Multimedia Systems (2026) 32:301
https://doi.org/10.1007/s00530-026-02334-6

REGULAR PAPER

Sonar images generation method based on improved stable diffusion combined with ControlNet model

Jier Xi[1,2] · Xiufen Ye[1] · Shuxiang Guo[2,3] · Chunying Li[2] · Hanjie Huang[1,2]

Received: 13 November 2025 / Accepted: 3 March 2026

Abstract

With development of sonar technology and deep learning, sonar images have been widely used to detect underwater targets. However, the shortage of sonar data sources has become one of the biggest challenges in underwater targets detecting with neural networks. For example, actual underwater experiments are complicated and high cost, and the un-known detection targets and wreck degree, which makes it challenging on sonar object detection and recognition algorithms. In order to solve these problems, we propose a sonar image generating method based on diffusion model combined with ControlNet and LoRA method. The sonar-style target sample images are generated through Stable Diffusion with real sonar images. By combining the optical targets and sonar targets, the proposed image can produce multiple sonar-style images. Use of LoRA method with few real sonar images, which can train sonar stylized model to generate images. In the meantime, by introducing fine-tuned ControlNet model via control wreck degree of optical image, which makes the specific target be expanded with more diversification. The model makes the degree of damage to sonar target wreckage controllable. The experimental results show that our method achieves features similar to those of real sonar targets and can be used for training detection tasks.

Keywords Sonar images · Sonar-style target · Stable diffusion · ControlNet

1 Introduction

With the development of neural networks, more and more researchers applied the technology to underwater object detection with sonar images [1, 2].And optical imaging can be used as a type of measurement technology and is widely applied in the simulation and recognition tasks of sonar images [3, 4].Some of the research simulates sonar images via generative adversarial networks (GAN) [5, 6]to generate high-resolution images. As one of the most popular methods, GAN has been applied in many fields, especially in generative images. The network of generators is challenging to control in the training phase, and the loss function has random fluctuation, making the model converges difficultly. And the method requires a large amount of real data. It has become one of the blockers in the generative sonar images. With the limited number of real sonar images, using optical images via style transfer learning to generate sonar-like images is one of the effective methods to simulate sonar targets [7–10]. The data with style transfer learning can be applied on training detection model phase. However, these

Communicated by Bing-kun Bao.

✉ Xiufen Ye
yexiufen@hrbeu.edu.cn

Jier Xi
xijier@hrbeu.edu.cn

Shuxiang Guo
guo.shuxiang@sustech.edu.cn

Chunying Li
licy@sustech.edu.cn

Hanjie Huang
836452567@hrbeu.edu.cn

1 College of Intelligent Systems Science and Engineering, Harbin Engineering University, Nantong Street, Harbin, Heilongjiang 150001, China

2 Advanced Institute for Ocean Research, Southern University of Science and Technology, Xueyuan Avenue, Shenzhen, Guangdong 518055, China

3 Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Xueyuan Avenue, Shenzhen, Guangdong 518055, China

Published online: 06 May 2026

methods still need to consider the sonar imaging mechanism and hard to applied on target-semantic.

Some of the research via improved style transfer learning and imaging mechanisms has addressed the issues. Bai et al. [7] using a context external-attention network to generate pseudo-side-scan sonar images for zero-shot classification, with optical and acoustic images serving as content and style inputs. A global context block merges style and content features, while an external attention block finds correlations to augmentation sonar images. Xi et al. [8] proposed an improved style transfer-based method via stain-like noise and shadow distribution enhancement on optical images to augment sonar image samples for zero-shot learning. Huang et al. [11] proposed a comprehensive sample augmentation method for side-scan sonar targets, backgrounds, textures, resolutions, and noise using a wreck as an example. However, style transfer-based methods, it has always spent enormous computer resources and time and the optical image with sonar style has a low representative with actual sonar.

To minimize the domain gap between real-world data and synthetic simulations, some of the research applied the diffusion model to simulated sonar images [12, 13]. Stable diffusion [14–16]has extensive applications in various fields, especially in the field of image generation. It is highly improved in the field of generative sonar images. The methods with fine-tuned diffusion models generate more realistic sonar images compared with other methods. However, these methods still cannot simulate wreck target and improve detection performance effectively. In our paper, we applied fine-tuned stable diffusion as the basic method to generate sonar images. The general process of applying stable diffusion to generate sonar image is shown in the Fig. 1. The overall idea is extracting source target x^{src} features and fusing these features to generate new target x^{target} via adding noising data z^{src} and reducing noise z^{target}.

Our paper introduces a fine-tuned stable diffusion combined with ControlNet method and its applications in sonar image generation with sonar features using optical image. Our contribution lies in exploring stable diffusion and LoRA effectiveness compared to other methods and its potential for various applications. The generated sonar images by our method can serve as training datasets for detection models and detected targets in real sonar images.

2 Related work

This work aims to generate high-fidelity and geometry-consistent sonar images from optical images to augment training data for both side-scan sonar (SSS) and forward-looking sonar (FLS) tasks. Existing studies on sonar data augmentation can be broadly categorized into physics-based simulation, GAN-based generation, and recent diffusion-based methods.

2.1 Sonar image synthesis and data augmentation

Physics-based sonar simulators and synthetic datasets [2] generate sonar-like images by modeling acoustic propagation and seabed scattering. While these approaches provide physically meaningful images, they often fail to reproduce the complex appearance and structural variability of real sonar targets, and they require expensive parameter tuning and scene modeling.

GAN-based and Style transfer methods [4–8, 10, 11] have been widely used to map optical or simulated images to sonar-like appearances. These methods focus on learning appearance-level transformations, but they often suffer from structure distortion, mode collapse, and limited controllability, which makes them unsuitable for generating task-oriented training data for detection and classification.

Recently, zero-shot and diffusion-based sonar generation methods [12] have shown improved visual realism. However, existing diffusion-based approaches mainly target image-level style synthesis and lack explicit mechanisms to preserve target geometry or enable controllable cross-modal mapping from optical to sonar domains.

