frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 12 November 2025
DOI 10.3389/fmars.2025.1687877

OPEN ACCESS

EDITED BY
Habib Hamam,
Université de Moncton, Canada

REVIEWED BY
M. Sohail Khan,
Gyeongsang National University, Republic of Korea
Kuo Jui Hu,
National Taiwan University of Science and Technology, Taiwan

*CORRESPONDENCE
Raheem Sarwar
r.sarwar@mmu.ac.uk

RECEIVED 18 August 2025
ACCEPTED 13 October 2025
PUBLISHED 12 November 2025

CITATION
Tiwari NK, Bajpai A, Yadav S, Bilal A, Darem AA, Sarwar R and Singh J (2025) DM-AECB: a diffusion and attention-enhanced convolutional block for underwater image restoration in autonomous marine systems. *Front. Mar. Sci.* 12:1687877. doi: 10.3389/fmars.2025.1687877

DM-AECB: a diffusion and attention-enhanced convolutional block for underwater image restoration in autonomous marine systems

Naveen Kumar Tiwari [1], Abhishek Bajpai [1], Shashank Yadav[1], Anas Bilal[2,3], Abdulbasit A. Darem[4,5], Raheem Sarwar[6]* and Jaibir Singh[7]

[1]Department of Computer Science and Engineering, Rajkiya Engineering College Kannauj, Kannauj, Uttar Pradesh, India, [2]College of Information Science and Technology, Hainan Normal University, Haikou, China, [3]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India, [4]Center for Scientific Research and Entrepreneurship, Northern Border University, Arar, Saudi Arabia, [5]Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia, [6]OTEHM, Manchester Metropolitan University, Manchester, United Kingdom, [7]School of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

Introduction: Effective underwater vision is critical for real-time marine ecosystem observation and conservation, especially for autonomous underwater vehicles (AUVs) operating in challenging oceanic environments.
Methods: We propose a novel underwater image enhancement framework tailored for smart robotic systems used in biodiversity monitoring, habitat mapping, and environmental sensing. Our method integrates a Denoising Diffusion Probabilistic Model (DDPM) for progressive image restoration with an Attention-Enhanced Convolutional Blocks (AECB) augmented Transformer backbone. The AECB modules provide dual channel and spatial attention, selectively amplifying features to enhance visual quality. Additionally, a lightweight architecture combined with a skip-sampling strategy is designed to optimize computational efficiency for onboard deployment in AUVs and underwater drones.
Results: Experimental evaluations demonstrate that our framework achieves superior image restoration performance while maintaining computational efficiency, outperforming existing transformer-diffusion approaches. The dual attention mechanism within AECB modules distinctly improves the clarity and detail of underwater images.
Discussion: This work advances AI-driven perception systems for intelligent ocean observation technologies, supporting improved marine biodiversity protection. The proposed model promises practical real-time applications in autonomous underwater exploration and monitoring. The model and code will be made publicly available on GitHub: https://github.com/ntiwari91/DM-AECB.

KEYWORDS

oceanic underwater images, underwater image enhancement, transformer-based denoising network, attention mechanism, channel attention, spatial attention, diffusion model

1 Introduction

The rapidly increasing deterioration of underwater ecosystems such as coral reefs, oyster reefs, and deep-sea environments has underlined the urgent necessity of intelligent, real-time monitoring systems. Among different sensing modalities, underwater vision is a crucial means to facilitate the autonomy and dependability of underwater robotic systems, particularly in applications such as offshore renewable energy exploration, aquaculture farming, marine conservation, and environmental monitoring. Notwithstanding the recent breakthroughs in robotic platforms and sensing hardware, underwater imaging with quality is a basic challenge because of the hostile physical conditions of aquatic environments.

The deterioration of underwater images is caused by several environmental factors. First, the absorption of light by suspended particles and water molecules creates a fast attenuation of illumination with depth, causing loss of image contrast and visibility. Second, scattering of light by particulate matter contributes to additional blurring and loss of details. Third, turbulence from water motion and currents adds geometric distortions and visual artifacts, which make the images hard to interpret for scientific and ecological use.

Identifying these challenges, researchers have suggested a range of techniques for improving underwater images (Iqbal et al. (2007); Zhao et al. (2016); Liu et al. (2019)). The techniques are designed to eliminate the impact of absorption, scattering, and turbulence and enhance the visual quality for future analysis (Dong et al. (2020); Islam et al. (2020b); Wang et al. (2021)). Most of these methods are plagued by high computational cost, reliance on handcrafted priors, and inflexibility to diversity in underwater conditions.

In an attempt to overcome these challenges and facilitate real-time deployment in Autonomous Underwater Vehicles (AUVs) and underwater drones, this paper introduces a novel underwater image enhancement framework based on recent advancements in diffusion models and Transformer-based neural networks. Unlike previous approaches such as the Transformer-diffusion model by Tang et al. (2023) that primarily focus on general diffusion techniques, our DM-AECB method distinctively integrates Attention-Enhanced Convolutional Blocks (AECB) Woo et al. (2018) within the Transformer architecture to provide dual channel and spatial attention. This targeted attention mechanism improves the denoising capability by emphasizing critical underwater scene features while effectively suppressing noise, addressing the complex and variable degradations inherent in underwater imagery.

We further leverage the Denoising Diffusion Probabilistic Model (DDPM) framework to iteratively reconstruct images by reversing a gradual noise-injection process. The combination of diffusion modeling with the AECB-empowered Transformer backbone enables more precise restoration of complex underwater scenes under challenging lighting and visibility conditions, going beyond the scope of prior transformer51 diffusion efforts.

Our core contributions are as follows:

- We propose a novel attention-guided Transformer backbone equipped with AECB modules. These blocks enhance the model's ability to capture spatial and spectral correlations in underwater scenes for effective denoising.
- We integrate diffusion-based modeling with attention mechanisms to build a robust underwater enhancement pipeline capable of removing noise artifacts while preserving critical structural details necessary for tasks like object detection and ecological assessment.
- We optimize the model architecture for embedded deployment through skip-sampling and lightweight Transformer design, facilitating real-time image processing onboard resource-constrained AUVs and smart underwater monitoring platforms.

Rest of the paper is organized as follows: Section 2 describes some recent literature, Section 3.1 explains the probabilistic diffusion model and basic mathematics, the architecture of the proposed model with diffusion model is covered in Section 3. Section 4 covers different types of experiments performed with ablation studies and result discussion, and finally conclusion is covered in Section 5.

2 Related works

2.1 Traditional approach

Drews et al. (2016) proposed a model aimed at enhancing underwater image accuracy, leading to the development of UDCP, an improved version of DCP. Constructing their dataset from outdoor landscape images, their findings demonstrate UDCP's superiority in improving underwater image quality compared to DCP, MDCP, and BP. However, UDCP exhibits limitations, particularly regarding reliability and robustness.

Song et al. (2018) trained the ULAP model to restore underwater images. Central to their approach is determining scene depth, vital for color and lighting correction. Their research indicates that scene depth correlates directly with the disparity between the maximum intensity of green-blue light and red light. They assert that their method offers a faster means of estimating scene depth compared to various CNN-based models.

2.2 Neural network based research

Several researchers have explored CNN-based technology (Fu et al. (2022); Anwar et al. (2018); Zamir et al. (2022); Zhang et al. (2018)). Anwar et al. (2018) developed an innovative CNN model focusing not only on minimizing mapping function objectives but also on learning discrepancies between degraded underwater images and their cleaned versions. This technique stimulates a diverse range of degraded underwater images for data

augmentation, achieving superior performance in diverse color and visibility conditions.

