Hero-Mamba: Mamba-based Dual Domain Learning for Underwater Image Enhancement

Tejeswar Pokuri [*], **Shivarth Rai** [*]

Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal - 576104, India
tejeswar.mitmpl2022@learner.manipal.edu, shivarth.mitmpl2022@learner.manipal.edu

Abstract

Underwater images often suffer from severe degradation, such as color distortion, low contrast, and blurred details, due to light absorption and scattering in water. While learning-based methods like CNNs and Transformers have shown promise, they face critical limitations: CNNs struggle to model the long-range dependencies needed for non-uniform degradation, and Transformers incur quadratic computational complexity, making them inefficient for high-resolution images. To address these challenges, we propose Hero-Mamba, a novel Mamba-based network that achieves efficient dual-domain learning for underwater image enhancement. Our approach uniquely processes information from both the spatial domain (RGB image) and the spectral domain (FFT components) in parallel. This dual-domain input allows the network to decouple degradation factors, separating color/brightness information from texture/noise. The core of our network utilizes Mamba-based SS2D blocks to capture global receptive fields and long-range dependencies with linear complexity, overcoming the limitations of both CNNs and Transformers. Furthermore, we introduce a ColorFusion block, guided by a background light prior, to restore color information with high fidelity. Extensive experiments on the LSUI and UIEB benchmark datasets demonstrate that Hero-Mamba outperforms state-of-the-art methods. Notably, our model achieves a PSNR of 25.802 and an SSIM of 0.913 on LSUI, validating its superior performance and generalization capabilities.

Introduction

Underwater imaging technology is a critical tool for studying and exploring the Earth's oceans and seas, providing essential data for understanding marine ecosystems, climate change, and human impact (Monterroso Muñoz et al. 2023). It plays a crucial role in a wide range of applications, including autonomous underwater vehicle (AUV) navigation, marine biology, seabed mapping, and underwater structure inspection. The purpose of Underwater Image Enhancement (UIE) is to improve the visual quality of this imagery, a step that is vital for the accuracy of these downstream computer vision tasks and for our scientific understanding of the underwater world.

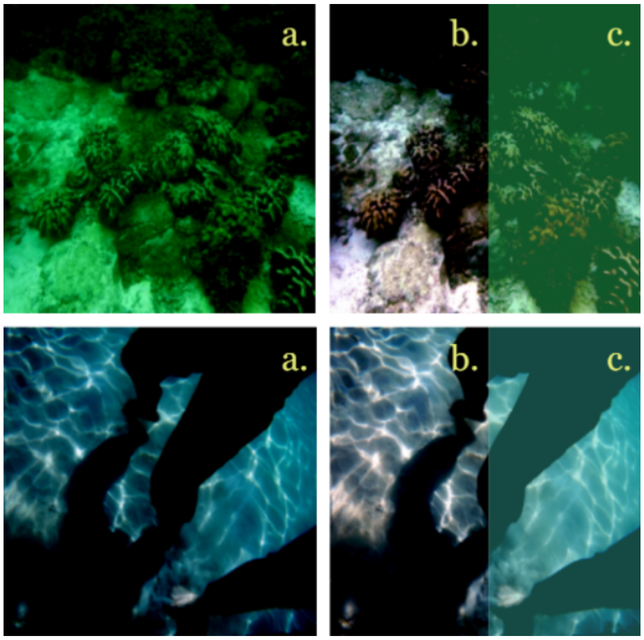


Figure 1: Visualizing the contribution of the background light prior on UIEB samples. (a) Original degraded images. (b) Ground Truth. (c) The background light prior, as estimated by our ColorFusion block, overlaid for visualization. The stark color difference between (b) and (c) highlights the severe color degradation caused by background light.

The quality of underwater images, however, is often severely compromised by the unique optical properties of water (Vlachos and Skarlatos 2021). The medium itself, along with dissolved impurities and suspended particles, causes strong light absorption and scattering. This degradation manifests as low contrast, blurred details, and significant color distortion. Light absorption is wavelength-dependent, with longer wavelengths (like red) attenuating much faster than shorter wavelengths (like blue and green). This phenomenon is the primary cause of the characteristic blue-green color cast seen in most underwater photographs, which obscures the true colors of the scene (Zhou et al. 2023).

For a mathematical description of this degradation, we

[*] These authors contributed equally.

can utilize the simplified underwater imaging model provided by (McGlamery 1980) and (Jaffe 2002). This model postulates that a degraded image I acquired by a camera comprises two primary constituents: direct illumination and backscattering.The direct illumination is the true scene radiance J (the clear image) attenuated by the medium transmission rate, t. The backscattering is the ambient background light B scattered by particles into the camera's path, veiling the scene. This relationship can be expressed as:

$$I_c(x) = J_c(x)t_c(x) + B_c(1 - t_c(x)), \tag{1}$$

here, $I_c(x)$ is the observed degraded image pixel, $J_c(x)$ is the clear image, $t_c(x)$ is the transmission map, and B_c is the background light.

Traditional physical-model based UIE methods (Huang, Cheng, and Chiu 2013; Drews Jr et al. 2013; Akkaynak and Treibitz 2019; Berman et al. 2021) aim to directly estimate the parameters of the imaging model to reverse the degradation. However, these models require manual tuning of parameters that perform sub optimally in diverse underwater environments leading to poor generalization. With the rise of deep learning, Convolutional Neural Networks (CNNs) became a dominant approach (Huang et al. 2023a; Islam, Xia, and Sattar 2020a; Li et al. 2021; Fu et al. 2022a). CNNs can automatically learn complex feature representations from large datasets for end-to-end image restoration. However, their reliance on static, local convolutional kernels limits their receptive field, making it difficult to capture long-range pixel dependencies required to accurately non-uniform degradation across an entire image.