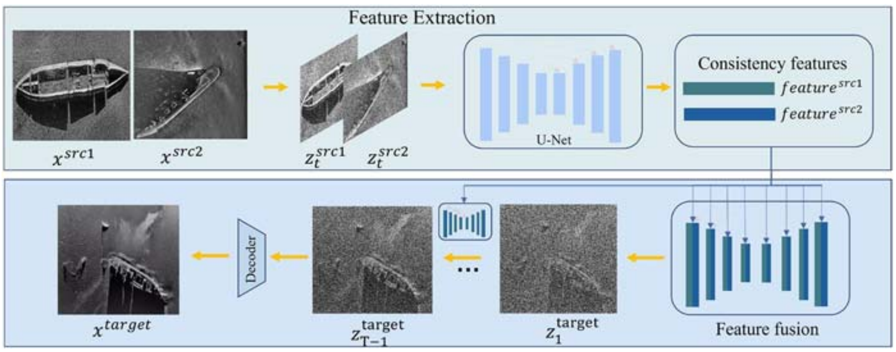


Fig. 1 Applying diffusion to generate Sonar image

2.2 Cross-modal image translation for sonar

In many practical scenarios, optical images provide accurate geometry and clear object boundaries, whereas sonar images encode acoustic reflectivity and shadow patterns. Most prior sonar translation methods perform appearance-level style transfer based method [4, 7, 8, 10, 11] without explicit geometric constraints, which leads to misalignment between the generated sonar image and the original object structure. Such misalignment limits the usability of generated samples for downstream detection and recognition tasks.

2.3 Diffusion models with controllable adaptation

Diffusion models [14, 15] provide a powerful generative framework by iteratively denoising random noise to synthesize images. Compared with GANs, diffusion models exhibit better stability and diversity. However, fine-tuning large diffusion models for domain-specific tasks is computationally expensive and often requires large-scale labeled data.

Low-rank adaptation (LoRA) [17] enables efficient fine-tuning of diffusion models with limited data, while ControlNet introduces structural constraints to guide the generation process. In this work, we integrate LoRA and ControlNet into a unified diffusion-based framework to achieve controllable optical-to-sonar translation for both SSS and FLS images. Unlike prior sonar generation methods, our approach explicitly preserves object geometry while adapting sonar-specific appearance, enabling the generation of task-consistent training samples for detection and classification.

As summarized in Table 1, existing sonar augmentation methods are either limited to appearance-level translation or lack geometric consistency, which restricts their applicability to detection and classification tasks. Our method enables geometry-preserving optical-to-sonar generation for both SSS and FLS, making it directly usable for downstream detection and classification, as will be detailed in the following sections.

Table 1 COMPARISON OF SONAR DATA AUGMENTATION METHODS

Method	**Optical→Sonar**	**Geometry Control**	**SSS**	**FLS**
Physics-based [2]	×	✓	✓	limited
GAN-based [4–7]	✓	×	✓	✓
Style transfer [8, 11]	✓	weak	✓	✓
Diffusion [12]	×	weak	✓	limited
Ours	✓	✓	✓	✓

3 Method

3.1 Overall idea of stable diffusion for generate sonar image

To meet the data requirements of detection model on sonar target detection tasks, the generated sonar images model should have the characters of similar feature distribution compare with real data, more data diversity than real data, generating synthetic data for rare classes to avoid the class imbalance problem, and the generated image with human visual perception.

Sonar image is an acoustic-based image, using the acoustic-based data as diffusion model input, the generated data has similar acoustic-based data distribution. The sonar image detection task needs a large amount of diverse data as training data to improve generalization ability of the detection model. Diffusion can generate images with different backgrounds, angles, lighting, and environments. Such of diversity expands the sonar data, its makes object recognition models from various perspectives and under changing conditions. Diffusion model restores complex structures and features from Gaussian noise when generating images gradually. The model not only extract features from the data but also the generated image must maintain semantic similarity with real target images in a low-dimensional feature space, ensuring that the generated data is structurally similar to real images. Another, in the sonar object detection tasks, part of classes of data may be scarce or need detect certain specific targets, such as wreck. Images generated by LoRA based on diffusion model can be used to synthesize samples for these classes, addressing the data imbalance issue and enhancing the model's ability to recognize specific targets. Since this method is based on a Stable Diffusion-based generation framework, image synthesis is performed by first extracting original image features and then recombining these features to form new target images. As a result, the generated sonar images demonstrate cross-sonar-type generality and are applicable to both Side-Scan Sonar (SSS) and Forward-Looking Sonar (FLS).

3.2 Our methods

Generally, our research goal is to overcome the limited amount of real sonar data and the lack of clear, detailed features in the sonar images generated by existing methods. Due to the less real sonar image, the target's background also has less diversity. To control the generated target with special sonar features and reduce fine-tuning costs, we propose a generative sonar image framework based on LoRA and stable diffusion model that combined ControlNet method which using optical images and wreck function as

