remote sensing

MDPI

Article

Diffusion-Model-Based Data Augmentation for Target Detection in Side-Scan Sonar Images

Yuanxu Yang [1,2,3] and Tao Zhang [1,2,3,4,5,*]

[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China; 220243700@seu.edu.cn
[2] The Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology, Ministry of Education, Nanjing 210096, China
[3] State Key Laboratory of Comprehensive PNT Network and Equipment Technology, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[4] Southeast University Shenzhen Research Institute, Shenzhen 518063, China
[5] Advanced Ocean Institute of Southeast University, Nantong 226010, China
* Correspondence: zhangtao22@seu.edu.cn

Highlights

What are the main findings?

- FLUX+LoRA-generated side-scan sonar images improve YOLOv8n target detection under a fixed-detector and real-only validation/test protocol.
- The screened FLUX+LoRA subset achieves higher mAP@0.5 and mAP@0.5:0.95 than real-only training and traditional augmentation baselines.

What are the implications of the main findings?

- Diffusion-based generation provides an effective data-augmentation route for sonar target detection when annotated real samples are limited.
- The released synthetic dataset and code repository support reproducible evaluation of generative augmentation for side-scan sonar imagery.

Abstract

Side-scan sonar images play an important role in underwater target detection, seabed mapping, and marine environment monitoring. However, the performance of deep learning-based detectors is often limited by the small scale of available sonar datasets, the high cost of data acquisition, and class imbalance among target categories. To address these issues, this paper proposes a diffusion-model-based data augmentation method for side-scan sonar target detection. A FLUX.1 diffusion model is adopted as the base generative framework and is fine-tuned using low-rank adaptation (LoRA) to adapt the pretrained model to the side-scan sonar image domain under limited training data conditions. The generated samples are further filtered and added only to the training set, while the validation and test sets are kept unchanged and contain only real sonar images. To ensure a fair evaluation of the augmentation strategy, all detection experiments are conducted using a fixed YOLOv8n (You Only Look Once version 8 nano) detector under the same training hyperparameters and three random seeds. Compared with training on the original dataset, the proposed FLUX+LoRA augmentation improves mean average precision (mAP)@0.5 from 0.7400 $\pm$ 0.0132 to 0.8582 $\pm$ 0.0328 and mAP@0.5:0.95 from 0.3994 $\pm$ 0.0187 to 0.5115 $\pm$ 0.0164. It also outperforms conventional augmentation methods under the same real-only validation/test protocol. In addition, Fréchet Inception Distance (FID)/Kernel Inception Distance (KID)-based image quality evaluation, generated-sample amount ablation, screening-strategy ablation, LoRA-rank sensitivity analysis, and a controlled 600-sample

Check for updates

Received: 15 June 2026
Revised: 24 June 2026
Accepted: 25 June 2026
Published: 4 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

diffusion-backbone comparison are conducted. The results show that the 600-sample manually annotated FLUX+LoRA subset selected from generated samples achieves better image quality and detection performance than FLUX-base and SD1.5+LoRA under the same annotation budget. These findings demonstrate that FLUX+LoRA-generated sonar images can provide useful structural diversity for detector training and improve target detection performance under limited-data conditions.

Keywords: side-scan sonar; data augmentation; diffusion model; LoRA; underwater target detection

1. Introduction

The ocean contains abundant resources and has significant application value in marine engineering, seabed exploration, underwater environmental monitoring, and underwater search and rescue. With the continuous development of marine sensing technologies, how to acquire seabed environmental information efficiently and accurately has become an important research topic in the field of ocean information acquisition. Side-scan sonar, owing to its wide detection coverage, high imaging efficiency, and strong adaptability to complex underwater environments, has been widely used in tasks such as seabed topographic mapping, underwater target detection, subsea pipeline inspection, and shipwreck search [1–3]. The images acquired by side-scan sonar systems can reflect the acoustic differences between targets and the seabed background, and characterize the geometric and scattering properties of targets through highlights, shadow regions, and texture structures. Therefore, automatic target detection in side-scan sonar images is of considerable theoretical significance and practical engineering value.

However, compared with natural optical images, side-scan sonar images typically exhibit stronger noise, lower contrast, more complex texture distributions, and more pronounced imaging distortions [1,2,4]. These factors substantially increase the difficulty of target detection. In recent years, with the rapid development of deep learning, object detection algorithms represented by You Only Look Once (YOLO) and Faster Region-Based Convolutional Neural Network (Faster R-CNN) have achieved remarkable progress in the field of computer vision and have gradually been introduced into underwater target detection tasks [1,4,5]. Related studies have shown that deep learning methods can automatically extract target features through end-to-end training and achieve better detection performance than traditional handcrafted feature-based methods under complex background conditions. However, such methods usually rely on large-scale, high-quality annotated datasets. For side-scan sonar images, data acquisition depends on specialized equipment and offshore operations, resulting in high collection costs and long acquisition cycles. In addition, the annotation process requires substantial domain expertise. Consequently, the currently available public datasets are generally limited in scale, and their class distributions are often imbalanced [6,7]. Under these conditions, deep learning models are prone to overfitting, which in turn limits further improvements in detection performance.

To address the problem of limited data availability, data augmentation is currently one of the most commonly used solutions. Conventional data augmentation methods typically expand the training set through operations such as image flipping, rotation, scaling, cropping, and brightness adjustment. These methods are straightforward to implement and can, to some extent, increase the amount of training data and improve model robustness. However, they essentially remain limited to geometric transformations or pixel-level perturbations of existing samples and are unable to introduce new structural

information. As a result, they have clear limitations in enhancing data diversity [1,7]. Particularly in the context of side-scan sonar imagery, the scattering characteristics of targets, shadow structures, and background textures exhibit strong imaging regularities. Relying solely on geometric transformations makes it difficult to effectively supplement the missing sample patterns in the training data.

To further enhance data expansion capability, researchers have begun to explore the use of generative models for image synthesis. Generative adversarial networks (GANs) have been widely applied to image generation and data augmentation tasks and have achieved promising results in fields such as medical imaging, remote sensing, and underwater image enhancement [8,9]. However, GANs are prone to problems such as mode collapse, training instability, and insufficient diversity of generated samples during training, which are particularly pronounced in high-quality sonar image synthesis tasks. In contrast, diffusion models progressively add noise to samples and learn the reverse denoising process, enabling a more stable approximation of the real data distribution. In recent years, they have demonstrated superior generation quality and sample diversity in the field of image synthesis [10]. Existing studies have shown that diffusion model-based sonar image synthesis methods can generate side-scan sonar images with high realism and improve the performance of downstream target detection tasks [8,11].

Motivated by the above considerations, this study investigates the use of diffusion models for data augmentation in side-scan sonar target detection. FLUX.1 is adopted as the base generative model because of its strong text-conditioned image generation capability, while low-rank adaptation (LoRA) fine-tuning is used to adapt the pretrained model to the sonar image domain with limited training samples [12,13]. Instead of focusing on the comparison of detector architectures, this work evaluates whether diffusion-generated sonar samples can improve downstream detection performance under a fixed detector architecture and a strict real-only validation/test protocol. In the revised experimental design, generated and augmented samples are added only to the training set, while the validation and test sets are kept unchanged and contain only real side-scan sonar images.

The main contributions of this work can be summarized as follows.

(1) A FLUX+LoRA-based side-scan sonar image augmentation pipeline is proposed. The method adapts a pretrained FLUX.1 diffusion model to the sonar image domain and generates training samples with sonar-style highlight, shadow, and seabed texture characteristics.

(2) A strict fixed-detector evaluation protocol is constructed for assessing the effectiveness of generated data. All augmentation methods are evaluated using the same YOLOv8n detector, the same training hyperparameters, three random seeds, and unchanged real-only validation/test sets. This protocol avoids introducing generated samples into validation or test data and provides a more reliable evaluation of downstream detection utility.

(3) A comprehensive experimental analysis is conducted, including comparison with conventional augmentation, generated-sample screening statistics, Fréchet Inception Distance (FID)/Kernel Inception Distance (KID)-based image quality evaluation, generated-sample amount ablation, screening-strategy ablation, LoRA-rank sensitivity analysis, and a controlled 600-sample diffusion-backbone comparison. The results show that the proposed FLUX+LoRA augmentation improves target detection performance and that the 600-sample manually annotated FLUX+LoRA subset selected from generated samples outperforms FLUX-base and SD1.5+LoRA under the same annotation budget.

2. Materials and Methods

This section describes the methodological background, the proposed FLUX+LoRA-based augmentation pipeline, and the experimental protocol used for evaluating downstream target detection. Specifically, Section 2.1 summarizes the relevant background of sonar target detection, image augmentation, diffusion models, and parameter-efficient fine-tuning. Sections 2.2–2.5 then present the proposed generation and screening procedure, dataset organization, experimental settings, and evaluation metrics.

2.1. Methodological Background

2.1.1. Research on Underwater Sonar Target Detection

With the growing demands in marine engineering, seabed resource exploration, and underwater security monitoring, underwater target detection technology has gradually become an important research direction in the field of marine information engineering [3]. In complex underwater environments, conventional optical imaging devices often struggle to acquire clear and reliable image information because of severe light attenuation, high turbidity, and pronounced scattering effects. In contrast, sonar imaging technology exploits the propagation characteristics of acoustic waves in water and can operate stably under low-visibility conditions; it is therefore widely used in underwater detection tasks.

Common sonar imaging devices include multibeam sonar and side-scan sonar. Among them, side-scan sonar transmits acoustic waves to both sides of the water column and receives the returned echoes, thereby acquiring large-area information on seabed topography and underwater targets. Side-scan sonar images typically exhibit distinct highlight and shadow regions, where the highlight regions correspond to areas of strong acoustic reflection, while the shadow regions reflect the geometric characteristics of the targets. Consequently, target recognition often relies on the analysis of intensity distributions, texture features, and shadow structures.

Early sonar target detection methods mainly relied on handcrafted features. For example, edge information, texture descriptors, or shape features were extracted from images and then combined with traditional machine learning algorithms, such as support vector machines or random forests, to perform target classification. However, these methods depend heavily on feature engineering, and their detection performance is often limited when sonar images are affected by noise interference, seabed clutter, or complex backgrounds [14]. In recent years, with the rapid development of deep learning, convolutional neural networks have achieved remarkable progress in image recognition and object detection. Deep learning-based detection methods are able to automatically learn multi-level feature representations from images, thereby significantly improving detection accuracy. For example, detection frameworks such as Faster R-CNN, YOLO, and SSD have been applied to underwater sonar target detection tasks and have achieved promising results [14–16].

