Journal of
Marine Science
and Engineering

Article

Side-Scan Sonar Image Generator Based on Diffusion Models for Autonomous Underwater Vehicles

Feihu Zhang *, Xujia Hou, Zewen Wang, Chensheng Cheng and Tingfeng Tan

School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China;
hxj1363947894@mail.nwpu.edu.cn (X.H.); wzw828@mail.nwpu.edu.cn (Z.W.);
chensheng.cheng@mail.nwpu.edu.cn (C.C.); ttf@mail.nwpu.edu.cn (T.T.)
* Correspondence: feihu.zhang@nwpu.edu.cn

Abstract: In the field of underwater perception and detection, side-scan sonar (SSS) plays an indispensable role. However, the imaging mechanism of SSS results in slow information acquisition and high complexity, significantly hindering the advancement of downstream data-driven applications. To address this challenge, we designed an SSS image generator based on diffusion models. We developed a data collection system based on Autonomous Underwater Vehicles (AUVs) to achieve stable and rich data collection. For the process of converting acoustic signals into image signals, we established an image compensation method based on nonlinear gain enhancement to ensure the reliability of remote signals. On this basis, we developed the first controllable category SSS image generation algorithm, which can generate specified data for five categories, demonstrating outstanding performance in terms of the Fréchet Inception Distance (FID) and the Inception Score (IS). We further evaluated our image generator in the task of SSS object detection, and our cross-validation experiments showed that the generated images contributed to an average accuracy improvement of approximately 10% in object detection. The experimental results validate the effectiveness of the proposed SSS image generator in generating highly similar sonar images and enhancing detection accuracy, effectively addressing the issue of data scarcity.

Keywords: side-scan sonar; diffusion model; image generation; deep learning; gain enhancement; object detection

Citation: Zhang, F.; Hou, X.; Wang, Z.; Cheng, C.; Tan, T. Side-Scan Sonar Image Generator Based on Diffusion Models for Autonomous Underwater Vehicles. *J. Mar. Sci. Eng.* **2024**, *12*, 1457. https://doi.org/10.3390/jmse12081457

Academic Editor: Lei Wan

Received: 4 July 2024
Revised: 31 July 2024
Accepted: 21 August 2024
Published: 22 August 2024

1. Introduction

The oceanic realm accounts for approximately two-thirds of the Earth's surface area, yet human exploration has only mapped about five percent of it [1]. Consequently, the majority of oceanic zones remain obscure, abysmal, and unknown to humankind. To gain further insights into the oceans, more in-depth research on underwater environmental perception technologies is imperative [2]. At present, underwater environment detection technology mainly includes acoustic-based methods and optical-based methods. Due to the significant attenuation of electromagnetic waves under water, optical signals only have the capability to propagate over several tens of meters subaqueously. Should water clarity be compromised, the effective range diminishes further, rendering optical detection principally suitable for proximate inspection and verification [3]. Acoustic detection, however, leverages the propagation laws of sound signals underwater for exploration and monitoring purposes. It boasts a notable advantage over optical detection in terms of range; thus, the bulk of underwater detection is predominantly conducted via acoustic means. A Sound Navigation and Ranging (sonar) device is a type of sensor engineered based on the transduction of sound waves and information processing exploiting the propagation and reflection properties of acoustics in aquatic environments. Within this spectrum of tools, SSS operates by projecting fan-shaped beams of acoustic pulses laterally downwards, and due to its high resolution and cost-effective nature, it constitutes a vital element of underwater acoustic detection. SSS is extensively applied in marine cartography [4,5], subaquatic geological surveying [6], underwater target detection and identification [7], and dam foundation inspection [8]. Moreover, with the continued evolution of scientific technology, deep learning-based SSS detection methods are increasingly superseding traditional expert system-based approaches, thereby consistently increasing the precision of detection [9]. Nevertheless, data-driven detection models require an abundance of sample data and labels to assure performance efficacy. During SSS data acquisition, the imaging principle dictates that a subsequent acoustic wave is emitted only upon receipt of the returning echo signal at the current moment, resulting in a data interval per ping typically exceeding 0.2 s. This entails that during the data collection process, approximately five pings of sonar data are generated per second. Acquiring a sonar image with a height of 400 pixels requires a minimum of 80 s, which is considerably longer compared with the time required for forward-looking sonars or optical cameras. Additionally, the carrier platform for a sonar sensor may also host other hydroacoustic devices, such as Doppler velocimeters or Ultra-Short Baseline (USBL) positioning systems. If the operational frequencies of these devices are integer multiples of the working frequency of the SSS sensor, interference may occur, resulting in striping noise artifacts on the sonar imagery.

In this study, we initially built an AUV-based SSS data acquisition device to achieve efficient data acquisition. The AUV possessed a streamlined shape to significantly diminish drag during subaquatic navigation and was powered by an electric propulsion system to further reduce noise emissions from the platform. An SSS sensor was mounted on the AUV, a vehicle used to perform constant-depth cruising missions in diverse marine areas to collect data. Subsequently, for the raw SSS data, we developed a nonlinear gain enhancement algorithm to compensate for the image features at the far end that are generally weakened due to the propagation loss of sound waves underwater, thereby obtaining sonar images with high contrast and clarity. Building on this, a sonar image dataset containing five categories of information, which can provide authentic raw data for subsequent research based on SSS imagery, was constructed. Furthermore, we developed a diffusion model-based controllable category sonar image generator that obviates the need for training additional classifiers. By balancing the labeled conditional diffusion model with the unlabeled non-conditional diffusion model, we achieved the generation of both high-quality and diverse images.Our study effectively addresses the present insufficiency of SSS image datasets, supplying numerous exemplary samples for data-driven detection models. The results of experiments conducted on publicly available target detection models demonstrate the efficacy of the proposed SSS image generator in enhancing object detection accuracy. In summary, the main contributions of this study are as follows:

- We established an SSS data collection platform based on an AUV and collected a large number of raw sonar data from different marine areas. We developed a nonlinear gain enhancement algorithm suitable for compensating for spherical wave propagation loss, achieving balanced sonar image processing and improving the quality of SSS imaging.
- We created a five-category SSS image dataset that includes common seabed backgrounds and targets. Based on this dataset, we established a controllable category SSS image generator that can generate images of specified categories without relying on additional classifiers, effectively expanding the SSS image dataset.
- We conducted both quantitative and qualitative evaluations of the SSS image generator, using the FID, the IS, and Haralick texture features to assess the model. The generative model was also applied to the task of target detection, and the cross-validation results demonstrate its positive impact on improving detection accuracy, providing data support for subsequent SSS image-based research.

2. Related Work

2.1. Basic Image Generation Model

The objective of image generation tasks is to obtain the probabilistic statistical distribution of the original data. Early image generation efforts predominantly relied on feature representation methods to generate images which were only capable of handling simple and regular image generation tasks [10]. The advent of deep neural networks, with their exceptional feature-learning capabilities and nonlinear expression abilities, has significantly advanced the development of generative models [11]. To date, the primary deep learning-based generative models include Variational Autoencoders (VAEs) [12], Generative Adversarial Networks (GANs) [13], autoregressive models [14], and denoising diffusion models.

VAEs, deep generative models based on the autoencoder structure [15], introduce constraints on the latent space during training, such as the Kullback–Leibler (KL) divergence, to ensure that the latent variables follow a Gaussian distribution. However, constraining the latent space to conform to a normal distribution is often overly challenging. Therefore, Oord et al. [16] developed the Vector-Quantized Variational Autoencoder (VQVAE), which constrains the VAE's latent space to satisfy a discrete distribution and employs an autoregressive model to model the discrete codes, thus partially addressing the issue of fitting the discrete space distribution. Subsequent work [17] developed hierarchical quantized VAEs to encode images, which can more comprehensively address the issue of fitting the discrete space distribution and diminish the decoder's burden in reconstructing the image, thereby enabling the generation of more realistic images.