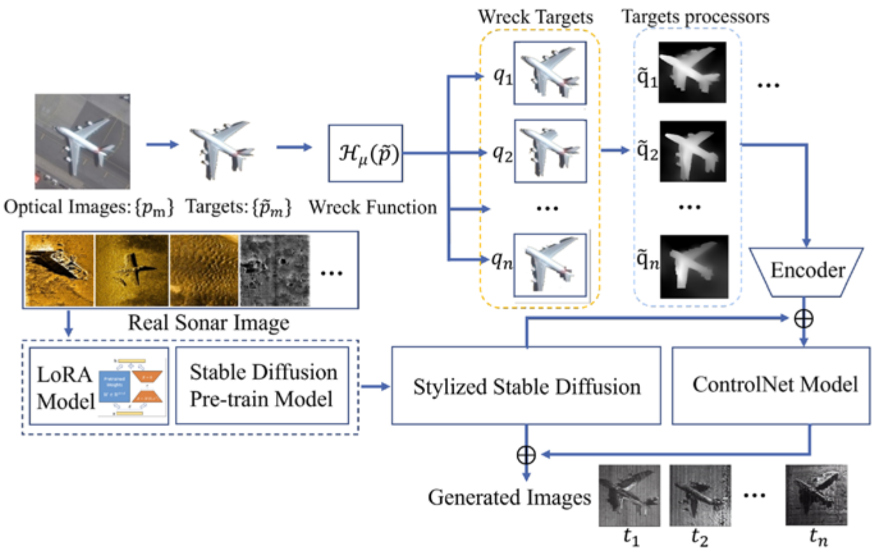


Fig. 2 Overall process of the proposed method

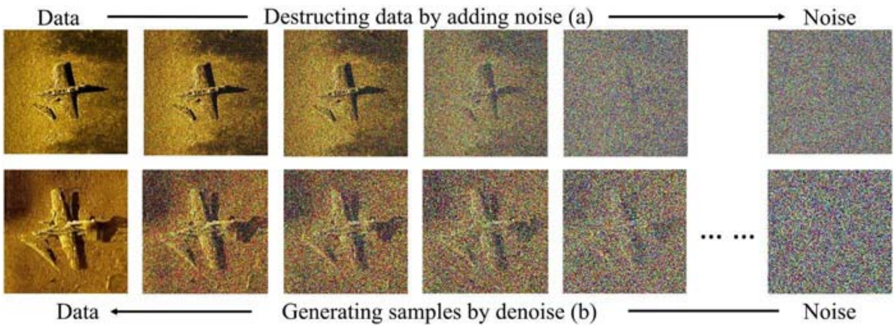


Fig. 3 Stable diffusion process of sampling via noises

input. The overall framework to generate sonar image of the proposed method is illustrated in Figure 2.

As described in Figure 2, our method using two types of models to generate sonar image. The first model is the stylized stable diffusion which using real sonar image as training dataset. In training phase, we using LoRA as the Parameter-Efficient Fine-Tuning method for rising training effective. The second is the ControlNet model to control the shapes and features. We propose a wreck function $\mathcal{H}_\mu(\widetilde{p})$ to control the degree of wreck for the optical images which used for control input. The stylized stable diffusion combined ControlNet can generate the sonar image with difference wreck and shapes. The general validation methods via FID, SSIM and LPIPS [18–20] to validate the generated images. The survey by sonar targets researchers (The survey is not present in the process).

3.3 Sonar image generated by LoRA and stable diffusion

In this paper, we applied stable diffusion model as the core method. Stable diffusion is the state-of-the-art field of generating images using deep learning methods. The method contains forward (Encoder) and reverse (Decoder) processes. The general process for our experiments is shown in Figure 3.

During the forward process, Gaussian noise is progressively added to the original image. Over T times, the steps of adding noise to the original image transformed the image into Gaussian noise with a standard normal distribution. The original data is assumed to satisfy the distribution $x_0 \sim q(x_0)$. As the number of steps increases, the noise also increases in image $x_1, x_2, \cdots x_t, \cdots, x_T$. Each noise on the moment of $step_t$ is only related to the previous status $step_{t-1}$. Therefore, the forward process can be treated as a Markov process [21]. $\alpha_t = 1 - \beta_t$ is the weight of noise intensity. β_t is ranged from 0.0001 to 0.002 in [14, 15]. We use the same parameters in our paper. The number of steps increases, and the noises become larger than that in the previous status. The status x_t can be presents as below:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} z_1 \tag{1}$$

The $z_1, z_2, \cdots, \sim \mathcal{N}(0, I)$ follows Gaussian distribution. Given x_{t-1}, x_t should follow a normal distribution. The resulting $q(x)$ can be presents as below:

$$\begin{cases} q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{\alpha_t} x_{t-1},\left(1-\alpha_t\right) \mathrm{I}\right), \\ q\left(x_{1: T} \mid x_0\right)=\prod_{t=1}^T q\left(x_t \mid x_{t-1}\right) \end{cases} \tag{2}$$

Essentially, the reverse denoising process is generating images step by step from a standard normal distribution through decoder. With a parameterized model p_θ (e.g. a neural network) to approximate $q\left(x_{t-1} \middle| x_t\right)$ which follows Gaussian distribution[16]. Choose p_θ to be Gaussian, and parameterize mean $\mu_\theta\left(x_t, t\right)$ and covariance matrix $\sum_\theta\left(x_t, t\right)$. The reverse process can be presents as below:

$$\begin{cases} p_\theta(X_{0:T}) = p(x_T)\prod_{t=1}^T p\left(x_{t-1} \mid x_t\right), \\ p\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \end{cases} \tag{3}$$

And the loss function as below [22]:

$$\mathcal{L}(\theta)=\mathbb{E}\left[\left\|\epsilon_\theta\left(x_t\right)-\epsilon\right\|^2\right] \tag{4}$$

ϵ_θ is a neural network designed to predict the diffusion noise ϵ. When using Stable Diffusion model to find optimized parameters, a large amount of fine-tuning and data processing is required, leading to increased computational costs and manual work. LoRA is widely used in conjunction with Stable Diffusion [22] for customized image generation. And the method has highly reduced the cost of computer cost and training time.

When we modify one of the parameters $W \in \mathbb{R}^{d \times d}$ in stable diffusion model, the LoRA model focus on the degree of change $\Delta W \in \mathbb{R}^{d \times d}$. $W+\Delta W$ as the changed parameters to presents the model. In the fine-tuning phase of LoRA, the model only needs to add one weight $\alpha \in[0,1]$ to generate the images. The parameter can be presented as $W+\alpha \Delta W$. For efficient parameters in the fine-tuning phase, representing the variation of parameters with less information ΔW, W can be split by two matrixes with low rank[17]. It can be presented as below:

$$\Delta W=B A, \quad A \in \mathbb{R}^{r \times d}, \quad B \in \mathbb{R}^{d \times r}, \quad r \ll d \tag{5}$$

In our practice, LoRA significantly reduces the number of optimizable parameters while the training memory consumption is reduced to 1/4 of the basic diffusion model, and the convergence speed during sonar image training is accelerated by approximately 10 times (from 48 h to 4 h approximately).

3.4 Sonar image generation based on ControlNet

Inspired by the work from Yang et al. [12] proposed the mask-based method to guided image generation, but the target diversity still has limitation when training sonar detection model in deep learning. Although the Stable Diffusion model can generate multiple sonar images from original sonar images, the lack of original sonar data limits the diversity of the generated sonar images. And some of control generate image method has been proposed which has significantly improved the field of control generative images [23–25]. To solve the issue, we applied fine-tuned ControlNet with wreck control on stable diffusion model to generate sonar images by using optical images as the condition. The general process is shown in Figure 4.