Recently, Transformers and their self-attention mechanism have been extended to vision tasks with great success (Peng, Zhu, and Bian 2023; Khan et al. 2024; Huang et al. 2022; Ren et al. 2022). Using self-attention mechanisms, Transformers can capture global context and model long-range dependencies effectively. However, their primary drawback is the quadratic computational complexity with respect to the image's spatial size, making them computationally expensive and difficult to apply to high-resolution underwater images.

To address the quadratic complexity issue of Transformers, state space models (Gu, Goel, and Ré 2022) with selective mechanisms such as Mamba have been developed. Mamba (Gu and Dao 2024) provides the ability to capture long-range dependencies and global features with linear complexity. These models have gained widespread attention and application in visual tasks such as semantic segmentation and object detection.

In this study, we wish to address the aforementioned complexities of underwater image enhancement, and make the following contributions:

- We propose Hero-Mamba, a novel U-shaped network that introduces a Mamba-based dual-domain learning paradigm. By processing spatial (RGB) and spectral (FFT) features in parallel, our model efficiently captures global long-range dependencies with linear complexity while simultaneously decoupling degradation factors for more effective enhancement.
- Leveraging a physical imaging model-based ColorFusion block, we incorporate background light prior in our network for accurate color restoration in underwater scenes.
- Extensive quantitative and qualitative experiments demonstrate that Hero-Mamba surpasses state-of-the-art methods on the public UIEB and LSUI benchmark datasets, showing improvements in both structural similarity and perceptual quality.

Related Work

Traditional methods. (Huang, Cheng, and Chiu 2013) presents a hybrid method for contrast enhancement using gamma correction and probability distribution of luminance pixels, to effectively enhance dimmed images by balancing high visual quality with low computational cost. (Drews Jr et al. 2013) introduces the underwater dark channel prior (UDCP), an adaptation of the dark channel prior for estimation of transmission in underwater scenes. (Berman et al. 2021) account for the wavelet-dependent attenuation of light in underwater environments, proposing a physics-based model for image restoration utilizing haze-lines prior.

CNN-based methods. (Huang et al. 2023a) introduces a novel semi-supervised learning framework. Combining a reliable pseudo-labeling mechanism with contrastive learning, the framework effectively exploits unlabeled data to improve model performance on underwater images. (Li et al. 2021) introduces Ucolor, a network combining multi-color space embedding with physics-inspired guidance. The network processes and adaptively fuses features from RGB, Lab and HSV inputs to capture characteristics from different color spaces. Using physical imaging model, the network computes a medium transmission map, guiding it to focus on regions with higher degradation. (Fu et al. 2022a) resolves UIE into distribution estimation and consensus process. Combining a conditional variational autoencoder (CVAE) with adaptive instance normalization (AdaIN), the network learns to model a distribution of possible enhanced outputs for a single input. Following this, a consensus process is utilized to predict a deterministic result from the distribution.

Transformer-based methods. (Peng, Zhu, and Bian 2023) introduces the U-Shape Transformer, integrated with a channel-wise multi-scale feature fusion transformer (CMSFFT) module and a spatial-wise global feature modeling transformer (SGFMT). These modules, designed specifically for UIE, target inconsistent attenuation across channels and spatial areas. (Khan et al. 2024) presents Spectroformer, a novel UIE transformer, integrating spatial and frequency domains using its Multi-Domain Query Cascaded Attention (MQCA) mechanism, which uses frequency-domain queries with spatial keys/values. A Spatio-Spectro Fusion Attention Block enhances feature propagation in skip connections by fusing both domains. Finally, a Hybrid Fourier-Spatial Upsampling Block combines upsampling techniques for superior feature resolution enhancement.(Ren et al. 2022) proposes a U-Net based network employing Swin Transformer blocks. This work introduces Reinforced Swin-Convs Transformer Block (RSCTB), which incorporates convolutions within the Swin Transformer's attention mechanism to cap-