Most research in underwater image enhancement uses Generative Adversarial Networks (GAN) (Fabbri et al. (2018); Wang et al. (2019); Islam et al. (2020b); Guo et al. (2019); Ye et al. (2018)). Fabbri et al. (2018) used CycleGAN to generate pairs of undistorted and distorted underwater images for training to enhance image accuracy. Wang et al. (2019) solved visibility problems in underwater images through the use of GAN to generate realistic images. Islam et al. (2020b) used Conditional GAN on the EUVP dataset to produce Funie-GAN. Guo et al. (2019) proposed a multi-scale GAN structure for UW image enhancement. Ye et al. (2018) concentrated on co-joint haze detection through stacked conditional GAN.

Iqbal et al. (2010) presented a new method to improve underwater images based on unsupervised color correction techniques. Also, Li et al. (2019) presented WaterNet, an all-round underwater image enhancing model, whereas Liu et al. (2019) constructed an undersea image capture system for real-world underwater image dataset generation.

Islam et al. (2020a) proposed SESR for super-resolution enhancement of UW images. Various color channels were also researched (Wang et al. (2021); Zhang et al. (2022); Iqbal et al. (2007)). Wang et al. (2021) applied red-green-blue and Hue-saturation-value color space methods to their UIEC2-Net architecture. Zhang et al. (2022) suggested MILLE based on the CIELAB color space to resolve color deviation issues in underwater images. Iqbal et al. (2007) presented a novel approach centered on slide stretching to improve UW images.

Ancuti et al. (2017) employed white balancing and image fusion techniques. Additionally, Ancuti et al. (2012) utilized inputs from degraded images for enhancement. Sahu et al. (2014) explored existing methods, including Forward Unsharp Masking (USM) and median filters, for image enhancement.

Tang et al. (2023) proposed an underwater image enhancement diffusion model, while Guo et al. (2020) developed a deep curve estimation method for low-light image enhancement. In addition, Zhuang et al. (2022) introduced a Retinex variation model inspired by hyper-Laplacian reflectance priors.

Experiments with the use of the Dehazing algorithm have been carried out (Dong et al. (2020); Chiang and Chen (2011). Chiang and Chen (2011)) aimed at restoring underwater images via a dehazing algorithm, taking into consideration attenuation differences and artificial lighting. Sun et al. (2019) set forth a deep pixel-to-pixel network structure for UW image improvement, whereas Zhao et al. (2016) investigated perceptually-driven losses for image restoration, specifically in super-resolution applications.

3 Materials and methods

3.1 Denoising Diffusion Probabilistic Model

The diffusion model is a probabilistic generative model that aims to generate samples from a given dataset by modeling the process of diffusion, where noise is gradually added to an initial input to generate the final output. The DDPM by Ho et al. (2020) is a new way to obtain enhanced images. It has two main parts: forward diffusion and reverse diffusion.

Forward diffusion is like adding layers of noise to an image, gradually making it harder to see. Reverse diffusion is like using a special filter to remove the noise and reveal the original image.

3.2 Forward diffusion process

In this process, Gaussian noise is gradually added to an initial image to create a sequence of intermediate images. Each step in this process involves adding a bit of noise to the previous image, creating a progression of images. The forward diffusion process can be expressed as Equation 1:

$$\mathcal{I}_t = \sqrt{1 - d_t} \cdot \mathcal{I}_0 + \sqrt{d_t} \cdot \mathcal{N}, \tag{1}$$

where $\mathcal{I}_O$ is original image, $\mathcal{I}_t$ is degraded image at time t, d_t is coefficient representing diffusion level at time t and $\mathcal{N}$ represents random noise vector

3.3 Reverse diffusion process

This process works in the opposite direction. Given an image from the sequence generated by the forward diffusion, the goal is to predict the original image without noise. It involves estimating the noise that was added at each step to remove it from the image.

The reverse diffusion process can be expressed in the following Equation 2.

$$\mathcal{I}_{t-1} = \frac{1}{\sqrt{1 - r_t}} (\mathcal{I}_t - r_t \sqrt{1 - d_t} \cdot \mathcal{E}(\mathcal{I}_t, t)) + s_t \cdot \mathcal{Z}, \tag{2}$$

Where $\mathcal{I}_{t-1}$ represents estimated original image at time $t-1$, r_t coefficient representing reverse diffusion level at time t, $\mathcal{E}(I_t, t)$ estimated noise by neural network, s_t standard deviation of noise at time t and $\mathcal{Z}$ represents random noise vector.

3.4 Proposed methodology

Our proposed method consists of two main stages: noise simulation using the Gaussian diffusion probabilistic model and noise reduction using the proposed transformer-based neural network which provide better results as compared to the existing model. In the first stage, we introduce Gaussian noise to the input image to simulate real-world noisy conditions as depicted in Figure 1. This noise addition process follows the principles of the Gaussian diffusion probabilistic model, which accurately models the distribution of noise in natural images. In the second stage, we utilize a transformer-based neural network to remove the added noise and restore the image to its original clarity. The architecture of our noise reduction network includes several essential components: convolutional layers for feature extraction, normalization layers to

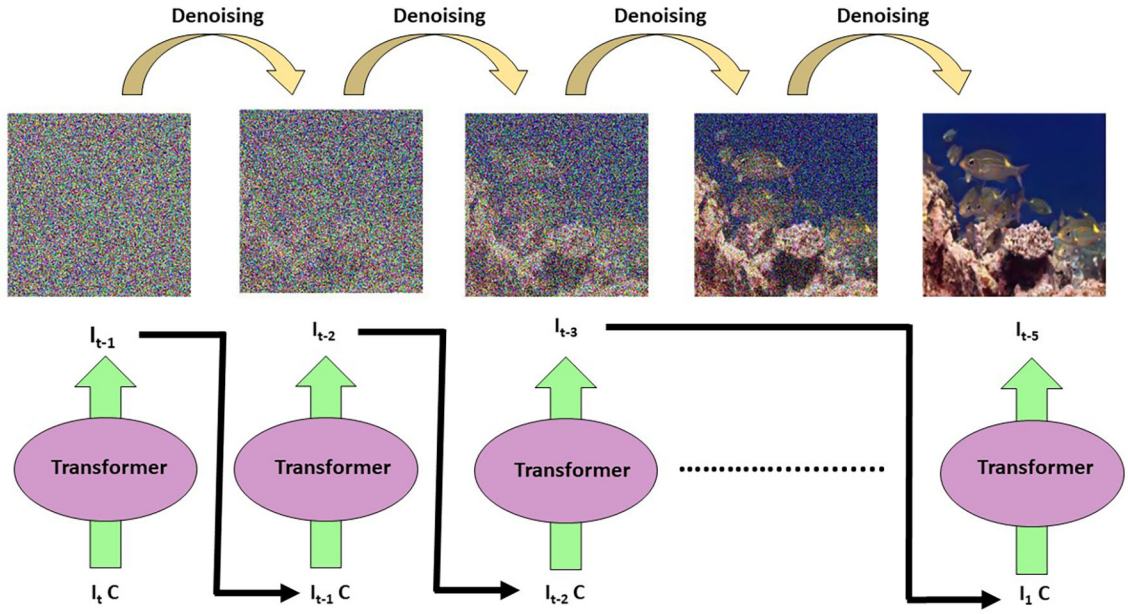

FIGURE 1
The figure representing noise reduction through transformer model.

enhance stability and convergence, attention blocks for capturing long-range dependencies, and feed-forward networks for refining features and producing the final denoised output.

