Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Segmentation-guided Diffusion Framework for High-Fidelity Sonar Image Generation

Zhihao Ma
Southeast University
Weiliang Meng
Institute of Automation, Chinese Academy of Sciences
Chenghanxue Tang
Southeast University
Longyu Jiang
JLY@seu.edu.cn

Southeast University

Research Article

Keywords: Sonar Image Generation, Data Augmentation, Diffusion Models, Segmentation-guided Image Generation, Hierarchical Architecture

Posted Date: October 31st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7493621/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hierarchical Segmentation-guided Diffusion Framework for High-Fidelity Sonar Image Generation

Zhihao Ma[1], Weiliang Meng[2,3], Chenghanxue Tang[1] and Longyu Jiang[*1]

[1]Southeast University, Nanjing, China.
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

Abstract

The scarcity of high-quality datasets significantly hinders sonar image research. Existing image generation methods often produce low-fidelity, detail-poor results due to uncontrollable randomness. To overcome this, we introduce the Hierarchical Segmentation-guided Diffusion Model (HSD-Sonar), a novel framework for optimized sonar image synthesis. HSD-Sonar employs a hierarchical approach: at the image level, it dynamically adjusts diffusion steps based on input characteristics and precisely controls target regions using masks, enhancing output detail and semantic consistency. At the feature level, it reinforces positional information with mask data and optimizes loss calculation via a region importance-based SG-Loss, effectively reducing background noise and improving perceptual realism. Extensive experiments on multiple public sonar segmentation datasets demonstrate HSD-Sonar's significant superiority in generating high-quality, semantically consistent sonar images. Our research offers a robust and flexible solution, advancing sonar imagery analysis and interpretation.The code is available at https://github.com/mzh529/HSD-Sonar.

[*]Corresponding author:JLY@seu.edu.cn

Keywords: Sonar Image Generation, Data Augmentation, Diffusion Models,Segmentation-guided Image Generation,Hierarchical Architecture

1 Introduction

Sonar imaging is crucial in military and civilian domains for subaqueous reconnaissance [1] [2], target detection [3] [4] [5], and navigation [6]. It supports military tasks like submarine detection and mine countermeasures, alongside civilian uses such as oceanographic exploration and marine infrastructure inspection. However, high-quality sonar image datasets are scarce due to military confidentiality and the high costs of specialized equipment, obstructing research progress.

To counter this, generative models like GANs [7] [8] and Diffusion Models [9] [10] [11] are employed. Yet, these methods are limited by the lack of high-quality, reliable outputs and their inability to accurately identify and reconstruct critical regions in sonar image features. This limitation hampers their ability to capture complex seafloor details, target echoes, and subtle textures, often introducing artifacts.

We introduce HSD-Sonar, a framework enhancing diffusion-based generation with segmentation guidance at image and feature levels. At the image level, segmentation adjusts diffusion steps in preprocessing, guides sampling for key region integrity, and refines details in post-processing. At the feature level, mask information integrates via attention weighting to focus feature extraction. Additionally, our SG-Loss uses masks to refine loss distribution, paired with LPIPS and VGG perceptual losses for improved quality.

HSD-Sonar merges diffusion model strengths with structural guidance, producing realistic, tailored sonar images for diverse applications, advancing research by overcoming data scarcity and control issues.

In summary, the main contributions of our contributions are:

- We propose HSD-Sonar, a novel Hierarchical Segmentation-guided Diffusion framework, enabling precise mask control for sonar image generation.
- We propose a three-tier progressive segmentation-guided optimization strategy at the image level to better utilize mask information for incorporating positional and other control information into the generated images.
- At the feature level, we introduce a mask information fusion strategy that integrates mask information into the prediction network via attention weighting, enabling the network to focus more effectively on the regions of interest marked by the mask.
- We design a mask-guided loss optimization strategy (SG-Loss), which dynamically optimizes loss computation based on regional importance to enhance foreground quality, reduce background noise, and maintain perceptual consistency, thereby achieving more realistic and structurally complete sonar image generation.

- We implement comprehensive testing on public sonar datasets, validating that our HSD-Sonar is superior to existing methods.

2 Related works

2.1 Basic Methods for Sonar Image Generation

Sonar image generation encompasses diverse methodological approaches. Traditional signal processing techniques like Synthetic Aperture Sonar (SAS) and beamforming achieve high-resolution imaging through spatial filtering and inverse problem-solving. Silkaitis et al. [12] and Chiang et al. [13] have significantly contributed to understanding and optimizing these techniques by improving array design and signal processing methods.

Machine learning and deep learning approaches have emerged as powerful alternatives, utilizing Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to enhance sonar image quality and enable advanced classification tasks. Notable works include [14], [15], [16], [17], and [18]. Despite these advancements, challenges persist in managing noise artifacts, resolution limitations, and reconstruction complexity.

2.2 Diffusion Models for Image Generation

Diffusion models have revolutionized image generation through their capacity to produce high-quality and diverse images. Denoising Diffusion Probabilistic Models (DDPM) [19] introduced a foundational noise-injection and reverse denoising mechanism. Subsequent improvements like Improved Denoising Diffusion Probabilistic Models (IDDPM) [20] and Score-Based Generative Models (NCSN) [21] refined sampling techniques and generation performance.

Recent innovations such as Latent Diffusion Models (LDM) [22] have reduced computational costs by operating in compressed latent spaces, while Classifier-Free Guidance (CFG) [23] enhanced conditional generation tasks. Specialized research has explored sonar image generation through various approaches [24], [25], [26], [27], [28], [29], [30], and [31].

Nonetheless, current diffusion models continue to face challenges in controllability and maintaining structural coherence during image generation.

2.3 Segmentation-guided Image Generation

Segmentation-guided image generation methods have made significant strides across multiple domains by leveraging semantic information to control image generation. Approaches like Surgery-GAN [32] integrate advanced normalization techniques to improve image quality and consistency with segmentation maps. SGGAN [33] and SGSFormer [34] have further enhanced spatial control and image reconstruction by incorporating semantic and instance-level segmentation information.

Diffusion-based methods such as SegGuidedDif [35] have demonstrated promising results in medical imaging by using masks as conditional inputs.

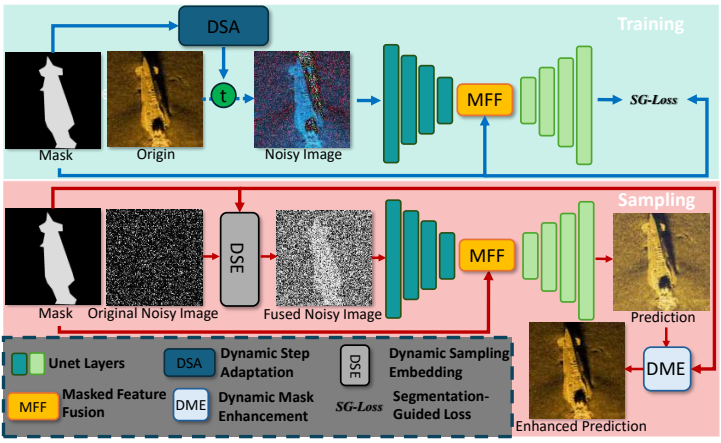


Fig. 1 The overall architecture of HSD-Sonar. The blue area in the figure represents the training phase, and the red area represents the sampling and testing phase. The gray area in the bottom left corner is the annotation area.The "t" in the training region denotes the timestep of the noise addition phase.The "Original Noisy Image" refers to the random initial noise image used in typical diffusion model sampling. The dark color in the unet layer represents the encoder part, while the light color represents the decoder part.

Additional research has explored various segmentation-guided generation approaches [24], [25], [36], [37], [38], and [39].

Despite remarkable progress, challenges remain in maintaining detail fidelity and global structural consistency throughout the generation process, indicating the need for continued research and refinement.

3 Method

3.1 Design Motivation

Our HSD-Sonar addresses five key limitations of conventional diffusion models, as shown in Fig. 1. First, diffusion steps are typically randomized to balance diversity and quality, but this lacks semantic alignment with target objects. To improve controllable generation, we introduce a Dynamic Step Adaptation(DSA)) module during preprocessing, which adaptively adjusts the corruption process based on object semantics. Second, standard sampling begins with purely random noise maps, which weakly constrain structural outputs. Existing mask-embedding methods often concatenate masks statically with noise inputs, overly restricting stochasticity. Instead, we develop a Dynamic Sampling Embedding(DSE) module that progressively integrates masks into initial noise maps, preserving randomness while ensuring spatial coherence. Third, although mask guidance improves positional accuracy, edge details and textures remain suboptimal in high-noise scenarios (e.g., sonar imaging). To address this, we give a Dynamic Mask Enhancement(DME) module, iteratively reinforcing mask-derived positional priors and edge-texture cues to suppress artifacts and enhance high-frequency fidelity. Fourth, image-level

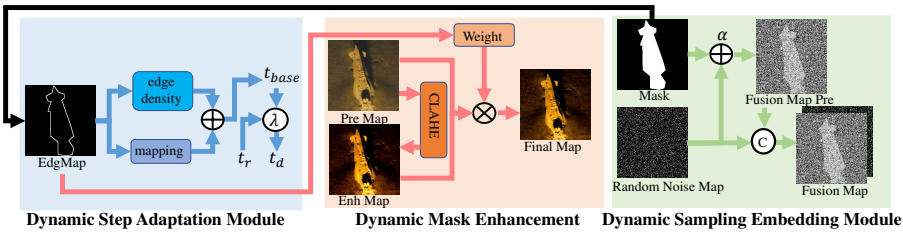


Fig. 2 Architecture diagram of the three-stage optimization module at the image level. The blue area represents Dynamic Step Adaptation (DSA), the yellow area represents Dynamic Mask Enhancement (DME), and the green area represents Dynamic Sampling Embedding (DSE). In DSA, λ denotes the fusion ratio. In DSE, α represents the weight ratio for image addition. "C" indicates concatenation.

mask fusion can effectively optimize edge information, but it has limited influence on the internal regions of the mask. To address this, we propose the Masked Feature Fusion (MFF) module, which guides feature attention toward the masked regions in the image, enhancing the model's focus on target areas.