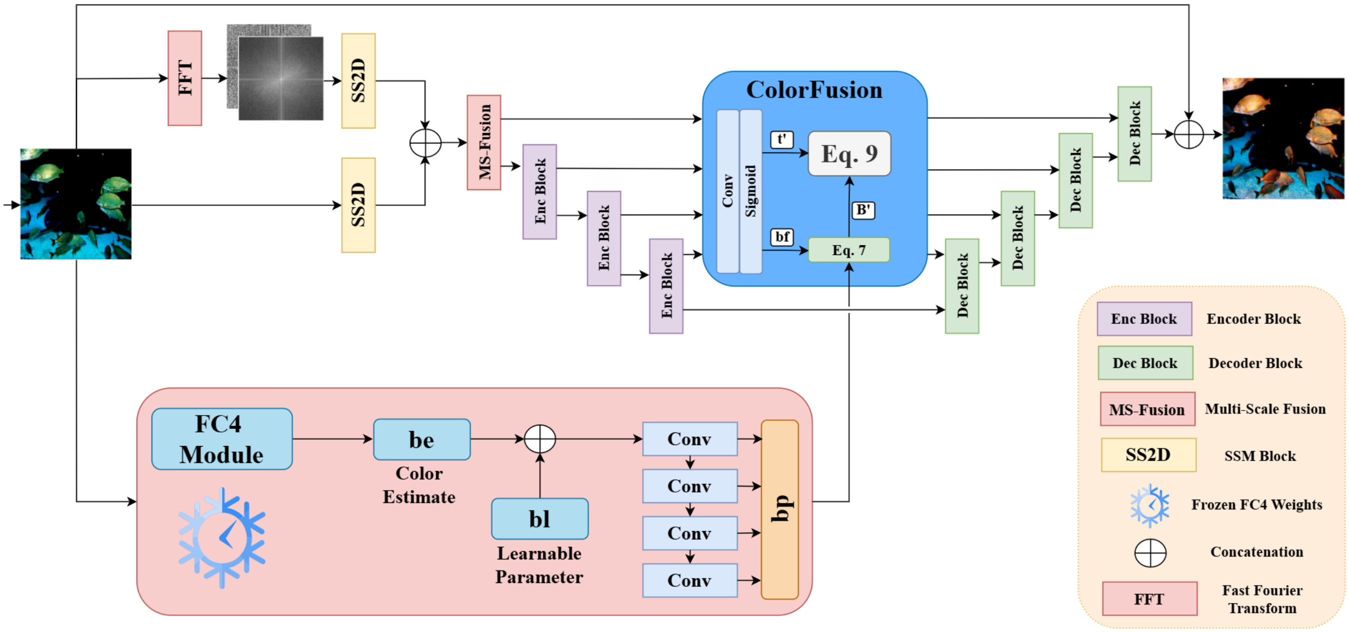


Figure 2: Architectural design of Hero-Mamba, utilizing spatial and spectral domains for accurate feature reconstruction, and ColorFusion block for enhanced color restoration. Using SS2D layers allows for long-range feature learning efficiently.

ture local attention alongside the Transformer's global dependency modeling.

SSM-based methods. (Peng and Bian 2025) propose SS-UIE, a Mamba-based network achieving spatial-spectral dual-domain adaptive learning with linear complexity, addressing limitations of prior CNN and Transformer methods. It introduces the Spatial-Spectral (SS) block, which combines a Mamba-inspired Multi-scale Cycle Selective Scan (MCSS) for global spatial modeling and an FFT-based Spectral-Wise Self-Attention (SWSA) for spectral modeling in parallel, allowing the network to model degradations of different spatial regions and spectral bands. (Zhang et al. 2024) introduces Mamba-UIE, a physical model constraint-based UIE framework. The network utilizes Mamba-In-Convolution Blocks (MIC) to capture long-range dependencies in the spatial and channel dimensions, while the CNN backbone extracts local features. (Guan et al. 2024) introduce WaterMamba, which incorporates the proposed Spatial-Channel Omnidirectional Selective Scan (SCOSS) blocks. SCOSS blocks model pixel-channel dependencies by processing spatial and channel information in four directions.

Methodology

Overall Architecture

The proposed Hero-Mamba network (Fig.2) primarily consists of encoder, decoder and ColorFusion blocks, following a U-shaped structure with residual connections for multi-scale feature fusion. Given a degraded underwater image $I \in R^{3\times H\times W}$, I is first passed through a two-dimensional selective-scan module (SS2D) to obtain encoded spatial features. Since underwater image degradation causes global losses, utilizing SS2D modules allows us to capture long-range spatial information efficiently. In parallel, we apply the Fast Fourier Transform (FFT) to obtain amplitude component and phase component of I, and concatenate them together to obtain $I_S \in R^{2\times H\times W}$, a spectral representation of the input image. As discussed in (Cheng et al. 2024), the amplitude and phase components obtained by applying FFT can allow the network to enhance overall brightness and color accuracy by processing the amplitude while separately enhancing details and reducing noise by processing the phase. I_S is passed through a SS2D layer, the output of which is concatenated to the output of the parallel SS2D with I as input, to get the initial feature representation, $I_F \in R^{5\times H\times W}$. This operation is represented as follows:

$$I_F = SS2D(I) \oplus SS2D(I_S), \tag{2}$$

where $\oplus$ represents the concatenation operation.

Following this, I_F passes through a MS-Fusion block and three encoder blocks, each down samples features by a factor of 2, to dimensions of $f_1 \in R^{32\times\frac{H}{2}\times\frac{W}{2}}$, $f_2 \in R^{64\times\frac{H}{4}\times\frac{W}{4}}$, $f_3 \in R^{128\times\frac{H}{8}\times\frac{W}{8}}$ and $f_4 \in R^{256\times\frac{H}{16}\times\frac{W}{16}}$ respectively. The encoder block, as shown in Fig. 3, a SS2D layer followed by MS-Fusion block, is designed to capture global dependencies in linear complexity and extract multi-scale features, providing a rich feature representation of the input.

Simultaneously, each of the multi-scale features, $f_i \forall i \in \{1, 2, 3, 4\}$, are fed through a ColorFusion block. In the architecture, we use ColorFusion blocks to preserve the color accuracy of the generated output. They are discussed in the following subsection. Let the output of each ColorFusion block be represented by $c_i \forall i \in \{1, 2, 3, 4\}$, then c_i can be