GANs, previously the most widely used generative models, consist of a generator and a discriminator. The former aims to fit the data distribution, while the latter distinguishes between real and generated data. However, due to the effective fitting capabilities of both the generator and the discriminator and the inability to guarantee optimal convergence during iterations, these networks do not reach the Nash equilibrium state during actual training [18]. Additionally, GANs often face issues such as training instability, mode collapse, and gradient vanishing problems. To address these issues, Mao et al. [19] developed the Least Squares GAN (LSGAN) to enhance the stability of adversarial training and improve the quality of the synthesized images. The authors in [20] established WGAN, further analyzed the reasons for training instability, and introduced the Lipschitz continuity constraint on the discriminator, effectively enhancing training stability. DCGAN [21] is the first network in which deep convolutional layers and batch normalization were introduced to improve network stability. The authors in [22] developed SAGAN by incorporating the self-attention mechanism into a GAN, which enhanced the network's representational capacity, improving the quality and diversity of the generated images. Furthermore, PGGAN [23] represents a progressive growth strategy for image generation, increasing the resolution, thereby making it possible to successfully generate 1024 × 1024-resolution images for the first time. Inspired by image style transfer, StyleGAN [24] is based on adjustable instance normalization to inject modulation signals, which greatly enhances the network's capabilities. StyleGAN2 [25], an improvement on the latter, enables end-to-end training and achieves high-quality, high-resolution image generation.

Autoregressive models are commonly used generative models based on the idea that subsequent variables in a sequence can be regressed from preceding variables. Different autoregressive models employ various network structures for generation. PixelRNN [26] uses the temporal concept of Recurrent Neural Networks (RNNs) for the prediction process, where the RNN model is composed of multiple Long Short-Term Memory (LSTM) layers. PixelCNN [27] approximates the RNN structure with masked convolutional layers and eliminates the pooling layers. Subsequent work on PixelCNN further enhanced its computational efficiency and generation results [28]. In recent years, the Transformer [29] structure has allowed for significant advances in natural language generation by introducing the attention mechanism, which significantly strengthens the fitting capabilities of autoregressive models. In the field of image generation, DALL-E [29] leverages the VQVAE to encode images into discrete latent codes and then uses an autoregressive model based on the Transformer structure to fit the distribution of these discrete codes. However, this approach faces the following issues: (1) The efficiency of serially generating conditional distribution probabilities is lower than directly generating joint distributions and sampling, resulting in significant time consumption for image generation. (2) The assumption that each position in the autoregressive model depends on the previous positions is unreasonable for images, as image patches are often related to various positions. Once a position is generated, it cannot be modified based on subsequent positions, causing any errors in earlier positions to propagate and affect subsequent results, leading to cumulative errors.

Denoising diffusion models are the latest advancement in image generation, comprising a forward noise diffusion Markov process and a reverse denoising Markov process [30]. By incrementally adding noise in multiple steps, this approach decomposes the complex distribution fitting problem into several simpler sub-problems. During the training phase, the denoising network updates its parameters to fit the data distribution of the noise-added process at each step. In the image generation phase, noise signals are first sampled from a known distribution and then progressively denoised through multiple steps by the denoising network, thereby sampling from the complex image distribution. This method has garnered widespread attention since its introduction, with researchers exploring its application in tasks such as unconditional image generation [31,32], class-conditional generation [33], and text-to-image translation [34], achieving state-of-the-art results across these domains.

2.2. SSS Image Generation Model

Due to the difficulty of data acquisition in side-scan sonar and the sparsity of the target distribution, an important research direction in the field of SSS is image generation. To mitigate the scarcity of SSS data, Jiang [35] proposed a GAN-based image generation model that rapidly generates side-scan sonar images. However, this method can only generate single-channel grayscale images, and since the generator and discriminator are not equal in fitting ability, it is difficult to reach the Nash equilibrium, resulting in model collapse. Bore [36] created a generation model with the environment as the input and the side-scan sonar image as the output based on the conditional generation of adversarial networks. However, this method has significant restrictions on implementation. Song [37] suggested using Extreme Learning Machines for sonar image segmentation and synthesis, employing a multi-path convolutional neural network to learn different image features, thereby progressively synthesizing detailed side-scan sonar images. Wang [38] employed transfer learning to introduce a style transfer approach between optical and SSS images, effectively enhancing classification accuracy. Ge [39] also proposed the use of synthetic data and transfer learning to convert optical images into side-scan sonar images for classification. In the case of zero samples, Xu [40] proposed a multi-feature fusion self-attention network (MFSANet) to generate SSS images of new categories, which transformed the problem into a traditional supervised learning problem. By taking the optical image as the input to the network, the Simplified Self-Attention Module (SSAM) is used to model the acoustic image, so as to effectively generate virtual SSS samples. However, this transfer learning-based method requires an additional photo, such as one taken on land, to be used as a specification. In many cases, this extra requirement limits the application of the method.

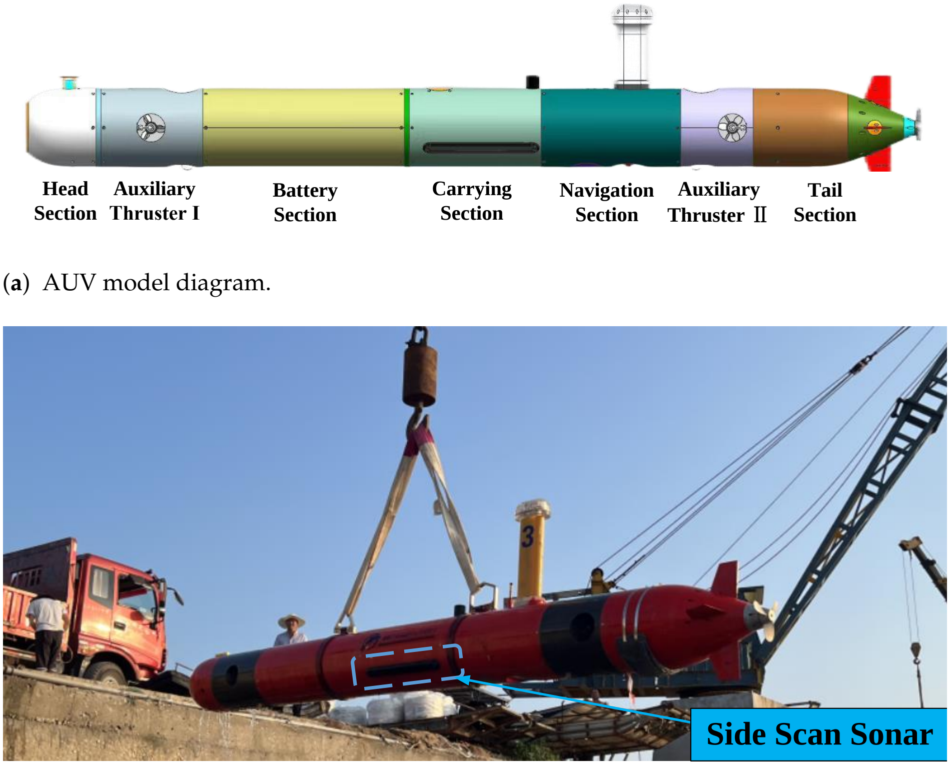
With recent breakthroughs in diffusion models [30–33], more and more researchers are turning to this approach for image generation. Yang [41] was the first to utilize a diffusion model for SSS image modeling: they transformed the sonar image into random noise with a Gaussian distribution and iteratively refined this noise in a reverse process to reconstruct new samples that matched the prior data distribution. Similarly, Zhang [42] and Cheng [43] applied diffusion models to expand the collected sonar data and created a hybrid dataset composed of real and generated samples. Upon experimenting with various mainstream target detection algorithms, they concluded that the generated samples could effectively improve the accuracy of the detection models considered.

However, most of the aforementioned methods have been applied to the SCTD dataset [44]. A significant portion of images in the SCTD dataset are not raw data (having undergone text editing, watermarking, etc.), which disrupts the statistical distribution of authentic data. In addition, previous work has not been able to achieve category-specific image generation, which means that the generated images often require further manual selection before they can be employed for subsequent detection tasks. Therefore, the primary limitations of the current side-scan sonar image generation task include the following: (1) the existing public datasets do not truly reflect the data distribution of sonar data, thus hindering the development of generation models; (2) generators based on diffusion models can only produce uncategorized images and necessitate extensive human effort to further filter the generated images.