Let $\mathcal{I}_0$ be an input image, $\mathcal{I}_t$ be the noisy image at step t, $\mathcal{N}(.)$ be a normal distribution function, $\sigma_t(t)$ be a time-dependent function representing noise variance. The mathematical equation can be represented as Equation 3:

$$\mathcal{I}_t = \mathcal{I}_0 + \mathcal{N}(\sigma_t(t)), \tag{3}$$

now the obtained image $\mathcal{I}_t$ is processed for the noise reduction i.e. denoising by using the proposed transformer-based network can be represented in the following Equation 4.

$$\mathcal{D}_t = \mathcal{T}(\mathcal{F}(\mathcal{I}_t, c, t)). \tag{4}$$

Where $\mathcal{D}_t$ represents the denoised image at step t, $\mathcal{T}(\dot{})$ signifies the transformer network, $\mathcal{F}(\dot{})$ function represents the feature extraction and processing steps before feeding the data to the transformer network, I_t noisy image at step t, c represents clean or partially denoised image, t represents time step information.

3.5 Loss function

The network is optimized during training using a loss function. Specifically, we employ the L1 loss or mean absolute error, which measures the difference between the predicted noisy image and the ground truth noisy image. This loss function effectively guides the training process to reduce the discrepancy between the output and target images, enabling the model to produce higher-quality restorations from degraded inputs.

Mathematically, the L1 loss is expressed as Equation 5:

$$L_s = \left\| \hat{\mathcal{X}}_t - \hat{\mathcal{H}}_\theta(\mathcal{X}_t, C, t) \right\|, \tag{5}$$

where L_s represents the loss function, $\hat{\mathcal{X}}_t$ denotes the actual noisy image at timestep t, $\hat{\mathcal{H}}_\theta(\mathcal{X}_t, C, t)$ signifies the predicted noisy image produced by the network given the input noisy image, and $\|\cdot\|$ denotes the L1 norm, computing the absolute difference between the predicted and actual noisy images.

While perceptual and SSIM-based loss functions are known to enhance visual fidelity in image restoration tasks, we excluded them in this work to maintain computational efficiency and stability during training. Our empirical evaluations showed that L1 loss alone sufficed to achieve a balance between qualitative and quantitative performance, especially when combined with the attention mechanisms and diffusion framework. Incorporating these additional losses could be explored in future work to potentially further improve perceptual quality.

3.6 Restoration through transformer-based network

In this paper, we present a novel transformer-based network tailored for noise reduction in noisy images which is given in Figure 2. Our network offers a more efficient noise reduction process compared to conventional methods by utilizing a shallower architecture, resulting in improved image quality. Inspired by the effectiveness of Transformer structures, we adopt a unique approach to computing attention, focusing on channels as well as spatial dimensions. In the attention block of the transformer,

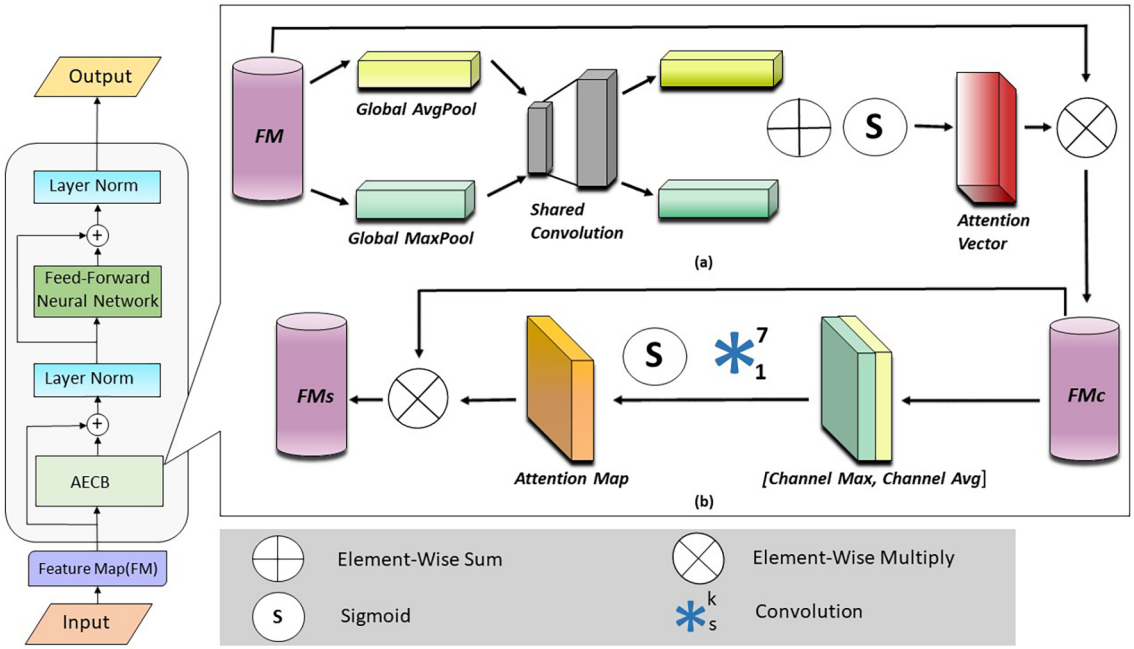

FIGURE 2
The proposed architecture is working of transformer using attention-enhanced convolutional. Block (AECB) with the sequential working of **(a)** Effective Channel Attention Module and **(b)** Spatial Attention Module.

we have incorporated both channel-wise and spatial attention mechanisms. We also prioritize practicality and efficiency by focusing on making the Transformer model lightweight and applying the skip sampling method in our approach. This modification reduces computational complexity and also enhances the network's ability to address color distortions common in low-quality underwater images.

To prepare the input for the network, we first combine the noisy image $\mathcal{I}_t$ and the conditional image $\mathcal{C}$ using channel concatenation. This creates a new feature map $\mathcal{F}_t$ with n channels, where n is a hyperparameter chosen based on the network architecture and the desired level of feature interaction between the two input images. The time step t is fed into a fully connected (dense) layer to encode temporal information represented as Equations 6–8:

$$\mathcal{X}_t = \otimes(||(\mathcal{I}_t, \mathcal{C})), \tag{6}$$

$$\mathcal{F}_t = \mathcal{R}e(\mathcal{F}C, t), \tag{7}$$

$$\mathcal{F}M = \mathcal{X}_t + \mathcal{F}_t. \tag{8}$$

Where $\otimes(.)$ represents convolution, $||||(.)$ signifies concatenation, $\mathcal{R}e$ represents reshape function and $\mathcal{F}C$ represents fully connected, $\mathcal{F}M$ represents feature map.

3.6.1 Network architecture

Let FM denote the feature map resulting from the above operation. The output achieved from the normalization block is fed to the convolution block attention module. The Attention-Enhanced Convolutional Block (AECB), is used to improve feature representation in convolutional neural networks (CNNs). It consumes an input feature map (FM) and produces an improved output (FMs) by integrating two attention mechanisms: channel attention (Mc) and spatial attention (Ms).

