vol. 189, pp. 48–55, 2025.

[3] X. Liu, J. Fan, C. Sun, Y. Yang, and J. Zhuo, "High-resolution and low-sidelobe forward-look sonar imaging using deconvolution," *Applied Acoustics*, vol. 178, p. 107986, 2021.

[4] P. Zhou, D. Chen, and X. Teng, "Structural noise removal for imaging sonar: A case study for blueview imaging sonar," in *2021 6th International Conference on Transportation Information and Safety (ICTIS)*. IEEE, 2021, pp. 1582–1586.

[5] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[6] T. Ye, X. Deng, X. Cong, H. Zhou, and X. Yan, "Parallelisation strategy of non-local means filtering algorithm for real-time denoising of forward-looking multi-beam sonar images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[7] A. Vishwakarma, "Denoising and inpainting of sonar images using convolutional sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–9, 2023.

[8] B. Zhao, Q. Zhou, L. Huang, and Q. Zhang, "Unpaired sonar image denoising with simultaneous contrastive learning," *Computer Vision and Image Understanding*, vol. 235, p. 103783, 2023.

[9] A. Agrawal, A. Sikdar, R. Makam, S. Sundaram, S. K. Besai, and M. Gopi, "Syn2real domain generalization for underwater mine-like object detection using side-scan sonar," *IEEE Geoscience and Remote Sensing Letters*, 2025.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[13] N. D. Mascarenhas, "An overview of speckle noise filtering in sar images," in *Image Processing Techniques, First Latino-American Seminar on Radar Remote Sensing*, vol. 407, 1997, p. 71.

[14] A. Lopes, R. Touzi, and E. Nezry, "Adaptive speckle filters and scene heterogeneity," *IEEE transactions on Geoscience and Remote Sensing*, vol. 28, no. 6, pp. 992–1000, 2002.

[15] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE transactions on image processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[16] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2. Ieee, 2005, pp. 60–65.

[17] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[18] Y. Ji, L. Xie, G. Xu, Z. Zhu, Y. Gu, J. Fei, Y. Li, Y. Chen, and K. Cai, "Sonar image denoising based on noise2void self-supervised learning," *Remote Sensing Letters*, vol. 16, no. 5, pp. 560–571, 2025.

[19] C. Si, S. Zhang, Q. Cai, T. Zhang, M. Zhang, X. Han, and J. Dong, "Wtcrnet: a wavelet transform and contrastive regularization network for sonar denoising by self-supervision," *Intelligent Marine Technology and Systems*, vol. 2, no. 1, p. 17, 2024.

[20] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, D.-P. Fan, R. Timofte, and L. V. Gool, "Practical blind image denoising via swin-conv-unet and data synthesis," *Machine Intelligence Research*, vol. 20, no. 6, pp. 822–836, 2023.

[21] Z. Hong, X. Fan, T. Jiang, and J. Feng, "End-to-end unpaired image denoising with conditional adversarial networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 4140–4149.

[22] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.

[23] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European conference on computer vision*. Springer, 2020, pp. 319–345.

[24] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, "Structure and illumination constrained gan for medical image enhancement," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3955–3967, 2021.

[25] Y. Huang, W. Xia, Z. Lu, Y. Liu, H. Chen, J. Zhou, L. Fang, and Y. Zhang, "Noise-powered disentangled representation for unsupervised speckle reduction of optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2600–2614, 2020.

[26] T. Lin, A. Hinduja, M. Qadri, and M. Kaess, "Conditional gans for sonar image filtering with applications to underwater occupancy mapping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1048–1054.

[27] X. Zhou, K. Tian, and Z. Zhou, "Stgan: Sonar image despeckling method utilizing gan and transformer," in *International Symposium on Artificial Intelligence and Robotics*. Springer, 2023, pp. 56–67.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[29] O. Bryan, T. Berthomier, B. D'Ales, T. Furfaro, T. S. Haines, Y. Pailhas, and A. Hunter, "A diffusion-based super resolution model for enhancing sonar images," *The Journal of the Acoustical Society of America*, vol. 157, no. 1, pp. 509–518, 2025.

[30] Z. Yang, J. Zhao, H. Zhang, Y. Yu, and C. Huang, "A side-scan sonar image synthesis method based on a diffusion model," *Journal of Marine Science and Engineering*, vol. 11, no. 6, p. 1103, 2023.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[32] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "Holoocean: An underwater robotics simulator," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3040–3046.

[33] W.-S. Choi, D. R. Olson, D. Davis, M. Zhang, A. Racson, B. Bingham, M. McCarrin, C. Vogt, and J. Herman, "Physics-based modelling and simulation of multibeam echosounder perception for autonomous underwater manipulation," *Frontiers in Robotics and AI*, vol. 8, p. 706646, 2021.

[34] M. M. Dos Santos, G. G. De Giacomo, P. L. Drews-Jr, and S. S. Botelho, "Cross-view and cross-domain underwater localization based on optical aerial and acoustic underwater images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4969–4974, 2022.

[35] N. Venkatanath, D. Praneeth, S. C. Sumohana, S. M. Swarup *et al.*, "Blind image quality evaluation using perception based features," in *2015 twenty first national conference on communications (NCC)*. IEEE, 2015, pp. 1–6.

[36] Y. Feng, W. Lu, H. Gao, B. Nie, K. Lin, and L. Hu, "Differentiable space carving for 3d reconstruction using imaging sonar," *IEEE Robotics and Automation Letters*, 2024.