In deep learning-based detection frameworks, object detection can generally be formulated as follows:

$$y = f(x; \theta), \tag{1}$$

where x denotes the input image, θ represents the model parameters, and y denotes the predicted target categories and locations. Through training on large-scale datasets, the model can learn the mapping relationship between image features and target categories.

However, compared with natural image datasets, sonar image datasets are usually much smaller in scale. On the one hand, sonar equipment is expensive, and data acquisition requires dedicated marine survey operations. On the other hand, the data annotation process also requires substantial domain expertise. Therefore, limited data availability has become an important factor restricting the performance of deep learning models [8,14].

To address this issue, researchers have begun to explore the use of data augmentation techniques to expand the training samples and thereby improve the generalization ability of the model [8,11].

2.1.2. Image Data Augmentation Methods

Data augmentation is an important technique for improving the performance of deep learning models. Its basic idea is to increase the diversity of training data by transforming existing samples or generating new ones. Traditional data augmentation methods mainly rely on image geometric transformations or pixel-level transformations, such as rotation, flipping, scaling, cropping, and brightness adjustment [8].

Traditional augmentation operations can be expressed as follows:

$$x' = T(x), \tag{2}$$

where x is the original image, $T(\cdot)$ denotes the image transformation operation, and x' is the augmented image.

These methods are simple to implement and have low computational cost, and they can alleviate overfitting to a certain extent. However, they only transform existing samples and cannot generate new structural information; therefore, their effectiveness is limited when the dataset size is small [8].

To further enlarge the dataset, researchers have begun to employ generative models for data augmentation. Among them, generative adversarial networks (GANs) constitute an important class of image generation models. GANs learn the real data distribution through adversarial training between a generator and a discriminator, and their objective function can be written as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[logD(x)] + E_{z\sim p_z}[log(1-D(G(z)))], \tag{3}$$

where G denotes the generator, D denotes the discriminator, and z represents a random noise vector.

GANs have achieved remarkable success in the field of image generation and have been applied to areas such as medical imaging, remote sensing, and underwater image enhancement. However, GAN models often suffer from training instability and are prone to issues such as mode collapse, which to some extent limits their practical application [8]. In recent years, diffusion models have gradually emerged as an important research direction in the field of image generation [17].

2.1.3. Diffusion Model-Based Image Generation

Diffusion models are a class of generative models based on a progressive denoising process. Their core idea is to gradually add noise to the data and then learn the reverse denoising process to recover the original data distribution. Representative diffusion models include Denoising Diffusion Probabilistic Models (DDPMs) and related variants [17].

The generation process of a diffusion model mainly consists of two stages: a forward diffusion process and a reverse denoising process. In the forward diffusion process, Gaussian noise is gradually added to the image so that it progressively approaches a pure noise distribution. This process can be mathematically expressed as follows:

$$q(x_t|x_{t-1}) = N\left(\sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \tag{4}$$

where x_t denotes the noisy image at diffusion step t, and β_t is the noise schedule parameter.

The reverse process, by contrast, progressively reconstructs the original image by training a neural network to predict the added noise. Its training objective is typically formulated as follows:

$$L = E_{x,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t)\|^2, \tag{5}$$

where ϵ denotes the true noise, and ϵ_θ denotes the noise predicted by the model.

In recent years, diffusion models have achieved breakthrough progress in image generation. For example, models such as Stable Diffusion are capable of generating high-quality and diverse images and have demonstrated excellent performance in text-to-image generation tasks [10,18].

Compared with GAN-based models, diffusion models offer advantages such as more stable training, higher generation quality, and better sample diversity. As a result, they have gradually become one of the mainstream approaches in contemporary image generation research [10,17].

However, diffusion models usually contain a large number of parameters, and training them on domain-specific datasets requires substantial computational resources. Therefore, researchers have begun to investigate parameter-efficient fine-tuning techniques [19].

2.1.4. Parameter-Efficient Fine-Tuning Methods

As deep learning models continue to grow in scale, reducing training cost while maintaining model performance has become an important research issue. Parameter-efficient fine-tuning (PEFT) methods enable pretrained models to adapt to new tasks or data distributions by training only a small number of additional parameters [12].

Common parameter-efficient fine-tuning methods include Adapter, Prefix Tuning, and LoRA. Among them, LoRA (Low-Rank Adaptation) is a widely used fine-tuning approach. By performing a low-rank decomposition of the weight matrices, LoRA trains only a small number of low-rank matrices while keeping the original model parameters fixed, thereby significantly reducing the number of trainable parameters [12].

The basic idea of LoRA can be expressed as

$$W' = W + BA, \tag{6}$$

where W is the original weight matrix, and A and B are low-rank matrices. In this way, only the two matrices A and B need to be updated during fine-tuning, while the original weight parameters remain unchanged [12].

Owing to its advantages of requiring only a small number of trainable parameters, offering high training efficiency, and being easy to deploy, LoRA has been widely applied to the fine-tuning of large-scale models, including both language models and diffusion models [19].

In summary, underwater sonar target detection is confronted with the problem of limited data availability, while diffusion models provide an effective approach for image generation. At the same time, parameter-efficient fine-tuning methods can substantially reduce model training costs, thereby creating new opportunities for applying diffusion models to domain-specific tasks. Against this background, this paper proposes a sonar image generation method that combines diffusion models with parameter-efficient fine-tuning techniques, and uses the generated data to expand the training set, thereby improving underwater target detection performance.

2.2. FLUX+LoRA-Based Data Augmentation Method

2.2.1. Overall Framework of the Proposed Method

To address the problems of high acquisition cost, annotation difficulty, and imbalanced sample distribution across target categories in side-scan sonar imagery, this paper proposes a side-scan sonar image generation method based on the FLUX.1 diffusion model and LoRA fine-tuning. The method is designed to construct new samples with realistic sonar imaging characteristics, thereby supporting subsequent data augmentation. The overall workflow of the proposed method is shown in Figure 1 and can be divided into two parts: the training stage and the inference stage.

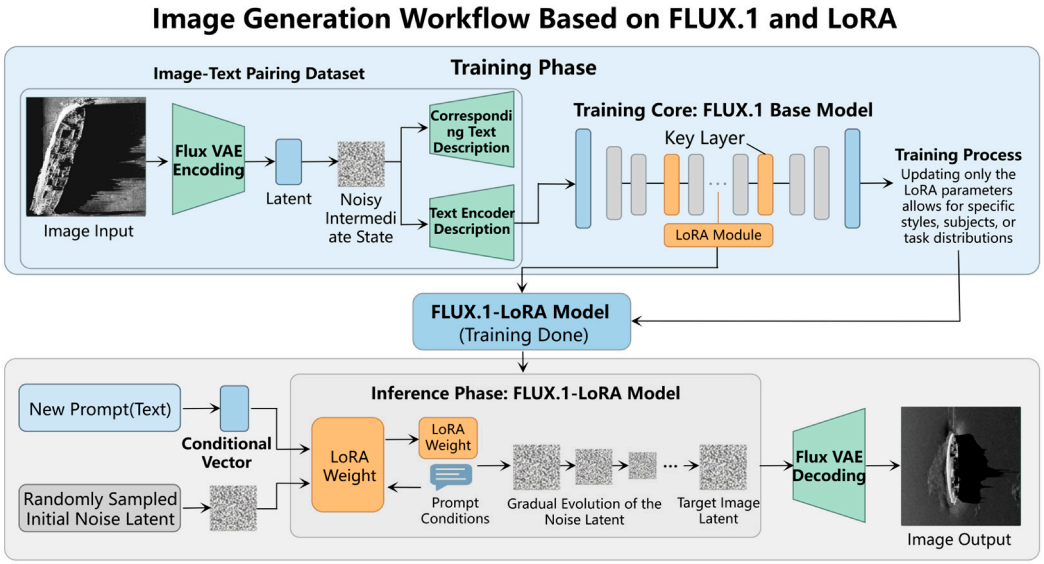


Figure 1. Overall framework of the proposed FLUX+LoRA-based side-scan sonar image generation and augmentation pipeline. LoRA denotes low-rank adaptation. The training stage constructs image–text pairs from real side-scan sonar samples and fine-tunes the FLUX.1 model using LoRA modules. The inference stage generates new sonar-style target samples under text-prompt guidance, which are subsequently processed by label auditing and quality screening before being added only to the training set for downstream detector training.

During the training stage, an image–text paired dataset is first constructed using side-scan sonar images and their corresponding textual descriptions. The input images are mapped into the latent space through the Flux VAE encoder to obtain the corresponding latent representations. Noise is then sampled in the latent space to construct intermediate noisy states. Meanwhile, the text descriptions corresponding to the images are fed into a text encoder to obtain conditional vectors that characterize the image semantics and imaging features. Subsequently, the noisy latent variables and the text conditions are jointly input into the FLUX.1 base model. With the parameters of the base model frozen, only the LoRA modules inserted into key layers are trained, enabling the model to gradually adapt to the highlight regions, shadow regions, background textures, and overall grayscale distribution characteristics of side-scan sonar images. After training, a FLUX.1-LoRA image generation model tailored to the side-scan sonar scenario is obtained.

During the inference stage, new text prompts and randomly sampled initial noisy latent variables are provided as inputs. Under the joint constraints of the LoRA weights and the text conditions, the model progressively evolves and denoises the noisy latent variables, ultimately producing latent target image representations that are consistent with the target semantics and exhibit side-scan sonar style characteristics. These latent representations are then decoded by the Flux VAE decoder to recover the final output

images. Through this process, the model can generate new side-scan sonar image samples, thereby supplementing the original dataset with a wider variety of target structural patterns and background combinations.

Unlike conventional data augmentation methods, the proposed approach does not simply perform rotations, scaling, or grayscale perturbations on existing images. Instead, it generates new target samples by learning the latent distribution of side-scan sonar images while preserving the realistic imaging style. Therefore, this method can increase the structural diversity of the training data to a certain extent and provide a richer sample basis for subsequent target detection tasks.

2.2.2. Diffusion Model Based on FLUX

Diffusion models are a class of generative models that have been widely used in image generation tasks in recent years. Their core idea is to gradually add noise to the original samples and learn the reverse denoising process, so as to recover new samples that conform to the real data distribution from random noise [17,20]. Compared with traditional generative models such as generative adversarial networks, diffusion models offer advantages including more stable training, higher generation quality, and better sample diversity, and therefore exhibit strong performance in image synthesis tasks [21].