3. Materials and Methods

3.1. Data Acquisition and Processing

The quality of SSS data is influenced by various factors, including the motion of the carrier, its speed, and its altitude above the seabed. Using an AUV equipped with an SSS sensor for data collection offers advantages such as constant speed, continuously adjustable altitude above the seabed, and minimal noise interference. Therefore, to obtain reliable and sufficient underwater sonar images, we selected an AUV with a diameter of 324 mm, equipped with the SS4590 module for SSS data collection, as shown in Figure 1. The AUV had a modular design, allowing each section to be detached and replaced. The front segment was equipped with a USBL for underwater communication. The forward and aft auxiliary propulsion sections enabled the AUV to move horizontally and vertically, respectively. The tail was fitted with rudders and a primary propulsor, providing the AUV with a maximum cruising speed of 12 knots. The battery section contained lithium-ion batteries, which could power the AUV for 10 h at a cruising speed of 3 knots. Additionally, the AUV was equipped with a Global Positioning System (GPS) module and inertial navigation components for self-localization. The sonar module section carried the SS4590 SSS sensor, which could emit both 450 kHz and 900 kHz frequency CW or Chirp signals, with a maximum single-side detection range of 150 m.



(**a**) AUV model diagram.

(**b**) AUV being deployed in the test area.

Figure 1. AUV overall layout structure and layout drawing.

To ensure the collection of representative and abundant data, we selected two marine areas for our data acquisition missions. Additionally, we placed cylindrical objects with a base diameter of 100 cm and a height of 150 cm on the seabed to serve as targets. Due to the side-scan sonar sensor's installation angle of 20 degrees horizontally downwards on the AUV, there were blind spots in its field of view; therefore, it was necessary to plan the AUV's route during data acquisition to ensure that the sonar beams could cover the blind spots to avoid missing any targets. The AUV performed back-and-forth patrolling detection tasks in the two preset marine areas (as shown in Figure 2); to collect clear images, its cruising speed was set to 3 knots and the depth to 10 m. The distance between the return paths was set to 100 m to guarantee that the sonar beams could cover the blind spots. The experimental areas selected in the marine environment had depths ranging from 30 to 50 m, ensuring the acquisition of seabed lines and background, along with seabed terrain that exhibited significant depth variations. During the collection process, Nvidia Jetson AGX Orin edge computing devices were utilized to control the SSS sensor and store data. The AUV navigated for 1 h and 40 min across the two marine areas, respectively, generating 2.29 GB of XTF files.

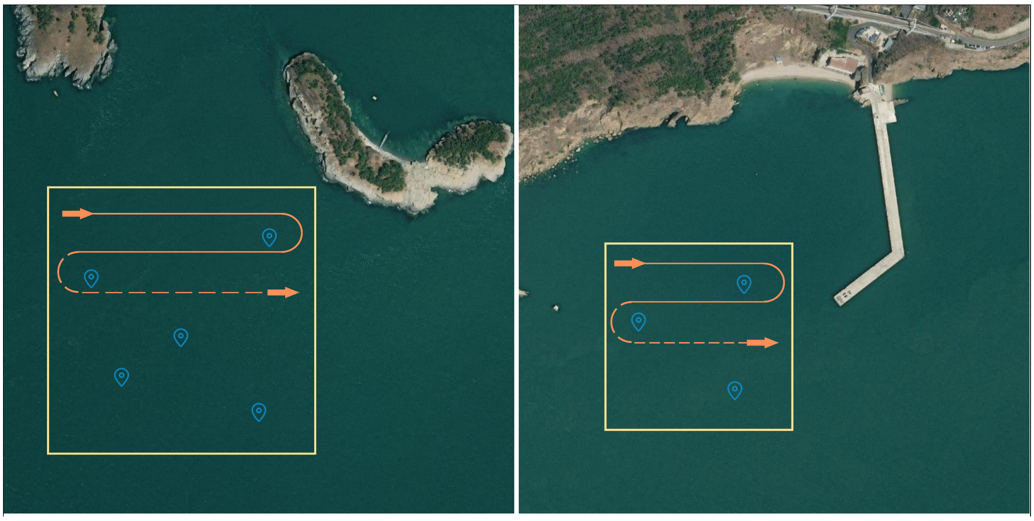


Figure 2. Schematic diagram of AUV path planning in experimental sea areas.

After parsing the XTF files, it was necessary to convert the acoustic signals into image signals. An SSS device emits sound waves from both sides of a moving platform, covering a broad area and receiving echo signals, which are then processed and transformed into intensity values. SSS images are created based on the variation in the grayscale intensity of the pixels within each scan line, forming a grayscale contrast [45]. Areas with stronger grayscale intensity depict the target image's geometric shapes, and the intensity of the image's grayscale directly corresponds to the amplitude variations of the echo signal, which are primarily associated with seabed topography, geomorphological features, and sediment types. The original data sampling precision was 16-bit, and sonar data imaging typically requires quantizing the received signal strength into a grayscale range of 0∼255. At this point, a model needed to be established to convert the acoustic intensity information into the grayscale information describing the image. The quantization formula for SSS data with 16-bit sampling precision is

$$G = \frac{GB - GB_{min}}{GB_{max} - GB_{min}}(G_{max} - G_{min}) \tag{1}$$

$$G = C \times e^{\frac{now-close}{far-close}} \times G \tag{2}$$

In Equation (1), G represents the quantized grayscale data; GB represents the pre-quantization echo data; G_{max} and G_{min} are the maximum and minimum values of the grayscale image, respectively, and GB_{max} and GB_{min} are the maximum and minimum values of the echo data, respectively. In Equation (2), C is a constant, and $e^{\frac{now-close}{far-close}}$ stands for the distance dimension, where the gain is normalized in the exponential domain.

Equations (1) and (2) enhance the echo intensity while converting it into grayscale levels, making it suitable for sonar data with weaker sampling echo intensity. The advantage of this technique lies in its ability to effectively compensate for the echo intensity in the far field area of the image, thus making the targets in the distant area of the image distinctly visible. Once the data were converted into image signals, the images were cropped into 256 × 256 pixel segments. These segments were then categorized into five classes based on their specific content: targets, seabed lines, seabed backgrounds, seabed terrain, and images with interference. The interference was caused by the AUV's DVL and Ultra-Short USBL, which can generate acoustic interference when their operating frequencies are harmonically related to the frequency of the SSS device. This interference appeared in the images as regular patterns of lines.

Following the aforementioned data collection and processing procedures, the final authentic dataset comprised images for each category, as indicated in Table 1. Figure 3 provides a partial visual display of the dataset.

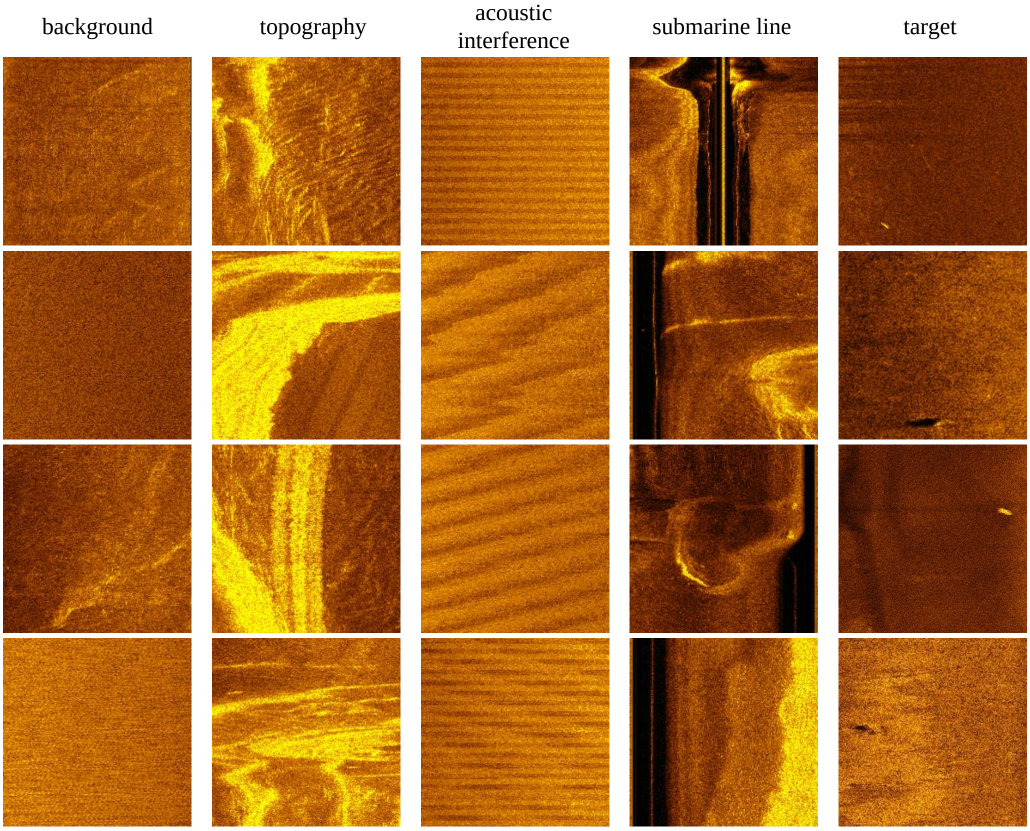


Figure 3. Five-category SSS dataset based on raw images.