Moreover, traditional loss functions overlook regional importance in complex scene generation, often causing an imbalance between foreground and background quality. To mitigate this, we design the Segmentation-Guided Loss(SG-Loss), which utilizes masks to dynamically adjust loss computation, enhancing foreground clarity while reducing background interference. Additionally, by integrating LPIPS and VGG perceptual loss, our SG-Loss improves texture and structural consistency, resulting in more realistic image generation.

3.2 Image-Level Approach

The image level refers to modules that operate directly on images without extracting image features. It typically focuses on the overall architecture of diffusion models, rather than just the prediction network within them.

3.2.1 Dynamic Step Adaptation

In classical diffusion models, the diffusion timesteps are randomly selected, which inevitably leads to excessive computational consumption for images that do not require many steps. Equation 1illustrates the standard noise addition process in conventional diffusion models.In this section, we propose the Dynamic Step Adaptation (DSA) module, which leverages mask information to adaptively determine the number of timesteps t.

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

As shown in the Dynamic Step Adaptation(DSA) part of Figure 2, we first extract edge texture information from the given mask to generate an edge map representing object boundaries in the image. Then, the process splits into two branches. On one hand, using the edge map, we compute the edge density, which quantifies the concentration of edge pixels within the given region. On the other hand, the edge map is mapped to a predefined range of

timesteps. The mapped value is then added to the edge density, and the result is re-scaled into a reasonable timestep range to derive the base timestep t_{base}. This mapping ensures that timestep allocation is adaptively aligned with the structural complexity of different regions in the image.

To introduce stochasticity into timestep selection, we generate a randomly perturbed timestep denoted as t_{rand}. This injection of randomness prevents the model from depending exclusively on deterministic timestep allocation, thereby improving its generalization capability. The final dynamic timestep t_{dyn} is computed by blending the base and random timesteps using a mixing coefficient λ, defined as:

$$t_{\text{dyn}} = \lambda t_{\text{base}} + (1 - \lambda) t_{\text{rand}} \tag{2}$$

We select the parameter λ based on experimental evaluation. We will conduct further tests on its value in Section 4.8.Substituting the above equation into the noise addition process yields the following:

$$\mathbf{x}_{t_{\text{dyn}}} = \sqrt{\alpha_{t_{\text{dyn}}}}\mathbf{x}_0 + \sqrt{1 - \alpha_{t_{\text{dyn}}}}\boldsymbol{\epsilon} \tag{3}$$

By substituting Equation (2) into Equation (3), we obtain:

$$\bar{\alpha} \triangleq \alpha_{\lambda t_{\text{base}} + (1-\lambda) t_{\text{rand}}} \tag{4}$$

$$\mathbf{x}_{t_{\text{dyn}}} = \sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}}\boldsymbol{\epsilon} \tag{5}$$

By comparing Equation (1) and Equation (5), we can see that the architecture of the diffusion-based noise addition process remains unchanged.; instead, we adaptively adjust the parameter t within it.This adaptive timestep selection strategy is essential for refining the noise injection process during training. Regions with high edge density, which correspond to complex textures and intricate structural details, are assigned larger timesteps, leading to stronger noise injection. This encourages the model to learn and reconstruct these critical details more effectively. Conversely, areas with low edge density, typically representing smooth or homogeneous regions, receive smaller timesteps, minimizing unnecessary perturbations and preserving their structural integrity. By dynamically adjusting timesteps based on local structural complexity, our DSA ensures a more efficient allocation of the models learning capacity, ultimately enhancing reconstruction quality and improving attention to key visual features.

To further illustrate our Dynamic Step Adaptation Module, Algorithm 1 presents the pseudocode for our HSD-Sonar training phase. Specifically, Steps 3 to 7 detail the generation of dynamic timesteps within the module. Here,The Sobel algorithm is used to extract the edge map; mean represents regularization; $t_{\text{base}} : (1, T) \leftarrow d$ represents mapping d to the range between 1 and T.

Algorithm 1 Training procedure of our HSD-Sonar

Input: mask m,adjust ratio λ
repeat
  $x_0 \sim q(x_0)$
  $e = Sobel(m)$
  $d = mean(e)$
  $t_{\text{base}} : (1, T) \leftarrow d$
  $t_{\text{dynamic}} = \lambda \cdot t_{\text{base}} + (1 - \lambda) \cdot t_{\text{random}}$
  $\epsilon \sim \mathcal{N}(0, I)$
  $m_0 \leftarrow normalize(m)$
  Take gradient descent step on

$$\mathbb{E}\left[m_0 \cdot \left(\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\ t_{\text{dynamic}}\right)\|^2\right)\right]$$

$$+\lambda_1 \cdot \text{LPIPS}(x_0, \hat{x}) + \lambda_2 \cdot \|\phi(x_0) - \phi(\hat{x})\|_1$$

until converged

3.2.2 Dynamic Sampling Embedding

As shown in the Dynamic Sampling Embedding(DSE) part of Fig 2, our DSE module first selects a dynamic mixing variable, denoted as α, to interpolate between the random noise map and the mask. The mixing process follows the equation:

$$\alpha(t) = 1.0 - t/T \tag{6}$$

where T denotes the total number of inference timesteps. The random noise map and the mask are then combined proportionally:

$$\mathbf{F}\text{pre}(t) = (1 - \alpha(t)) \cdot \mathbf{N}\text{random} + \alpha(t) \cdot \mathbf{M} \tag{7}$$

Here, $\mathbf{N}_{\text{random}}$ represents the random noise map, and $\mathbf{M}$ denotes the mask. Based on this, we concatenate the intermediate fused image with the random noise map to obtain the final fused image:

$$\mathbf{F}\text{fusion}(t) = [\mathbf{N}\text{random} \parallel \mathbf{F}_{\text{pre}}(t)] \tag{8}$$

As shown in the equations above, the mixing coefficient α gradually decreases with increasing timestep t. At the beginning of the inference process ($t = 0$), α is close to 1.0, meaning the mixed image relies heavily on the mask, strongly constraining the generation process. As t approaches the final inference step, α tends toward 0, reducing the influence of the mask while increasing the contribution of the noise map.

This dynamic adjustment allows the generation process to transition smoothly from a structure-constrained phase in the early timesteps to a more flexible and free-form synthesis in the later stages. By employing this adaptive mixing strategy, the model effectively leverages the structural information from the mask in the initial phase to guide the formation of meaningful spatial

arrangements. As noise gradually dominates, the model gains greater flexibility to refine details, thereby enhancing the realism and diversity of the generated outputs.

3.2.3 Dynamic Mask Enhancement

As shown in the Dynamic Mask Enhancement(DME) part of Fig. 2, our DME module first extracts and normalizes edge features from the mask to ensure consistency across image regions. To avoid abrupt artifacts, we apply Gaussian smoothing to reduce sharp boundary discontinuities that could affect visual coherence. Next, we use Contrast Limited Adaptive Histogram Equalization (CLAHE) [40] to enhance local contrast, improving detail while preventing excessive amplification of noise. For color images, CLAHE is applied to the luminance (L) channel in the LAB color space to maintain color balance, while for grayscale images, CLAHE is applied directly:

$$\mathbf{E} = \text{CLAHE}(\mathbf{P}) \tag{9}$$

where $\mathbf{P}$ represents the predicted map and $\mathbf{E}$ represents the enhanced map. Finally, the processed edge map is used as a weighting function to blend the enhanced region with the original image to obtain the final result map:

$$\mathbf{F} = (1 - \mathbf{M}\text{edge}) \odot \mathbf{P} + \mathbf{M}\text{edge} \odot \mathbf{E} \tag{10}$$

where $\mathbf{M}_{\text{edge}}$ represents the normalized edge map. This selective enhancement avoids over-sharpening and excessive contrast, ensuring a natural and visually appealing result. By optimizing edge-rich areas while preserving non-edge regions, our method enhances structural details with clarity and definition, avoiding unwanted artifacts.