For an input image xt0, the diffusion model first progressively adds noise to it through the forward diffusion process, yielding a noisy sample xt at the t time step. This process can generally be expressed as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \tag{7}$$

where β_t denotes the noise schedule parameter at step t, and I is the identity matrix. As the number of diffusion steps increases, the noise in the sample gradually intensifies, and the sample eventually approaches a standard Gaussian distribution.

For convenience in training, the above process is often rewritten as follows:

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon, \ \ \epsilon \sim \mathcal{N}(0,1), \tag{8}$$

where $\overline{\alpha_t}$ is the cumulative noise coefficient, and ϵ denotes Gaussian noise. The objective of model training is to learn a parameterized function $\epsilon_\theta(x_t, t, c)$, which predicts the noise term at a given time step t under the condition information c, thereby approximating the reverse denoising process. The corresponding loss function can typically be written as follows:

$$L = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2], \tag{9}$$

where c denotes the conditional information, which in this paper mainly corresponds to the conditional vector obtained by encoding the text descriptions.

In the reverse diffusion process, the model gradually reconstructs the original image from the noisy image by learning the noise prediction function. This process can be expressed as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)\right), \tag{10}$$

where $\epsilon_\theta(x_t, t)$ is the noise predicted by the neural network.

In this study, FLUX.1 is adopted as the base generative model. FLUX.1 is a conditional image generation framework based on latent-space modeling. Its overall generation process does not perform diffusion directly in the pixel space; instead, the image is first mapped into a low-dimensional latent space through a variational autoencoder (VAE), and noise modeling as well as progressive generation are then carried out in the latent

space [18,22]. Compared with performing operations directly in the original image space, latent-space generation can significantly reduce computational complexity, improve training and inference efficiency, and at the same time preserve the main structural information of the image [23]. For side-scan sonar images, target recognition depends not only on local intensity variations but also heavily on the spatial relationships among the target highlight regions, shadow regions, and background textures. For example, ships, aircraft wreckage, and human targets in sonar images are all typically characterized by strong highlight scattering regions followed by shadows, yet different target categories exhibit clear differences in length, shape, shadow extent, and coupling patterns with the background. Therefore, a generative model must learn not only the grayscale distribution but also the structural relationship among the target, shadow, and background. FLUX.1 possesses strong conditional modeling capability and latent-space structural representation ability, making it well suited to such image generation tasks with pronounced structural characteristics.

During the training stage, side-scan sonar images are first mapped into the latent space through the Flux VAE encoder to obtain latent image representations. These latent representations compress the information in the original images while preserving, as much as possible, the main target contours, light–dark relationships, and background texture features. Noise is then added to the latent representations, which are subsequently combined with the conditional vectors output by the text encoder and fed into the FLUX.1 model for training. By learning noise prediction, the model realizes the mapping from "noisy latent representation + text condition" to "denoised latent representation", thereby gradually capturing the distribution characteristics of side-scan sonar images.

During the inference stage, the model starts from randomly sampled initial noisy latent variables and progressively performs the reverse denoising process under the guidance of newly provided text prompts, ultimately obtaining target latent representations, which are then restored into images by the Flux VAE decoder. The generated images are semantically controlled by the prompts, while their style is constrained by the sonar-domain characteristics introduced through LoRA fine-tuning. As a result, the model can produce new samples that simultaneously contain target semantic information and side-scan sonar style characteristics.

2.2.3. LoRA Fine-Tuning Strategy

Although FLUX.1 has strong image generation capability, its parameter scale is relatively large. If full-parameter fine-tuning is performed directly on a small-scale domain dataset such as side-scan sonar imagery, it not only requires substantial computational and storage resources, but also easily leads to overfitting due to the limited number of training samples, thereby reducing the stability of the generated results. To address these issues, this paper adopts the LoRA (Low-Rank Adaptation) method to perform parameter-efficient fine-tuning of FLUX.1.

The basic idea of LoRA is to introduce a low-rank incremental term on top of the original weight matrix, rather than directly updating all the parameters of the original model [12,24]. Let the original weight matrix of a linear mapping layer be W ∈ Rm × n. Its fine-tuned weight can then be expressed as

$$W' = W + \Delta W = W + BA, \tag{11}$$

where $A \in \mathrm{R}^{r \times n}$, $B \in \mathrm{R}^{m \times r}$, and r is a low-rank dimension much smaller than m and n. During training, the original weight matrix W is kept fixed, and only the low-rank matrices A and B are updated, so that model adaptation can be achieved with a relatively small number of trainable parameters.

Compared with full-parameter fine-tuning, LoRA offers several advantages. First, it requires training only a small number of additional parameters, which significantly reduces the GPU memory usage and computational resources needed for training, making it suitable for experiments under limited hardware conditions. Second, LoRA preserves the general knowledge already contained in the large-scale pretrained model and avoids excessively disrupting the original model distribution under small-sample training conditions, thereby improving training stability. Third, LoRA modules are highly pluggable, and the resulting weight files are relatively compact, which facilitates rapid switching and comparison across different tasks or data conditions.

In this study, the LoRA modules are mainly inserted into the key linear transformation layers of the FLUX.1 model, enabling it to gradually learn the highlight structures, shadow distributions, and background texture patterns specific to side-scan sonar images while retaining the original image generation capability of the base model [12]. It should be noted that the purpose of employing LoRA fine-tuning in this study is not to enable the model to learn image generation from scratch, but rather to achieve lightweight adaptation to the sonar image domain style on the basis of the pretrained model's original generation capability [25]. Therefore, LoRA is particularly well suited to the present side-scan sonar image generation task, where the sample size is limited but the requirement for domain-specific style adaptation is clear.

From the perspective of the training pipeline, the image latent variables, noisy states, and text conditions are jointly fed into the FLUX.1 base model. During noise prediction, the model mainly relies on the LoRA modules to capture the domain differences between side-scan sonar images and natural images. After training, the LoRA weights encode the model's adaptation to the sonar image style. During inference, it is only necessary to load the trained LoRA weights into the base FLUX.1 model, after which images with sonar imaging characteristics can be generated under new text prompts.

2.2.4. Construction of Image–Text Paired Data and Prompt Design

Diffusion models are conditional generative models, and their generation performance depends not only on the domain adaptation capability of the model itself but also closely on the quality of the image–text paired data used during training. To enable the model to learn both the target semantics and the imaging characteristics of side-scan sonar images, this study constructs an image–text paired dataset composed of side-scan sonar images and their corresponding textual descriptions, on the basis of which LoRA fine-tuning is performed.

During dataset construction, real side-scan sonar image samples are first collected, followed by sample screening and preprocessing. Samples with excessively low quality, blurred targets, or severely missing structural information are removed. The retained images mainly cover typical target categories such as ships, aircraft wreckage, and human bodies, while also attempting to account for variations in target scale, orientation, and background conditions. Unlike ordinary natural images, side-scan sonar images are typically grayscale, and target structures often depend on the combined relationship between strong highlight echo regions and the shadow regions behind them. Therefore, during data selection, particular attention is paid to whether the target and shadow structures are clear and whether the background textures are representative.

With regard to textual description construction, this study does not use only simple category labels. Instead, by considering the imaging characteristics of side-scan sonar images, a text description containing both category semantics and imaging features is designed for each image. Specifically, the text descriptions usually include the target category, the sonar image type, and necessary structural cues, such as "side-scan sonar

image of a shipwreck with bright highlight and dark shadow" or "side-scan sonar image of a human-shaped target on seafloor background". In this way, the text condition conveys not only what the target is, but also, to some extent, guides the model to learn the expected light-dark relationships and spatial structures of the target in sonar imagery.

To further strengthen the model's ability to recognize the domain style, a unified trigger word is introduced into the text descriptions to explicitly indicate that the image belongs to the side-scan sonar style domain. For example, a style cue such as "sonar image" can be included in the description, allowing the model to establish a mapping between this phrase and the side-scan sonar imaging style during fine-tuning. After training, this trigger word can also be used during inference to improve the consistency of the generated results in terms of sonar style.

In addition to the text descriptions used during training, prompt design in the inference stage is equally important. Inference prompts determine not only the target category of the generated image, but also influence the target size, pose, and background representation in the generated image. For the task considered in this study, a suitable prompt should simultaneously contain two types of information: one is target semantic information, such as ship, plane, and human; the other is sonar imaging information, such as side-scan sonar, highlight, shadow, and seafloor texture. If only the target name is provided, the model is more likely to rely on its original pretrained semantic knowledge while neglecting the sonar style constraints. By contrast, when sonar imaging descriptions are included, the model is more likely to generate images consistent with the visual characteristics of side-scan sonar imagery.

Therefore, this study adopts a prompt design strategy based on category semantics + imaging structure + style trigger words to enhance both the controllability of target generation and the consistency of sonar style in the generated samples. This design is beneficial not only for establishing a more stable image–text alignment during training, but also for providing a foundation for generating richer samples under different prompting conditions in subsequent stages.

2.2.5. Sonar Image Generation and Sample Screening

After training the FLUX.1-LoRA model, the trained generative model is further used to generate new side-scan sonar image samples in order to expand the original dataset. The generation process starts from randomly initialized noisy latent variables. Under the joint action of the input prompts and LoRA weights, the model progressively removes noise and evolves the latent structure in the latent space, ultimately obtaining the latent representation of the target image, which is then restored into a side-scan sonar image through the Flux VAE decoder.

From the perspective of the generation mechanism, the inference stage can be divided into four steps. First, the corresponding text conditional vector is obtained according to the predefined prompt. Second, an initial noisy latent variable is randomly sampled from a Gaussian distribution as the starting point for generation. Third, under the joint constraints of the text condition and the LoRA weights, the model iteratively denoises the noisy latent variable over multiple steps, so that the latent variable gradually evolves from a random noise distribution into a latent representation with target structure and sonar style. Finally, the resulting target latent representation is fed into the Flux VAE decoder to recover the final image.

To improve the diversity of the generated samples, this study appropriately varies the random seeds and prompt formulations during generation, enabling the model to produce differences in target scale, orientation, shadow length, and background texture. Since targets in side-scan sonar images are typically represented by the combination of

highlight regions and shadow regions, special attention must be paid during generation to the correspondence between the target body and its shadow region. If the target body is relatively clear but the shadow structure is missing, or if the shadow position is obviously unreasonable, such samples are usually of limited value for subsequent tasks.