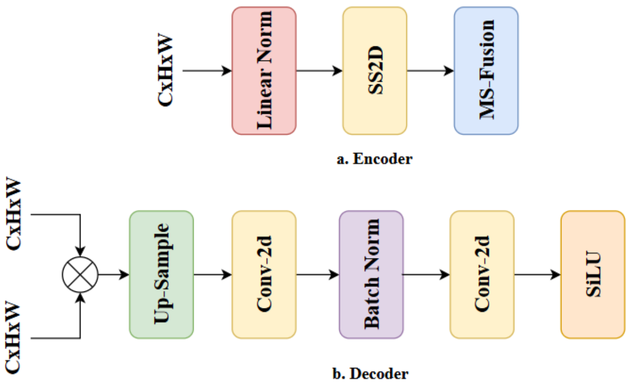


Figure 3: Overview of the Encoder Block

represented as:

$$c_i = ColorFusion(f_i) \tag{3}$$

Then, feature f_4 passes through the decoder network, consisting of four decoder blocks. As in Fig. 3, each decoder consists of an up sampling layer, followed by Conv layer, BatchNorm layer, Conv layer and SiLU activation function. Each decoder receives decoded features and ColorFusion features, c_i of corresponding size concatenated together. Let the output of each decoder block be represented by $d_i \forall i \in \{1, 2, 3, 4\}$, then, in general, d_i can be represented as:

$$d_i = SiLU(Conv(BN(Conv(Up(d_{i-1} \oplus c_i))))) \tag{4}$$

where d_{i-1} is the decoded feature from preceding decoder block and $\oplus$ represents the concatenation operation.

ColorFusion Block

Motivated by the success of Joint Prior Module in (Fan et al. 2024a), we adapt a simplified version in the form of ColorFusion block in our network. As discussed earlier, severe color degradation is prevalent in underwater imagery due to selective absorption of longer wavelength signals cause images to have a predominant blue/green hue. In the underwater imaging model described in Eq. 1, background light heavily affects color and visibility (see Fig.1). Thus, using it as ”prior knowledge” provides a stable, theory-based lighting model that helps guide the restoration process.

The blue/green hue for each degraded image can be estimated as background light prior using the theory of color constancy (Cheng, Prasad, and Brown 2014; Gehler et al. 2008). An initial background light prior, b_e is estimated using a FC4 (Hu, Wang, and Lin 2017) model pre-trained on color-constant data (Shi 2000). Since the pre-training data is captured on land, estimated b_e is an approximate, and to further tune the prior for underwater environments, a learnable parameter, b_l, is computed along with b_e. The resultant estimated background light prior, b_p, then, can be expressed as:

$$b_p = Conv(b_e + b_l), \tag{5}$$

where $+$ is the expansion operation. Together with b_p, a background light feature is extracted directly from the input image features. This feature, b_f, is represented as:

$$b_f = Sigmoid(Conv(f_i)), \tag{6}$$

where f_i is the feature extracted in the encoder at corresponding scale. Then, b_p and b_f are dynamically mixed in order to get an accurate background light prior estimate for the input image. This is represented by:

$$B' = \omega b_f + (1 - \omega) b_p, \tag{7}$$

where ω is a dynamic weight coefficient ranging from 0 to 1. Now, using the underwater image model in Eq.1, we can approximate (Fan et al. 2024b) the undegraded image J as:

$$J = It + B(1 - t) \tag{8}$$

In terms of feature representations in the network, Eq.8 can be written as:

$$c_i = f_i t' + B'(1 - t'), \tag{9}$$

where c_i is the output of the ColorFusion block, f_i is the encoded input feature, B' is the background light prior, and t' is the feature transmission map, estimated as follows:

$$t' = Sigmoid(Conv(f_i) \tag{10}$$

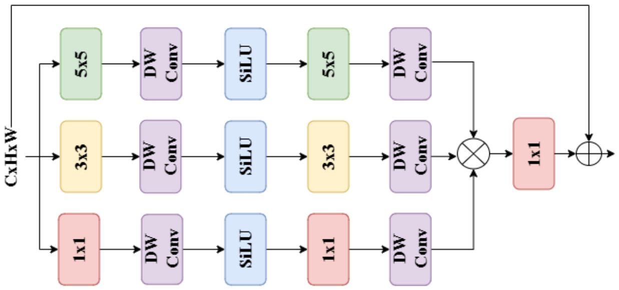


Figure 4: Overview of the MS-fusion block. Parallel branches of 1×1, 3×3 and 5×5 kernel sizes facilitate comprehensive feature extraction.

MS-Fusion Block

As shown in Fig. 4, MS-Fusion block simultaneously processes features at three spatial scales, providing a comprehensive feature representation required for restoring local details and sharp edges. The MS-Fusion block has a multi-branch structure, where the first branch can be defined as:

$$B_1 = D_W^{1\times1} C^{1\times1}(SiLU(D_W^{1\times1} C^{1\times1}(F))), \tag{11}$$

where $D_W^{1\times1}$ denotes depthwise convolution of size 1×1, $C^{1\times1}$ represents 1×1 convolution, and F is the input feature. Similarly, the second and third branches perform the same set of operations with 3×3 and 5×5 kernel sizes, respectively. Following this, the features from all three branches, B_1, B_2 and B_3, are concatenated and passed through a 1×1 convolution layer to reduce feature channels. This feature aggregate is then added back to the input feature to get the MS-Fusion block output. This is represented as:

$$F' = C^{1\times1}(B_1 \oplus B_2 \oplus B_3) + F, \tag{12}$$

here $\oplus$ denotes concatenation and, $+$ denotes element-wise addition.

Loss Function

In order to holistically address the multi-faceted nature of degradation in underwater images, such as low contrast, blurred details and color distortion, we use a composite loss function with L1 loss, SSIM loss and contrastive loss components.

L1 Loss: calculates the average of absolute difference between every single pixel in the model's output image and the corresponding pixel in the ground truth. This component promotes pixel-level accuracy, making sure the colors and brightness values of the enhanced image are as close as possible to the ground truth. L1 loss is defined as follows:

$$\mathcal{L}_1 = \frac{\sum_{i=1}^{n} |f(X_i) - Y_i|}{n}, \tag{13}$$

where $f(X_i)$ represents generated output and Y_i is ground truth for the i^{th} sample, and, n is the number of samples.

