Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 3, 2025

Underwater Fish Image Generation Based on Diffusion Model

Xinyu Wang *

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture Beijing, China.
* **Corresponding author**

Abstract: Fish play a vital role in the ecosystems of rivers and lakes, and the application of deep learning-based intelligent identification methods offers an efficient and accurate approach to water environment regulation. However, the random spatial distribution of underwater fish complicates the collection of extensive real-image datasets. This scarcity of images diminishes the generalization capability of deep learning models, thereby limiting their practical applicability. To address this challenge, we propose a diffusion model-based approach that involves pre-training on an open-source underwater fish dataset and subsequently generating realistic, diverse underwater fish images from pure noise. The generated images achieved 6.8070, 9.1829 and 26.3132 on the evaluation metrics NIQE, PIQE and BRISQUE, respectively.

Keywords: Underwater fish; Image generation; Diffusion model; U-Net.

1. Introduction

Deep learning techniques have been extensively applied to underwater fish recognition tasks. However, the scarcity of underwater image data significantly undermines the models' generalization capabilities, thus limiting high-precision recognition in real-world environments. First, factors such as water turbidity, variable light intensity, and inconsistent shooting conditions often result in low-quality underwater images, which necessitate pre-processing operations—such as denoising—to enhance image clarity. Second, despite the rich diversity of fish species and the complexity of their habitats, existing datasets are unable to capture all species across various environments. Finally, class imbalance causes common fish species to be overrepresented, while rare species are underrepresented, leading to model overfitting due to the limited image data available for these species. Data augmentation methods, including the addition of noise, rotation, and adjustments to contrast and brightness, are commonly employed to address the image shortage. However, such techniques do not inherently enrich the dataset's diversity, as they fail to capture the underlying pixel feature distributions and stylistic attributes of the images. In contrast, generative models can learn the intrinsic data distribution from existing samples, enabling them to comprehend image structure and distribution patterns and to generate new samples that are similar yet not identical to the original data. Their unsupervised nature, which obviates the need for labeled samples, makes these models especially useful in scenarios where image availability is limited.

Generative models have achieved significant progress in deep learning, notably in image generation tasks. These include autoregressive models [1], generative adversarial networks [2], normalized streaming models [3], variational autoencoders [4], and diffusion models [5]. Among these, the denoising diffusion model is characterized by its parameterization through a Markov chain, comprising two primary components: the forward diffusion process and the backward denoising process. During the forward process, noise is progressively introduced into the image until it transforms entirely into a Gaussian noise distribution. In the subsequent backward process, the noise injected at each step is estimated and incrementally removed to recover the original, clean image. Despite their effectiveness, denoising diffusion models typically require extensive sampling time and encounter challenges in conditional generation tasks. To accelerate the sampling process, DDIM [6] introduced a non-Markovian forward process that enables Gaussian diffusion under varying step sizes. Guided Diffusion [7] incorporates a classification network at each step of the backward process to steer the generation toward the desired outcome, while Classifier-free Guidance [8] circumvents the need for an auxiliary classifier, mitigating both computational overhead and the risk of erroneous gradient estimation inherent in classifier training. Building on these developments, Semantic Guided Diffusion [9] enhances classifier performance by generating descriptive reference maps in a text- and reference map-guided manner. Furthermore, GLIDE [10] presents a text-to-image generation approach based on diffusion models, with empirical evidence suggesting that the classifier-free method generates images that are more realistic and text-consistent than those produced using CLIP guidance.

Accordingly, this study employs a denoising diffusion model to generate fish images in complex underwater environments, thereby augmenting the underwater fish sample dataset. Furthermore, evaluation metrics for image generation are utilized to assess the realism of the synthesized images.

2. Methods

2.1. Denoising Diffusion Probabilistic Model

The denoising diffusion probabilistic model (DDPM) is widely applied in image generation tasks. As illustrated in Fig. 1, the green arrow denotes the forward process in which noise is progressively introduced, whereas the red arrow represents the reverse process where noise is removed. This methodology is based on a Markov chain process: during the forward process, noise is iteratively added to the original image until its structural features are entirely obliterated, and then a deep learning model is employed to learn the reverse process, gradually eliminating noise to restore the image's

original structure. For underwater fish images, the term "image structure" refers to the fish's morphology, texture, and color, in addition to the contextual background information of the aquatic environment.

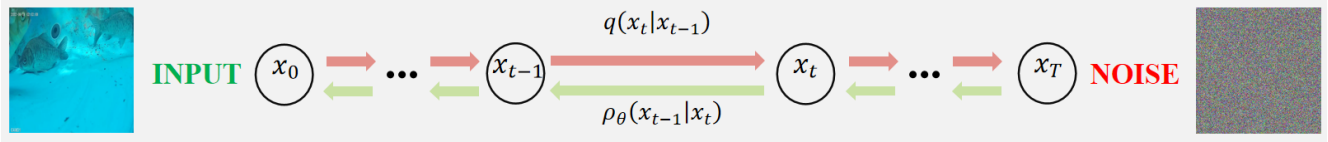

Figure 1. The denoising diffusion probabilistic model process

Noise Addition Process: A clean image, x_0, is incrementally corrupted by the successive addition of Gaussian noise, ϵ_t, over T steps until it becomes pure Gaussian noise, x_T. This process conforms to the following distribution:

$$x_t \sim q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}; (1-\alpha_t)\mathrm{I}\right). \quad (1)$$

where I is the unit matrix and x_t obeys a normal distribution with mean $\mu_t = \sqrt{\alpha_t}x_{t-1}$ and variance $\Sigma_t = (1-\alpha_t)\mathrm{I}$.