- Channel Attention: This module focuses on identifying "which" information within a feature map is most important by emphasizing informative channels and suppressing less useful ones. To achieve this, AECB first reduces the spatial dimensions (height and width) of the feature map using both average pooling and max pooling. These operations yield two separate spatial context descriptors, Fc_{avg} and Fc_{max}, which capture the average and maximum responses for each channel. Both descriptors are then fed into a shared Multi-Layer Perceptron (MLP) to generate the channel attention map M_c. The MLP typically includes a single hidden layer with a reduction ratio r, controlling the size of the hidden representation. A sigmoid activation function σ is applied at the MLP's output, assigning each channel a weight between 0 and 1. Greater weights point to more informative channels. The mathematical expression for the same is Equation 9:

$$\mathcal{F}M_c = M_c(\mathcal{F}M), \tag{9}$$

where FM_c is the feature map through channel attention and M_c is the function for channel attention applied over the feature map ($\mathcal{F}\mathcal{M}$).

- Spatial Attention: This module is concerned with "where" informative features are positioned in the spatial domain of the feature map. Here, AECB employs average pooling and max pooling along the channel dimension to pool channel-wise information. This yields two feature maps (Fs_avg and Fs_max) that correspond to the average and maximum activations at a given

spatial position. These maps are concatenated and passed to a typical convolution layer having a filter size of 7×7. The sigmoid activation function (σ) is once again applied to the output of the convolutional layer, which produces a 2D spatial attention map (Ms). The regions needing greater attention are represented by values closer to 1 in Ms, and those needing suppression are represented by values closer to 0. The mathematical formula for the same is Equation 10:

$$\mathcal{F}M_s = M_s(\mathcal{F}M_c), \tag{10}$$

where $\mathcal{F}M_s$ is output from the spatial attention and M_s is function to which $\mathcal{F}M_c$ is applied.

• Overall Attention Process: The channel attention map (M_c) is transmitted along the spatial axes in order to perform element-wise multiplication with the initial feature map ($\mathcal{F}M$). This enhances the feature map according to channel-wise importance. The ensuing attention-weighted feature map (F') is then element-wise multiplied with the spatial attention map (M_s). This again refines the feature map by concentrating on informative spatial locations. The resulting output (F'') is refined feature map expressed as Equations 11, 12:

$$F' = Mc(FM_c) \dot{o} FM, \tag{11}$$

$$F'' = Ms(F') \dot{o} F', \tag{12}$$

where $Mc(F)$ refers to the channel attention function acted on the feature map FM, . is the element-wise multiplication symbol FM is the original feature map, F' is the feature map refined after channel attention, F'' is the final feature map refined after channel and spatial attention, $Ms(F')$ refers to the spatial attention function acted on the attention-weighted feature map F', . is an element-wise multiplication symbol, F' is the attention-weighted feature map of the previous equation.

After obtaining the refined feature map F'' from the Attention-Enhanced Convolutional Block (AECB), the output is passed through an addition operation and layer normalization block to enhance feature representation and stability further. The resulting feature map is then fed into a feedforward neural network ($FFNN$) for additional refinement. Mathematically, the process can be represented as follows Equations 13–15:

$$F_{add} = F_{FM} + F'', \tag{13}$$

$$FLN = LayerNorm(F_{add}), \tag{14}$$

$$FFNN = FFNN(FLN)\,. \tag{15}$$

Where F_{FM} is the initial feature map resulted from the normalization block, LayerNorm denotes the layer normalization operation that acts on the feature map F_{add}, and $FFFNN$ denotes the output of the feedforward neural network.

The resulting final output image is achieved through applying another layer normalization block. This step helps ensure that the feature map goes through extra refinement and normalization prior to use for additional processing, further helping the overall efficiency of the transformer-based network structure with augmented attention blocks.

3.7 Enhanced skip-sampling technique

This section introduces the concept of Enhanced Probabilistic Skip-Sampling (EPSS), a novel technique designed to improve the efficiency of the inference process in diffusion models. It aims to address the trade-off between theoretical optimality (large time steps) and computational cost associated with iterative inference. We utilize a lightweight network architecture to reduce the computational burden within each iteration. We propose a modification to the standard iterative diffusion process that eliminates the random term. This modification leverages an alternative non-Markovian process introduced by the Efficient Method of the Iterative Implicit Probabilistic Model (DDIM) introduced by Song et al. (2020). This allows for a deterministic sampling approach during inference.

4 Results & discussions

This section discuss a range of experiments with dataset description and experimental settings as follows:

4.1 Datasets

In this article, we employ two recently published datasets for training and testing of networks i.e. Underwater Image Enhancement Benchmark (UIEB) provided by Li et al. (2019) and Large-Scale Underwater Image (LSUI) presented by Peng et al. (2023). UIEB dataset consists of 890 pair images. The underwater images are downloaded from the Internet, and the ground truth images are created by a combination of some earlier enhancement techniques and manual choice. In particular, several enhancement methods are used to enhance the underwater images gathered to create diverse improved versions of the LSUI dataset, Since the number of training images used in earlier datasets was small, the LSUI dataset contained a higher number of images. The LSUI dataset contains 5004 underwater images and their respective high-quality images, which provide a diversity of underwater views, object types, as well as deep-sea and cave images. Here, we use the training set of LSUI, containing 4500 pairs of images, to train the diffusion model. We use the remaining 504 images to check our proposed method.

4.2 Evaluation metrics

Earlier methods typically depend on subjective evaluation metrics, including UCIQE and UIQM. These measures, however, cannot be used to accurately evaluate performance in all scenes. In this paper, we mainly employ two full-reference assessment

measures: Peak Signal Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR reflects how much the image content approximates the reference, whereas SSIM computes the structural and texture similarity.

4.3 Implementation details

The suggested strategy in this paper is utilized through PyTorch, where the Adam optimizer is used for minimizing the objective function. The learning rate is 1.0×10^{-4}. When training, the batch size and image size are utilized at 8 and 128×128, respectively, in order to trade off computational efficacy and image quality. Pixel values of images are normalized to $[-1, 1]$. The diffusion model is run at a time step of 2000, and β is linearly sampled over the range $[10^{-6}, 10^{-2}]$. During testing, according to the configuration provided by Lugmayr et al. (2022), the size of the input image is set to 256×256. For balancing performance and computation runtime, skip sampling strategy is utilized with 10 sampling times. In both the training and testing phases, the hardware used is a workstation with an NVIDIA RTX 3080 GPU.

4.4 Training performance analysis

To evaluate the training performance and convergence behavior of the different model variants, we plotted the loss values over the training epochs (steps). Figure 3 illustrates the loss curves for DM-AECB Large (larger model variant), DM-AECB, DM-Trans (Tang et al. (2023)), DM-CA Only Generative (generative variants of the proposed model having only channel attention), and DM-AECB Generative (generative variants of the proposed model). It should be noted that the generative diffusion variants of the proposed models may not produce extremely satisfying results but they are much more immune to additive Gaussian noise. The loss curves provide valuable insights into the training dynamics of each model variant.

Ideally, the loss should decrease steadily over epochs, indicating effective learning and convergence. However, the observed loss curves exhibit varying behaviors.

DM-AECB demonstrates a consistent and rapid decline in loss, reaching a plateau relatively early in training. This suggests efficient convergence and optimization. DM-AECB Large also shows a decreasing trend but with more fluctuations and a slower convergence rate compared to DM-AECB.

DM-Trans exhibits a more erratic loss curve with several plateaus and spikes, indicating potential challenges in optimization. DM-CA Only Generative and DM-AECB-Generative show relatively high and stable loss values throughout training, suggesting difficulties in learning the target task.

These observations indicate that DM-AECB exhibits the most promising training behavior, followed by DM-AECB Large. DM-Trans, DM-CA Only Generative, and DM-AECB Generative encounter challenges during training, as evidenced by their loss curves.