Table 1. Categories and quantities of SSS dataset images.

Background	Topography	Acoustic Interference	Submarine Line	Target
1317	342	402	1712	399

3.2. SSS Diffusion Model

Diffusion models have recently become popular machine learning algorithms, demonstrating powerful representational learning capabilities. They have been widely applied in bioinformatics [46], object detection [47], and image reconstruction [48]. A diffusion model consists of a forward diffusion process and a reverse diffusion process, as shown in Figure 4. In the forward process, noise is progressively added to the initial real data, and in the reverse process, the original image is gradually recovered from the noise. Our diffusion model can be divided into two components: the objective modeling of the diffusion process and the parameter estimation network.

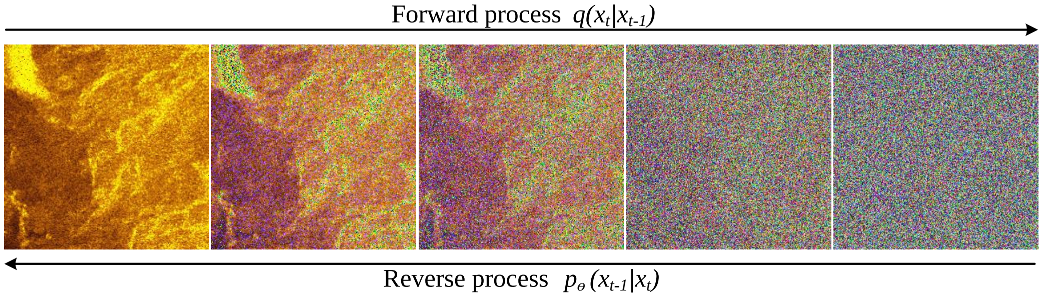


Figure 4. Schematic of the principle of diffusion model.

3.2.1. Target Modeling

Given the initial SSS raw image data ($x \sim q(x)$), the forward diffusion process consists of T steps, with each step adding Gaussian noise to the data from the previous step (x_{t-1}), and this process constitutes a Markov chain. According to the modeling of denoising diffusion probabilistic models (DDPMs) [30], the key equation for the forward process is as follows:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{3}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{4}$$

$$q(x_t|x_0) = N\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right) \tag{5}$$

In the equation, the standard deviation of the noise added at each step is determined by a fixed value, β_t, and the mean is determined by β_t and the current data, x_t, at time t. Moreover, we specify that $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod \alpha_i$.

In the reverse process, the goal is to recover the original data from the Gaussian noise. It is assumed that this is also a Gaussian distribution, but since it is not feasible to sequentially fit the distribution, a parameterized distribution is constructed for estimation. The reverse process is still a Markov chain, and its core formulas are as follows:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{i=1}^{T} p_\theta(x_{t-1}|x_t) \tag{6}$$

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t,t), \sum_\theta(x_t,t)\right) \tag{7}$$

$$q(x_{t-1}|x_t,x_0) = N\left(x_{t-1}; \tilde{\mu}(x_t,x_0), \tilde{\beta}_t I\right) \tag{8}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{9}$$

$$\tilde{\mu}_t(x_t,x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) \tag{10}$$

It can be observed that the variance in the posterior diffusion conditional probability $q(x_{t-1}|x_t,x_0)$ does not contain any unknown parameters, while the mean includes a term with a stochastic parameter, ϵ. Therefore, by continuously fitting a neural network to make distribution $p_\theta(x_{t-1}|x_t)$ as close to $q(x_{t-1}|x_t,x_0)$ as possible, $p_\theta(x_{t-1}|x_t)$ can be used for generation. It is important to note that the target of neural network estimation can be noise $\epsilon_\theta(x_t)$ or the initial x_0, or it can be predicted score $\nabla_{x_t} \log p_\theta(x_t)$ [49].

To achieve controllable category image generation, in addition to the given original data, x, there is also the corresponding category information, y. Assuming that the forward noise addition process is still a Markov chain, its data distribution is $x \sim q(x)$. According to the derivation by Dhariwal [33], it is evident that the forward process of the conditional diffusion model is completely identical to that of a DDPM, which is

$$\hat{q}(x_t|x_{t-1}, y) = \hat{q}(x_t|x_{t-1}) = q(x_t|x_{t-1}) \tag{11}$$

During the reverse denoising process, according to Bayes' theorem, the following relationship holds:

$$p(y|x_t) = \frac{p(x_t|y)p(y)}{p(x_t)} \tag{12}$$

We note that $p(y)$ serves as the prior distribution; by taking the logarithm of the above equation, deriving with respect to x_t, and then incorporating the score function, we obtain

$$\nabla_{x_t} \log p(y|x_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}(\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t)) \tag{13}$$

By substituting the aforementioned equation into Dhariwal's classifier gradient, the estimated parameters can be derived:

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t) + s[\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t)] \tag{14}$$

Thus, two models are trained: one is an unconditional generation model, like DDPMs, and the other is a conditional generation model. The final results can be obtained by using linear extrapolation from both the conditional and the unconditional generation models. The quality of the generated output can be adjusted with the guidance coefficient, s, which controls the balance between the realism and the diversity of the generated samples.

3.2.2. Parameter Estimation Network

We previously discussed controllable category generation based on diffusion models, where the core is to train a noise prediction model. Since noise and the original data are in the same dimensionality, we can opt for an autoencoder architecture [50] for the noise prediction model. Specifically, we utilized a U-Net model [51] that incorporates attention mechanisms and residual blocks, as shown in Figure 5.

The U-Net is composed of an encoder, a decoder, and their cross-layer connections. The encoder consists of four stages, each of which includes two residual blocks, a linear attention layer, and a downsampling module to reduce the size of the feature space. The linear attention mechanism was utilized because its time and memory consumption are linearly related to the sequence length, which is much more efficient than traditional attention mechanisms with quadratic complexity. After passing through the linear attention layer, the data are normalized with RMSNorm [52] to further improve computational efficiency. With each stage in the encoder, the dimensions of the height and width of the feature space are halved, while the channels are doubled. The structure of the decoder mirrors the encoder, except for the final downsampling module, which is replaced with an upsampling module to reverse the operation. In the decoder module, U-Net also introduces skip connections, which concatenate the features of the same dimensionality obtained from the encoder. This facilitates network optimization by allowing for the flow of information from previous layers to the following layers, which helps to reconstruct the finer details in the output image.

To differentiate among different time steps, inspired by the concept of positional encoding in Transformers [29], we employed sinusoidal position embeddings to encode the time (t). This enables the model to recognize which time step's noise it is predicting in the batch. Furthermore, to achieve controllable category generation, it is necessary to input the category information (y). Hence, an embedding is utilized to embed y, and during the training process, it is optimized to maximize the encoding differences among different categories. As a result, we can train a single shared U-Net model to generate images of different categories. Specifically, in the various residual blocks of U-Net, both the time (t) and category information (y) are introduced.