Denoising process: Train the denoising model and then de-predict the noise ϵ_θ to approximate the true noise ϵ_t added by the forward process step. obeying the following distribution:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t); \sum_\theta(x_t, t)). \quad (2)$$

where $p_\theta(x_{t-1}|x_t)$ denotes the posterior probability of x_{t-1} under the condition that x_t is known. θ represents the parameters of the neural network, $\mu_\theta(x_t, t)$ is the mean, and $\sum_\theta(x_t, t)$ is the variance.

2.2. U-Net model

Based on the aforementioned mathematical framework, the underwater fish image dataset is utilized for training. The model employs the U-Net [11] architecture from the original denoising diffusion model, integrated with a self-supervised learning approach, to develop a pre-trained diffusion model for image generation. During training, the diffusion model learns to progressively restore the original image from various levels of noise, thereby capturing both the global structure and intricate details. Specifically, it gradually assimilates the morphology, texture, and color characteristics of underwater fish, as well as the contextual information of the surrounding aquatic environment. This self-supervised mechanism enables the model to autonomously extract features from the raw data without the need for manual labeling.

In order to correspond the time step t of the diffusion model to the image noise enhancement and denoising process, a time coding approach is used. This includes Time Embedding and Time Linear. Time Embedding maps the current moment t to a vector form that the model can handle. Fixed position encoding based on trigonometric functions is used to satisfy the following equation:

$$PE(pos, 2t) = \sin\left(\frac{pos}{10000^{\frac{2t}{d_{model}}}}\right). \quad (3)$$

$$PE(pos, 2t+1) = \cos\left(\frac{pos}{10000^{\frac{2t}{d_{model}}}}\right). \quad (4)$$

where PE stands for Positional Encoding, $2t+1$ and $2t$ stand for odd and even moments respectively, and d_{model} denotes the number of channels. Then it is processed by time linearization to correspond to the feature map dimension.

The loss function of the model is used to evaluate the difference between the noise predicted by the model and the noise added during the forward noise addition process, and is computed using the mean-square error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \quad (5)$$

2.3. Dataset

An open-source underwater fish motion target detection dataset consisting of 355 images was selected. To ensure compatibility with computer processing algorithms, all images were standardized to a resolution of 256×256 pixels.

Figure 2. Underwater fish images

2.4. Experiments

2.4.1. Evaluation metrics.

Natural Image Quality Evaluator (NIQE) is a reference-free image quality assessment metric based on the statistical features of natural scenes; the lower the metric, the higher the quality of the image.

$$D(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{\left((v_1 - v_2)^T \left(\frac{\Sigma_1 \Sigma_2}{2}\right)^{-1} (v_1 - v_2)\right)}. \quad (6)$$

Where $v_1, v_2, \Sigma_1, \Sigma_2$ denote the mean vector and covariance matrix of natural and distorted images respectively.

Perception based Image Quality Evaluator (PIQE) is a reference-free image quality evaluation metric based on perceptual features, which utilizes the block structure and noise characteristics of an image to calculate the quality score of the image.

$$PIQE = \alpha BM + \beta NM. \quad (7)$$

BM is a block effect metric that measures the block structure in an image, which is mainly affected by artifacts caused by image compression. NM is a noise metric that

measures the noise level in an image, which is mainly affected by the noise caused by image loss and transmission errors.

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a reference-free image quality evaluation metric based on statistical features of natural scenes.

2.4.2. Experiment setup.

All experiments in this paper were implemented on an NVIDIA 3070 GPU with 8G of video memory. Pytorch version 2.1.0 and CUDA version 11.8. The epochs is 200, the batch size is 2, the learning rate is 0.0001, and the preheat scheduler and cosine function decay are used. Optimizer is AdamW, time step T is 1000.

2.4.3. Results

The fish images of underwater scenes generated from random noisy images using a pre-trained diffusion model are shown in Figure 3. Its ability to generate all parts of the fish, including the head, body, tail and fins, is good. In most of the cases, the fish shows realistic motion patterns. Meanwhile, since the generation process is randomized, the brightness, resolution, background, and the number of fish in the image are also randomized, which fully demonstrates that the generative model can greatly expand the richness of the dataset, rather than simply enhancing the original image.