For a detailed overview of the Dynamic Sampling Embedding Module and Dynamic Mask Enhancement Module, we present the HSD-Sonar sampling pseudocode in Algorithm 2, where steps 45 correspond to the Dynamic Sampling Embedding Module and steps 1015 to the Dynamic Mask Enhancement Module.Here,the expression: $edge_norm \leftarrow e/(max(e)+1e-6)$ represents normalizing e; GB denotes GaussianBlur, which prevents over-sharpening; reshape refers to adjusting the shape of the result from GaussianBlur to match y; y_0 corresponds to P in equation 10; y_e corresponds to E in equation 10.

3.3 Masked Feature Fusion

The aforementioned image-level methods are ineffective in influencing the internal regions of the masked target areas. To address this limitation, we propose the Masked Feature Fusion (MFF) module. As illustrated in Fig. 3, we first extract an edge map from the mask. Both the mask and the edge map are then passed through convolutional operations to obtain feature representations.

After convolution, the mask features are split into two parts: one part undergoes global spatial pooling to extract one-dimensional channel-wise

Algorithm 2 Sampling procedure of HSD-Sonar

Input: mask m
$test \leftarrow True/False$
$x_{\mathrm{T}} \sim \mathcal{N}(0, I)$
$\alpha = 1.0 - \left(\frac{t}{\mathrm{T}}\right)$
$y_{\mathrm{T}} = (1 - \alpha) \cdot x_{\mathrm{T}} + \alpha \cdot m$
for $t = T, \ldots, 1$ **do**
    $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
    $y_{t-1} = 1/\sqrt{\alpha_t}\left(y_t - (1 - \alpha_t)/\sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(y_t, t)\right) + \sigma_t z$
end for
if *test* **then**
    $e = Sobel(m)$
    $edge_norm \leftarrow e/(max(e) + 1e - 6)$
    $\mathbf{M}\text{edge} \leftarrow reshape(GB(edge_norm))$
    $y_e = CLAHE(y_0)$
    $y_0 = y_0 \cdot (1 - \mathbf{M}\text{edge}) + y_e \cdot \mathbf{M}\text{edge}$
end if
return y_0

features, while the other part applies global channel pooling to obtain two-dimensional spatial features. Similarly, the convolved edge map also undergoes global pooling to extract spatial two-dimensional features. These three feature sets are subsequently processed using self-attention weighting.

First, they are linearly projected to generate the query, key, and value matrices:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{11}$$

Next, the self-attention mechanism is applied as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(QK^\top/\sqrt{d_k}\right)V \tag{12}$$

For one-dimensional features, our self-attention mechanism aligns with the original Transformer architecture [41]. However, for two-dimensional features, we adopt the approach found in Vision Transformers (ViT) [42]. After attention weighting, all three feature sets are expanded back to their original spatial and dimensional shapes, then summed to create a weighted feature representation. This representation is then integrated into the prediction network's (UNet) intermediate layers via a residual connection, allowing for modulation of intermediate features.

In this way, the spatial information from the masked target regions is effectively fused into the prediction network via attention-based weighting. The inclusion of edge map features in this fusion process further enhances boundary localization, reinforcing region-aware feature refinement.

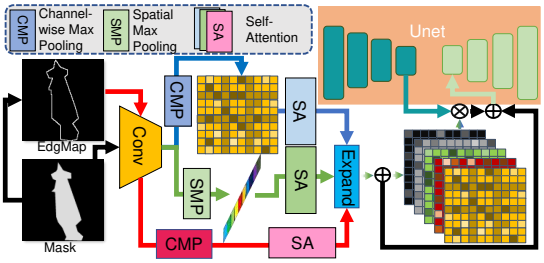


Fig. 3 Architecture of the Masked Feature Fusion (MFF) module. The blue, green, and red lines indicate spatial attention weighting of the original image, channel attention enhancement, and edge information weighting, respectively.

3.4 Segmentation-Guided Loss

In generative model training, the design of the loss function directly impacts the quality of the generated results. To mitigate the learning bias of traditional methods when handling complex scenesespecially in distinguishing important regions from backgrounds, we design the Segmentation-Guided Loss (SG-Loss). By leveraging segmentation information from the target scene, our SG-Loss dynamically adjusts loss computation based on the perceptual significance of different regions, enhancing structural fidelity and perceptual quality.

Traditional diffusion models typically rely on global Mean Squared Error (MSE) loss, which fails to account for perceptual differences across various scene regions, often leading to imbalanced foreground and background generation. To address this, our SG-Loss extends MSE loss by integrating LPIPS [43] perceptual loss and VGG feature perceptual loss, incorporating the mask m of the target scene. Through dynamic weight adjustments, it refines the generation process. The mathematical formulation is as follows:

$$\begin{aligned} L = \mathbb{E}\Big[m_0 \cdot \big(\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\ t_{\text{dynamic}}\big)\|^2\big)\Big] \\ + \lambda_1 \cdot \text{LPIPS}(x_0, \hat{x}) + \lambda_2 \cdot \|\phi(x_0) - \phi(\hat{x})\|_1 \end{aligned} \tag{13}$$

where ϵ_θ and ϵ denote the model-generated and real noise, respectively, while $\hat{x}$ and x represent the generated and real images. The mask m is normalized to obtain m_0, which values in the range $[0, 1]$, assigns different loss weights to different regions, amplifying the loss contribution in important areas (e.g., the target foreground) while suppressing it in the background. The function $\phi(\cdot)$ extracts feature representations using a VGG network to measure perceptual differences in feature space. The final "1" in the formula represents the L1 norm.The hyperparameters λ_1 and λ_2 control the balance between LPIPS loss and VGG perceptual loss.

Compared to traditional uniformly weighted MSE loss, our SG-Loss offers several advantages. By assigning higher weights to foreground regions, it improves clarity and structural integrity in key target areas, leading to more precise foreground generation. Simultaneously, lower weights in background

regions reduce the influence of irrelevant information, mitigating artifacts and visual distortions. Additionally, integrating LPIPS and VGG perceptual losses helps preserve texture and structural consistency, enhancing the realism and naturalness of the generated images.

4 Experiments

4.1 Datasets

We utilized multiple publicly available sonar image datasets to validate the versatility and generalizability of the proposed HSD-Sonar method across various sonar imaging modalities. Specifically, the following datasets were used:

Marine Debris Forward-Looking Sonar Dataset (FLSMDD) [44]: This dataset, collected using the ARIS Explorer 3000 sonar in a controlled water tank, comprises 1,868 images. The images are resized to 320×480 pixels and annotated into 11 classes (excluding the background).

Seafloor Salvage Side-Scan Sonar Dataset (SSSD) [45]: A custom-labeled dataset designed for sonar image segmentation tasks, containing 1,200 side-scan images primarily featuring targets such as shipwrecks, aircraft, and human remains.

AUV Navigation Narrow-Beam Side-Scan Sonar Dataset (NNSSS) [46]: This dataset consists of 9 narrow-beam side-scan images in a long-strip format, categorized into three classes: rocks, sand, and others.

Waterfall Stream AUV Navigation Side-Scan Sonar Dataset (WSSSS) [47]: Following segmentation using the waterfall stream method, this dataset comprises 152 images categorized into five classes: water, mountains, land, shadow areas, and unlabeled regions (background).

AI4Shipwrecks Shipwreck Side-Scan Sonar Dataset (AI4Shipwrecks) [48]: Collected from Lake Huron, this dataset includes 286 images, divided into two classes: shipwrecks and background, with only 161 images containing shipwreck targets.

Although NNSSS contains a limited number of images, WSSSS has undergone waterfall processing, and some images in AI4Shipwrecks lack target objects, the three major datasetsNNSSS, WSSSS, and AI4Shipwrecksstill hold significant importance in their respective fields due to their high authority and accuracy.Specifically, the NNSSS dataset, provided by the Spanish National Research Council's Department of Marine Technology in collaboration with Heriot-Watt University experts and collected using real underwater equipment, guarantees its exceptional authority and professionalism. The WSSSS dataset was acquired by a team from Northwestern Polytechnical University's College of Oceanography and Civil Engineering using their independently developed AUV and a HYDRO 3060 side-scan sonar system in real environments; it underwent meticulous manual annotation and data augmentation, fully demonstrating its professionalism and practical utility in underwater image segmentation research. Furthermore, the AI4Shipwrecks dataset, a collaboration between teams from Michigan Technological University and the University

of Michigan with support from the NOAA Ocean Exploration Program, had its data systematically surveyed in the Thunder Bay National Marine Sanctuary using advanced sonar-equipped AUVs. The annotation process involved consultation with marine archaeology experts and thorough review by the research team, making it the first publicly available pixel-level segmentation sonar image dataset of underwater shipwrecks with immense professional and practical value. Therefore, we've chosen to utilize these datasets to further validate and showcase our model's outstanding generalization capabilities in real-world environments.

In the experiments, all datasets were split into training and testing sets with an 8 : 2 ratio. Through these datasets, we systematically evaluated the robustness and generalizability of HSD-Sonar across diverse sonar imaging scenarios.

4.2 Implementation Details

We employ a U-Net-based architecture as the core of our diffusion model, with input images standardized to 256×256 pixels. The implementation is built on the PyTorch framework, and all training and testing are performed on a single NVIDIA A6000 GPU with 48GB VRAM.