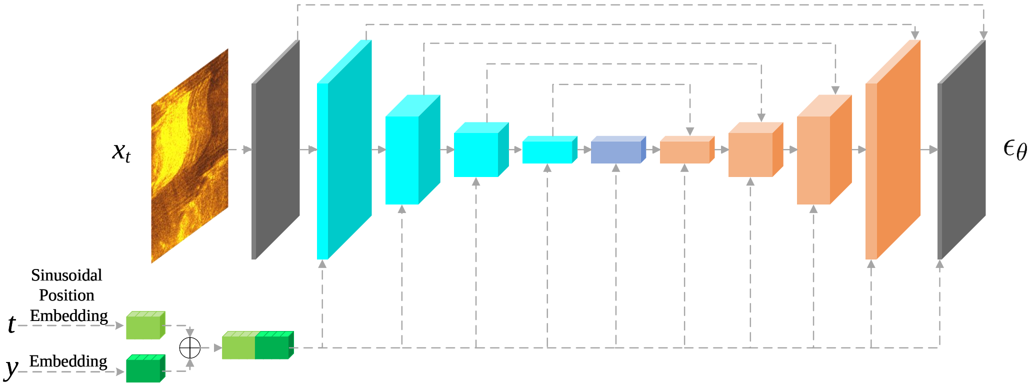


Figure 5. Schematic diagram of parameter estimation network structure of controllable category SSS image generator.

4. Experiments and Analysis

In this section, we present the experimental validation of the proposed method. First, we compared the imaging results of the SSS device, analyzing the impact of different gain factors on the nonlinear gain enhancement algorithm to determine suitable parameters. Next, we tested the proposed SSS image generator, examining the effects of different scale factors on the FID and IS. Additionally, we used Haralick texture features to further statistically assess the similarity between the real and generated images. Finally, we applied the generator to an object detection task to verify the extent to which the generator improved detection performance. For the experiments described in this chapter, the hardware used included four RTX 3090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA), 256 GB of RAM, and an Intel Xeon Silver 4210 processor (Intel Corporation, Santa Clara, CA, USA). The software environment consisted of the Ubuntu 20.04 LTS operating system, CUDA 11.6, Python 3.9, PyTorch 1.12.1, and Matlab 2018. The data in the experiment were derived from the sea trial data collected as described in Section 3.1.

4.1. SSS Imaging Results

The sonar images were obtained through the quantization of the original acoustic signals, and the quality of imaging significantly impacted the subsequent experiments. Therefore, an initial analysis of the imaging results was conducted, presenting the original image and amplitude images with different C values, as shown in Figure 6.

By parsing the original data according to the Extended Triton Format (XTF) [53] protocol and converting them into image signals by using Equation (1), we obtained the initial image, as shown in Figure 7a. It was observed that at longer propagation distances, the echo intensity decreased, resulting in darker images in distant areas, which hindered the identification of detailed information within the images. To address this issue, we applied nonlinear gain enhancement to the images based on Equation (2), experimenting with five different values of C. As shown in the remaining subfigures of Figure 7, the brightness of the images increased with larger values of C. However, the images exhibited an overexposure-like effect when $C = 4e$, leading to reduced contrast. At a C value of $2e$, the intensity in distant areas was effectively compensated, making targets in those areas more discernible. When the C value was less than e, it was difficult to identify any effective information in the images. Based on this analysis, we selected $C = 2e$ as the default value for image conversion in the subsequent experiments.

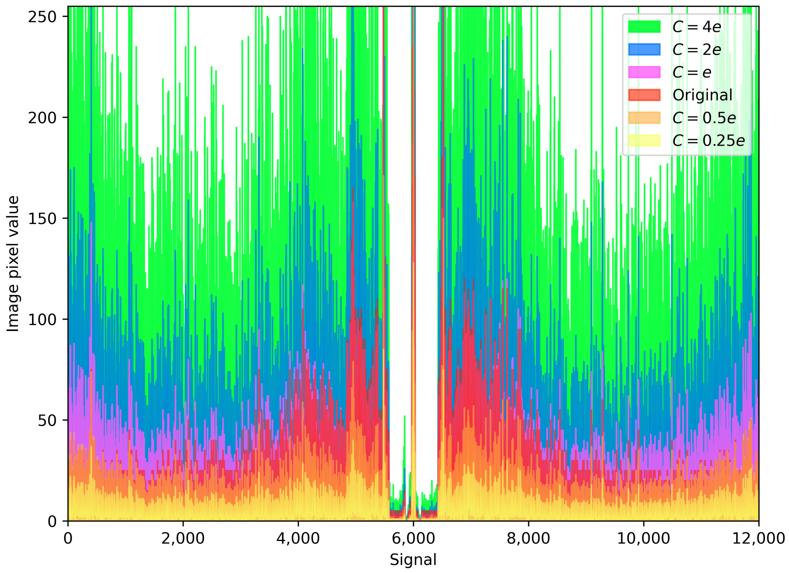
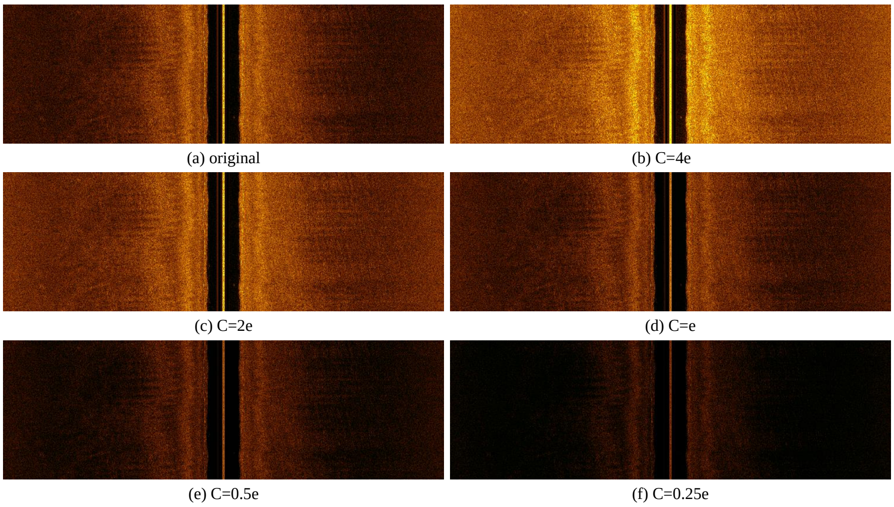


Figure 6. Amplitude comparison plots for different values of C and raw data.


(a) original

(b) C=4e

(c) C=2e

(d) C=e

(e) C=0.5e

(f) C=0.25e

Figure 7. Comparison of imaging results with different gain values.

4.2. SSS Image Generation

Based on the analysis of the original image in the previous section, training was conducted according to the method described in Section 3.2. Using the data in Table 1 as the dataset, the images were randomly selected into training and testing sets with a ratio of 0.8:0.2. After 40,000 iterations, we obtained the weight files for the model. By utilizing this trained model, we generated 20,000 images each for the experiments by using different scale factors (s); Figure 8 shows some samples of the generated images. To quantitatively assess the performance of our generative model, we first selected the evaluation metrics. In the domain of generative models, the FID and the IS [54] are commonly used as evaluation criteria. The IS evaluates the model by measuring the diversity and quality of the generated images, using a pre-trained Inception network [55] for classifying the generated images and calculating the entropy of their distribution. The FID measures the statistical difference between the generated and real images in the feature space, extracting image feature vectors with a pre-trained Inception network and calculating the Fréchet distance between these features for real and generated images. The FID not only considers pixel-level differences but also considers the overall statistical distribution in the feature space, making it more sensitive to capturing subtle differences in the generated images. However, considering that our study focused on the generation of side-scan sonar images, using the original Inception pre-trained network directly would have rendered the results meaningless. Therefore, we fine-tuned the existing pre-trained models to obtain classification networks based on InceptionV3 and ResNet [56], which were used as feature extraction networks to calculate the FID and the IS; the results are shown in Table 2.

Table 2. Generation evaluation metric scores for the model. A lower FID score and a higher IS are better.

Scale	**FID ↓**		**IS ↑**	
	InceptionV3	**ResNet**	**InceptionV3**	**ResNet**
$s = 0.0$	17.5257	27.7368	2.6612	2.5764
$s = 0.2$	17.5456	28.2699	2.8712	2.8424
$s = 0.4$	17.6471	29.3880	3.1743	3.1884
$s = 0.5$	17.5502	31.2412	3.4934	3.6395
$s = 0.6$	17.5741	32.4902	3.6551	3.9083
$s = 0.8$	17.7355	32.6387	3.7988	4.0885
$s = 1.0$	17.7328	33.2389	3.8120	4.1173
$s = 2.0$	18.2058	32.7953	3.8921	4.1597
$s = 4.0$	19.6987	33.3931	3.8971	4.2231
$s = 6.0$	20.0235	33.4207	3.9059	4.2294
$s = 8.0$	20.1990	32.9504	3.9128	4.2487

According to Equation (14), the essence of controlled category generation is achieved through the linear extrapolation of conditional and unconditional generative models. It was observed that as the value of s increased, both the FID and IS metrics exhibited a gradual increase. The increase in the FID was primarily due to the increase in the s value making the model more inclined towards conditional generation. As the s value increased, the model's controllability was enhanced and its uncertainty was reduced, leading to a decrease in the diversity of the generated images. The increase in the IS was due to the increase in the s value allowing the model to generate images of specified categories more reliably, resulting in a more balanced distribution of generated image categories, thereby increasing entropy. Additionally, it was noted that when $S = 0$, the model degenerated into an unconditional generative model; when S was between 0 and 1, the model essentially combined conditional and unconditional generative models; when S was greater than 1, it manifested as the conditional generative model minus the unconditional generative model.