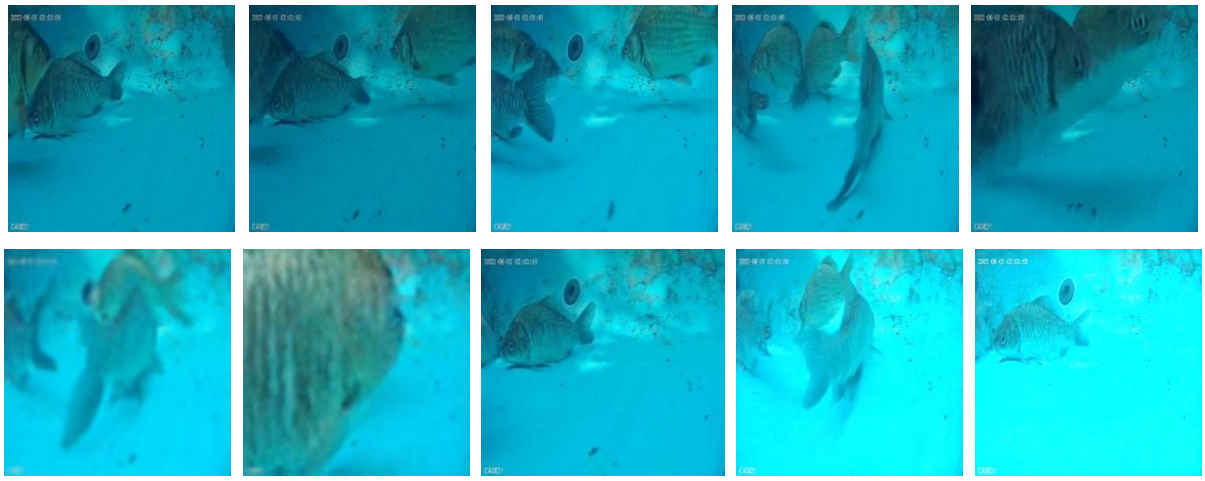

Figure 3. Diffusion model generation results

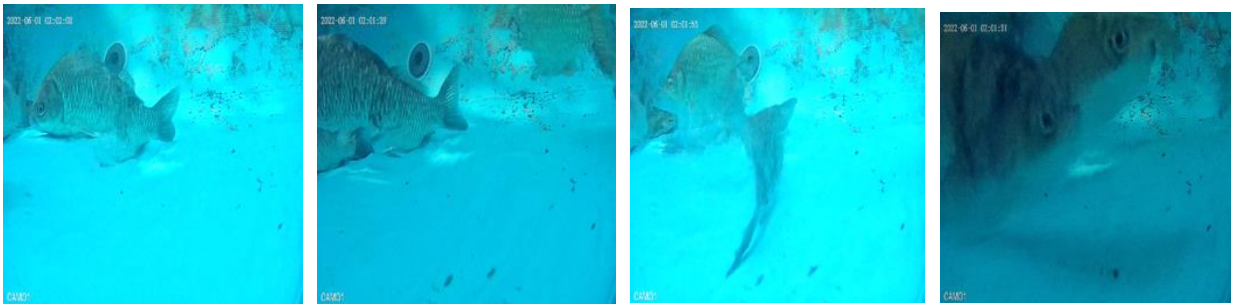

Figure 4. Some less desirable results

In this paper, a total of 150 randomized underwater fish activity images were generated, which reached 6.8070, 9.1829 and 26.3132 in NIQE, PIQE and BRISQUE, respectively.

Table 1. Accuracy evaluation results

NIQE	PIQE	BRISQUE
6.8070	9.1829	26.3132

However, since the generation process is not constrained by any geometrical form, sometimes the images also show distortion and blurring. These can be used as complementary images for complex underwater scenes to verify the robustness of the recognition model. However, the information like the acquisition time contained in the images in the original dataset can be generated very accurately by the diffusion model.

2.5. Conclusions

In this paper, we propose a method for capturing complex data distribution patterns of fish and their underwater environments from sample datasets. Our approach utilizes self-supervised training of diffusion models to learn the intrinsic structure of the images, thereby enhancing model robustness against noise, interference, and uncertainty. Noise is progressively eliminated from initially random images until diverse and realistic underwater fish images are generated. This method not only enriches the dataset but also addresses the scarcity of available underwater fish images.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] Van Den Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016, 12.

[2] Goodfellow I J, Pouget-Abadie J, Mirza M, et al.Generative Adversarial Nets[J]. Nips'14, 2014: 2672–2680.

[3] Dinh L, Krueger D, Bengio Y.Nice: Non-linear independent components estimation [J].arXiv preprint arXiv:1410.8516, 2014.

[4] Kingma D P, Welling M.Auto-encoding variational bayes [J].arXiv preprint arXiv:1312.6114, 2013.

[5] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models [J]. ArXiv, 2020, abs/2006.11239.

[6] Song J, Meng C, Ermon S. Denoising Diffusion Implicit Models [J]. ArXiv, 2020, abs/2010.02502.

[7] Dhariwal P, Nichol A. Diffusion Models Beat GANs on Image Synthesis [J]. ArXiv, 2021, abs/2105.05233.

[8] Ho J. Classifier-Free Diffusion Guidance [J]. ArXiv, 2022, abs/2207.12598.

[9] Liu X, Park D H, Azadi S, et al. More Control for Free! Image Synthesis with Semantic Diffusion Guidance [J]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 289-99.

[10] GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

[11] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [J]. ArXiv, 2015, abs/1505.04597.