SSIM Loss: based on the Structural Similarity Index (SSIM), a metric designed to measure the perceptual quality of an image, which aligns better with human judgment. SSIM Loss is defined as follows:

$$\mathcal{L}_{ssim} = 1 - \frac{(2\mu_X\mu_Y + \varepsilon_1)(2\sigma_{XY} + \varepsilon_2)}{(\mu_X^2 + \mu_Y^2 + \varepsilon_1)(\sigma_X^2 + \sigma_Y^2 + \varepsilon_2)}, \tag{14}$$

where μ_X and μ_Y represent the respective means of X and Y, while σ_X^2 and σ_Y^2 are their respective variances. σ_{XY} denotes the covariance between X and Y. ε_1 and ε_2 are small constants included for numerical stability.

Contrastive Loss: helps the model learn a better feature representation, ensuring that the features of the generated image are mathematically closer to the features of the ground truth than to the degraded image. This teaches the model to learn the features of a high-quality underwater image, leading to a more robust and perceptually accurate restoration. This loss is defined as:

$$\min \|J - \phi(I, w)\| + \beta \cdot \rho(\varphi(p), \varphi(n), \varphi(\phi(I, w))), \tag{15}$$

where $\min$ represents the minimization objective, J is the reference image, and $\phi(I, w)$ is the predicted undegraded image J', derived through network parameters w from the degraded image I, i.e., the anchor sample. $\|J - \phi(I, w)\|$ measures the difference between J and J', using the L1 norm. In the second term, β is used to adjust the weight of different features of the samples, and ρ measures the similarity of features between samples, also using the $\mathcal{L}_1$ norm. $\phi(p)$, $\phi(n)$, and $\phi(\phi(I, w))$ represent the common intermediate features extracted for all samples through the same pretrained model, where p is the positive sample, n is the negative sample, and $\phi(I, w)$ is the prediction from the anchor sample.

Our overall composite loss function is expressed as:

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_{ssim} + \gamma\mathcal{L}_{contrastive}, \tag{16}$$

where α, β and γ are the weighting coefficients. After extensive experimentation, they were set to 0.3, 0.8 and 0.1 respectively.

Experiments

Experimental Setup

Implementation Details. The proposed model was implemented using the PyTorch 2.4.1 framework and was trained and tested on an NVIDIA Tesla T4 GPU. The network was trained end-to-end using the AdamW (Loshchilov and Hutter 2019) optimizer with a learning rate of 3×10^{-4} and $\beta_1 = 0.9$, $\beta_2 = 0.999$. All images were resized to a fixed size of 256×256 pixels. A training batch size of 4 was used, and the network was trained for 250 epochs. The learning rate was dynamically adjusted using a cosine annealing strategy (Zamir et al.).

Datasets. We utilize two publicly available and widely bench marked underwater image datasets, UIEB (Li et al. 2020) and LSUI (Peng, Zhu, and Bian 2023), for our study. UIEB (Underwater Image Enhancement Benchmark) dataset contains 890 real underwater images with reference. We divide the 890 paired images into a training set of 800 image pairs and validation set of 90 image pairs. LSUI (Large-Scale Underwater Image) dataset contains 4279 real underwater image pairs. We partition 3851 pairs for the training set and 428 pairs for the test set. These datasets cover various underwater scenes and imaging conditions, such as coral reefs, underwater organisms, and underwater terrains.

Evaluation Metrics. For a comprehensive assessment of the performance of our proposed method, we utilize SSIM (Wang et al. 2004),PSNR (Korhonen and You 2012), FSIM (Zhang et al. 2011) and LPIPS (Zhang et al. 2018) metrics. Peak Signal to Noise Ratio (PSNR) computes the ratio of image signal to noise and reflects the overall quality of the generated image. Structural Similarity Index Measure (SSIM) assesses the perceived similarity between two images based on structural information. Feature Similarity Index Measure (FSIM) assesses image quality by comparing low-level features between two images. Learned Perceptual Image Patch Similarity (LPIPS) evaluates the perceptual similarity between two image patches using features learned by deep neural networks (eg. VGG19).

Comparison Methods. To validate the performance of our proposed method, we perform a comparative analysis between our Hero-Mamba and 10 SOTA methods for UIE. These methods include U-GAN (Fabbri, Islam, and Sattar 2018), FUnIE-GAN (Islam, Xia, and Sattar 2020b), U-Shape Transformer (Peng, Zhu, and Bian 2023), SS-UIE (Peng and Bian 2025), CE-VAE (Pucci and Martinel 2024), WaterMamba (Guan et al. 2024), Semi-UIR (Huang et al. 2023b), PUIE-Net (Fu et al. 2022b), WF-Diff (Zhao et al. 2023) and NU2Net (Guo et al. 2022). Wherever possible, the results for these methods have been obtained using their publicly available codes.