Considering the stochastic nature of diffusion-based generation, not all generated samples are directly suitable for downstream detector training. Therefore, after image generation, a two-stage filtering procedure consisting of label auditing and quality screening is adopted. First, each generated image is checked for the existence and validity of its corresponding YOLO-format label file. Samples with missing label files, empty labels, invalid class indices, or invalid bounding-box coordinates are removed. Second, the remaining candidates are screened according to their visual usability for side-scan sonar target detection. The screening mainly considers whether the generated sample contains a clear target structure, whether the target-highlight and acoustic-shadow patterns are visually plausible, whether the seabed background is consistent with the sonar imaging style, and whether the target can be reliably annotated with a bounding box.

In the final generation process, 2317 raw images were produced by the FLUX+LoRA model. Among them, 2302 images had corresponding label files, 15 images had missing labels, and 223 images contained empty labels. After label auditing, 2079 valid candidates remained. The final screened subset contained 1780 generated images, corresponding to an acceptance rate of 76.82% relative to all raw generated images and 85.62% relative to the valid candidates. The selected generated samples contained 635 body objects, 309 plane objects, and 836 ship objects. These samples were added only to the training set in the downstream detection experiments, while the validation and test sets were kept unchanged and contained only real side-scan sonar images.

This screening strategy helps reduce the influence of low-quality generated samples and ensures that the augmented training set contains sonar-style samples with usable target structures. The real-image split and augmentation composition used for evaluation are described in Section 2.3, and the effect of the screened generated samples on downstream detection performance is evaluated in Section 3. The screening statistics of the FLUX+LoRA-generated samples are summarized in Table 1.

Table 1. Screening statistics of FLUX+LoRA-generated side-scan sonar samples.

Item	Number
Raw generated images	2317
Images with label files	2302
Missing label files	15
Empty-label images	223
Valid candidates after label audit	2079
Final screened samples	1780
Acceptance rate from raw generation	76.82%
Acceptance rate from valid candidates	85.62%
Selected body objects	635
Selected plane objects	309
Selected ship objects	836

2.3. Dataset

The side-scan sonar image dataset used in this study contains three target categories: body, plane, and ship. All images were annotated in YOLO format using bounding boxes and category labels. To evaluate the effect of different augmentation strategies in a fair and leakage-free manner, the dataset was reorganized into a canonical real-image split. Specifically, 1340 real images were used for training, 383 real images were used for

validation, and 192 real images were used for testing. The validation and test sets were kept unchanged for all experiments and contained only real side-scan sonar images.

Different augmentation strategies were applied only to the training set. For the original setting, only the 1340 real training images were used. For conventional augmentation, two training sets were constructed: a target-region augmentation setting and a full-image augmentation setting. The target-region augmentation setting added 1832 augmented samples to the training set, while the full-image augmentation setting added 2680 augmented samples. For the proposed diffusion-based augmentation, FLUX+LoRA-generated samples were filtered through label auditing and quality screening, and 1780 screened generated samples were added to the training set. In all settings, no augmented or generated samples were added to the validation or test sets.

This protocol differs from a random split after augmentation. By adding augmented or generated samples only to the training set and keeping the validation/test sets real-only and unchanged, the evaluation more directly reflects whether the generated samples improve detector generalization to real sonar images. The detailed dataset composition is shown in Table 2.

Table 2. Dataset split and augmentation composition under the real-only validation/test protocol.

Method	Real Train Images	Generated/Augmented Train Images	Validation Images	Test Images	Total Train Images	Note
Original	1340	0	383	192	1340	Real images only
Traditional-target-only	1340	1832	383	192	3172	Target-region augmentation only
Traditional-all	1340	2680	383	192	4020	Full-image traditional augmentation
FLUX+LoRA auto-screened	1340	1780	383	192	3120	Generated samples added only to train

Validation and test sets are identical across all settings and contain only real side-scan sonar images.

2.4. Experimental Settings

To evaluate the effect of different augmentation strategies on side-scan sonar target detection, all detection experiments were conducted using a fixed YOLOv8n detector [26]. The purpose of this design is to isolate the influence of the training data source and avoid confounding the evaluation with differences among detector architectures. The compared training settings include the original real-image training set, two conventional augmentation settings, and the proposed FLUX+LoRA-generated augmentation setting.

For all experiments, the validation and test sets were kept unchanged and contained only real side-scan sonar images. Augmented or generated samples were added only to the training set. This setting ensures that the reported performance reflects the generalization ability of the detector on real sonar images rather than on augmented or synthetic validation/test samples.

The input image size was set to 640 $\times$ 640, the number of training epochs was set to 100, and the batch size was set to 24. The SGD optimizer was used with an initial learning rate of 0.01. To reduce the influence of random initialization and training stochasticity, each experiment was repeated with three random seeds, namely 42, 43, and 44, and the results are reported as mean $\pm$ standard deviation. Built-in random augmentation operations in the detector training pipeline were disabled so that the comparison mainly reflected the effect of the explicitly constructed augmentation datasets. The experiments were implemented using Python 3.10, PyTorch 2.5.1 with CUDA 12.4, and Ultralytics YOLOv8.

In addition to the main augmentation comparison, generated-sample amount ablation, screening-strategy ablation, LoRA-rank sensitivity analysis, and a controlled diffusion-backbone comparison with FID/KID-based image quality evaluation were also conducted under the same fixed-detector evaluation protocol.

2.5. Evaluation Metrics

To objectively evaluate the influence of different data augmentation methods on side-scan sonar target detection performance, commonly used evaluation metrics for object detection tasks were adopted in this study, including Precision, Recall, mAP@0.5, and mAP@0.5:0.95.

Among them, Precision is used to measure the proportion of predicted positive bounding boxes that are truly positive, while Recall indicates the proportion of real targets that are successfully detected by the model. mAP@0.5 represents the mean average precision calculated under an intersection over union (IoU) threshold of 0.5, whereas mAP@0.5:0.95 denotes the mean average precision calculated over IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, which more comprehensively reflects the overall performance of the model under different localization accuracy requirements.

Compared with any single metric, the joint use of Precision, Recall, and mAP under different thresholds can provide a more comprehensive evaluation of the detection model in terms of classification accuracy, recall capability, and bounding box localization accuracy. Therefore, the comparative analysis of different augmentation strategies in this study is mainly based on the above metrics.

2.6. Use of Generative AI Tools

Generative AI tools were used during manuscript preparation for language polishing, grammar checking, and improving the clarity and organization of the text. These tools were not used to determine the study design, collect data, perform experiments, analyze data, interpret results, or draw scientific conclusions. All AI-assisted text was carefully reviewed, edited, and verified by the authors, who take full responsibility for the content of this manuscript.

In addition, the use of FLUX+LoRA as a generative model for synthetic side-scan sonar image generation is part of the research methodology and is described in detail in Sections 2.2.1–2.2.5.

3. Results

3.1. Qualitative Analysis of FLUX+LoRA-Generated Sonar Samples

Following the generation and screening procedure described in Section 2.2, the generated samples were first qualitatively examined to verify whether they preserved the main visual characteristics of side-scan sonar imagery. Representative examples were selected to compare real sonar images, conventionally augmented images, and FLUX+LoRA-generated images in terms of target structure, acoustic-shadow extent, and seabed texture.

Before conducting downstream detection experiments, the generated samples were qualitatively examined to verify whether they preserved the main visual characteristics of side-scan sonar imagery. In this study, the FLUX.1 model was fine-tuned with LoRA using representative side-scan sonar samples and text prompts containing both target semantics and sonar imaging cues. During inference, prompts were designed to describe the target category and the expected sonar-style appearance, including grayscale acoustic texture, bright highlight regions, dark acoustic shadows, and seabed background patterns.

Figure 2 shows representative visual examples of real sonar images, conventionally augmented images, and FLUX+LoRA-generated images for the three target categories. These examples are used to qualitatively illustrate the differences in target structure, acoustic-shadow extent, and seabed texture among different sample sources.