Training is conducted with a learning rate of 1×10^{-5}, a batch size of 16, and an evaluation batch size of 8. We optimize the model using the Adam optimizer, with momentum parameters set to $(0.9, 0.999)$.Since our experiments involve multiple datasets with varying numbers of images, we did not train for a fixed number of epochs. Instead, we adopted a unified training-time strategy: all models were trained for 15 hours to ensure convergence across datasets. For datasets with fewer images, each epoch finishes quickly, but more epochs are required due to slower convergence. Conversely, for datasets with more images, each epoch takes longer, but convergence is typically faster. As a result, all models are able to converge within the same total training time.

4.3 Visualization of comparative results

To validate the effectiveness of our HSD-Sonar, we conducted comparative experiments on five datasets, selecting six mask-guided image generation networks for comparison: SPADE [49], ControlNet [50], Surgery-GAN [32], SGSFormer [34], SegGuidedDif [35], and our HSD-Sonar. These networks employ different core mechanisms, mask guidance types, and levels of model complexity, as detailed in Fig. 1.

4.3.1 Comparison on the FLSMDD

As shown in Fig. 4, the FLSMDD images are grayscale and relatively blurry, which leads to generally mediocre generation performance across different networks. The red circles in the figure highlight comparisons between boundary regions and internal features of the targets. For instance, in the third row, the features of the bottle clearly demonstrate that our HSD-Sonar produces

Table 1 Analysis of the networks used for subsequent comparative experiments. Here, "Yr" denotes the publication year, "Mask G. Type" indicates the type of mask guidance employed, "Model Para." represents the model parameters, and "Core Mech." stands for the core mechanism.

Method	Yr	Mask G. Type	Model Para.(M)	Core Mech.
SPADE [49]	2019	Single-layer	78	Convolution
SGSFormer [34]	2021	Single-layer	135	Transformer
ControlNet [50]	2023	Single-layer	257	Diffusion Model
Surgery-GAN [32]	2024	Single-layer	118	Convolution
SeaGuidedDif [35]	2024	Single-layer	167	Diffusion Model
Our HSD-Sonar	**2025**	**Multi-layer**	**210**	**Diffusion Model**

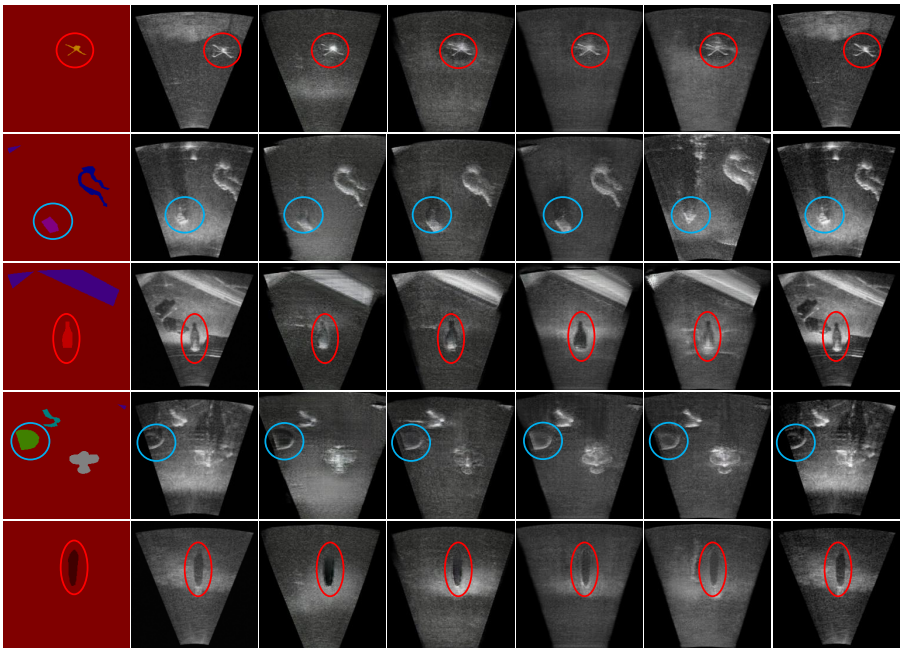


Fig. 4 Visual comparison on the Marine Debris Forward-Looking Sonar Dataset (FLSMDD): (a) Mask, (b) Input image, (c) SGSFormer [34], (d) Surgery-GAN [32], (e) ControlNet [50], (f) SegGuidedDif [35], and (g) Our HSD-Sonar. The areas highlighted in red demonstrate the alignment between the edges and the mask. In contrast, the regions encircled in blue illustrate the consistency of internal texture features within the mask when compared to the corresponding areas in the original image.

results that align more closely with the mask and the original image, especially along the edges. Similar comparisons suggest that HSD-Sonar is capable of generating sonar images that better match both the mask boundaries and the original content, even when dealing with blurry grayscale forward-looking sonar (FLS) images.



Fig. 5 Visual comparison on the Seafloor Salvage Side-Scan Sonar Dataset (SSSD): Both (a)(g) and the circles share the same meanings as defined in Fig 4.

4.3.2 Comparison on the SSSD

SSSD consists of colored sidescan sonar images that are generally clearer, leading to better overall generation performance. As shown in Fig. 5, the red-circled areas highlight the effectiveness of our HSD-Sonar. In the first and third rows, the generated target boundaries are more closely aligned with both the mask and the original image. In the second, fourth, and fifth rows, our HSD-Sonar benefits from the feature-level attention weighting introduced by the Masked Feature Fusion(MFF) module, which allows the model to focus more on the internal features within the mask. As a result, the generated content more accurately reflects the original image. In summary, HSD-Sonar maintains strong controllability and high-quality generation even when applied to clearer sidescan sonar image datasets.

4.3.3 Comparison on the NNSSS

NNSSS is a typical horizontally elongated image format, with only 9 images in the dataset. Normally, such irregularly shaped images are better handled by first cropping them before processing. However, to fairly compare the performance of different networks, we treated them as standard images and resized the outputs back to their original dimensions after generation. As shown in Fig. 6, even when dealing with irregular image structures, our HSD-Sonar



Fig. 6 Visual comparison on the AUV Navigation Narrow-Beam Side-Scan Sonar Dataset (NNSSS):Both (a)(g) and the circles share the same meanings as defined in Fig 4.The relative positions of the targets within the red circles in the figure indicate that the results generated by our HSD-Sonar are the most closely aligned with the mask positions.

still produces results that align more closely with both the original images and the masks in many detailed regions, demonstrating superior precision and controllability.

4.3.4 Comparison on the WSSSS

WSSS is a sidescan sonar dataset processed using a waterfall flow method. As shown in Fig. 7, the images in this dataset are relatively abstract, leading to significant differences between the generated content and the original images across all networks. However, the generated images still maintain strong correlations with the corresponding masks. In this regard, our HSD-Sonar demonstrates superior consistency between the generated object boundaries and the masks, reflecting more precise control. This indicates that even when faced with abstract and structurally complex waterfall-style sonar images, HSD-Sonar is capable of delivering more controlled and accurate generation results.

4.3.5 Comparison on the AI4Shipwrecks

The AI4Shipwrecks dataset contains a large number of target-free images and varies in original image size. The presence of masks in target-free samples can interfere with the generation process, especially for networks that rely on mask guidance. To ensure a fair comparison, we selected the first two rows to represent target-free generation results and the last three rows for target-containing examples, while resizing all images to a uniform resolution. As shown in Fig. 8, it is difficult to evaluate the quality of generation for target-free images across all networks. However, for images with clear targets, our HSD-Sonar consistently outperforms other networks in terms of detail reconstruction and boundary alignment, demonstrating superior generative quality and controllability.

4.4 Quantification of comparative results

To comprehensively assess the performance of different methods, we adopted two widely used image quality evaluation metricsFrchet Inception Distance

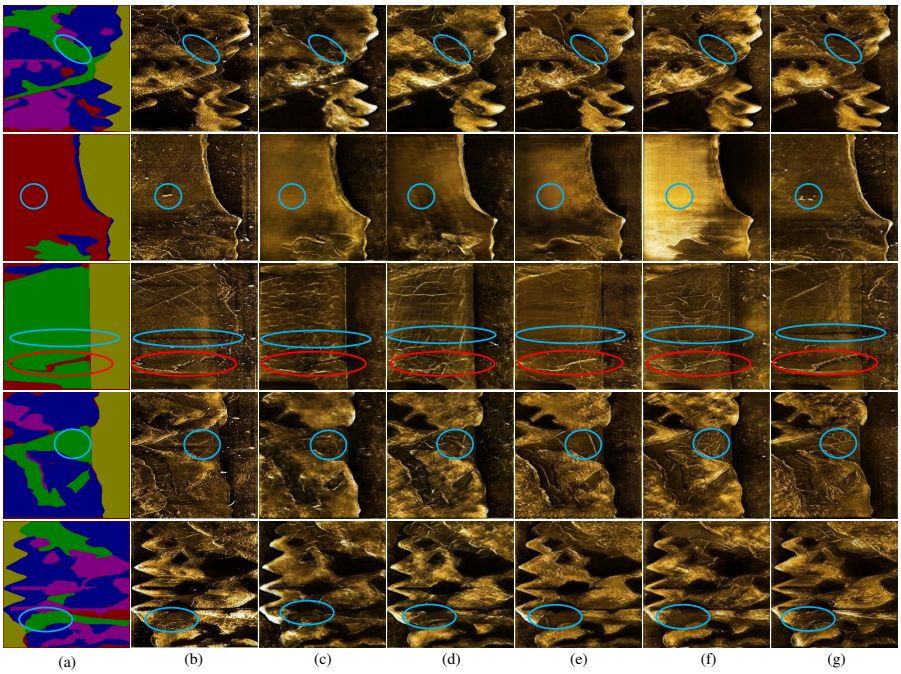


Fig. 7 Visual comparison on the Waterfall Stream AUV Navigation Side-Scan Sonar Dataset (WSSSS): Both (a)(g) and the circles share the same meanings as defined in Fig 4.