After considering the FID and IS results comprehensively, we chose $S = 0.6$ as the scale factor for the generative model. Based on this, further statistical analysis was conducted on the generated images. We calculated eight Haralick texture features [57] between the real and generated images to quantify the texture differences in the two-dimensional space between image pairs. Specifically, we first established the Gray-Level Co-occurrence Matrices (GLCMs) [58] for the two sets and then calculated angular second moment, contrast, correlation, inverse difference moment, sum entropy, entropy, difference variance, and difference entropy. Subsequently, we employed the Multidimensional Scaling (MDS) method [59] to measure the texture dissimilarities in two dimensions. To minimize the impact of outliers in both sets on the overall results, we fitted the results with ellipses in a 95% confidence interval. The final results are shown in Figure 9. It can be observed that almost all ellipses representing the set of generated images encompass the ellipses representing the set of real images and their centroids are very close. This indicates that the generated images and the real images had high similarity in Haralick texture features.

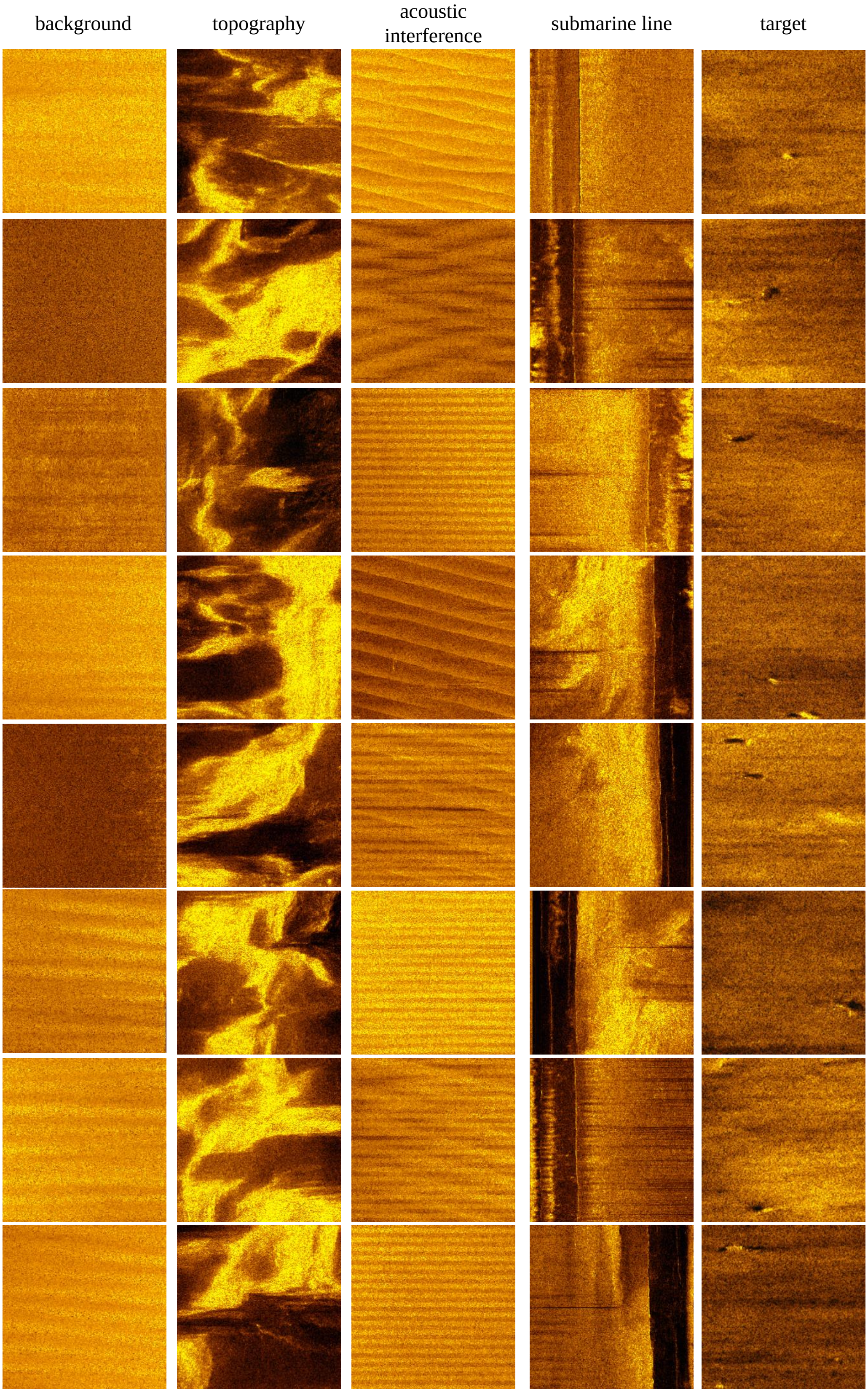
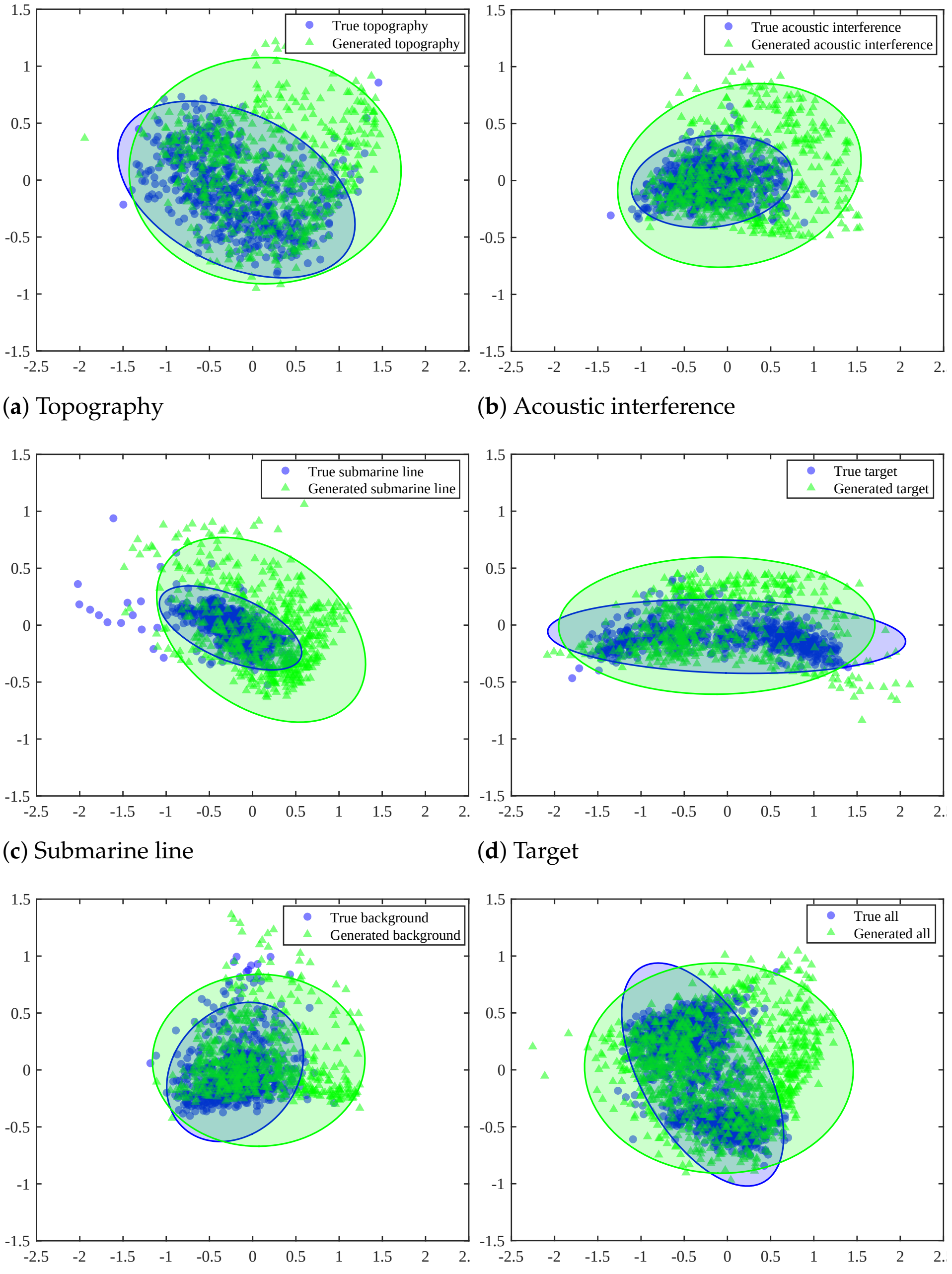