Further analysis, including additional metrics and visualizations, is necessary to gain deeper insights into the underlying reasons for these performance differences and to identify potential areas for improvement in the respective models.

4.5 Progressive enhancement through AECB blocks

The progressive enhancement of images through successive AECB stages vividly demonstrates the cumulative impact of our model's architectural design. By iteratively applying the AECB module, we observe a systematic improvement in image quality, as quantified by the increasing PSNR and SSIM values. This quantitative evidence underscores the critical role of multiple AECB stages in achieving superior underwater image enhancement.

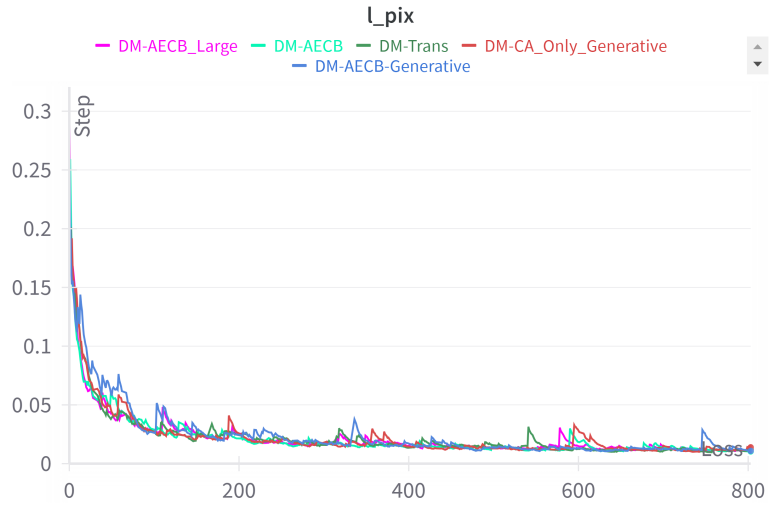

FIGURE 3
Comparison of loss curves training epochs for all model variants.

Each AECB stage contributes uniquely to the overall image restoration process. The initial stages are primarily focused on noise reduction and initial feature enhancement. As the process progresses, subsequent AECB stages refine these enhancements, targeting more subtle details and color corrections. This multi stage approach ensures that the model comprehensively addresses the challenges posed by underwater image degradation, resulting in a robust and effective image restoration pipeline.

Furthermore, by decomposing the complex task of underwater image enhancement into a series of more manageable sub-tasks, our model exhibits improved generalization capabilities. This is evident in its ability to handle diverse underwater imaging conditions, including varying levels of turbidity, color distortion, and low light.

As visualized in Figure 4, the progressive enhancement of image quality is visually apparent. The initial input image is characterized by significant noise, color distortion, and reduced visibility. With each successive AECB stage, these artifacts are progressively mitigated, culminating in a restored image that closely resembles the ground truth. This visual corroboration reinforces the efficacy of our proposed multi-stage AECB architecture.

4.6 Quantitative comparison with leading methods

In this paper, we compare our method with eight existing techniques in Table 1, which include both traditional methods and deep learning models. Traditional methods such as Fusion (Ancuti et al. (2017)), MMLE (Zhang et al. (2022)), and HLRP (Zhuang et al. (2022)) generally exhibit lower performance compared to their deep learning counterparts, with HLRP performing the weakest among them. Among the deep learning models, TACL (Liu et al. (2022)) and Water-Net (Li et al. (2019)) show competitive results, but Ushape (Peng et al. (2023)) achieves the highest PSNR and SSIM scores. Notably, our method surpasses all these techniques, delivering PSNR values of 28.78 and 29.56, and SSIM scores of 0.91 and 0.98 on the LSUI and UIEB datasets, respectively, while maintaining a processing time and parameter count comparable to leading models. This demonstrates that our approach provides the best overall enhancement in underwater image quality.

4.7 Statistical analysis of quantitative results

To validate the robustness of our findings, we report the mean and standard deviation of PSNR and SSIM (across three trials) for all methods. Additionally, paired t-tests were performed between the proposed DM-AECB and the strongest baseline, DM-Trans, on both benchmarks.

As shown in Table 2, DM-AECB consistently outperformed DM-Trans, with all paired t-tests indicating statistically significant improvements ($p< 0.05$).

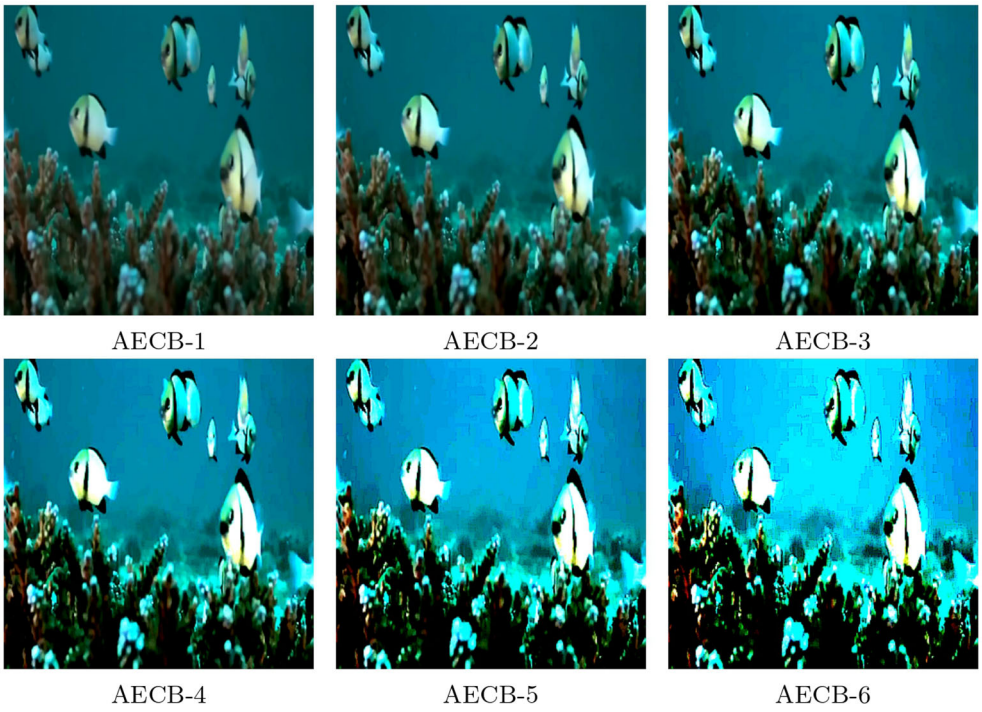

FIGURE 4
Progressive enhancement of underwater images through each transformer stage.

TABLE 1 Comparison of PSNR and SSIM values for different underwater image enhancement methods on the LSUI and UIEB datasets.