(FID) and structured FID (sFID)to quantify the distributional differences between generated and real images. In addition, we introduced Precision (Pre.) to evaluate the realism of generated images, and Recall (Rec.) to assess their diversity.

As shown in Table 2, performance varies across the NNSSS, WSSSS, and AI4Shipwrecks datasets, which can be attributed to limitations such as limited sample size, cascading processing pipelines, and missing target objects in some masks. Nevertheless, our proposed HSD-Sonar method consistently achieves the best results across all five datasets, significantly outperforming other state-of-the-art approaches in both FID/sFID and Precision/Recall metrics.

Specifically, on the FLSMDD dataset, HSD-Sonar achieved a FID score of 101.49, representing a 28.4% reduction compared to SegGuidedDif (FID=141.67). On the SSSD dataset, the FID improved even more significantly, with a 32.9% reduction (from 55.29 to 37.12). Furthermore, on the AI4Shipwrecks dataset, HSD-Sonar improved the sFID by 16.5% (from 61.528 to 51.392), and showed increases of approximately 23.6% in Precision and 26.0% in Recall, highlighting its overall advantage in both realism and diversity.

In addition, we further employed PSNR and SSIM, two traditional image reconstruction metrics, to evaluate the fidelity and structural similarity of the generated images at the pixel level. PSNR primarily reflects differences

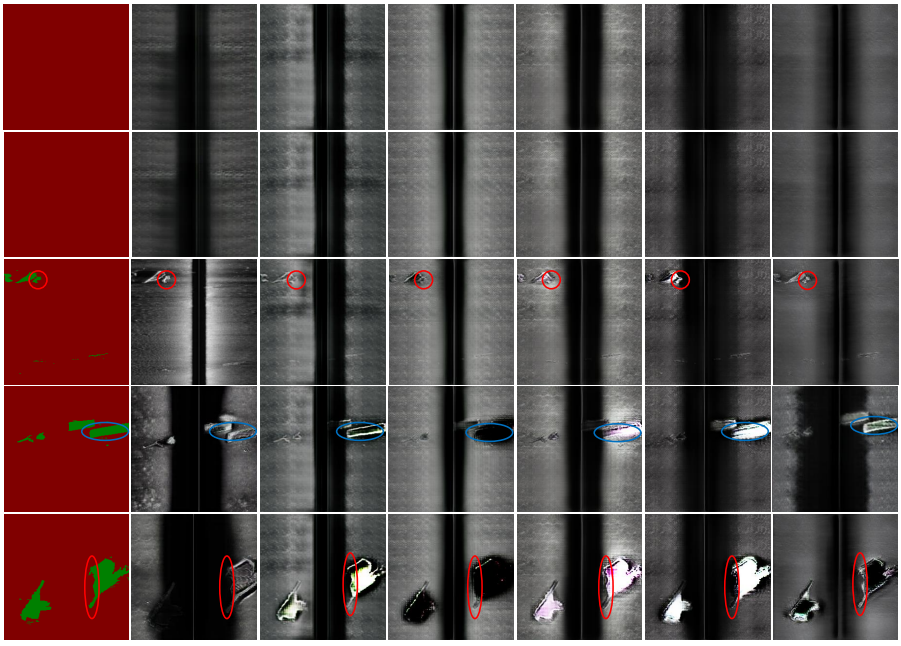


Fig. 8 Visual comparison on the AI4Shipwrecks Shipwreck Side-Scan Sonar Dataset (AI4Shipwrecks): The masks in the first two rows of the figure contain no targets, while the last three rows include shipwreck targets.

in brightness between the generated and real images, while SSIM focuses more on structural, textural, and contrast-level consistency. These indicators complement the performance evaluation from both detail and structural perspectives.

As presented in Table 3, HSD-Sonar outperforms all competing methods in terms of PSNR and SSIM across all five datasets. For instance, on the FLSMDD dataset, HSD-Sonar achieved a PSNR of 21.85 dB, a 5.5% improvement over SegGuidedDif (20.71 dB), and improved SSIM by 14.5% (from 0.55 to 0.63). On the SSSD dataset, PSNR and SSIM increased by 6.8% and 12.2%, respectively. Particularly in the challenging NNSSS and AI4Shipwrecks datasets, SSIM improvements reached 19.% and 6.3%, respectively, demonstrating the superior capability of HSD-Sonar in reconstructing fine-grained structures.

These results further validate the effectiveness and robustness of our hierarchical segmentation-guided mechanism in the task of high-fidelity sonar image generation.

Given the limitations of the NNSSS, WSSSS, and AI4Shipwrecks datasetssuch as inconsistent annotations, insufficient sample sizes, and missing labelsour subsequent experiments will focus on the more mature and widely adopted FLSMDD and SSSD datasets. These datasets are well-recognized

within the research community and offer a more representative and comparable basis for evaluating sonar image generation and processing techniques. By focusing on these core datasets, we aim to more accurately assess model performance under generalizable conditions, thereby enhancing the relevance and credibility of our findings for both current and future research.

Method	FLSMDD		SSSD		NNSSS		WSSSS		AI4Shipwrecks	
	FID↓/sFID↓	Prec.↑/Rec.↑	FID↓/sFID↓	Prec.↑/Rec.↑	FID↓/sFID↓	Prec.↑/Rec.↑	FID↓/sFID↓	Prec.↑/Rec.↑	FID↓/sFID↓	Prec.↑/Rec.↑
SPADE [49]	35.03/1.253	0.123/0.044	15.82/0.258	0.296/0.138	89.13/4.538	0.099/0.038	185.29/8.995	0.014/0.007	133.47/6.765	0.040/**0.039**
Surgery-GAN [32]	15.50/0.504	0.391/0.196	6.84/0.081	0.574/0.420	81.70/3.817	0.110/0.051	163.91/7.561	0.025/0.014	119.00/6.033	0.050/0.019
SGSFormer [34]	22.34/0.767	0.265/0.106	10.52/0.132	0.453/0.322	96.33/4.900	0.087/0.032	193.02/9.760	0.019/0.010	141.03/7.170	0.031/0.016
ControlNet [50]	27.37/1.034	0.185/0.070	12.72/0.171	0.349/0.204	103.84/5.301	**0.138**/0.024	200.00/10.500	0.011/0.006	148.58/7.648	0.026/0.013
SegGuidedDif [35]	10.49/0.291	0.535/0.421	4.09/0.043	0.801/0.639	66.75/3.054	0.062/0.033	**75.85/3.391**	0.038/0.021	88.99/4.555	**0.073**/0.032
Our HSD-Sonar	**7.51**/**0.130**	**0.681**/**0.558**	**2.75**/**0.035**	**0.812**/**0.673**	**53.38**/**2.099**	0.074/**0.073**	100.00/4.762	**0.049**/**0.026**	**66.26**/**3.398**	0.061/0.025

Table 2 Compare the performance of different methods in terms of FID, sFID, Precision (Pre.), and Recall (Rec.) on five datasets. The best values are highlighted in **red**.

Method	FLSMDD		SSSD		NNSSS		WSSSS		AI4Shipwrecks	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SPADE [49]	18.73	0.41	22.58	0.21	10.16	0.17	20.95	0.15	16.27	0.29
Surgery-GAN [32]	27.71	0.63	34.14	0.50	13.63	0.29	23.85	0.26	19.24	0.45
SGSFormer [34]	24.19	0.54	29.81	0.38	11.92	0.23	22.85	0.23	18.23	0.39
ControlNet [50]	21.79	0.47	26.23	0.29	11.12	0.20	22.05	0.20	17.30	0.35
SegGuidedDif [35]	30.49	0.83	37.47	0.62	14.80	0.32	24.79	0.27	20.46	0.48
Our HSD-Sonar	**32.17**	**0.95**	**40.00**	**0.69**	**15.87**	**0.38**	**25.19**	**0.29**	**21.32**	**0.51**

Table 3 Comparison of different methods in terms of PSNR and SSIM on five datasets. Best results are highlighted in **red**.

4.5 Mask Fidelity Comparison

To evaluate the fidelity of the generated images with respect to the original semantic masks, we adopted the evaluation strategy used in SegGuidedDif [35]. As shown in Table 4, we employed the Dice Similarity Coefficient (DSC) to measure how well the semantic information in the generated images aligns with the input masks or real images.