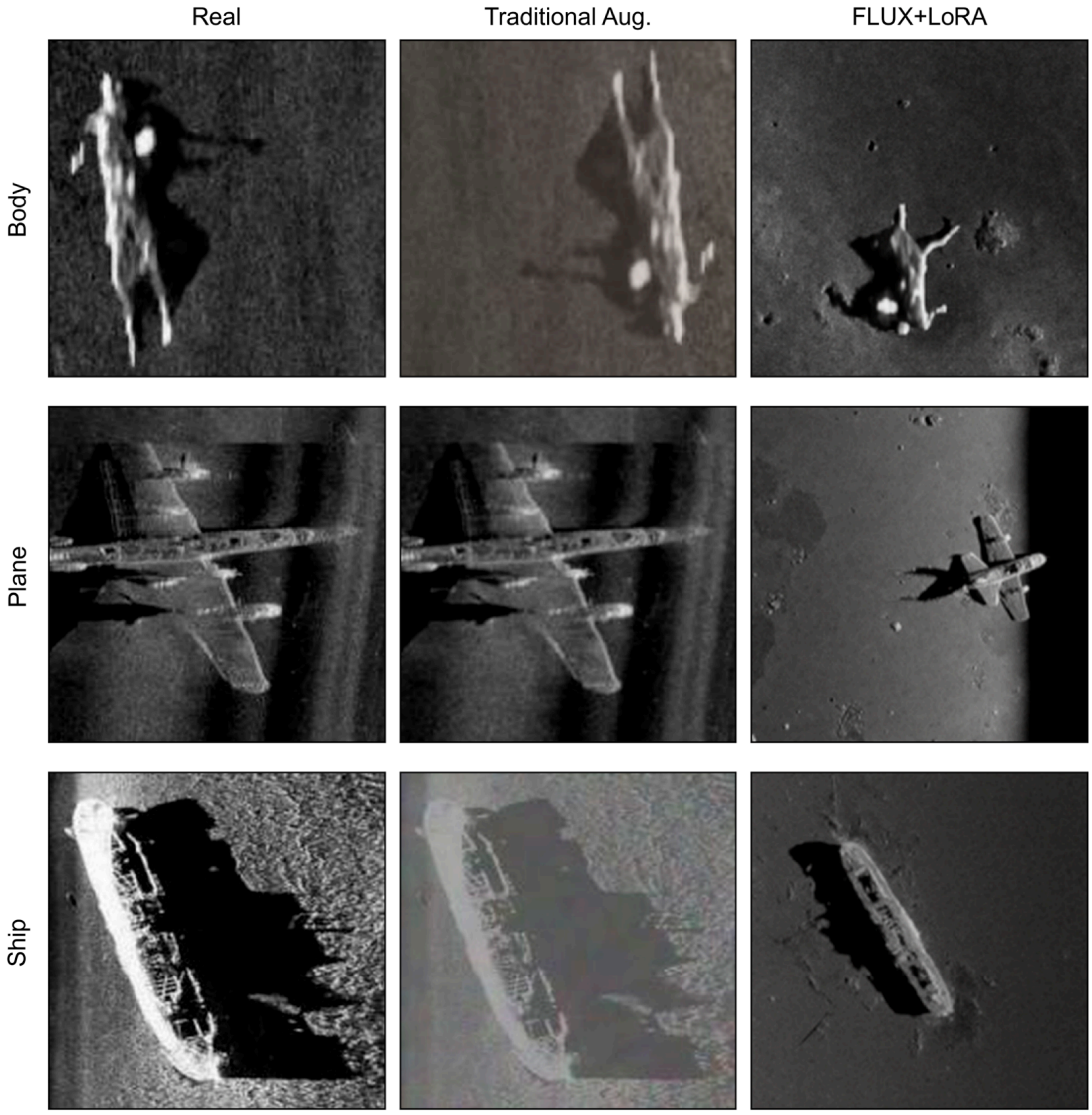


Figure 2. Representative visual examples of real, conventionally augmented, and FLUX+LoRA-generated side-scan sonar samples for body, plane, and ship targets. Rows correspond to target categories, and columns correspond to sample sources. The examples are used for qualitative illustration of target structure, acoustic-shadow extent, and seabed texture variation.

From the visual examples, it can be observed that conventional augmentation mainly produces variations through geometric or pixel-level transformations of existing sonar samples, such as flipping, rotation, and intensity perturbation. In contrast, the FLUX+LoRA-generated samples introduce additional variations in target appearance, acoustic-shadow extent, and local seabed texture while preserving the grayscale style of side-scan sonar imagery. These generated samples generally exhibit side-scan sonar characteristics, including speckle-like background texture, target highlight regions, and acoustic shadows, which are important for downstream detection because sonar targets are often recognized not only by target brightness but also by the spatial relationship between the target body and its shadow. However, generated samples may still contain artifacts, ambiguous target boundaries, or unreasonable highlight-shadow relationships. Therefore, the generated

images were further processed through label auditing and quality screening before being added to the training set, as described in Section 2.2.5.

It should be emphasized that this study focuses on whether FLUX+LoRA-generated samples can improve detection performance for the annotated target categories under a strict real-only validation/test protocol. Generation beyond the annotated categories is not used as a primary claim in this paper and is left for future investigation.

3.2. Detection Results

To evaluate the effectiveness of different augmentation strategies, the original training set, two conventional augmentation settings, and the proposed FLUX+LoRA-generated augmentation setting were compared under the same fixed YOLOv8n detector. For all methods, the validation and test sets were identical and contained only real side-scan sonar images. The results averaged over three random seeds are shown in Table 3.

Table 3. Main detection results under the fixed YOLOv8n and real-only validation/test protocol.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Original	0.7156 ± 0.0909	0.6305 ± 0.0535	0.7400 ± 0.0132	0.3994 ± 0.0187
Traditional-target-only	0.7524 ± 0.1669	0.7042 ± 0.0606	0.7428 ± 0.0731	0.4388 ± 0.0540
Traditional-all	0.8425 ± 0.0835	0.7248 ± 0.0629	0.7862 ± 0.0343	0.4495 ± 0.0363
FLUX+LoRA auto-screened	**0.8505 ± 0.0674**	**0.8185 ± 0.0445**	**0.8582 ± 0.0328**	**0.5115 ± 0.0164**

Bold values indicate the best performance in each column.

As shown in Table 3, training with only the original real images achieved an mAP@0.5 of 0.7400 ± 0.0132 and an mAP@0.5:0.95 of 0.3994 ± 0.0187. After introducing conventional augmentation, the detection performance improved to different degrees. The target-region augmentation setting achieved an mAP@0.5 of 0.7428 ± 0.0731 and an mAP@0.5:0.95 of 0.4388 ± 0.0540, while the full-image augmentation setting achieved an mAP@0.5 of 0.7862 ± 0.0343 and an mAP@0.5:0.95 of 0.4495 ± 0.0363. These results indicate that conventional augmentation can improve detector generalization to some extent, especially in terms of the stricter mAP@0.5:0.95 metric.

Compared with both the original training setting and conventional augmentation, the proposed FLUX+LoRA-generated augmentation achieved the best overall performance. Specifically, it obtained a Precision of 0.8505 ± 0.0674, a Recall of 0.8185 ± 0.0445, an mAP@0.5 of 0.8582 ± 0.0328, and an mAP@0.5:0.95 of 0.5115 ± 0.0164. Compared with the original setting, the mAP@0.5 increased by 0.1182 and the mAP@0.5:0.95 increased by 0.1121. Compared with the stronger conventional full-image augmentation setting, the FLUX+LoRA augmentation still improved mAP@0.5 by 0.0720 and mAP@0.5:0.95 by 0.0620.

These results demonstrate that the FLUX+LoRA-generated samples provide additional training information beyond geometric or pixel-level transformations of existing images. Although the generated samples are not used in validation or testing, adding them to the training set improves the detector's performance on real sonar test images. This suggests that the proposed diffusion-based augmentation strategy can introduce useful target-structure and background-texture diversity for side-scan sonar target detection under limited-data conditions.

The main mAP comparison among different training settings is further visualized in Figure 3, where FLUX+LoRA denotes the auto-screened FLUX+LoRA augmentation setting.

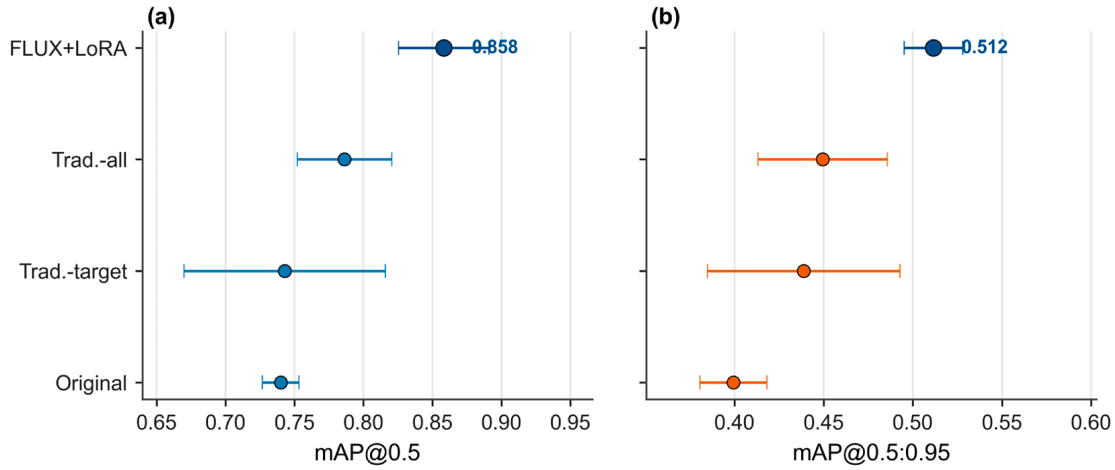


Figure 3. Main detection performance comparison under the fixed YOLOv8n and real-only validation/test protocol. (**a**) Comparison of mAP@0.5 among different training settings. (**b**) Comparison of mAP@0.5:0.95 among different training settings. Points denote mean values, and error bars denote the standard deviation over three random seeds. In this figure, FLUX+LoRA refers to the auto-screened FLUX+LoRA augmentation setting.

In addition to the quantitative comparison, Figure 4 provides qualitative detection visualizations on representative real test images. The examples include body, plane, and ship targets and compare the detection results obtained under the original training setting, the conventional augmentation baseline, and the proposed FLUX+LoRA augmentation setting. These visualizations provide an intuitive illustration of the detection behavior of different training strategies on real side-scan sonar samples.

3.3. Ablation Study

To further analyze the influence of key design choices in the proposed augmentation pipeline, three ablation analyses were conducted under the same fixed YOLOv8n evaluation protocol, including generated-sample quantity, screening strategy, and LoRA-rank sensitivity. In all ablation settings, the validation and test sets were kept unchanged and contained only real side-scan sonar images. The generated-sample quantity and screening-strategy results are summarized in Table 4.

Table 4. Ablation study on generated sample quantity and screening strategy.

Ablation Type	Setting	Generated Samples	mAP@0.5 ↑	mAP@0.5:0.95 ↑
Amount	25%	445	0.8094 ± 0.0896	0.4457 ± 0.0460
Amount	50%	890	0.8116 ± 0.0181	0.4599 ± 0.0365
Amount	75%	1335	0.8549 ± 0.0117	0.4683 ± 0.0227
Amount	100%	1780	**0.8582 ± 0.0328**	**0.5115 ± 0.0164**
Screening	Valid-only	2079	0.8512 ± 0.0439	0.5039 ± 0.0216
Screening	Auto-screened	1780	**0.8582 ± 0.0328**	**0.5115 ± 0.0164**

All ablation experiments were conducted using the fixed YOLOv8n detector under the same real-only validation/test protocol. Results are reported as mean ± standard deviation over three random seeds. ↑ indicates that higher values are better. Bold values indicate the best performance in each column.

First, the number of FLUX+LoRA-generated training samples was varied by using 25%, 50%, 75%, and 100% of the screened generated set. When 25% of the generated samples were used, the detector achieved an mAP@0.5 of 0.8094 ± 0.0896 and an mAP@0.5:0.95 of 0.4457 ± 0.0460. As the number of generated samples increased, the detection performance generally improved. Using the full screened generated set achieved the best overall performance, with an mAP@0.5 of 0.8582 ± 0.0328 and an mAP@0.5:0.95 of 0.5115 ± 0.0164.

This indicates that increasing the number of screened generated samples can provide more useful structural variations for detector training.