Figure 8. Generated SSS images of five categories ($s = 0.6$).



(**a**) Topography (**b**) Acoustic interference

(**c**) Submarine line (**d**) Target

(**e**) Background (**f**) All

Figure 9. Relative texture dissimilarity between generated and original data.

4.3. SSS Image Detection

An important application of generative models is supplementing the original dataset to address the issue of insufficient data volume. Therefore, in this section, we describe the use of the generated images to augment the dataset for object detection purposes. Our generative model is capable of specifically generating sonar images of five categories, among which the "target" category is the most commonly used for detection. Accordingly, we first generated a substantial number of "target" category images by using the generative model. To explore the impact of varying quantities of generated images on object detection accuracy in real images, we prepared six different datasets, as shown in Table 3. All the original data in the table were derived from the target category in Table 1. The test and validation sets for all datasets were sourced from real images in order to genuinely reflect the effect of different training sets on enhancing detection accuracy.

Table 3. The six kinds of datasets used for the object detection experiments. The data of dataset B were all real data. The training set of dataset A was entirely derived from the generated data. The training sets of the remaining datasets were mixed datasets.

Dataset	Training Set		Test Set		Validation Set	
	Original	Generated	Original	Generated	Original	Generated
A	0	1500	103	0	153	0
B	143	0	103	0	153	0
C	143	300	103	0	153	0
D	143	500	103	0	153	0
E	143	800	103	0	153	0
F	143	1000	103	0	153	0

In our object detection experiments, we utilized the state-of-the-art (SOTA) YOLOv10 detection model [60]. Due to its exceptional performance and low computational power consumption, the YOLO series [61] has been a significant paradigm in the field of object detection, now updated to its tenth version. YOLOv10 introduces a novel real-time object detection method by improving the model architecture and eliminating Non-Maximum Suppression (NMS). Based on varying computational power requirements, YOLOv10 is further divided into six distinct model variants. We conducted cross-experiments on these six model variants and six datasets under the same experimental conditions. The detection accuracy results of the experiments are presented in Table 4.

Table 4. Average precision scores for different models and different datasets.

Dataset	YOLOv10n	YOLOv10s	YOLOv10m	YOLOv10b	YOLOv10l	YOLOv10x	Average
A	0.629	0.597	0.625	0.633	0.669	0.602	0.626
B	0.723	0.718	0.743	0.739	0.728	0.777	0.738
C	0.785	0.759	0.793	0.777	0.761	0.781	0.776
D	0.838	0.813	0.851	0.850	0.831	0.834	0.836
E	0.757	0.812	0.788	0.805	0.805	0.800	0.795
F	0.815	0.820	0.801	0.802	0.806	0.805	0.808

The results for dataset A reveal that even in the absence of any real images in the training set, the average precision of the six models is still high, 0.626, further confirming the high consistency between the generated and real images. With the incremental addition of generated images, the average precision could be enhanced by up to approximately 10%, significantly improving detection accuracy. However, when more than 800 generated images were incorporated into the training set, a slight decline in the model's average detection precision was observed, but this metric was still higher than that for dataset B. This phenomenon may be attributed to the inclusion of images located in the green area but not in the blue area of Figure 9, introducing some noise. Overall, incorporating the generated images into the real dataset effectively enhanced detection accuracy, especially when the volume of added images was 3.5 times that of the original images, yielding the most pronounced effect. Furthermore, we visualized the detection results based on YOLOv10n (as shown in Figure 10). The results in the second row of Figure 10 were caused by the training set being composed entirely of real images. Five targets were missed due to insufficient data. Notice that there is one missed detection instance in the first row, but its training set was entirely derived from the generated images. According to the results from dataset C to dataset F, the model performed better in reducing missed detection instances and false alarms, achieving better detection results.

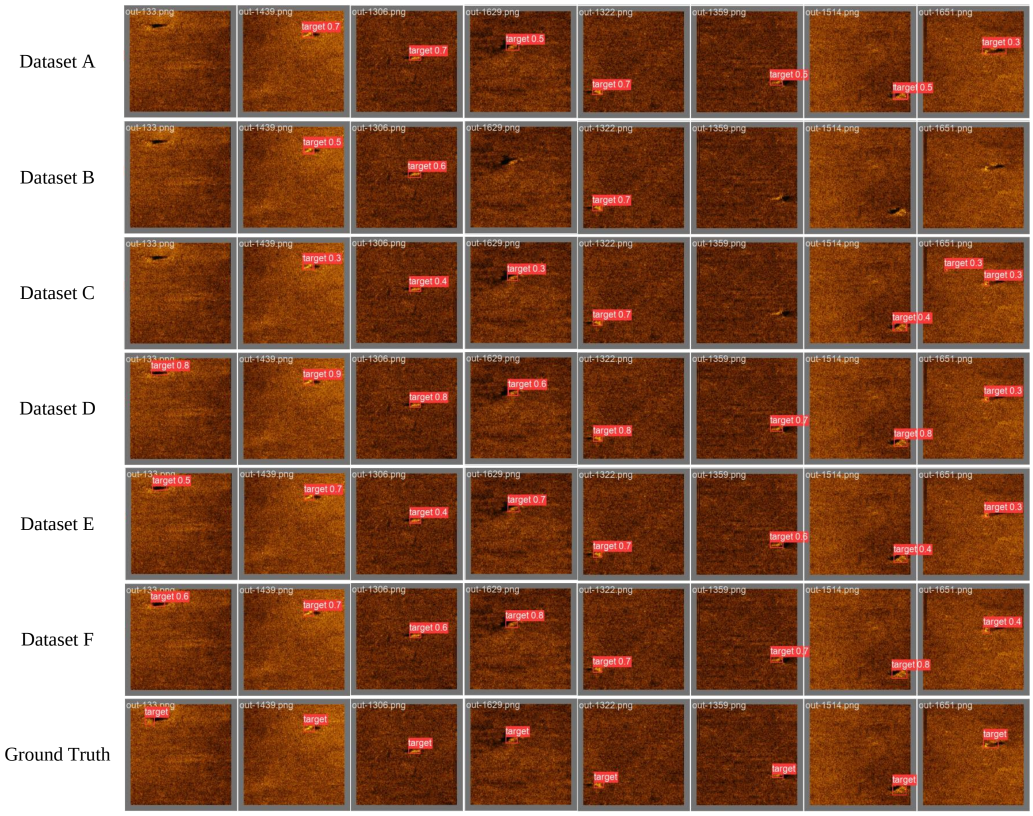


Figure 10. The detection results of the model trained on six datasets using YOLOv10n on the validation set.

5. Conclusions

In this study, we developed a controllable category SSS image generator capable of producing diverse and high-quality SSS images. To achieve this, we first established a data collection platform based on an AUV and gathered raw datasets from multiple marine areas. We also introduced a nonlinear gain enhancement algorithm for converting acoustic signals into image signals, effectively compensating for the signal strength at the far end of the sonar and enhancing the representation of target features. Based on diffusion models, we performed controllable five-category image generation without the need for additional classifiers and conducted comprehensive quantitative and qualitative evaluations. Furthermore, we validated the practical application of our generative model by using the YOLOv10 target detection algorithm. The experimental results indicate that the generated images can effectively complement the original images, improving target detection accuracy. Our work marks a significant milestone in the field of SSS data generation, providing data support for subsequent SSS-based developments.