Specifically, $\text{Dice}(m^{\text{pred}}\text{gen}, m)$ reflects the fidelity between the semantic structure in the generated image and the guided mask. $\text{Dice}(m^{\text{pred}}\text{gen}, m^{\text{pred}}_{\text{real}})$ measures the consistency of the semantic structure between the generated image and the real image.

We used the MONAI UNet as the segmentation network, with a 1-channel input and an output of (number of target object classes + 1) channels. The sequence of intermediate UNet channels was configured as (16, 32, 64, 128, 256). Each model was trained for 100 epochs with a batch size of 8, and the model with the lowest validation loss was selected. The training used an initial learning rate of 10^{-3} and a cosine annealing learning rate scheduler.

As shown in the table, our proposed HSD-Sonar achieves the highest fidelity to both the input masks and the semantic content of real images, demonstrating its superior ability to preserve and reproduce semantic structures.

Model	FLSMDD		SSSD	
	$\text{Dice}(m_{\text{gen}}^{\text{pred}}, m)$	$\text{Dice}(m_{\text{gen}}^{\text{pred}}, m_{\text{real}}^{\text{pred}})$	$\text{Dice}(m_{\text{gen}}^{\text{pred}}, m)$	$\text{Dice}(m_{\text{gen}}^{\text{pred}}, m_{\text{real}}^{\text{pred}})$
ControlNet [50]	0.2135	0.2247	0.2351	0.2482
Surgery-GAN [32]	0.3614	0.3721	0.3845	0.3927
SPADE [49]	0.6238	0.5632	0.7342	0.7286
SGSFormer [34]	0.6478	0.5879	0.7598	0.7524
SegGuidedDif [35]	0.6777	0.6328	0.8114	0.7601
Our HSD-Sonar	**0.7577**	**0.6828**	**0.8614**	**0.8001**

Table 4 Mask fidelity comparison. $\text{Dice}(m_{\text{gen}}^{\text{pred}}, m)$ represents the Dice score between the predicted mask from the generated image and the original input mask. $\text{Dice}(m_{\text{gen}}^{\text{pred}}, m_{\text{real}}^{\text{pred}})$ represents the Dice score between the predicted masks of the generated image and the real image.

4.6 Ablation Study

We conducted ablation experiments to validate the individual contributions of each module within our HSD-Sonar framework. The experimental results are presented in Table 5.

Firstly, the removal of Segmentation-Guided Loss(SG-Loss) significantly impacted most evaluation metrics. This highlights its crucial role in guiding the model to generate high-quality segmentation results. SG-Loss integrates segmentation information into the training process, ensuring that generated images are semantically consistent with the target content, thereby preventing blurry or semantically incoherent outputs. Its removal leads to a substantial drop in metrics, demonstrating that segmentation guidance is indispensable for our synthesis task.

Secondly, the removal of Dynamic Sampling Embedding(DSE) led to a notable drop in the Recall metric, although its effect on other metrics was secondary. Upon deeper analysis, we found that DSE dynamically adjusts the sampling strategy during generation, enabling the model to more effectively explore and produce diverse samples, especially those representing less common or complex features within the target data distribution. When DSE is removed, the model may tend to generate more "average" or "common" samples, thus missing out on a wider range of possibilities within the data distribution. This directly leads to a decrease in recall. In essence, DSE enhances the model's ability to capture the complexity and diversity of real-world data, ensuring that generated samples cover a broader spectrum of the target distribution.

Masked Feature Fusion(MFF) showed a more pronounced effect on the FID and sFID metrics. MFF introduces a masking mechanism during the feature fusion stage, ensuring that information from different feature layers can be selectively integrated. This better preserves detailed information and suppresses potential artifacts. This is crucial for the quality and realism of generated images, as FID and sFID primarily measure the similarity and diversity between generated and real images. MFF's effectiveness lies in its ability to more precisely fuse contextual and local information, thereby generating higher-quality and perceptually more realistic images.

Dynamic Mask Enhancement(DME) primarily affected the Precision metric. DME's role is to dynamically optimize the quality and accuracy of the masks, ensuring that the generation process can more precisely focus on critical regions of the image. This refined mask processing enables the model to generate semantically more accurate images with sharper boundaries, thereby reducing inaccurate or blurry regions and directly improving the precision of the generated results.

Finally, the impact of Dynamic Step Adaptation(DSA) was relatively small, but its main role is to accelerate network convergence. DSA dynamically adjusts the optimization step size based on the training state, allowing the model to reach an optimal solution more quickly without sacrificing final performance. Although its improvement on the ultimate evaluation metrics is not significant, in practical training, DSA substantially reduced training time and improved training efficiency, which is particularly important for training large models.

In conclusion, while each module affects different metrics to varying degrees and focuses, our in-depth experimental analysis reveals that each plays an indispensable role within the HSD-Sonar framework. Each module contributes uniquely and critically to the overall performance from different aspects (e.g., semantic consistency, sample diversity, image quality, local precision, and training efficiency), collectively ensuring the superior performance of the framework.

Variant	FLSMDD				SSSD			
	FID	sFID	Prec.	Rec.	FID	sFID	Prec.	Rec.
w/o DSA	117.35	2.380	0.5721	0.4390	48.72	0.570	0.6835	0.5304
w/o DSE	198.63	5.148	0.3550	0.1263	92.45	1.032	0.4126	0.1879
w/o DME	128.76	2.825	0.5105	0.3924	57.30	0.623	0.6379	0.5015
w/o MFF	163.82	3.954	0.4621	0.2835	76.41	0.892	0.5238	0.3642
w/o SG-Loss	212.47	6.239	0.2964	0.1074	121.56	1.485	0.3724	0.1752
Complete	**101.49**	**1.763**	**0.6017**	**0.4928**	**37.12**	**0.467**	**0.7523**	**0.6234**

Table 5 Ablation study conducted on FLSMDD and SSSD.

4.7 SG-Loss Effectiveness Validation

To validate the effectiveness of Segmentation-Guided Loss (SG-Loss), we trained our model for 300 epochs on the FLSMDD dataset using both the traditional MSE loss and our proposed SG-Loss. As depicted in Figure 7, the purple curve (representing SG-Loss) exhibits a smoother descending trend in the early stages of training and reaches a lower, more stable value in the later stages. This indicates a clearer optimization direction and more efficient convergence. The cyan curve (representing MSE loss) shows greater fluctuations in the later stages, likely due to interference from background noise or less significant regions.This demonstrates that SG-Loss leverages the mask m to weight perceptually salient areas, thereby focusing more on important regions and effectively reducing interference from irrelevant areas.

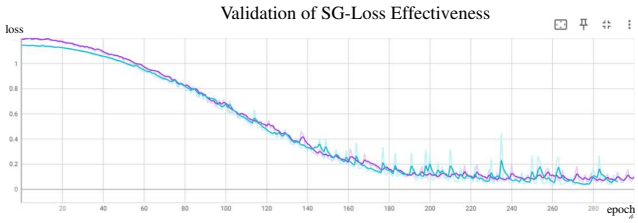


Fig. 9 Validation of SG-Loss Effectiveness. The purple curve represents training with SG-Loss, while the cyan curve represents training with traditional MSE loss. The x-axis denotes the number of epochs, and the y-axis indicates the loss value.

Furthermore, the purple curve is generally more stable with no prominent abnormal spikes, indicating that SG-Loss successfully guides the model to concentrate on critical regions and suppresses background instability. The frequent spikes observed in the cyan curve may reflect the model being hindered by less important regions during the optimization process. This further confirms that by concentrating loss weights on the foreground, SG-Loss helps to reinforce the structural fidelity of target regions during training.

4.8 Adjusting the ratio λ

To determine the optimal value of λ and to validate the efficacy of Dynamic Step Adaptation(DSA) in accelerating network convergence, we conducted a series of comparative experiments for λ. As shown in Fig 10, on the FLSMDD dataset, when λ=0, the network relied entirely on random timesteps, resulting in the slowest convergence at 243 epochs. This indicates that timesteps lacking semantic guidance lead to inefficient exploration of the search space by the model. Conversely, when λ=1, the network depended solely on deterministic timesteps from mask mapping, which also led to suboptimal convergence speed, revealing that excessive reliance on deterministic information can limit the model's exploratory capabilities. The experimental results further demonstrated that: lower values of λ caused greater fluctuations in the loss function during training and an unstable convergence path; whereas higher values of λ yielded a more stable training process but potentially slowed down convergence in later stages. Overall, intermediate λ values, particularly λ=0.8, achieved the best balance between stability and convergence speed.

For the SSSD dataset, characterized by its high image quality, each sample provides richer information, enabling the model to learn efficiently from fewer samples. Consequently, the optimal λ value that facilitated the fastest convergence on this dataset was adjusted to 0.7.