"$\mathcal{E}$" and "$\mathcal{D}$" denote the encoder and decoder of the Stable Diffusion. Based on the Eq.(4), for the conditional image generation noise prediction network as follows [26]:

$$\epsilon_\theta\left(x_t, t\right)=\mathcal{D}\left(\mathcal{E}\left(x_t\right), \mathcal{C}_\theta\left(x_t\right), \mathcal{F}_\theta(c)\right) \tag{6}$$

$\mathcal{C}_\theta$ denotes the ControlNet, $\mathcal{F}_\theta$ denotes the condition embedding network. In this paper, we not only use original sonar image target as generation model input, but also control the degree of wreck on optical image target as the dataset to makes the expended targets diversity. Our method using original image as input dataset $\{p_j,\ j \in m\}$. Using the existing segmentation method in python get the targets $\{\widetilde{p}_m,\ j \in m\}$. m which denotes the m^{th} image. We define the function of $\mathcal{H}_\mu(\widetilde{p}), \mu \in[0,1]$ to control the degree of wreck on target. μ denotes the degree of wreck target, e.g. (μ=0 denotes without wreck target, μ=1 denotes no target

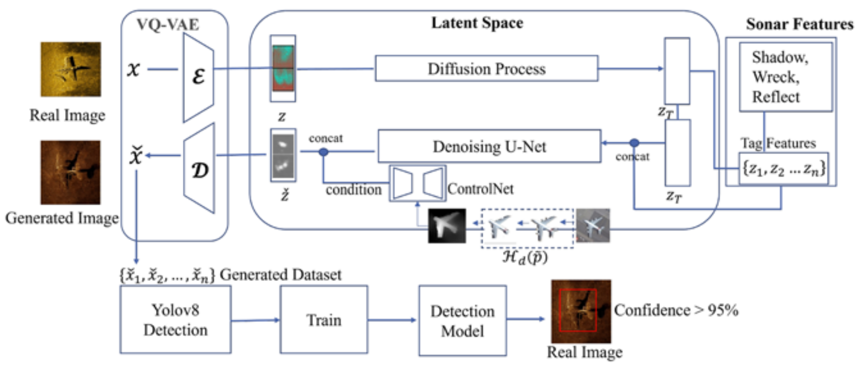


Fig. 4 General process of sonar image generation by stable diffusion and ControlNet

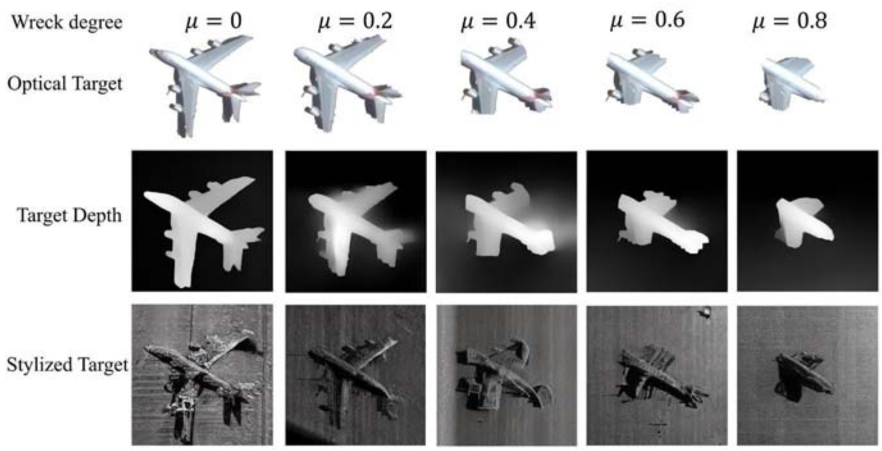


Fig. 5 Ablation study on airplane wreckage control with different values of μ, where larger μ produces more severe wreckage while preserving the object structure

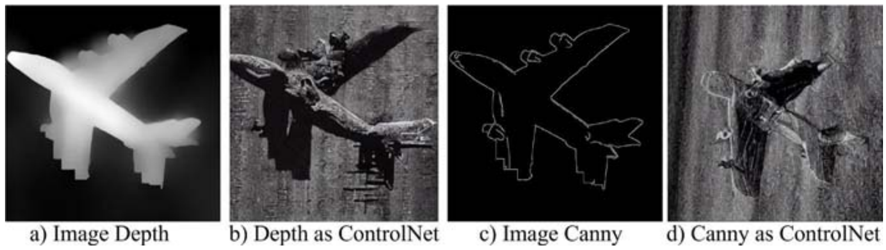


Fig. 6 Different target processor to generate images

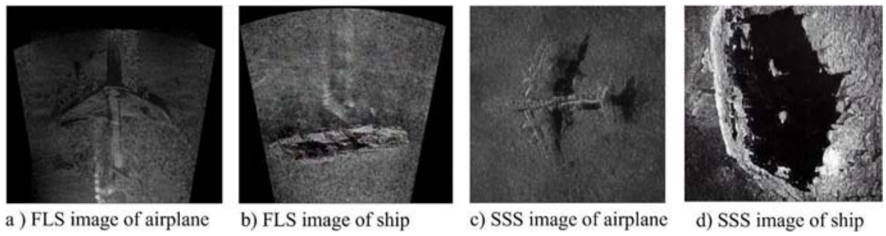


Fig. 7 Example of generated SSS and FLS target images

generated on image). Our proposed method which adapts the noise prediction network from Eq.(6) as follow:

$$\epsilon_\theta (x_t, t) = \mathcal{D}(\mathcal{E}(x_t), \mathcal{C}_\theta (x_t, \mathcal{H}_\mu(\widetilde{p}_j)), \mathcal{F}_\theta(c)) \quad (7)$$

The stylized stable diffusion model only acts on style of sonar image and the degree of wreck deepened on parameter of μ. The impact of parameter of μ on sonar feature of wreckage is shown in Figure 5 which presents an ablation study on the wreckage control parameter μ. By gradually increasing μ, the amount of visible wreckage on the generated sonar images increases monotonically while preserving the underlying object geometry.

Due to the stochastic nature of diffusion-based generation, shadow patterns are not explicitly constrained by ControlNet and may exhibit variability. Therefore, a semi-automatic quality control step is adopted, where physically implausible samples (e.g., inconsistent shadow directions or unrealistic cast shadows) are manually removed before being used for downstream training.