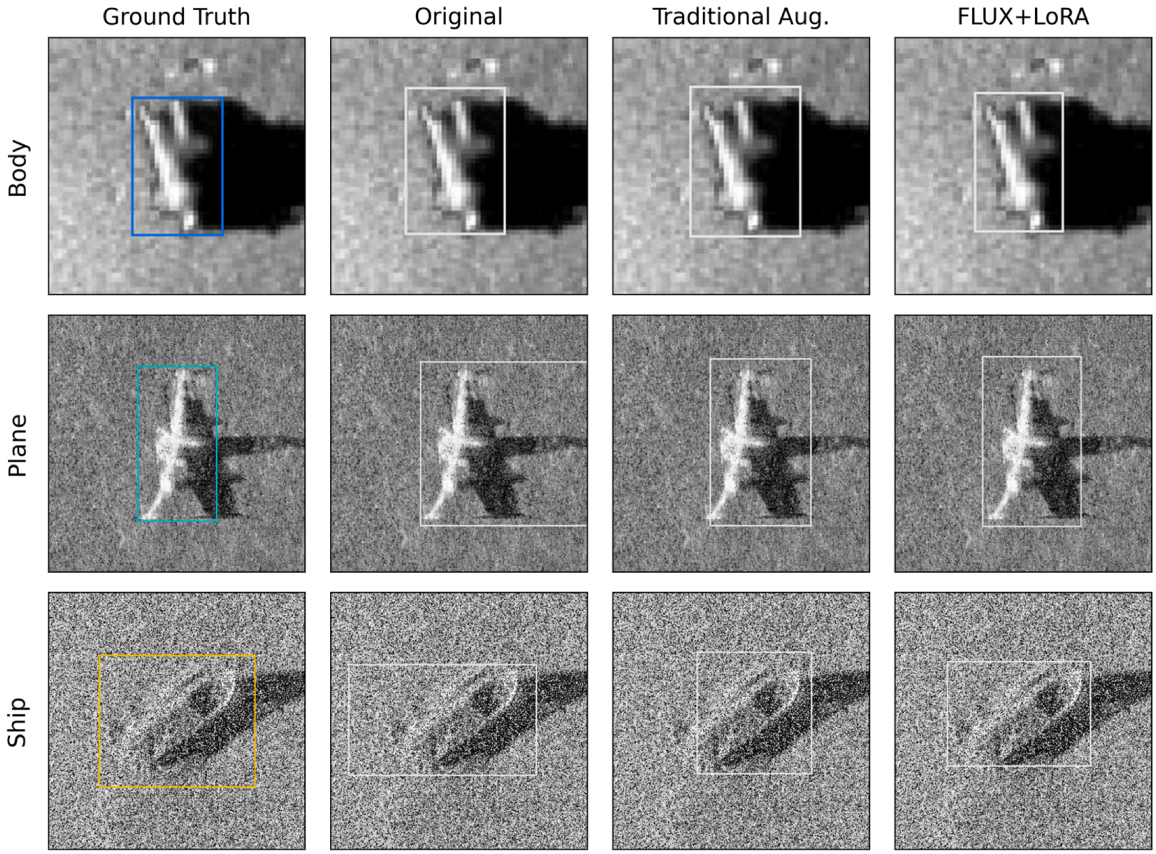


Figure 4. Qualitative detection comparison on representative real side-scan sonar test images. Each row shows one selected test image from a different target category, including body, plane, and ship. The columns show the ground-truth annotation and the prediction results obtained using the original training setting, the conventional augmentation baseline, and the proposed FLUX+LoRA augmentation setting. The visualizations are provided for qualitative illustration of detection behavior under different training settings.

Second, the effect of the screening strategy was evaluated by comparing the valid-only generated set and the final auto-screened generated set. The valid-only setting used 2079 generated samples that passed label validity checking, while the auto-screened setting retained 1780 higher-quality samples after additional screening. Although the auto-screened setting used fewer generated images, it achieved slightly better performance, improving mAP@0.5 from 0.8512 $\pm$ 0.0439 to 0.8582 $\pm$ 0.0328 and mAP@0.5:0.95 from 0.5039 $\pm$ 0.0216 to 0.5115 $\pm$ 0.0164. This suggests that removing less suitable generated candidates can provide a moderate but consistent benefit for downstream detection.

To further illustrate the screening procedure, representative non-selected and selected FLUX+LoRA-generated samples are shown in Figure 5. The non-selected samples denote generated candidates that were not retained in the final screened training set, whereas the selected samples were retained after label auditing and quality screening and were subsequently added to the training set.

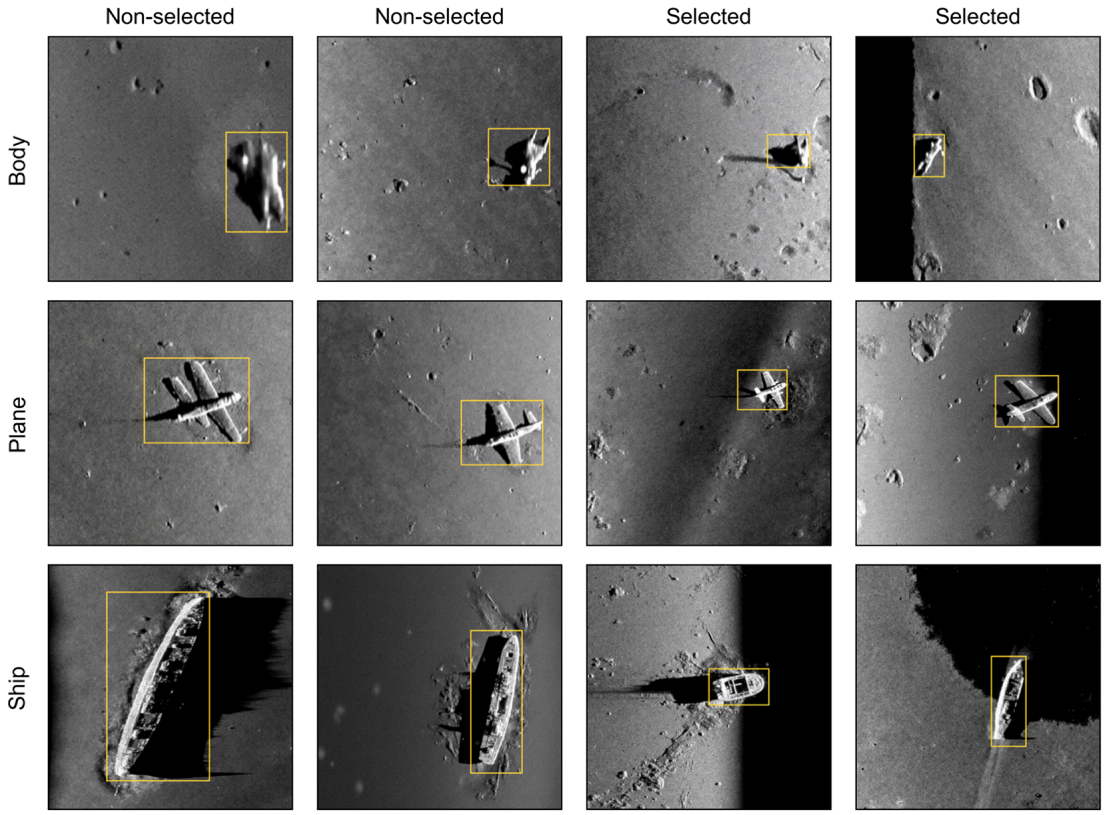


Figure 5. Representative examples of non-selected and selected FLUX+LoRA-generated side-scan sonar samples after the screening procedure. Rows correspond to body, plane, and ship targets. The non-selected examples refer to generated candidates that were not retained in the final screened training set, while the selected examples were retained after label auditing and quality screening. The examples qualitatively illustrate the role of the screening procedure in selecting generated samples with more usable target structures and sonar-style appearance. The yellow boxes indicate target regions.

The generated-sample quantity and screening-strategy ablation results are further visualized in Figure 6. Figure 6a shows that using more screened generated samples generally improves the two mAP metrics, while Figure 6b shows that the auto-screened subset achieves slightly better performance than the valid-only generated set despite using fewer generated images.

Overall, the ablation results show that both the quantity and quality of generated samples affect detection performance. The best results are obtained when the full screened FLUX+LoRA-generated set is used for training.

In addition to the generated-sample quantity and screening-strategy ablations, a LoRA-rank sensitivity analysis was conducted to evaluate the influence of LoRA adaptation capacity on sonar-style generation quality and downstream detection performance. Specifically, rank-4 and rank-8 FLUX+LoRA models were additionally evaluated and compared with the original rank-16 FLUX+LoRA manual-600 setting. For each rank, 600 manually annotated generated samples were used. The generated samples were added only to the training set, while the validation and test sets remained unchanged and contained only real side-scan sonar images. The same fixed YOLOv8n detector and real-only validation/test protocol were used for all rank settings.

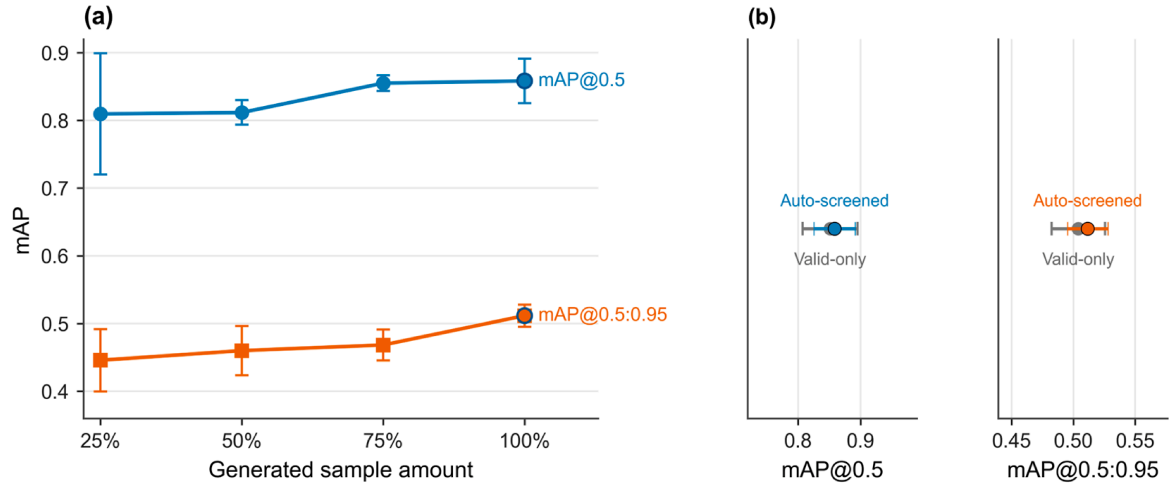


Figure 6. Ablation study on generated-sample quantity and screening strategy. (**a**) Detection performance obtained using different proportions of screened FLUX+LoRA-generated samples. (**b**) Paired comparison between the valid-only generated set and the auto-screened generated set. All experiments were conducted using the fixed YOLOv8n detector and unchanged real-only validation/test sets. Points denote mean values, and error bars denote the standard deviation over three random seeds.