Qualitative Analysis

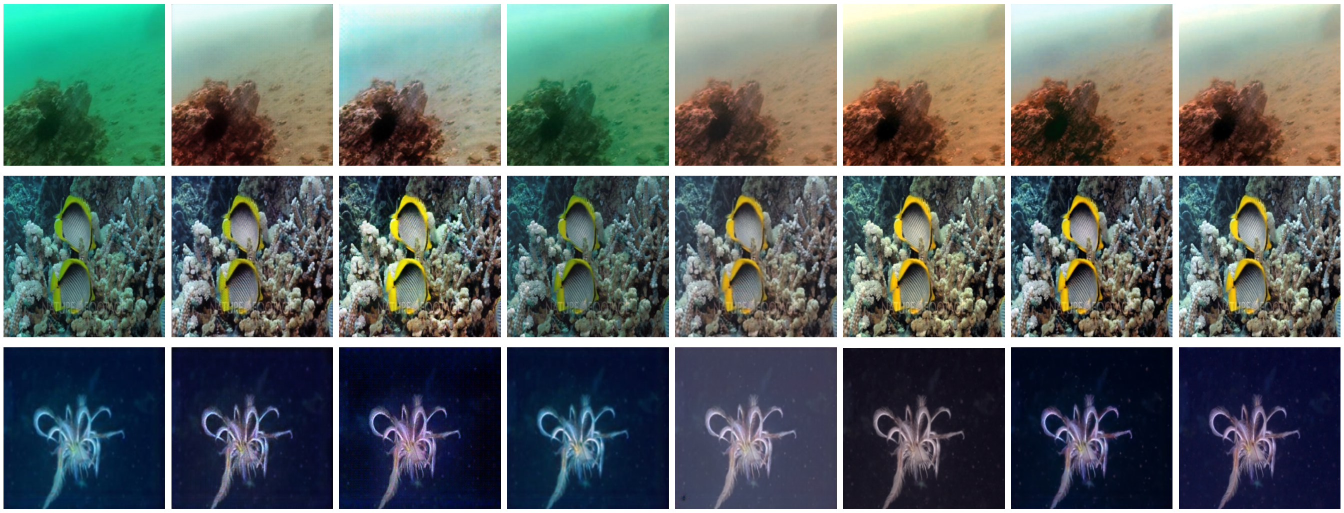
Figures 5 and 6 show visual comparison between Hero-Mamba and competing models. We select 3 images from both UIEB and LSUI datasets to cover a wide range of scenes. It is observed in both figures that Hero-Mamba outputs are the closest to the reference images in both datasets,

Table 1: Quantitative comparison on LSUI Dataset.The best result is highlighted in red and the second best result is in blue.

Paper Name	Venue	SSIM ↑	PSNR ↑	LPIPS ↓	FSIM ↑
U-Gan	ICRA 2018	0.772	19.423	0.374	0.763
FUnIE-Gan	RAL 2020	0.798	20.783	0.234	0.833
U shape Transformer	TIP 2023	0.821	21.623	0.298	0.847
SS-UIE	AAAI 2025	0.816	19.093	0.270	0.892
CE-VAE	WACV 2024	0.832	22.638	0.127	0.932
Water Mamba	Arxiv 2024	0.877	23.463	0.134	0.937
Hero-Mamba (Ours)		0.913	25.802	0.117	0.958

Table 2: Quantitative comparison on UIEB Dataset. The best result is highlighted in red and the second best result is in blue.

Paper Name	Venue	SSIM ↑	PSNR ↑	LPIPS ↓	FSIM ↑
U-Gan	ICRA 2018	0.805	19.676	0.197	0.912
FUnIE-Gan	RAL 2020	0.814	18.781	0.163	0.919
Semi-UIR	CVPR 2023	0.821	23.400	0.157	0.932
PUIE-Net	ECCV 2022	0.854	21.501	0.132	0.863
WF-Diff	CVPR 2024	0.873	27.260	0.139	0.897
NU2Net	AAAI 2023	0.907	22.633	0.100	0.949
Water Mamba	Arxiv 2024	0.931	24.751	0.143	0.973
Hero-Mamba (Ours)		0.942	24.526	0.125	0.945



a) Input b) U-Gan c) FUnIE-Gan d) SS-UIE e) CE-VAE f) Water Mamba g) Hero-Mamba h) Ground Truth

Figure 5: Visual comparison of enhancement results by various models on LSUI dataset.

with high contrast, minimal blurriness, natural-looking colors and high-fidelity local details.

Quantitative Analysis

We conducted comprehensive quantitative comparison of our proposed Hero-Mamba against 10 state-of-the-art (SOTA) methods, with the results detailed in Table 1 and Table 2. As shown in Table 1, on the LSUI dataset, our Hero-Mamba model outperforms all other competing methods across all four evaluation metrics. Notably, Hero-Mamba achieves an SSIM of 0.913 and a PSNR of 25.802, significantly surpassing the second-best method, WaterMamba, which scored 0.877 and 23.463, respectively. Furthermore, our model obtains the best (lowest) LPIPS score of 0.117 and the highest FSIM score of 0.958, confirming that its enhanced images are not only mathematically accurate but also perceptually closer to the ground truth.

On the UIEB dataset, presented in Table 2, our method achieves the highest SSIM score of 0.942, outperforming all competing methods. While WF-Diff achieves a higher PSNR and NU2Net a lower LPIPS, our Hero-Mamba maintains strong performance across all metrics, positioning it as a competitive solution.

To assess our model's generalization abilities, we per-

Table 3: Cross-dataset study results for Hero-Mamba.