We input "airplane wreck" with different value of μ, there have different results. From our experience, the model is hard to present a target when $\mu \geq 0.8$ and without contribution on detection model. In the target processing stage, as illustrated in figure 2, the input images are optical images rather than sonar images. Various target processors can be applied to these optical images, including depth estimation, Canny edge detection, and other segmentation or line-based operators. Some examples of optical target processors are shown in figure 6. In our experiments, we primarily adopt depth-based processors for better visual clarity and readability, although we observe no significant difference in the final recognition performance when using other optical target processors.It should be emphasized that no target extraction is performed directly on sonar images.

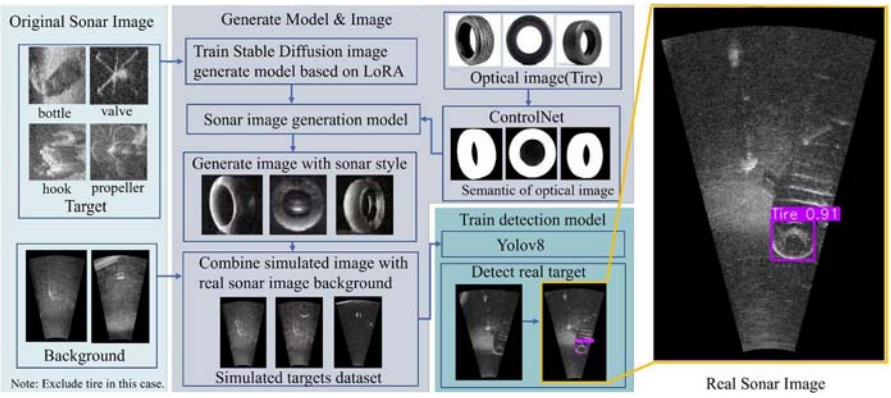


Fig. 8 Process of sonar target detection via simulate sonar style target

We select some of examples of generated on SSS and FLS target image as shown in Figure 7.

In the practice, the generated image supports resolution from (64x64) to (512x512) pixels based on Stable Diffusion version 1.5. We have selected some generated images as shown in Appendix Figure 14. We conducted experiments using Stable Diffusion versions 2.0, 3.0 and XL, which can generate images with a resolution of (2048x2048) pixels. However, since real sonar image recognition tasks do not require such high-resolution images and it is not beneficial for training detection models, we only use Stable Diffusion version 1.5 for the discussion in this paper. The detail compared data is shown in below session.

4 Experiments and results

To validate the effectiveness of the images generated by our method, we using generated images as input for training detection model. And we using real sonar dataset in our experiments [23], the experiments design process is shown in Figure 8.

The experiment consists of three phases, includes data preparation, model training and validation pipeline. In data preparation: collect and preprocess the raw sonar dataset by separating targets from backgrounds through image segmentation techniques. In model training: train a LoRA-adapted diffusion model using the isolated target data to develop a sonar image generator. Simultaneously, implement semantic segmentation on optical targets using ControlNet, then integrate this with the sonar generator through cross-modal fusion to synthesize sonar-style targets from optical inputs. In validation pipeline: synthesize diverse targets and combine them with background images to construct an augmented training dataset. Afterwards, train a recognition model on this hybrid dataset and evaluate its performance in identifying real sonar targets. In the end, we have survey from sonar researchers for the generated images as reference. The parts of generated images are shown in Appendix Figure 14, and the real sonar image data is shown in Appendix Figure 15.

Table 2 Comparison of sonar image generation methods in terms of image quality and efficiency

Method	FID↓	SSIM↑	LPIPS↓	Generated Time(s)
Improved style transfer [8]	245.2	0.12	0.53	0.12
GAN for sonar image [5]	180.5	0.21	0.35	2.4
Stable diffusion (v1.5)	165.6	0.25	0.31	2.1
Stable diffusion (v2.0)	163.2	0.23	0.30	2.9
Stable diffusion (v3.0)	163.5	0.21	0.29	4.4
Ours (fine-tuned diffusion)	**152.8**	**0.32**	**0.24**	**1.6**

4.1 Evaluation generates sonar image

Considering the sonar feature diversity and structural similarity, and combining evaluation indicators of image generation, we select Fréchet inception distance (FID) [18], learned perceptual image patch similarity (LPIPS) [19]and structural similarity (SSIM) [20] as the mainly evaluated methods. The FID is used to compute the distance between the real image set and the synthetic image set in the feature space. SSIM is used for the structural similarity. LPIPS calculates the perceptual similarity between two images. LPIPS essentially computes the similarity between the activations of two image patches for some pre-defined network. (Note: The lower the value of LPIPS, FID, the more similar the two images are. The SSIM value closer to 1 indicates greater similarity between images, while the value of 0 means no similarity in luminance, contrast, or structure.)

Table 3 Evaluation indicators on difference real targets of generate sonar image

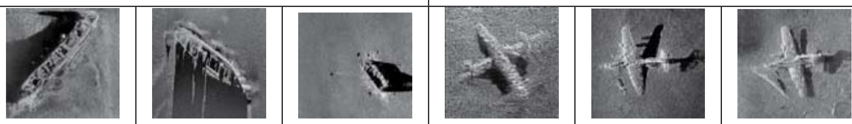
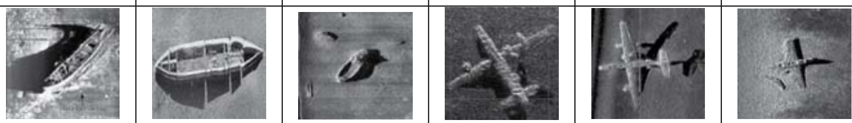
Class	Ship			Airplane		
Generated Image						
Real Sonar Image						
FID	15.6	18.3	20.2	14.4	15.2	16.2
SSIM	0.85	0.76	0.72	0.87	0.92	0.79
LPIPS	0.13	0.18	0.15	0.12	0.09	0.14

Fig. 9 Examples of colored real and generated sonar image

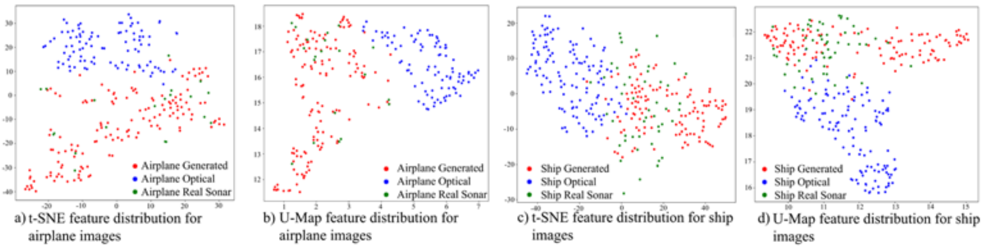


Fig. 10 Feature distribution of optical, real sonar, and generated images

Due to differing criteria for measuring image similarity, we adapt the common scores above to evaluate the results of our experiment on the sonar image. We compared the real sonar images (120 images download from public internet) with generated images using different methods, as shown in the table 2.