4.9 Adjusting the ratio α

To determine the optimal value of α, we conducted a series of comparative experiments. When $\alpha = 0$, only pure Gaussian noise is used without any mask guidance; when $\alpha = 1$, the initial noise is equally mixed with guidance noise derived from the mask. A small α leads to insufficient semantic guidance, while

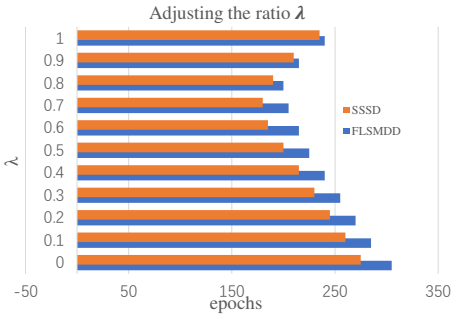


Fig. 10 Comparison of convergence speeds for different λ values. The horizontal axis represents the epochs required for training convergence, and the vertical axis shows λ values (ranging from 0.0 to 1.0). Due to the ambiguity in determining convergence, we use a minimum difference of 5 to select the closest epochs value.

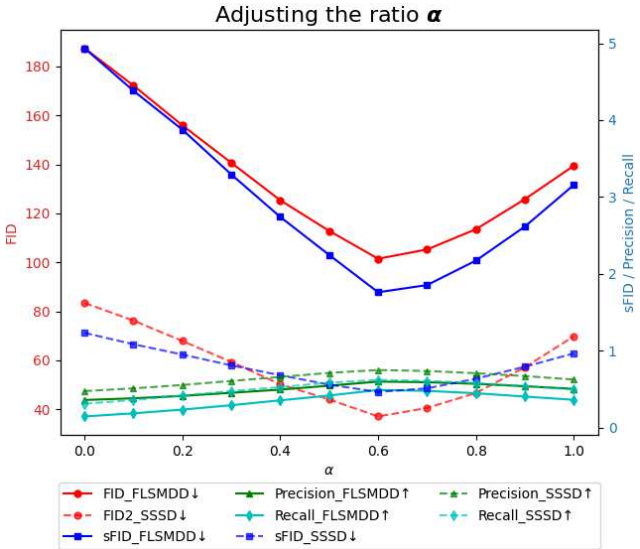


Fig. 11 The quantitative metrics comparison of different α values. The horizontal axis represents α ranging from 0.0 to 1.0, while the vertical axis displays various evaluation metrics. Since FID values differ significantly from the other three metrics, they are plotted on a separate left-side vertical axis, while the remaining three metrics share the right-side vertical axis for better readability.

a large α may cause the model to overly rely on mask information, limiting its creativity and diversity. As shown in Figure 11, the network achieves the best performance across various metrics when $\alpha = 0.6$. Notably, introducing any degree of mask guidance consistently outperforms using purely random noise, highlighting the effectiveness of mask-based guidance in improving model performance.

4.10 Analysis of Failure Cases in Generation

During the experiments, we observed that HSD-Sonar generates errors in certain specific situations, as shown in Figure 12. For example, the first row, when processing data from the FLSMDD dataset, shows significant noise at the edges of white objects (red circle area) in the generated results(FID: 254 PSNR: 6.58 SSIM: 0.087). This is primarily because the boundaries in this area of the original image are too blurry, making it difficult for the model to reconstruct accurately. The second row displays an image from the AUV dataset, which has been processed by a waterfall flow. The generated result within the green mask area shows significant discrepancies compared to the original image(FID: 340 PSNR: 6.26 SSIM: 0.074). This is because if the mask annotation is inaccurate, the model arbitrarily fills this area. Furthermore, the third row, an image from the AI4Shipwrecks dataset, indicates that the model does not actively generate targets that are not annotated by a mask(FID: 321 PSNR: 6.32 SSIM: 0.076). In summary, these failure cases reveal two main limitations of the HSD-Sonar model: its ability to handle blurry boundaries needs improvement, and the model is highly sensitive to the accuracy of masks, with imprecise annotations directly leading to a decrease in the quality of generated images.

5 Discussion

While our HSD-Sonar effectively leverages masks to generate high-quality sonar images, Fig 4to 8 reveal noticeable differences in quality and clarity compared to the ground truth(Input image). These discrepancies stem from two key factors: (i)Noise interference. Real sonar data contain environmental noise, whereas masks are typically derived from idealized annotations. This gap reduces the models adaptability to complex environments, necessitating further improvements. (ii)mask quality. Low-resolution or inaccurate masks can lead to blurred boundaries, missing details, or distorted target shapes, affecting the fidelity of generated images.

Additionally, mask guidance has inherent limitations. Future research should explore integrating target motion trajectories, material properties, or historical data to reduce reliance on single-mask guidance and enhance model robustness.

The future research directions will focus on further optimizing HSD-Sonar, aiming to integrate multi-source information such as historical data and target motion trajectories, and explore advanced techniques like multi-modal learning and real-time video tracking to significantly enhance model performance. In practical applications, HSD-Sonar is expected to empower fields such as underwater detection and resource exploration through high-fidelity image generation. Potential challenges, such as real-time performance and adaptability to complex environments, will be addressed by optimizing algorithms and deployment strategies. The proposed framework not only deepens the interdisciplinary research between sonar image processing and machine learning but

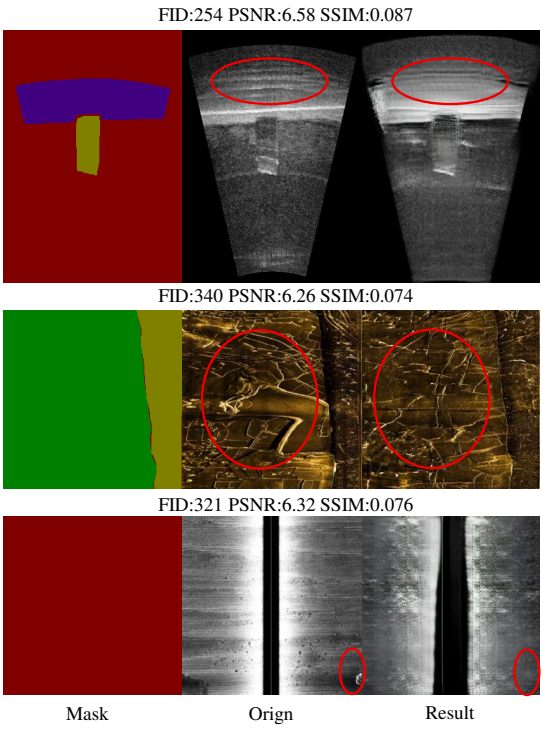


Fig. 12 Presentation of Generation Failure Cases. The left column shows the masks, the middle column displays the original images, and the right column presents the images generated by HSD-Sonar

also provides a new paradigm for high-fidelity sonar image synthesis in marine science and underwater robotics, demonstrating broad application prospects and positive societal benefits.

6 Conclusion

We propose a hierarchical segmentation-guided diffusion model for generating high-quality sonar images HSD-Sonar (Hierarchy Segmentation-guided Diffusion). This model introduces modules at three image hierarchy levels: Dynamic Step Adaptation (DSA), Dynamic Sampling Embedding (DSE), and Dynamic Mask Enhancement (DME), as well as a feature-level module called Masked Feature Fusion (MFF). By leveraging masks to guide the diffusion process, HSD-Sonar effectively enhances the restoration capability of target structures and the visual quality of images.

Experimental results demonstrate that, compared to other mask-guided diffusion models, HSD-Sonar outperforms existing state-of-the-art methods in terms of contour preservation, detail restoration, and image clarity. Future

work could further incorporate prior information such as target motion trajectories, material properties, or historical data to reduce discrepancies caused by environmental noise and mask quality.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Munnelly, R.T., Castillo, J.C., Handegard, N.O., Kimball, M.E., Boswell, K.M., Rieucau, G.: Applications and analytical approaches using imaging sonar for quantifying behavioural interactions among aquatic organisms and their environment. ICES Journal of Marine Science **81**(2), 207–251 (2023)

[2] Huang, P., Yang, P.: Synthetic aperture imagery for high-resolution imaging sonar. Frontiers in Marine Science **9** (2022)

[3] Tang, Y., Wang, L., Jin, S., Zhao, J., Huang, C., Yu, Y.: Auv-based side-scan sonar real-time method for underwater-target detection. Journal of Marine Science and Engineering **11**(4) (2023)

[4] Wang, K., Liu, P., Zhang, C.: Pronet: Underwater forward-looking sonar images target detection network based on progressive sensitivity capture. Computers, Materials Continua, 1–10 (2025)

[5] 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 Sensing **16**(13) (2024)

[6] Park, J., Lee, S., Ryu, Y., Jeong, D., Paik, J.: Enhancing side-scan sonar imaging: A comparative study of deep learning super-resolution techniques. In: 2024 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–3 (2024)

[7] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794 (2021)

[8] Yeom, T., Gu, C., Lee, M.: Dudgan: Improving class-conditional gans via dual-diffusion. IEEE Access **PP**, 1–1 (2024)

[9] Deshmukh, P., Ambulkar, P., Sarjoshi, P., Dabhade, H., Shah, S.A.: Advancements in generative modeling: A comprehensive survey of gans and diffusion models for text-to-image synthesis and manipulation. In: 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1–8 (2024)