The rank-16 setting corresponds to the original FLUX+LoRA manual-600 configuration used in the controlled diffusion-backbone comparison. Rank-4 and rank-8 settings were newly evaluated under the same fixed YOLOv8n and real-only validation/test protocol.

As shown in Table 5, the rank-16 setting achieves the best image quality and downstream detection performance among the evaluated LoRA ranks, with the lowest FID and KID values and the highest mAP@0.5 and mAP@0.5:0.95 scores. The rank-8 setting improves mAP@0.5 compared with rank-4, but its mAP@0.5:0.95 is slightly lower than that of rank-4, and the difference is small relative to the standard deviation. These results indicate that a very low LoRA rank may limit the sonar-domain adaptation capacity of the generative model, while rank 16 provides a better trade-off between image quality and downstream detection utility in this study.

Table 5. LoRA rank sensitivity analysis under the 600-sample manually annotated setting.

LoRA Rank	Generated Samples	FID ↓	KID ↓	mAP@0.5 ↑	mAP@0.5:0.95 ↑
4	600	260.8833	0.1703	0.7624 ± 0.0670	0.4450 ± 0.0358
8	600	263.0484	0.1789	0.7989 ± 0.0353	0.4427 ± 0.0453
16	600	**258.0063**	**0.1601**	**0.8611 ± 0.0611**	**0.4751 ± 0.0211**

↓ indicates that lower values are better, and ↑ indicates that higher values are better. Bold values indicate the best performance in each column.

The LoRA-rank sensitivity results are further summarized in Figure 7. The detection results in Figure 7a show that rank 16 achieves the best downstream detection performance among the evaluated ranks, while the image quality results in Figure 7b show that rank 16 also obtains the lowest FID and KID values.

3.4. Controlled Diffusion Backbone Comparison

To further evaluate the influence of different diffusion-based generation backbones, a controlled comparison was conducted under the same 600-sample manually annotated setting. Three generated subsets were compared: FLUX-base, SD1.5+LoRA, and the proposed FLUX+LoRA. For each method, 600 generated samples were used, with balanced

target categories and manually verified YOLO-format annotations. The generated samples were added only to the training set, while the validation and test sets were kept unchanged and contained only real side-scan sonar images.

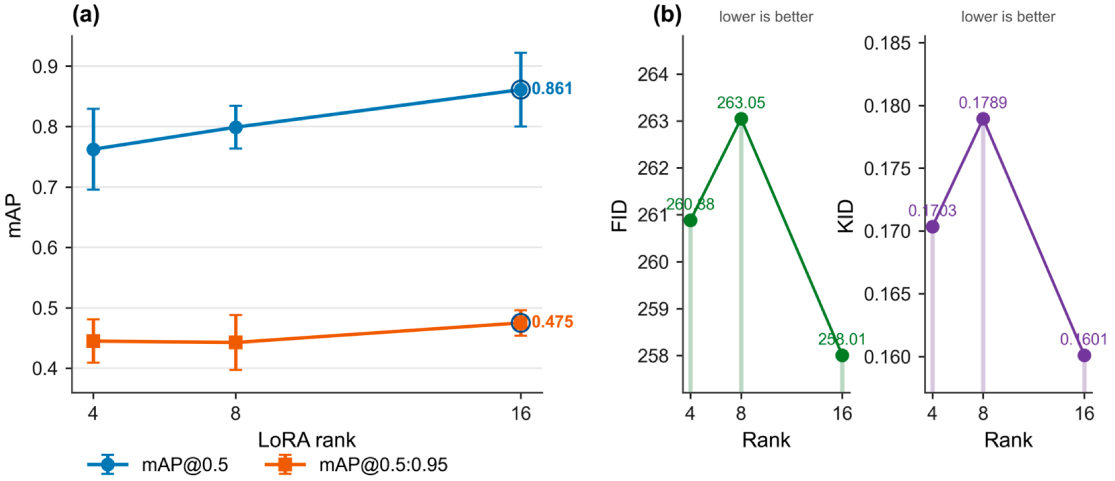


Figure 7. LoRA-rank sensitivity analysis under the 600-sample manually annotated setting. (**a**) Downstream detection performance of rank-4, rank-8, and rank-16 FLUX+LoRA settings. (**b**) Image quality comparison using FID and KID, where lower values indicate better distributional consistency with real training images. Points denote mean values, and error bars in the detection panel denote the standard deviation over three random seeds. The rank-16 setting corresponds to the original FLUX+LoRA manual-600 configuration.

In addition to downstream detection performance, FID and KID were calculated between each generated subset and the real training images to provide a quantitative reference for image quality. The results are shown in Table 6. Among the three diffusion-based generation settings, the FLUX+LoRA subset achieves the lowest FID and KID values, indicating better distributional consistency with real side-scan sonar images. It also obtains the best detection performance, with an mAP@0.5 of 0.8611 $\pm$ 0.0611 and an mAP@0.5:0.95 of 0.4751 $\pm$ 0.0211.

Table 6. Controlled comparison of diffusion-based generation backbones under a 600-sample manually annotated setting.

Method	Generated Samples	Annotation Protocol	FID ↓	KID ↓	mAP@0.5 ↑	mAP@0.5:0.95 ↑
FLUX-base	600	Manual	311.6181	0.2193	0.8173 $\pm$ 0.0657	0.4666 $\pm$ 0.0505
SD1.5+LoRA	600	Manual	290.9504	0.2043	0.7867 $\pm$ 0.0233	0.4423 $\pm$ 0.0157
FLUX+LoRA manual-600 subset	600	Manual	**258.0063**	**0.1601**	**0.8611 $\pm$ 0.0611**	**0.4751 $\pm$ 0.0211**

All generated subsets contain 600 manually annotated samples and are evaluated under the same fixed YOLOv8n and real-only validation/test protocol. FID and KID are computed between each generated subset and the real training images. ↓ indicates that lower values are better, and ↑ indicates that higher values are better. Bold values indicate the best performance in each column.

Compared with FLUX-base, the FLUX+LoRA subset improves mAP@0.5 from 0.8173 to 0.8611 and mAP@0.5:0.95 from 0.4666 to 0.4751. Compared with SD1.5+LoRA, it also achieves clear improvements in both image quality metrics and detection accuracy. These results suggest that LoRA adaptation on the FLUX backbone is beneficial for generating side-scan-sonar-style samples with higher downstream detection utility under the same annotation budget.

Overall, the controlled backbone comparison supports the effectiveness of the proposed FLUX+LoRA generation strategy. As further visualized in Figure 8, the FLUX+LoRA manual-600 subset achieves the best overall downstream detection performance and the lowest FID/KID values under the same annotation budget, which is consistent with the quantitative results in Table 6.

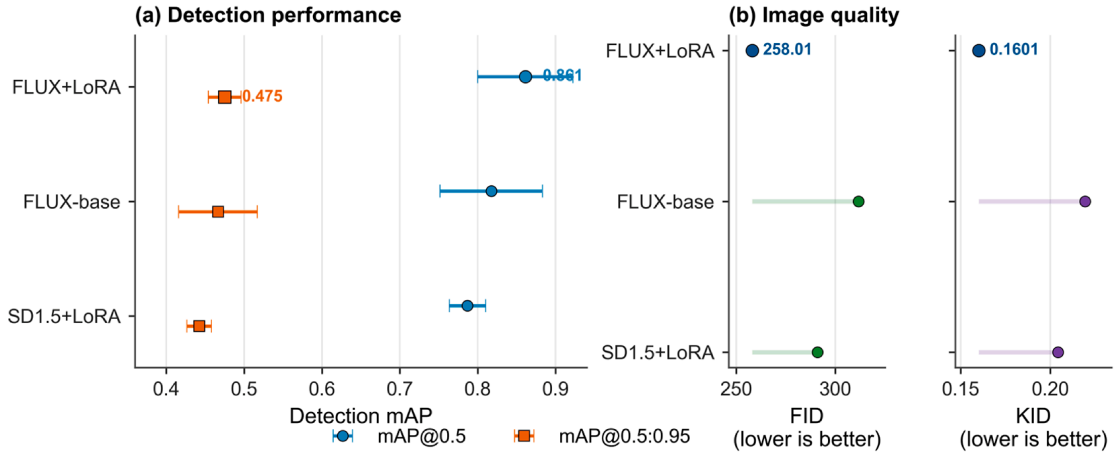


Figure 8. Controlled comparison of diffusion-based generation backbones under the 600-sample manually annotated setting. (**a**) Downstream detection performance of FLUX-base, SD1.5+LoRA, and FLUX+LoRA. (**b**) Image quality comparison using FID and KID, where lower values indicate better distributional consistency with real training images. Points denote mean values, and error bars in the detection panel denote the standard deviation over three random seeds. All generated subsets were added only to the training set, while the validation and test sets remained unchanged and real-only. In this figure, FLUX+LoRA refers to the manually annotated 600-sample FLUX+LoRA subset used for the controlled backbone comparison.

4. Discussion

These results indicate that the benefit of diffusion-based augmentation is not determined solely by the number of generated samples, but also by the visual plausibility, annotation reliability, and domain adaptation capacity of the generation pipeline.

The above experimental results collectively demonstrate the effectiveness of the proposed FLUX+LoRA-based augmentation strategy for side-scan sonar target detection under limited-data conditions. First, the main fixed-detector comparison shows that adding screened FLUX+LoRA-generated samples only to the training set improves the detector's performance on unchanged real sonar validation/test images. Compared with the original training setting and conventional augmentation settings, the proposed FLUX+LoRA auto-screened augmentation achieves the best overall detection performance. This indicates that the generated samples provide useful training information rather than merely improving performance on synthetic or augmented validation/test data.

Second, the generated-sample quantity and screening-strategy ablations show that both the amount and quality of generated samples affect downstream detection performance. Increasing the proportion of screened generated samples generally improves detection accuracy, suggesting that additional generated sonar samples can provide useful target-structure and background-texture variations for detector training. Meanwhile, the comparison between the valid-only generated set and the auto-screened generated set shows that further quality screening provides a moderate but consistent improvement. This result indicates that generated data should not be used without quality control, especially for side-scan sonar images, where the plausibility of target-highlight and acoustic-shadow relationships is important for detection.