Our fine-tuned stable diffusion mode got higher scores in generated images. The average FID is 152.8, average SSIM is 0.32 and the LPIPS is 0.24. The evaluation metrics in table 2 were computed as an aggregate average across all target categories, comparing the entire set of generated data against real sonar datasets. Although the proposed diffusion-based generator does not operate in real time, it is designed as an offline data augmentation module rather than an online component of the detection system. Once the synthetic sonar images are generated, the detection or classification models operate on real sonar data in real time without any additional computational overhead. With an average generation time of 1.6s to generate per image, more than 2,000 training samples can be synthesized within one hour,

Table 4 Quantitative embedding-space analysis based on t-sne and umap for airplane and ship categories

Category	Embedding	Gen–Real↓	Gen–Optical↓	Dispersion (Gen / Real)
Airplane	t-SNE	**7.30**	32.93	19.57 / 19.53
Airplane	UMAP	**0.79**	3.54	2.07 / 1.79
Ship	t-SNE	**11.06**	46.23	16.51 / 14.63
Ship	UMAP	**0.95**	3.31	1.79 / 1.01

which is sufficient to support effective model training and domain adaptation. To refine this approach, we conducted a per-category performance analysis and selected the optimal values from the complete assessment results as table 3.

We also obtained an impressive score from selected generated image compared with the real Pseudo color sonar image shown in Figure 9.

We divided the images into two categories: ships and airplanes, and performed UMAP [27] and T-SNE [28] feature distribution on both real and generated images. Within the same category, we obtained very similar distributions. And the feature distribution is shown in Figure 10.

As shown in Fig. 10, the generated images exhibit a distribution closely aligned with real sonar images, while optical images demonstrate a more independent distribution pattern. And a quantitative analysis is conducted based on figure 10. as shown in table 4.

The generated sonar samples are consistently much closer to real sonar samples than to optical images for both airplane and ship categories, with Gen-Real centroid distances of 7.30/0.79 (airplane) and 11.06/0.95 (ship) under t-SNE/UMAP, compared to much larger Gen-Optical distances. Moreover, the intra-class dispersion of generated sonar data remains comparable to that of real sonar data across both embeddings, indicating that the observed domain alignment is robust to embedding choices and generalizes across target categories.

4.2 Detection and classify task with generated image

Although the proposed framework is applicable to both side-scan sonar (SSS) and forward-looking sonar (FLS) images, different downstream tasks are adopted for evaluation based on their typical imaging characteristics and practical usage. Specifically, SSS images usually capture large-area seabed scenes where targets are weakly localized and commonly annotated at the image or patch level, making classification a representative evaluation task. In contrast, FLS images focus on near-field perception with clear spatial target distributions, where object detection better reflects real-world operational requirements. Therefore, the use of different downstream tasks does not aim to compare task difficulty across sonar types, but to demonstrate the generality and practical relevance of the proposed framework.

4.2.1 Detection task on forward-looking sonar images

To further evaluate the impact of our generated image in the sonar detection task, we used the generated images as training data to train the detection model under the Yolov8 framework [29]. The real sonar data will be used as validation data to evaluate the result. And our implemented detection work based on real sonar dataset [23]. To replicate the real experimental scenario, we using simulated target instead of real target and excludes the real sonar target in training dataset when training detection model. In this dataset, we selected 500 images with total number of object classes being 5. The classes include, tire, bottle, hook, propeller, valve.

Table 5 Class-wise recall results on the fls detection task

Class	Simulated Training Data	Real Sample Size (Test)	Class Recall (%)
Tire	200	100	91
Bottle	200	100	89
Hook	200	100	93
Propeller	200	100	90
Valve	200	100	93

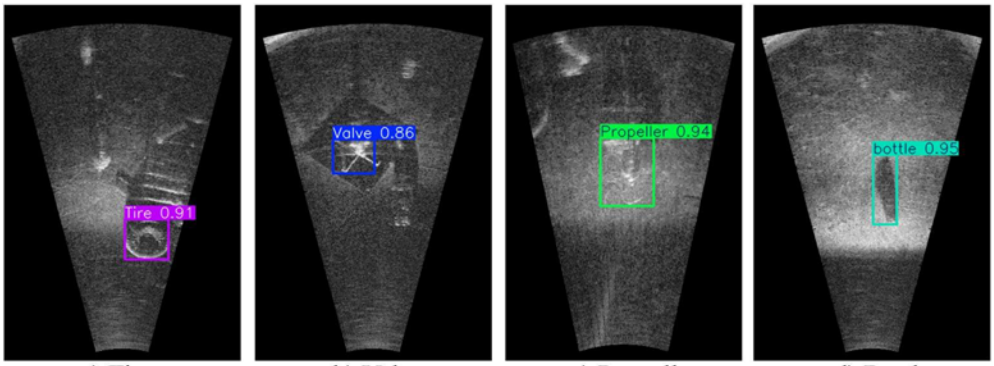
Table 6 Detection performance with different sonar image simulation method (FLS)

Method	Precision	Recall	mAP@0.5
Detection model based on basic stable diffusion model [12]	0.918	0.884	0.904
Improved style transfer [8]	0.892	0.861	0.876
Whitening and Coloring Transform [31]	0.871	0.842	0.855
GAN for sonar image generation [5]	0.905	0.872	0.890
Our fine-tuned diffusion-based method	**0.935**	**0.902**	**0.921**

$$\text{Class-wise Recall} = \frac{\text{Number of True Positives}}{\text{Number of Ground-Truth Objects}}$$

Table 5, class-wise recall is adopted to evaluate the detection performance of each object category, defined as the ratio between correctly detected objects and the total number of ground-truth objects of that class. And we applied the intersection over union ($\text{IoU} > 0.5$) metric [30] for each of class.