Method	Param. (Millions)	Time (sec.)	LSUI		UIEB	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑
Ancuti et al. (2012) (Fusion)	–	1.23s	17.69	0.644	18.79	0.792
Zhang et al. (2022) (MMLE)	–	0.30	17.70	0.725	19.30	0.830
Zhuang et al. (2022) (HLRP)	–	0.32	12.64	0.192	12.56	0.251
Liu et al. (2022) (TACL)	11	0.1	20.69	0.822	23.09	0.883
Li et al. (2019) (WaterNet)	25	0.55	22.99	0.789	20.48	0.789
Islam et al. (2020b) (FUnIE)	7	0.02	18.78	0.619	17.61	0.595
Fabbri et al. (2018) (UGAN)	57	0.06	22.79	0.754	20.59	0.682
Uplavikar et al. (2019) (UIE-DAL)	19	0.04	21.12	0.723	17.00	0.755
Li et al. (2021) (Ucolor)	157	1.87	22.91	0.890	20.78	0.872
Peng et al. (2023) (Ushape)	66	0.04	24.16	0.932	22.91	0.910
Tang et al. (2023) (DM-Trans)	10	0.13	27.65	0.8867	28.20	0.9429
Ours (DM-AECB)	10	0.14	28.78	0.91	29.56	0.98

4.8 Visual comparison

As shown in Figure 5, our method demonstrates superior performance in enhancing underwater images compared to several existing techniques, including both traditional methods like Fusion (Ancuti et al. (2017)) and MILLE (Zhang et al. (2022)), and deep learning models such as WaterNet (Li et al. (2019)), FUnIE (Islam et al. (2020b)), Ucolor (Li et al. (2021)), and Ushape (Peng et al. (2023)). Traditional methods, such as Fusion and MILLE, generally fail to restore color balance effectively, often leaving images with unnatural hues or excessive blur, as seen in the first and second columns. Among the deep learning models, WaterNet and Ucolor produce significant improvements in color correction and contrast but still struggle with preserving fine details, particularly in complex scenes, as observed in the third and fifth rows. Ushape shows competitive results, with enhanced contrast and sharper details, particularly in the middle rows. However, it still exhibits some over-saturation and slight loss of texture in certain areas. Our approach, illustrated in the second-to-last column, consistently provides the most balanced enhancement across all sample images. It effectively restores natural colors, enhances contrast without over-saturation, and preserves fine textures, closely matching the ground truth (GT) in the last column. This consistent performance across diverse underwater scenes highlights the robustness and effectiveness of our method in underwater image enhancement.

4.9 Ablation study

The ablation study unequivocally establishes DM-AECB as the superior model variant, as detailed in the following sections.

4.9.1 Ablation study quantitative analysis

Table 3 quantitatively summarizes the individual contributions of key components within the DM-AECB architecture by reporting the final average PSNR values for different model variants.

DM-AECB consistently delivers significantly higher PSNR values across all validation steps as training progresses. This marked performance differential underscores the model's exceptional capability to preserve image quality while effectively mitigating noise. In contrast, DM-AECB Large, while yielding commendable results, slightly underperforms compared to DM-

TABLE 2 Statistical comparison between DM-AECB and DM-Trans on LSUI and UIEB datasets.

Method	Dataset	PSNR (mean ± std)	SSIM (mean ± std)	*p*-value (t-test)
DM-Trans	LSUI	27.65 ± 0.05	0.887 ± 0.001	–
DM-AECB	LSUI	28.77 ± 0.04	0.912 ± 0.008	3.5×10^{-5} (PSNR) 0.025 (SSIM)
DM-Trans	UIEB	28.20 ± 0.02	0.943 ± 0.0002	–
DM-AECB	UIEB	29.54 ± 0.02	0.979 ± 0.001	7.4×10^{-5} (PSNR)0.00024 (SSIM)

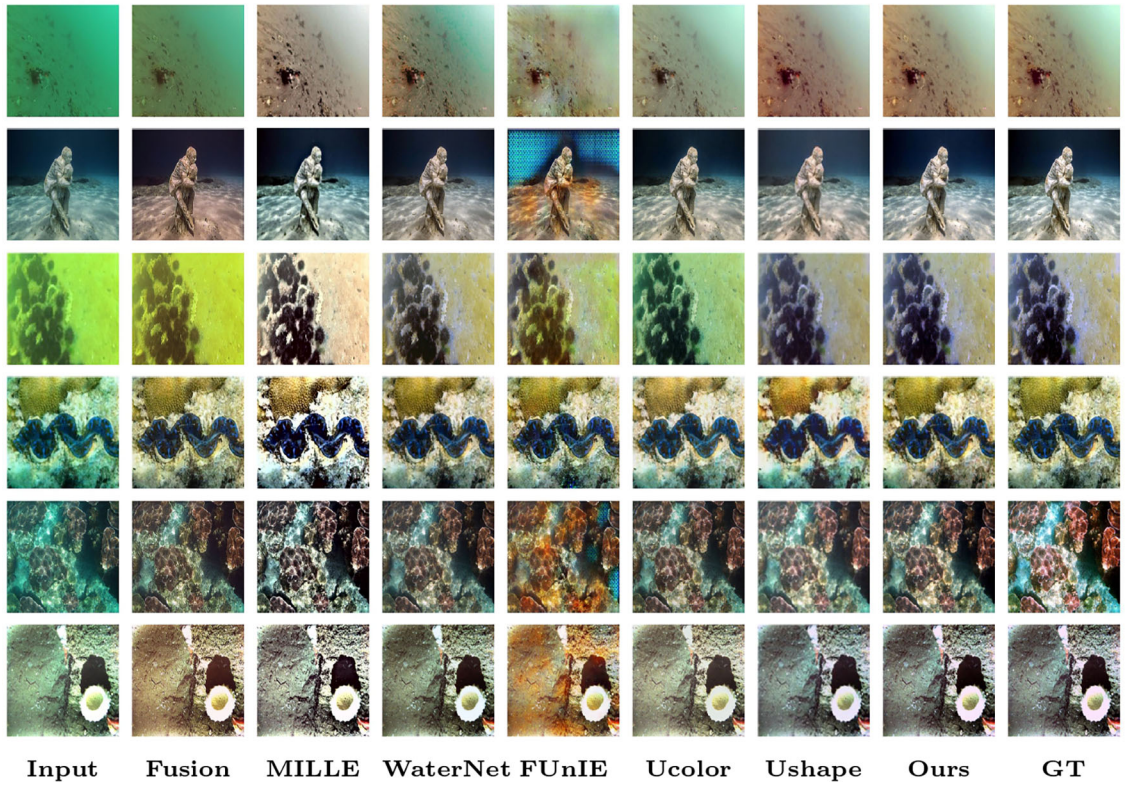

FIGURE 5
A visual comparison of underwater images and their corresponding enhanced results is presented. The ground truth images are shown in the second-to-last column for reference.

AECB. The performance gap between these top-tier models and the remaining variants, including DM Trans and generative models, is substantial, highlighting the efficacy of the DM-AECB architecture in optimizing image restoration.

The data in Table 3 clearly indicates that the diffusion process contributes a notable +2.0 dB boost in PSNR over the base model, illustrating its effectiveness in noise reduction and image restoration. Addition of Attention-Enhanced Convolutional Blocks (AECB) further advances performance by +3.5 dB, demonstrating the critical role of targeted dual channel and spatial attention in emphasizing important underwater features.

TABLE 3 Performance improvements from diffusion, AECB, and skip-sampling modules measured by final validation PSNR (dB).

Model variant	Final PSNR (dB)	PSNR gain over previous variant
DM AECB 02 (base)	22.5	Baseline
DM AECB Generative	24.5	+2.0 (incorporates diffusion)
DM CA Only Generative	28.0	+3.5 (adds AECB attention)
DM AECB 06 (full)	29.5	+1.5 (adds skip-sampling)

Finally, the skip-sampling technique enhances PSNR by an additional +1.5 dB, showing its utility in refining the image quality while optimizing computational efficiency.

These quantified improvements complement the PSNR curves shown in Figure 6 and the visual comparisons in Figure 7, providing robust numerical evidence of the effectiveness of each component. This comprehensive evaluation reinforces the design rationale and benefits of the DM-AECB model for underwater image enhancement.