However, it is worth noting that our current generative model only generates images and requires manual annotation when expanding the dataset. Therefore, our future research will focus on label-based generative models to achieve image–label matching generation.

Author Contributions: Conceptualization, F.Z. and X.H.; methodology, F.Z., X.H. and Z.W.; validation, Z.W., T.T. and X.H.; data collection, Z.W. and C.C.; writing—original draft preparation, X.H.; writing—review and editing, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program of China (2023QYXX).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author upon request.

Acknowledgments: We would like to acknowledge the Key Laboratory of Unmanned Underwater Transport Technology for the facilities and technical assistance provided.

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

Abbreviations

The following abbreviations are used in this manuscript:

SSS	side-scan sonar
AUV	Autonomous Underwater Vehicle
FID	Fréchet Inception Distance
IS	Inception Score
Sonar	Sound Navigation and Ranging
USBL	Ultra-Short Baseline
VAEs	Variational Autoencoders
GANs	Generative Adversarial Networks
KL	Kullback–Leibler
VQVAE	Vector-Quantized VAE
LSGAN	Least Squares GAN
RNNs	Recurrent Neural Networks
LSTM	Long Short-Term Memory
MFSANet	multi-feature fusion self-attention network
SSAM	Simplified Self-Attention Module
GPS	Global Positioning System
DDPM	denoising diffusion probabilistic model
XTF	Extended Triton Format
GLCM	Gray-Level Co-occurrence Matrices
MDS	Multidimensional Scaling
SOTA	state-of-the-art
NMS	Non-Maximum Suppression

References

1. Wang, Y.; Chu, H.; Ma, R.; Bai, X.; Cheng, L.; Wang, S.; Tan, M. Learning-Based Discontinuous Path Following Control for a Biomimetic Underwater Vehicle. *Research* **2024**, *7*, 0299. [CrossRef]
2. Huy, D.Q.; Sadjoli, N.; Azam, A.B.; Elhadidi, B.; Cai, Y.; Seet, G. Object perception in underwater environments: A survey on sensors and sensing methodologies. *Ocean Eng.* **2023**, *267*, 113202. [CrossRef]
3. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
4. Burguera, A.; Oliver, G. High-resolution underwater mapping using side-scan sonar. *PLoS ONE* **2016**, *11*, e0146396. [CrossRef]
5. Fallon, M.F.; Kaess, M.; Johannsson, H.; Leonard, J.J. Efficient AUV navigation fusing acoustic ranging and side-scan sonar. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 2398–2405.
6. Coiras, E.; Petillot, Y.; Lane, D.M. Multiresolution 3-D reconstruction from side-scan sonar images. *IEEE Trans. Image Process.* **2007**, *16*, 382–390. [CrossRef] [PubMed]
7. Rhinelander, J. Feature extraction and target classification of side-scan sonar images. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6–9 December 2016; pp. 1–6.
8. de Souza, L.A.P.; Azevedo, A.A.; da Silva, M. Side Scan Sonar Applied to Water Reservoir. In Proceedings of the 2013 IEEE/OES Acoustics in Underwater Geosciences Symposium, Rio de Janeiro, Brazil, 24–26 July 2013; pp. 1–7.
9. Tang, Y.; Wang, L.; Jin, S.; Zhao, J.; Huang, C.; Yu, Y. AUV-based side-scan sonar real-time method for underwater-target detection. *J. Mar. Sci. Eng.* **2023**, *11*, 690. [CrossRef]
10. Yan, X.; Yang, J.; Sohn, K.; Lee, H. Attribute2image: Conditional image generation from visual attributes. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part IV 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 776–791.
11. Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognit.* **2023**, *137*, 109347. [CrossRef]

12. Ehrhardt, J.; Wilms, M. Autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 129–162.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
14. Esser, P.; Rombach, R.; Blattmann, A.; Ommer, B. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3518–3532.
15. Pinaya, W.H.L.; Vieira, S.; Garcia-Dias, R.; Mechelli, A. Autoencoders. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 193–208.
16. Oord, A.V.D.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6307–6316.
17. Peng, J.; Liu, D.; Xu, S.; Li, H. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10775–10784.
18. Farnia, F.; Ozdaglar, A. Do GANs always have Nash equilibria? In Proceedings of the International Conference on Machine Learning. PMLR, Virtual Event, 13–18 July 2020; pp. 3029–3039.
19. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
20. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 214–223.
21. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
22. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
23. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
24. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
25. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
26. Van Den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 19–24 June 2016; pp. 1747–1756.
27. Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A. Conditional image generation with pixelcnn decoders. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4797–4805.
28. Salimans, T.; Karpathy, A.; Chen, X.; Kingma, D.P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv* **2017**, arXiv:1701.05517.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
30. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
31. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 8162–8171.
32. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
33. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
34. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3836–3847.
35. 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.* **2020**, *18*, 1505–1509. [CrossRef]
36. Bore, N.; Folkesson, J. Modeling and simulation of sidescan using conditional generative adversarial network. *IEEE J. Ocean. Eng.* **2020**, *46*, 195–205. [CrossRef]
37. Song, Y.; He, B.; Liu, P.; Yan, T. Side scan sonar image segmentation and synthesis based on extreme learning machine. *Appl. Acoust.* **2019**, *146*, 56–65. [CrossRef]
38. Wang, J.; Li, H.; Huo, G.; Li, C.; Wei, Y. Multi-modal multi-stage underwater side-scan sonar target recognition based on synthetic images. *Remote Sens.* **2023**, *15*, 1303. [CrossRef]
39. 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. [CrossRef]
40. Xu, H.; Bai, Z.; Zhang, X.; Ding, Q. Mfsanet: Zero-shot side-scan sonar image recognition based on style transfer. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1503105. [CrossRef]
41. 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]

42. 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. [CrossRef]
43. Cheng, C.; Hou, X.; Wen, X.; Liu, W.; Zhang, F. Small-Sample Underwater Target Detection: A Joint Approach Utilizing Diffusion and YOLOv7 Model. *Remote Sens.* **2023**, *15*, 4772. [CrossRef]
44. Zhang, P.; Tang, J.; Zhong, H.; Ning, M.; Liu, D.; Wu, K. Self-trained target detection of radar and sonar images using automatic deep learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
45. Lee, B.; Ku, B.; Kim, W.; Kim, S.; Ko, H. Feature sparse coding with coordconv for side scan sonar image enhancement. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]
46. Guo, Z.; Liu, J.; Wang, Y.; Chen, M.; Wang, D.; Xu, D.; Cheng, J. Diffusion models in bioinformatics and computational biology. *Nat. Rev. Bioeng.* **2024**, *2*, 136–154. [CrossRef]
47. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19830–19843.
48. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. Diffir: Efficient diffusion model for image restoration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 13095–13105.
49. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456.
50. Zhai, J.; Zhang, S.; Chen, J.; He, Q. Autoencoder and its various variants. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; pp. 415–419.
51. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
52. Zhang, B.; Sennrich, R. Root mean square layer normalization. *Adv. Neural Inf. Process. Syst.* **2019**, *32.*
53. Rodríguez, O.C.; Silva, A.J.; Hughes, A.P.; Moreira, A.C. Underwater Sonar as a Ray Tracing Problem. In *INCREaSE 2019: Proceedings of the 2nd International Congress on Engineering and Sustainability in the XXI Century, Faro, Portugal, 9–11 October 2019*; Springer: Cham, Switzerland, 2020; pp. 255–264.
54. Chong, M.J.; Forsyth, D. Effectively unbiased fid and inception score and where to find them. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6070–6079.
55. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
58. Partio, M.; Cramariuc, B.; Gabbouj, M.; Visa, A. Rock texture retrieval using gray level co-occurrence matrix. In Proceedings of the 5th Nordic Signal Processing Symposium, Trondheim, Norway, 4–7 October 2002; Volume 75.
59. Hout, M.C.; Papesh, M.H.; Goldinger, S.D. Multidimensional scaling. *Wiley Interdiscip. Rev. Cogn. Sci.* **2013**, *4*, 93–103. [CrossRef] [PubMed]
60. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
61. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]

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.