Third, the LoRA-rank sensitivity analysis further confirms that the adaptation capacity of LoRA affects both generated image quality and downstream detection utility. Under the same 600-sample manually annotated setting, rank 16 achieves the lowest FID and KID values as well as the highest mAP@0.5 and mAP@0.5:0.95 among the evaluated ranks. Rank 8 improves mAP@0.5 compared with rank 4, but its mAP@0.5:0.95 is slightly lower than that of rank 4, and the difference is small relative to the standard deviation. These results suggest that a very low LoRA rank may limit the sonar-domain adaptation capacity of the generative model, while rank 16 provides a better trade-off between image quality and downstream detection performance in this study.

Fourth, the controlled diffusion-backbone comparison supports the effectiveness of using FLUX+LoRA for sonar-domain image generation. Under the same 600-sample manually annotated setting, the FLUX+LoRA subset achieves lower FID/KID values and higher detection accuracy than FLUX-base and SD1.5+LoRA. This result indicates that sonar-domain LoRA adaptation on the FLUX backbone improves both the distributional consistency and downstream utility of generated samples. Compared with using the FLUX base model without LoRA adaptation or using SD1.5+LoRA, the proposed FLUX+LoRA setting is more effective in generating detection-useful side-scan sonar samples under the experimental conditions considered in this study.

Overall, the proposed method improves side-scan sonar target detection by increasing the structural diversity of the training set while maintaining a strict evaluation protocol based on unchanged real validation and test images. The results also show that the utility of generated sonar samples depends not only on the generative backbone, but also on sample screening, generated-sample quantity, and LoRA adaptation capacity. Therefore, for practical sonar data augmentation, both generation quality and downstream detection utility should be considered jointly.

5. Conclusions

This paper presents a diffusion-model-based data augmentation method for side-scan sonar target detection under limited-data conditions. A FLUX.1 diffusion model is fine-tuned with LoRA to adapt the pretrained generative model to the side-scan sonar image domain. The generated samples are further processed through label auditing and quality screening, and are added only to the training set, while the validation and test sets remain unchanged and contain only real side-scan sonar images.

To provide a fair evaluation of the augmentation strategy, all detection experiments are conducted under a fixed YOLOv8n detector and a real-only validation/test protocol. Compared with training on the original dataset, the proposed FLUX+LoRA augmentation improves mAP@0.5 from 0.7400 ± 0.0132 to 0.8582 ± 0.0328 and mAP@0.5:0.95 from 0.3994 ± 0.0187 to 0.5115 ± 0.0164. It also outperforms conventional augmentation methods, indicating that the generated samples provide useful target-structure and seabed-texture diversity beyond geometric or pixel-level transformations of existing images.

The ablation results further show that both the quantity and quality of generated samples influence downstream detection performance. Increasing the proportion of screened generated samples generally improves detection accuracy, while the additional screening step provides a moderate but consistent gain over using all valid generated candidates. In addition, the LoRA-rank sensitivity analysis shows that rank 16 achieves the best trade-off among the evaluated ranks, yielding the lowest FID/KID values and the highest downstream detection performance under the 600-sample manually annotated setting. These results indicate that the adaptation capacity of LoRA is an important factor affecting sonar-domain generation quality and detection utility.

The controlled 600-sample diffusion-backbone comparison further shows that the FLUX+LoRA subset selected from generated samples achieves lower FID/KID values and better detection performance than FLUX-base and SD1.5+LoRA under the same manual annotation budget. These results support the effectiveness of sonar-domain LoRA adaptation for generating detection-useful side-scan sonar samples.

Future work will focus on expanding the scale and diversity of real sonar datasets, improving the physical consistency of generated target–highlight–shadow relationships, and developing more controllable generation mechanisms for target category, scale, orientation, and seabed background conditions. More systematic evaluation protocols for generated sonar images will also be explored to better connect image-level quality metrics with downstream detection utility. In addition, further studies will investigate more LoRA configurations, additional parameter-efficient fine-tuning strategies, and physically constrained generation mechanisms for side-scan sonar image synthesis.

Author Contributions: Conceptualization, Y.Y.; Methodology, Y.Y.; Validation, Y.Y.; Resources, T.Z.; Data curation, Y.Y.; Writing—original draft, Y.Y.; Writing—review & editing, T.Z.; Visualization, Y.Y.; Supervision, T.Z.; Project administration, T.Z.; Funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Shenzhen Science and Technology Program under Grant JCYJ20241202123731040 and Research Fund for Advanced Ocean Institute of Southeast University, Nantong under Grant KP202403.

Data Availability Statement: The synthetic side-scan sonar images generated in this study are publicly available on Kaggle at: https://www.kaggle.com/datasets/yangyuanxu/sonar-flux-synthetic (accessed on 15 June 2026). The released dataset includes the FLUX+LoRA rank-16 auto-screened subset and the 600-sample manually annotated subset used for controlled comparison experiments, together with YOLO-format annotations, class mappings, file manifests, and metadata files. The original real side-scan sonar images used for detector training, validation, and testing are not redistributed due to data-licensing restrictions. The accompanying code, configuration files, result tables, plotting scripts, screening utilities, prompt templates, and reproduction documentation are available on GitHub at: https://github.com/yanzhi0526/sonar-flux-lora-augmentation (accessed on 15 June 2026). The repository is intended to support transparency and reproducibility of the data-processing, screening, evaluation, and figure-generation procedures. The FLUX base model, trained LoRA adapter weights, detector checkpoints, and original real sonar images are not included in the public release due to base-model licensing and data-governance considerations.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5 Thinking, OpenAI) for language polishing, grammar checking, and improving the clarity and organization of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

LoRA	Low-Rank Adaptation
GANs	Generative Adversarial Networks
YOLO	You Only Look Once
FID	Fréchet Inception Distance
KID	Kernel Inception Distance
mAP	mean Average Precision
IoU	Intersection over Union
VAE	Variational Autoencoder

PEFT	Parameter-Efficient Fine-Tuning
SD	Stable Diffusion

References

1. Zhang, F.; Zhang, W.; Cheng, C.; Hou, X.; Cao, C. Detection of small objects in side-scan sonar images using an enhanced YOLOv7-based approach. *J. Mar. Sci. Eng.* **2023**, *11*, 2155.
2. Steiniger, Y.; Groen, J.; Stoppe, J.; Kraus, D.; Meisen, T. A study on modern deep learning detection algorithms for automatic target recognition in sidescan sonar images. *Proc. Meet. Acoust.* **2021**, *44*, 070010. [CrossRef]
3. Wen, X.; Wang, J.; Cheng, C.; Zhang, F.; Pan, G. Underwater side-scan sonar target detection: YOLOv7 model combined with attention mechanism and scaling factor. *Remote Sens.* **2024**, *16*, 2492. [CrossRef]
4. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555. [CrossRef]
5. Huo, G.; Wu, Z.; Li, J. Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. *IEEE Access* **2020**, *8*, 47407–47418. [CrossRef]
6. Yule, C.; Bo, L.I.; Changsheng, Y. Research on sonar image few-shot classification based on deep learning. *Xibei Gongye Daxue Xuebao/J. Northwest. Polytech. Univ.* **2022**, *40*, 739–745. [CrossRef]
7. Ma, Q.; Jin, S.; Bian, G.; Cui, Y.; Liu, G. DBnet: A lightweight dual-backbone target detection model based on side-scan sonar images. *J. Mar. Sci. Eng.* **2025**, *13*, 155. [CrossRef]
8. Zhu, J.; Li, H.; Qing, P.; Hou, J.; Peng, Y. Side-scan sonar image augmentation method based on CC-WGAN. *Appl. Sci.* **2024**, *14*, 8031. [CrossRef]
9. 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. *Electronics* **2021**, *10*, 1823.
10. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 10684–10695.
11. Yang, Z.; Zhao, J.; Zhang, H.; Yu, Y.; Huang, C. A side-scan sonar image synthesis method based on a diffusion model. *J. Mar. Sci. Eng.* **2023**, *11*, 1103. [CrossRef]
12. 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. *ICLR* **2022**, *1*, 3.
13. Lu, J.; She, M.; Mao, W.; Wang, Z. CDM-QTA: Quantized Training Acceleration for Efficient LoRA Fine-Tuning of Diffusion Model. In *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*; IEEE: New York, NY, USA, 2025; pp. 1–5.
14. Yulin, T.; Jin, S.; Bian, G.; Zhang, Y. Shipwreck target recognition in side-scan sonar images by improved YOLOv3 model based on transfer learning. *IEEE Access* **2020**, *8*, 173450–173460. [CrossRef]
15. Mun, J.; Park, S.; Kim, J.; Kim, C.; Kim, D.; Sohn, C. Deep Learning based Mine Detection using Side-Scan Sonar Image. *J. Korea Inst. Mil. Sci. Technol.* **2025**, *28*, 146–154. [CrossRef]
16. Cao, Y.; Cui, X.; Gan, M.; Wang, Y.; Yang, F.; Huang, Y. MAL-YOLO: A lightweight algorithm for target detection in side-scan sonar images based on multi-scale feature fusion and attention mechanism. *Int. J. Digit. Earth* **2024**, *17*, 2398050. [CrossRef]
17. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
18. Greenberg, O. Demystifying flux architecture. *arXiv* **2025**, arXiv:2507.09595.
19. Shi, Y.; Xue, C.; Liew, J.H.; Pan, J.; Yan, H.; Zhang, W. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 8839–8849.
20. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2256–2265.
21. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
22. Fu, X.; He, M.; Lu, Y.; Wang, W.Y.; Roth, D. Commonsense-T2I challenge: Can text-to-image generation models understand commonsense? *arXiv* **2024**, arXiv:2406.07546.
23. Labs, B.F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv* **2025**, arXiv:2506.15742.
24. Stracke, N.; Baumann, S.A.; Susskind, J.; Bautista, M.A.; Ommer, B. Ctrloralter: Conditional loradapter for efficient 0-shot control and altering of t2i models. In *Proceedings of the European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 87–103.

25. Li, S.; Pan, Y. An Efficient and Harmonized Framework for Balanced Cross-Domain Feature Integration. *arXiv* **2024**, arXiv:2403.18461.
26. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 29 March 2026).

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