4.9.2 Visual comparison

Visual analysis of underwater image enhancement results in Figure 7 indicates that DM-AECB consistently generates the most natural and visually appealing images. This variant effectively restores colors, preserves fine details, and minimizes artifacts across various underwater scenes. Although DM-AECB-Gen shows potential in color correction, it introduces more artifacts and less natural appearance compared to DM-AECB. DM-Trans improves visibility but often causes color distortions and blurriness. DM-CA-Gen struggles with both color correction and detail preservation, resulting in less pleasing outputs. Generative variants (DM-AECB-Gen and DM-CA-Gen) suffer from color distortion but are noted for robustness to noisy inputs, as visible in column 01. Further quantitative evaluation using metrics such as PSNR and SSIM could provide additional insights into their comparative performance.

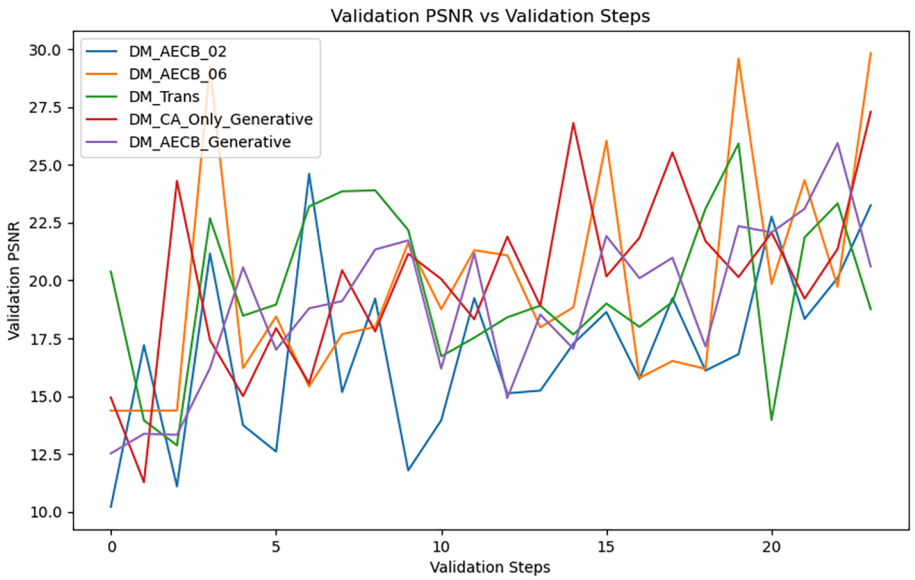

FIGURE 6
Comparison of PSNR values over training epochs for all model variants.

4.9.3 GPU utilization

GPU power consumption analysis from Figure 8 reveals that DM-AECB is the most energy-efficient model, rapidly stabilizing at a low power level without requiring additional GPU resources compared to DM-Trans. DM-AECB Large also demonstrates controlled power usage but consumes slightly more energy. DM-Trans exhibits fluctuating power consumption, and both generative models show significantly higher energy demands. These findings underscore DM-AECB's efficacy for cost-effective and environmentally friendly deployment.

5 Conclusion

In this paper, we present a novel approach for underwater image enhancement for marine robotic systems by integrating diffusion models with attention-enhanced convolutional blocks. Our model incorporates both channel and spatial attention mechanisms within the Attention-Enhanced Convolutional Block (AECB), leading to significant improvements in image quality metrics such as PSNR and SSIM. Extensive experiments, including ablation studies, demonstrate

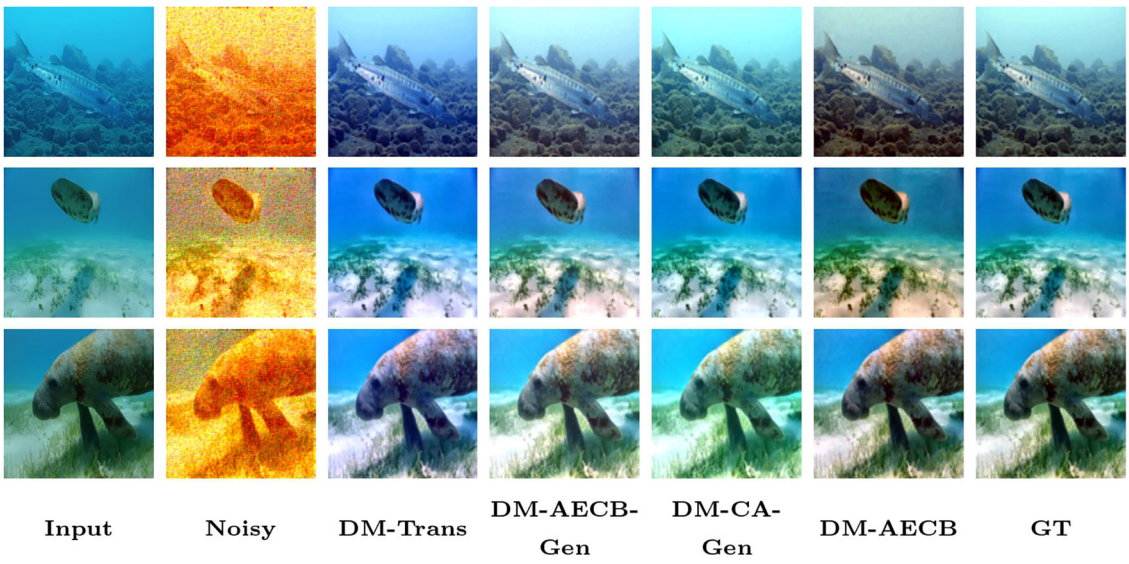

FIGURE 7
Visual comparison of underwater images and enhanced outputs produced by different model variants.

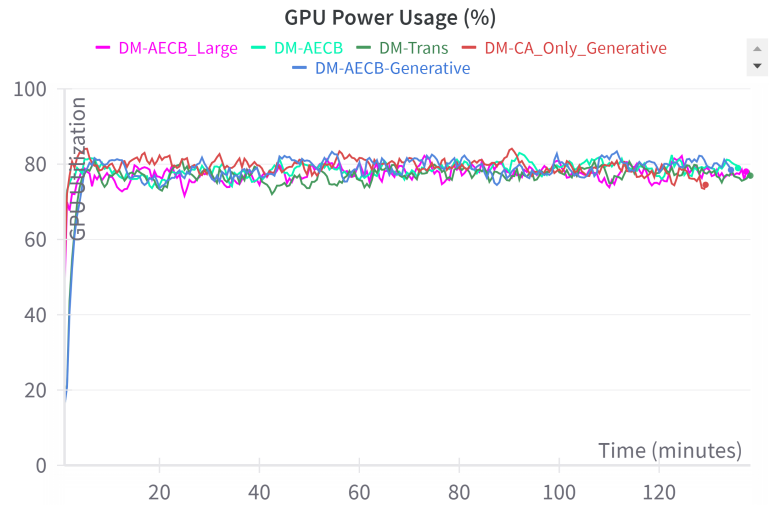

FIGURE 8
GPU utilization during training process.

that the inclusion of these attention mechanisms results in substantial enhancements over existing methods. A key aspect of our work is the addition of denoising capabilities through generative variants of the proposed model. It helps in achieving further clarity by effectively reducing noise in challenging underwater environments. The ablation studies underscore the importance of this generative component, revealing that it plays a crucial role in improving both the perceptual quality and the quantitative metrics of the enhanced images. This work sets a new benchmark in underwater image enhancement, providing a robust, practical solution for improving image quality in difficult underwater conditions.