Based on the example shown in figure. 8, all real samples of one target category (e.g., tire) are excluded from the training process. Similarly, each category listed in Table 5 is treated as an unseen-object detection scenario, where the detection model is trained without real samples from that category and evaluated on real sonar images. This protocol evaluates the generalization ability of the proposed image generation method under unseen-object conditions. The result of different generation methods and detection is shown in table 6 which reports the detection performance of different sonar image generation methods under the same



a) Tire b) Valve c) Propeller d) Bottle

Fig. 11 Examples of detection task on real sonar image

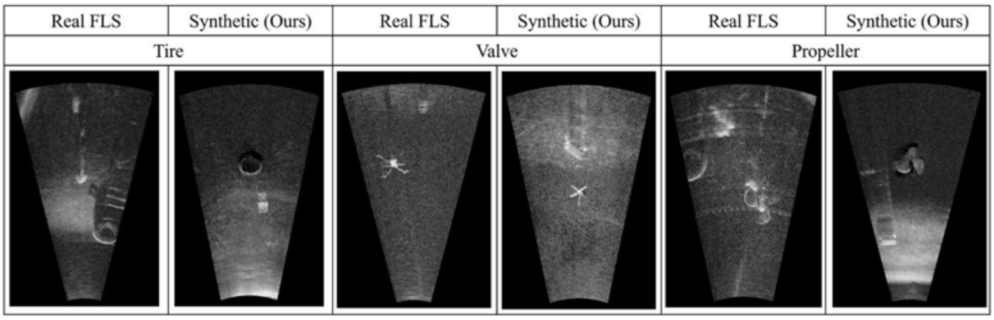


Fig. 12 Qualitative comparison between real FLS images and corresponding synthetic sonar samples for the same target categories. The comparison highlights similarities in target geometry and texture patterns, while also illustrating differences in acoustic highlights and shadow variability

Table 7 Classification accuracy trained on generated sonar images and tested on real SSS images

Class	Generated Images (Train)	Real Sonar Images (Test)	Class Accuracy
Ship	400	80	96.25%
Airplane	400	40	97.5%

detection architecture (YOLOv8) and evaluation protocol. Under identical detection settings, the proposed fine-tuned diffusion-based generation method achieves the highest mAP value of 0.921. Specifically, In our experiments, the training set consists of 100% synthetic sonar images, while all the test set consists of real FLS images. In the unseen-object setting, training set for the excluded category consists solely of synthetic sonar images generated by our methods. This design ensures that any detection performance on the excluded category is solely attributed to the effectiveness of the synthetic data generation strategy rather than real sample leakage.

The examples of detection task based on marine debris dataset [23] are shown in figure 11.

Figure 12 presents a comparison between real and synthetic FLS samples of the same target categories. It can be observed that the generated images exhibit a high degree of consistency with real sonar data in terms of texture patterns and shape distributions.

Since our ControlNet method is employed to constrain target geometry rather than explicitly modeling acoustic shadows or material-dependent scattering effects. A target-mask-based guidance strategy is adopted to highlight specific targets (e.g., metallic objects). Currently, our proposed ControlNet method mainly constrains target geometry, while acoustic highlights are implicitly learned by the diffusion model. Future work will incorporate explicit highlight-aware guidance to further improve physical consistency.

4.2.2 Classification task on side-scan sonar images

To further validate the generality of the proposed sonar image generation method, we conduct a classification task using side-scan sonar (SSS) images. Similar to the detection experiments, generated sonar images are used as the training data, while real sonar images are reserved exclusively for testing. As shown in figure 8, a classification model based on YOLOv8 is employed instead of a detection model. The training dataset consists of 400 generated ship images and 400 generated airplane images. The testing dataset includes 80 real ship images and 40 real airplane images collected from publicly available sources. Due to the limited number

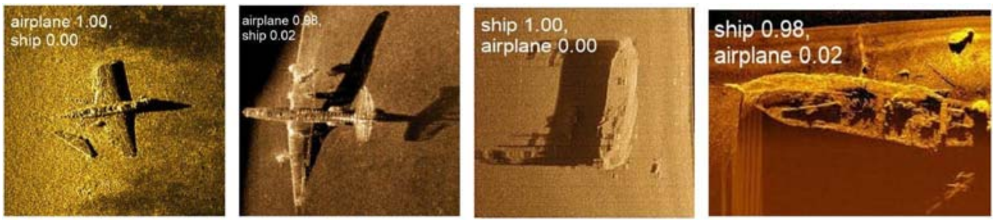


Fig. 13 Examples of classify task on real sonar image

Table 8 Survey of sonar image researchers

Question 1.	**Is the generated sonar image helpful for your field of research?**				
Feedback	Strong agree	Agree	Normal	Disagree	Strong disagree
	100%	0	0	0	0
Question 2.	**Is there a difference between the generated image and the real image?**				
Feedback	Very Similar	Similar	Normal	Different	Very different
	83%	17%	0	0	0
Question 3.	**Can you distinguish which images are generated and which are real?**				
Feedback	Very Hard	Hard	Normal	Easy	Very easy
	83%	17%	0	0	0
Question 4.	**Can the generated images be used for research purposes?**				
Feedback	Most of	Parts of	Normal	Few	No
	83%	17%	0	0	0

of real sonar samples, class-wise accuracy is adopted as the evaluation metric, defined as:

$$class\ of\ accuracy = \frac{correct\ number\ of\ classified\ images}{total\ sum\ of\ a\ class}$$

The experiment results are shown in table 7.

As shown in Table 7, 77 out of 80 real ship images and 39 out of 40 real airplane images are correctly classified, corresponding to class-wise accuracies of 96.25% and 97.5%, respectively. All reported results satisfy a confidence threshold of $\mathrm{confidence} > 0.95$. Representative classification examples are illustrated in Fig. 13. It should be noted that the classification task is designed as a target-focused evaluation. Only target categories are considered, while background is excluded in order to isolate the effect of cross-domain feature transfer. Although this two-class setting is relatively simple, it serves as a controlled diagnostic task rather than a full-scale sonar scene classification problem.

4.3 Survey on sonar image researchers

Although technical indicators can evaluate generated images in terms of scores, in sonar imagery, where the images are non-optically generated, the technical indicators have certain limitations without human perception. For this reason, we have surveyed 6 sonar image researchers for our generated images, and the survey results are shown in Table 8.

From this survey results, we can know that the generated sonar image by our method with positive feedback from researchers. The generated images have minor differences from real sonar images. (E.g., 83%, which means 83% select the answer.)