Train Dataset	Test Dataset	SSIM ↑	PSNR ↑	FSIM ↑	LPIPS ↓
LSUI	UIEB	0.868	19.655	0.934	0.129
UIEB	LSUI	0.846	20.380	0.919	0.206

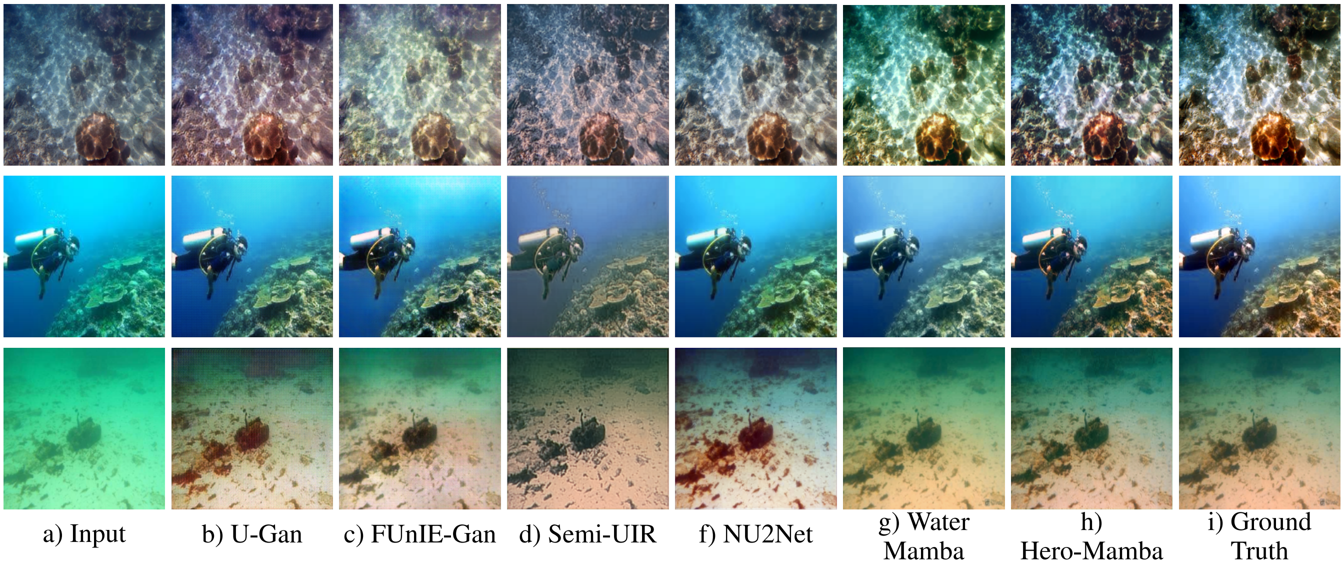


Figure 6: Visual comparison of enhancement results by various methods on UIEB dataset.

Table 4: Break-down ablation study for Hero-Mamba.

Model	SSIM
Base	0.847
Base + MS-Fusion	0.872
Base + MS-Fusion + SS2D	0.890
Base + MS-Fusion + SS2D + FFT	0.914
Base + MS-Fusion + SS2D + FFT + ColorFusion	0.942

form a cross-dataset evaluation, as shown in Table 3. When trained on the LSUI dataset and tested on UIEB, the model achieves a strong SSIM of 0.868 and FSIM of 0.934. Similarly, when trained on UIEB and tested on LSUI, it achieves an SSIM of 0.846 and PSNR of 20.380. These results demonstrate the robust generalization capability of Hero-Mamba.

Ablation Study

To demonstrate the effectiveness of each component in our proposed Hero-Mamba, we conducted a series of ablation studies with the UIEB dataset. The quantitative results of progressively adding each module are presented in Table 4.

We start with a ”Base” model, a simple encoder-decoder network with skip connections, which yields a baseline SSIM of 0.847. By incorporating the MS-Fusion block, the performance improves to 0.872, which demonstrates the effectiveness of processing features at multiple scales to reconstruct details. Subsequently, adding the Mamba-based SS2D blocks further boosts the SSIM to 0.890, validating the importance of global receptive fields and capturing long-range dependencies.

A performance jump to 0.914 is observed when we introduce FFT input, proving that our parallel, dual-domain learning strategy is effective for decoupling degradation factors. Finally, the full Hero-Mamba model, with the addition of ColorFusion block, achieves the highest SSIM of 0.942. This highlights the role of the background light prior in achieving accurate color restoration.

Conclusion

In this paper, we proposed Hero-Mamba, a novel dual-domain network for underwater image enhancement that leverages the linear complexity and long-range modeling capabilities of Mamba-based state space models. Our core contribution is a parallel architecture that processes both spatial (RGB) and spectral (FFT) features from the initial input. This dual-domain approach, powered by SS2D blocks, allows the network to efficiently model global dependencies and decouple complex degradation factors—addressing color, contrast, and detail simultaneously. Furthermore, we utilize a physics-guided ColorFusion block to accurately restore color by incorporating a background light prior, and an MS-Fusion block to reconstruct fine-grained local details. Extensive quantitative and qualitative experiments on the UIEB and LSUI datasets validate that Hero-Mamba achieves state-of-the-art performance, outperforming previous methods in key metrics like SSIM and PSNR.

References

Akkaynak, D.; and Treibitz, T. 2019. Sea-Thru: A Method for Removing Water From Underwater Images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1682–1691.

Berman, D.; Levy, D.; Avidan, S.; and Treibitz, T. 2021. Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2822–2837.

Cheng, D.; Prasad, D. K.; and Brown, M. S. 2014. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America A*, 31(5): 1049–1058.

Cheng, Z.; Fan, G.; Zhou, J.; Gan, M.; and Chen, C. P. 2024. FDCE-Net: underwater image enhancement with embedding frequency and dual color encoder. *IEEE Transactions on Circuits and Systems for Video Technology*.

Drews Jr, P.; do Nascimento, E.; Moraes, F.; Botelho, S.; and Campos, M. 2013. Transmission Estimation in Underwater Single Images. In *2013 IEEE International Conference on Computer Vision Workshops*, 825–830.