[10] Peng, Y.: A comparative analysis between gan and diffusion models in image generation. Transactions on Computer Science and Intelligent Systems Research **5**, 189–195 (2024)

[11] Karn, A., Kumar, S., Kushwaha, S.K., Katarya, R.: Image synthesis using gans and diffusion models. In: 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), vol. 1, pp. 1–6 (2023)

[12] Silkaitis, J.M., Douglas, B.L., Lee, H.: Synthetic-aperture sonar imaging: system analysis, image formation, and motion compensation. In: Conference Record of The Twenty-Ninth Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 423–427 (1995)

[13] Chiang, A., Broadstone, S.: Sonar beamforming system. Journal of The Acoustical Society of America **117**, 3348–3348 (2005)

[14] Lee, S., Park, B., Kim, A.: Deep learning from shallow dives: Sonar image generation and training for underwater object detection. ArXiv **abs/1810.07990** (2018)

[15] Li, L., Li, Y., Wang, H., Yue, C., Gao, P., Wang, Y., Feng, X.: Side-scan sonar image generation under zero and few samples for underwater target detection. Remote Sensing **16**(22) (2024)

[16] Wang, Z., Guo, Q., Lei, M., Guo, S., Ye, X.: High-quality sonar image generation algorithm based on generative adversarial networks. In: 2021 40th Chinese Control Conference (CCC), pp. 3099–3104 (2021)

[17] Sung, M., Yu, S.-C.: Sonar image generation from underwater optic images utilizing sensor models and depth estimation with neural network. In: OCEANS 2021: San Diego Porto, pp. 1–5 (2021)

[18] Sung, M., Kim, J., Kim, J., Yu, S.-C.: Realistic sonar image simulation using generative adversarial network. IFAC-PapersOnLine **52**(21), 291–296 (2019)

[19] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models.

Advances in neural information processing systems **33**, 6840–6851 (2020)

[20] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8162–8171 (2021)

[21] Jo, J., Lee, S., Hwang, S.J.: Score-based generative modeling of graphs via the system of stochastic differential equations. In: Proceedings of the 39th International Conference on Machine Learning, vol. 162, pp. 10362–10383 (2022)

[22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2022)

[23] Baykal, G., Karagoz, H.F., Binhuraib, T., Unal, G.: ProtoDiffusion: Classifier-free diffusion guidance with prototype learning. In: Proceedings of the 15th Asian Conference on Machine Learning, vol. 222, pp. 106–120 (2024)

[24] Sung, M., Kim, J., Kim, J., Yu, S.-C.: Realistic sonar image simulation using generative adversarial network. IFAC-PapersOnLine **52**, 291–296 (2019)

[25] Bryan, O., Berthomier, T., D'Ales, B., Furfaro, T., Haines, T.S.F., Pailhas, Y., Hunter, A.: A diffusion-based super resolution model for enhancing sonar images. The Journal of the Acoustical Society of America **157**(1), 509–518 (2025)

[26] Hu, X., Jin, Y., Liang, J., Liu, J., Luo, R., Li, M., Peng, T.: Diffusion model for image generation - a survey. In: 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), pp. 416–424 (2023)

[27] Oguz, I., Dinc, N.U., Yildirim, M., Ke, J., Yoo, I., Wang, Q., Yang, F., Moser, C., Psaltis, D.: Optical implementation of denoising diffusion models for efficient image generation. In: Frontiers in Optics + Laser Science 2024 (FiO, LS), pp. 7–3 (2024)

[28] Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(1) (2022)

[29] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., *et al.*: Photorealistic text-to-image diffusion models with deep language

understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)

[30] Miao, Z., Wang, J., Wang, Z., Yang, Z., Wang, L., Qiu, Q., Liu, Z.: Training diffusion models towards diverse image generation with reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10844–10853 (2024)

[31] Kulikov, V., Yadin, S., Kleiner, M., Michaeli, T.: SinDDM: A single image denoising diffusion model. In: Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 17920–17930 (2023)

[32] Colleoni, E., Sanchez Matilla, R., Luengo, I., Stoyanov, D.: Guided image generation for improved surgical image segmentation. Medical Image Analysis **97**, 103263 (2024)

[33] Jiang, S., Tao, Z., Fu, Y.: Segmentation guided image-to-image translation with adversarial networks. In: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1–7 (2019)

[34] Xue, J., Wang, T., Wang, J., Zhang, K., Luo, W., Ren, W., Liu, Z., Park, H., Cao, X.: Segmentation guided sparse transformer for under-display camera image restoration. arXiv preprint arXiv:2403.05906 (2024)

[35] Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 88–98 (2024)

[36] Chen, S., Ma, D., Lee, S., Yu, T.T.L., Xu, G., Lu, D., Popuri, K., Ju, M.J., Sarunic, M.V., Beg, M.F.: Segmentation-guided domain adaptation and data harmonization of multi-device retinal optical coherence tomography using cycle-consistent generative adversarial networks. Computers in Biology and Medicine **159**, 106595 (2023)

[37] Liu, M., Yan, X., Wang, C., Wang, K.: Segmentation mask-guided person image generation. Applied Intelligence **51**(2), 1161–1176 (2021)

[38] Aakerberg, A., Johansen, A.S., Nasrollahi, K., Moeslund, T.B.: Semantic segmentation guided real-world super-resolution. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), pp. 449–458 (2022)

[39] Yang, Y., Chen, Q., Li, Y., Wang, F., Han, X.-H., Iwamoto, Y., Liu, J., Lin, L., Hu, H., Chen, Y.-W.: Segmentation guided crossing dual decoding generative adversarial network for synthesizing contrast-enhanced computed tomography images. IEEE Journal of Biomedical and Health

Informatics **28**(8), 4737–4750 (2024)

[40] Pizer, S.M., Johnston, R.E., Ericksen, J.P., Yankaskas, B.C., Muller, K.E.: Contrast-limited adaptive histogram equalization: speed and effectiveness. In: Proceedings of the First Conference on Visualization in Biomedical Computing, pp. 337–345 (1990)

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017)

[42] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020)

[43] Zhang, R., Isola, P., Efros, A.A.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018)

[44] Singh, D., Valdenegro-Toro, M.: The marine debris dataset for forward-looking sonar semantic segmentation. In: Proceedings of the Ieee/cvf International Conference on Computer Vision, pp. 3741–3749 (2021)

[45] Ma, Z., Meng, W., Zhao, X., Jiang, L.: Enhancing sonar image segmentation with random fusion in a diffusion model framework. The Visual Computer, 1–15 (2025)

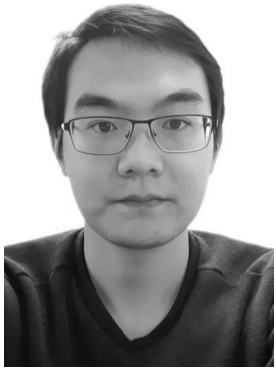
[46] Burguera, A., Bonin-Font, F.: On-line multi-class segmentation of side-scan sonar imagery using an autonomous underwater vehicle. Journal of Marine Science and Engineering **8**(8), 557 (2020)

[47] Yang, D., Cheng, C., Wang, C., Pan, G., Zhang, F.: Side-scan sonar image segmentation based on multi-channel cnn for auv navigation. Frontiers in Neurorobotics **16**, 928206 (2022)

[48] Sethuraman, A.V., Sheppard, A., Bagoren, O., Pinnow, C., Anderson, J., Havens, T.C., Skinner, K.A.: Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. The International Journal of Robotics Research **44**(3), 341–354 (2025)

[49] Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2332–2341 (2019)

[50] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV, pp. 3813–3824 (2023)



Zhihao Ma received the B.S. degree in Computer Science and Technology from Harbin Institute of Technology, China, in July 2017. He joined the State Key Laboratory of Multimodal Artificial Intelligence Systems at the Institute of Automation, Chinese Academy of Sciences in 2020 and received the M.S. degree in Electronic Information from the University of Chinese Academy of Sciences in July 2023. He is currently a Ph.D. student at Southeast University. His research interests include computer vision and image processing.

E-mail: 230238565@seu.edu.cn



Weiliang Meng received his Ph.D. degree in Computer Application from the State Key Laboratory of Computer Science at the Institute of Software, Chinese Academy of Sciences in 2010. He is currently an Associate Professor in the State Key Laboratory of Multimodal Artificial Intelligence Systems at the Institute of Automation, Chinese Academy of Sciences. His main research fields include artificial intelligence, computer vision, 3D scene analysis, 3D geometry processing, and computer graphics.

E-mail: weiliang.meng@ia.ac.cn



Chenghanxue Tang received her Bachelor's degree from Xidian University and is currently pursuing a Master's degree at Southeast University. His main research fields include artificial intelligence, computer vision.

E-mail: 220242385@seu.edu.cn



Longyu Jiang received her bachelor's degree from Wuhan University, her master's degree from Southeast University, and her Ph.D. from Grenoble Alpes University, France. She is currently a Professor at the School of Computer Science and Engineering, Southeast University. Her main research areas include underwater acoustic signal and image processing, artificial intelligence, and big data.

E-mail: JLY@seu.edu.cn

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HSDSonarMS.mp4