5 Conclusions

Few samples of sonar image limited the underwater object detection, recognition and segmentation tasks based on deep learning methods. In this paper, our proposed sonar image generating method via optical images and sonar images based on stable diffusion and LoRA model can generate multiple sonar style images and makes one object with more various forms. The generated sonar images possess both the diversity of optical images and the characteristics of sonar images. The fine-tuned LoRA method with few real sonar images, compared to existing methods, our approach allows for the controllability of the wreckage features on specific targets. At the same time, by applying ControlNet on optical images based on fine-tuned stable diffusion model, the sonar image target can be expanded with more diversification. From evaluation indicators and researchers' views, the generated sonar images through our method are closer to real images. Based on the public real sonar dataset, we use the generated image as the training dataset which can detect the real sonar images. Through the detail comparison experiment results show that our method can effectively improve the performance of the underwater object detection model.

Appendix A

See Figure 14 and 15

Fig. 14 Samples of generate image by our method. Row 1–4: Airplane; Row 5–8: Ship

Fig. 15 Samples of real sonar image from public resources

Acknowledgements We would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions that greatly improve the quality of this paper. Thanks to the researchers who have published datasets of sonar and optical images. https://www.edgetech.com/underwater-technology-gallery and https://sandysea.org/projects/visualisierung

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

Funding This work was supported by the National Natural Science Foundation of China under Grant No.42276187.

Data availability used data in experiments: DOI:10.1109/ICCVW54120.2021.00417 other samples: https://www.edgetech.com/underwater-technology-gallery/ and https://sandysea.org/projects/visualisierung All the experiments generated data by ourselves can be found in https://github.com/xijier/SonarDetection

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Elmezain, M., Saoud, L.S., Sultan, A., Heshmat, M., Seneviratne, L., Hussain, I.: Advancing underwater vision: a survey of deep learning models for underwater object recognition and tracking. IEEE Access (2025)

2. Jian, M., Yang, N., Tao, C., Zhi, H., Luo, H.: Underwater object detection and datasets: a survey. Intell. Marine Technol. Syst. **2**(1), 9 (2024)
3. Chen, L., Huang, Y., Dong, J., Xu, Q., Kwong, S., Lu, H., Lu, H., Li, C.: Underwater optical object detection in the era of artificial intelligence: Current, challenge, and future. ACM Comput. Surv. **58**(3), 1–34 (2025)
4. Fan, J., Wang, X., Zhou, C., Ou, Y., Jing, F., Hou, Z.: Development, calibration, and image processing of underwater structured light vision system: A survey. IEEE Trans. Instrum. Meas. **72**, 1–18 (2023)
5. Sung, M., Kim, J., Kim, J., Yu, S.-C.: Realistic sonar image simulation using generative adversarial network. IFAC-Papers OnLine **52**(21), 291–296 (2019)
6. Jiang, Y., Ku, B., Kim, W., Ko, H.: Side-scan sonar image synthesis based on generative adversarial network for images in multiple frequencies. IEEE Geosci. Remote Sens. Lett. **18**(9), 1505–1509 (2020)
7. Bai, Z., Xu, H., Ding, Q., Zhang, X.: Side-scan sonar image classification with zero-shot and style transfer. IEEE Trans. Instrum. Meas. (2024)
8. Xi, J., Ye, X.: Sonar image target detection based on simulated stain-like noise and shadow enhancement in optical images under zero-shot learning. J. Marine Sci. Eng. **12**(2), 352 (2024)
9. Sivachandra, K., Kumudham, R.: A review: Object detection and classification using side scan sonar images via deep learning techniques. Mod. Approaches in Mach. Learn. Cogn. Sci. : A Walkthrough **4**, 229–249 (2024)
10. Ge, Q., Ruan, F., Qiao, B., Zhang, Q., Zuo, X., Dang, L.: Side-scan sonar image classification based on style transfer and pre-trained convolutional neural networks. Electron. **10**(15), 1823 (2021)
11. Huang, C., Zhao, J., Yu, Y., Zhang, H.: Comprehensive sample augmentation by fully considering sss imaging mechanism and environment for shipwreck detection under zero real samples. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2021)
12. Yang, Z., Zhao, J., Yu, Y., Huang, C.: A sample augmentation method for side-scan sonar full-class images that can be used for detection and segmentation. IEEE Trans. Geosci. Remote Sens. (2024)
13. Yang, Z., Zhao, J., Zhang, H., Yu, Y., Huang, C.: A side-scan sonar image synthesis method based on a diffusion model. J. Marine Sci. Eng. **11**(6), 1103 (2023)
14. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. **88**, 102846 (2023)
15. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Adv. Neural. Inf. Process. Syst. **34**, 21696–21707 (2021)
16. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
18. Yu, Y., Zhang, W., Deng, Y.: Frechet inception distance (fid) for evaluating gans. China Univ. Min. Technol. Beijing Graduate Sch. **3** (2021)
19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit, pp. 586–595. (2018)
20. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: A new image similarity index. IEEE Trans. Image Process. **18**(11), 2385–2401 (2009)
21. Petrov, E., Kharina, N.: Markov processes in image processing. In: J. Phys: Conf. Ser., vol. 1015, p. 032059 (2018). IOP Publishing
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 10684–10695. (2022)
23. Singh, D., Valdenegro-Toro, M.: The marine debris dataset for forward-looking sonar semantic segmentation. In: Proceedings of the Ieee/cvf Int. Conf. Comput. Vis., pp. 3741–3749. (2021)
24. Xu, Y., He, Z., Shan, S., Chen, X.: Ctrlora: An extensible and efficient framework for controllable image generation. arXiv preprint arXiv:2410.09400 (2024)
25. Li, F., Wu, H., Zhang, J.: Lightweight diffusion model for synthesizing malicious network traffic. In: NAECON 2024-IEEE Nat. Aerosp. and Electron. Conf., pp. 409–413. (2024). IEEE
26. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
27. Healy, J., McInnes, L.: Uniform manifold approximation and projection. Nature Reviews Methods Primers **4**(1), 82 (2024)
28. Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
29. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics
30. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3213–3223. (2016)
31. Li, C., Ye, X., Cao, D., Hou, J., Yang, H.: Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples. Appl. Acoust. **173**, 107691 (2021)