Fabbri, C.; Islam, M. J.; and Sattar, J. 2018. Enhancing Underwater Imagery using Generative Adversarial Networks. arXiv:1801.04011.

Fan, J.; Xu, J.; Zhou, J.; Meng, D.; and Lin, Y. 2024a. See through water: Heuristic modeling towards color correction for underwater image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.

Fan, J.; Xu, J.; Zhou, J.; Meng, D.; and Lin, Y. 2024b. See through water: Heuristic modeling towards color correction for underwater image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.

Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; and Ma, K.-K. 2022a. Uncertainty Inspired Underwater Image Enhancement. arXiv:2207.09689.

Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; and Ma, K.-K. 2022b. Uncertainty Inspired Underwater Image Enhancement. arXiv:2207.09689.

Gehler, P. V.; Rother, C.; Blake, A.; Minka, T.; and Sharp, T. 2008. Bayesian color constancy revisited. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Gu, A.; Goel, K.; and Ré, C. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396.

Guan, M.; Xu, H.; Jiang, G.; Yu, M.; Chen, Y.; Luo, T.; and Song, Y. 2024. WaterMamba: Visual State Space Model for Underwater Image Enhancement. arXiv:2405.08419.

Guo, C.; Wu, R.; Jin, X.; Han, L.; Chai, Z.; Zhang, W.; and Li, C. 2022. Underwater Ranker: Learn Which Is Better and How to Be Better. arXiv:2208.06857.

Hu, Y.; Wang, B.; and Lin, S. 2017. FC^4 : $FullyConvolutionalColorConstancywithConfidence-WeightedPooling.330--339.$

Huang, S.; Wang, K.; Liu, H.; Chen, J.; and Li, Y. 2023a. Contrastive Semi-supervised Learning for Underwater Image Restoration via Reliable Bank. arXiv:2303.09101.

Huang, S.; Wang, K.; Liu, H.; Chen, J.; and Li, Y. 2023b. Contrastive Semi-supervised Learning for Underwater Image Restoration via Reliable Bank. arXiv:2303.09101.

Huang, S.-C.; Cheng, F.-C.; and Chiu, Y.-S. 2013. Efficient Contrast Enhancement Using Adaptive Gamma Correction With Weighting Distribution. *IEEE Transactions on Image Processing*, 22(3): 1032–1041.

Huang, Z.; Li, J.; Hua, Z.; and Fan, L. 2022. Underwater Image Enhancement via Adaptive Group Attention-Based Multiscale Cascade Transformer. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–18.

Islam, M. J.; Xia, Y.; and Sattar, J. 2020a. Fast Underwater Image Enhancement for Improved Visual Perception. arXiv:1903.09766.

Islam, M. J.; Xia, Y.; and Sattar, J. 2020b. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robotics and Automation Letters*, 5(2): 3227–3234.

Jaffe, J. S. 2002. Computer modeling and the design of optimal underwater imaging systems. *IEEE journal of oceanic engineering*, 15(2): 101–111.

Khan, M. R.; Mishra, P.; Mehta, N.; Phutke, S. S.; Vipparthi, S. K.; Nandi, S.; and Murala, S. 2024. Spectroformer: Multi-Domain Query Cascaded Transformer Network For Underwater Image Enhancement. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1443–1452.

Korhonen, J.; and You, J. 2012. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 37–38. Melbourne, VIC, Australia: IEEE.

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 Transactions on Image Processing*, 30: 4985–5000.

Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2020. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Transactions on Image Processing*, 29: 4376–4389.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

McGlamery, B. 1980. A computer model for underwater camera systems. In *Ocean optics VI*, volume 208, 221–231. SPIE.

Monterroso Muñoz, A.; Moron-Fernández, M. J.; Cascado-Caballero, D.; Diaz-Del-Rio, F.; and Real, P. 2023. Autonomous Underwater Vehicles: Identifying Critical Issues and Future Perspectives in Image Acquisition. *Sensors*, 23(10): 4986.

Peng, L.; and Bian, L. 2025. Adaptive Dual-domain Learning for Underwater Image Enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6): 6461–6469.

Peng, L.; Zhu, C.; and Bian, L. 2023. U-Shape Transformer for Underwater Image Enhancement. *IEEE Transactions on Image Processing*, 32: 3066–3079.

Pucci, R.; and Martinel, N. 2024. CE-VAE: Capsule Enhanced Variational AutoEncoder for Underwater Image Enhancement. arXiv:2406.01294.

Ren, T.; Xu, H.; Jiang, G.; Yu, M.; Zhang, X.; Wang, B.; and Luo, T. 2022. Reinforced Swin-Convs Transformer for Simultaneous Underwater Sensing Scene Image Enhancement and Super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Shi, L. 2000. Re-processed version of the gehler color constancy dataset of 568 images. *http://www. cs. sfu. ca/˜ color/data/*.

Vlachos, M.; and Skarlatos, D. 2021. An Extensive Literature Review on Underwater Image Colour Correction. *Sensors*, 21(17): 5690.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. ???? Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8): 2378–2386.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924.

Zhang, S.; Duan, Y.; Li, D.; and Zhao, R. 2024. Mamba-UIE: Enhancing Underwater Images with Physical Model Constraint. arXiv:2407.19248.

Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2023. Wavelet-based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration. arXiv:2311.16845.

Zhou, J.; Liu, D.; Zhang, D.; and Zhang, W. 2023. Light Attenuation and Color Fluctuation for Underwater Image Restoration. 374–389. ISBN 978-3-031-26312-5.