While demonstrating promising efficiency and quality in controlled experiments, real-world deployment on autonomous underwater vehicles (AUVs) requires further consideration of inference speed, onboard computational limits, and power consumption. Future work will optimize the model for embedded platforms and validate performance in actual marine environments to ensure suitability for marine conservation applications.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Ethics statement

The requirement of ethical approval was waived for the studies involving animals because authors used available data. The studies were conducted in accordance with the local legislation and institutional requirements.

Author contributions

NT: Formal Analysis, Supervision, Methodology, Conceptualization, Software, Writing – original draft, Resources. ABa: Validation, Investigation, Writing – review & editing, Formal Analysis, Data curation. SY: Conceptualization, Validation, Data curation, Writing – review & editing, Supervision, Visualization. ABi: Data curation, Writing – review & editing, Investigation, Resources. AD: Writing – review & editing, Funding acquisition, Project administration. RS: Supervision, Writing – review & editing, Conceptualization, Methodology. JS: Validation, Data curation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number “NBU-FFR-2025-2903-18".

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ancuti, C. O., Ancuti, C., De Vleeschouwer, C., and Bekaert, P. (2017). Color balance and fusion for underwater image enhancement. *IEEE Trans. image Process.* 27, 379–393. doi: 10.1109/TIP.2017.2759252

Ancuti, C., Ancuti, C. O., Haber, T., and Bekaert, P. (2012). "Enhancing underwater images and videos by fusion," in *2012 IEEE conference on computer vision and pattern recognition*. 81–88 (IEEE).

Anwar, S., Li, C., and Porikli, F. (2018). Deep underwater image enhancement. *arXiv preprint arXiv:1807.03528*. doi: 10.48550/arXiv.1807.03528

Chiang, J. Y., and Chen, Y.-C. (2011). Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* 21, 1756–1769. doi: 10.1109/TIP.2011.2179666

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., et al. (2020). "Multi-scale boosted dehazing network with dense feature fusion," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, WA, USA: IEEE Computer Society), 2157–2167.

Drews, P. L., Nascimento, E. R., Botelho, S. S., and Campos, M. F. M. (2016). Underwater depth estimation and image restoration based on single images. *IEEE Comput. Graph. Appl.* 36, 24–35. doi: 10.1109/MCG.2016.26

Fabbri, C., Islam, M. J., and Sattar, J. (2018). "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE international conference on robotics and automation (ICRA)*. 7159–7165 (IEEE).

Fu, Z., Wang, W., Huang, Y., Ding, X., and Ma, K.-K. (2022). "Uncertainty inspired underwater image enhancement," in *Computer Vision - ECCV 2022, 17th European Conference, Tel Aviv, Israel, October 23 -27, 2022, Proceedings, Part IV*, (Springer Cham) 465–482.

Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., et al. (2020). "Zero-reference deep curve estimation for low-light image enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, WA, USA: IEEE Computer Society), 1780–1789.

Guo, Y., Li, H., and Zhuang, P. (2019). Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE J. Ocean. Eng.* 45, 862–870. doi: 10.1109/JOE.2019.2911447

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33, 6840–6851. doi: 10.48550/arXiv.2006.11239

Iqbal, K., Odetayo, M., James, A., Salam, R. A., and Talib, A.Z. H. (2010). "Enhancing the low quality images using unsupervised colour correction method," in *2010 IEEE International Conference on Systems, Man and Cybernetics*. 1703–1709 (IEEE).

Iqbal, K., Salam, R. A., Osman, A., and Talib, A. Z. (2007). Underwater image enhancement using an integrated colour model. *IAENG Int. J. Comput. Sci.* 34.

Islam, M. J., Luo, P., and Sattar, J. (2020a). Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv preprint arXiv:2002.01155*.

Islam, M. J., Xia, Y., and Sattar, J. (2020b). Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* 5, 3227–3234. doi: 10.1109/LRA.2020.2974710

Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., and Ren, W. (2021). Underwater image enhancement *via* medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.* 30, 4985–5000. doi: 10.1109/TIP.2021.3076367

Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., et al. (2019). An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* 29, 4376–4389. doi: 10.1109/TIP.2019.2955241

Liu, R., Fan, X., Zhu, M., Hou, M., and Luo, Z. (2019). "Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light," in *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12), 4861–4875. doi: 10.1109/TCSVT.2019.2963772

Liu, R., Jiang, Z., Yang, S., and Fan, X. (2022). Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Trans. Image Process.* 31, 4922–4936. doi: 10.1109/TIP.2022.3190209

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). "Repaint: Inpainting using denoising diffusion probabilistic models," in 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos: IEEE Computer Society), 11461–11471.

Peng, L., Zhu, C., and Bian, L. (2023). U-shape transformer for underwater image enhancement. *IEEE Trans. Image Process.* doi: 10.1109/TIP.2023.3276332

Sahu, P., Gupta, N., and Sharma, N. (2014). A survey on underwater image enhancement techniques. *Int. J. Comput. Appl.* 87. doi: 10.5120/15268-3743

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*. doi: 10.48550/arXiv.2010.02502

Song, W., Wang, Y., Huang, D., and Tjondronegoro, D. (2018). "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia*, Hefei, China, September 21-22, 2018. 678–688 (Springer), Proceedings, Part I 19.

Sun, X., Liu, L., Li, Q., Dong, J., Lima, E., and Yin, R. (2019). Deep pixel-to-pixel network for underwater image enhancement and restoration. *IET Image Processing* 13, 469–474. doi: 10.1049/iet-ipr.2018.5237

Tang, Y., Kawasaki, H., and Iwaguchi, T. (2023). "Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy," in *Proceedings of the 31st ACM International Conference on Multimedia*. (New York: Association for Computing Machinery), 5419–5427.

Uplavikar, P. M., Wu, Z., and Wang, Z. (2019). "All-in-one underwater image enhancement using domain-adversarial learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Long Beach, CA, USA: IEEE).

Wang, Y., Guo, J., Gao, H., and Yue, H. (2021). Uiecˆ 2-net: Cnn-based underwater image enhancement using two color space. *Signal Processing: Image Communication* 96, 116250. doi: 10.1016/j.image.2021.116250

Wang, N., Zhou, Y., Han, F., Zhu, H., and Yao, J. (2019). Uwgan: underwater gan for real-world underwater color restoration and dehazing. *arXiv preprint arXiv:1912.10269*. doi: 10.48550/arXiv.1912.10269

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *ECCV 2018: 15th European Conference*, (Munich, Germany: Springer Cham), 3–19.

Ye, X., Xu, H., Ji, X., and Xu, R. (2018). "Underwater image enhancement using stacked generative adversarial networks," in *Advances in Multimedia Information Processing - PCM 2018*, (Springer Cham).

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). "Restormer: Efficient transformer for high-resolution image restoration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos: IEEE Computer Society), 5728–5739.

Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R. W., et al. (2018). "Dynamic scene deblurring using spatially variant recurrent neural networks," in *2018 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, (Salt Lake City, Utah: IEEE Computer Society), 2521–2529.

Zhang, W., Zhuang, P., Sun, H.-H., Li, G., Kwong, S., and Li, C. (2022). Underwater image enhancement *via* minimal color loss and locally adaptive contrast enhancement," in *IEEE Transactions on Image Processing*, Vol. 31. 3997–4010.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging*. 3, 47–57. doi: 10.1109/TCI.2016.2644865

Zhuang, P., Wu, J., Porikli, F., and Li, C. (2022). Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Trans. Image Process*. 31, 5442–5455. doi: 10.1109/TIP.2022.3196546