PixMamba: Leveraging State Space Models in a Dual-Level Architecture for Underwater Image Enhancement

Wei-Tung Lin[1], Yong-Xiang Lin[1], Jyun-Wei Chen[1], and Kai-Lung Hua[1,2]

[1] Dept. of Computer Science and Information Engineering,
National Taiwan University of Science and Technology
{m11115116,d10915006,m11215034,hua}@mail.ntust.edu.tw
[2] Microsoft
kai.hua@microsoft.com

Abstract. Underwater Image Enhancement (UIE) is critical for marine research and exploration but hindered by complex color distortions and severe blurring. Recent deep learning-based methods have achieved remarkable results, yet these methods struggle with high computational costs and insufficient global modeling, resulting in locally under- or over-adjusted regions. We present PixMamba, a novel architecture, designed to overcome these challenges by leveraging State Space Models (SSMs) for efficient global dependency modeling. Unlike convolutional neural networks (CNNs) with limited receptive fields and transformer networks with high computational costs, PixMamba efficiently captures global contextual information while maintaining computational efficiency. Our dual-level strategy features the **patch-level** Efficient Mamba Net (EMNet) for reconstructing enhanced image feature and the **pixel-level** PixMamba Net (PixNet) to ensure fine-grained feature capturing and global consistency of enhanced image that were previously difficult to obtain. PixMamba achieves state-of-the-art performance across various underwater image datasets and delivers visually superior results. Code is available at `https://github.com/weitunglin/pixmamba`.

Keywords: Underwater Image Enhancement · State Space Models

1 Introduction

Underwater environments pose unique challenges for image acquisition due to factors such as severe blurring, color distortion [1], low contrast, and complex light scattering [22,32] caused by wavelength-dependent absorption. These issues impede the quality and clarity of underwater images, making effective enhancement methods critical for various applications in marine archaeology, ecological and biological research. Therefore, Underwater Image Enhancement (UIE) is a crucial step in improving the underwater images quality. This enhancement facilitates improved understanding of the underwater world and enables the successful execution of high-level oceanography tasks [4, 8].

Traditional image enhancement methods [1, 2, 45] have relied on statistical properties and physical assumptions about the image and environment. These methods often attempt to correct color distortions and improve contrast using hand-crafted priors. However, they typically struggle with dynamic scenes and often fall short in restoring texture information and handling extensive blurring. Recent advancements in deep learning have introduced new approaches to underwater image enhancement. Methods utilizing convolutional neural networks (CNNs) are widely used for UIE due to their capability to learn visual representations end-to-end [6, 9, 15, 20, 21], which is more efficient and effective compared to traditional UIE methods. However, CNN-based methods have limitations: a small receptive field hinders modeling long-range pixel dependencies, and fixed convolutional kernels cannot adapt to the images across various underwater scenarios. The Transformer-based model, initially proposed for natural language processing [39] and further applied to vision tasks [26], could overcome the limitations of CNNs and archives remarkable performance results. However, quadratic complexity with respect to sequence length of Transformer poses a serious problem for its application in real-world underwater image enhancement (UIE) scenarios that may requires processing high-resolution images in real-time efficiency.

Recently, State Space Models (SSM) and their improved variants, Mamba [10] and Mamba-2 [7], have emerged as efficient and effective backbones for long-sequence modeling. This evolution hints at a potential solution for balancing global receptive fields and computational efficiency for computer vision tasks. The discretized state-space equations in Mamba can be formalized into a recursive form, enabling the modeling of very long-range dependencies through specially designed structured reparameterization. This capability allows Mamba-based restoration networks learns and interprets the images context better, thereby enhancing reconstruction quality [14]. Additionally, Mamba's parallel scan algorithm facilitates the parallel processing of each token, making efficient use of modern hardware like GPUs. These promising properties motivate us to explore the potential of Mamba-based architecture for achieving both efficient and effective in image restoration tasks.

Given the challenges in underwater image enhancement, we present PixMamba, a novel approach that utilizes the linear complexity and long-range modeling capabilities of State Space Models (SSMs). PixMamba is tailored for efficient and effective underwater image enhancement, consisting of two key components operating at different levels: the Efficient Mamba Net (EMNet) and the PixMamba Net (PixNet). EMNet combines the Efficient Mamba Block (EMB) for efficient patch-level feature extraction and dependency modeling with the Mamba Upsampling Block (MUB) for detail-preserving upsampling. However, relying solely on patch-level processing can lead to inconsistencies and fail to capture long-range dependencies that govern overall clarity, color balance, and global consistency.

To address limitations mentioned above, PixMamba introduces a novel dual-level architecture that integrates both pixel-level and patch-level processing. At

the pixel level, PixMamba Net (PixNet) leverages State Space Models (SSMs) to capture long-range dependencies with linear complexity, making it the first UIE approach to process the entire image at the pixel level. This design preserves fine-grained pixel-level features, addressing the shortcomings of prior methods that either apply patchification or rely on alternate domain projections, which compromise detail and global consistency. PixNet further incorporates a Block-wise Positional Embedding (BPE) technique to manage varying input resolutions, ensuring efficient pixel-level processing without sacrificing image quality. At the patch level, PixMamba employs the Efficient Mamba Block (EMB) and the Mamba Upsampling Block (MUB) to enhance the efficiency and accuracy of image reconstruction. These components optimize the performance of traditional patch-level processing by preserving intricate textures while mitigating common issues such as noise introduction and loss of detail during upsampling, as seen in traditional U-Net architectures. By synergistically combining these two levels of granularity, PixMamba achieves state-of-the-art performance in underwater image enhancement, ensuring both microscopic detail preservation and macroscopic image clarity. This dual-level approach enables PixMamba to outperform existing methods that rely solely on patch-level processing or simplified skip connections, establishing a new standard in the UIE field with state-of-the-art results across multiple underwater datasets.

Compared to existing SSM-based UIE method [12] only utilizes patch-level processing, which may potentially loss detailed features and lead to global inconsistency of the enhanced image. Our proposed dual-level processing architecture enables PixMamba to capture fine-grained feature and ensure overall consistency and clarity.

In this paper, we present the following key contributions:

- **PixMamba:** PixMamba presents a novel dual-level architecture designed for efficient and detailed image restoration. By integrating local patch-level processing through the Efficient Mamba Net (EMNet) with global pixel-level processing via the innovative PixMamba Net (PixNet), this framework effectively enhances the quality of underwater images.
- **EMNet:** The Efficient Mamba Net (EMNet) adeptly integrates the Efficient Mamba Block (EMB) and the Mamba Upsampling Block (MUB). The EMB excels at capturing essential image features with enhanced memory efficiency, while the MUB specializes in preserving intricate details during the upsampling process. This synergistic combination markedly enhances the quality of restored images and improves overall processing efficiency."
- **State-of-the-art Performance**: The integration of PixNet's pixel-level feature extraction with EMNet's robust patch-level processing allows PixMamba to achieve a more refined and enhanced underwater image restoration process, yielding impressive results across a variety of UIE datasets.

2 Related Works

Traditional physical-based and prior-based methods for underwater image enhancement are increasingly being replaced by deep learning-based approaches due to their superior ability to learn feature representations from underwater images through the deep learning process. Compared to traditional handcrafted algorithms, deep learning-based methods have gained more interests. Deep learning-based image enhancement approaches have three main categories: CNN-based, Transformer-based, and the most recent SSM-based. Each category will be discussed in the following section respetively.

2.1 CNN-based Image Enhancement

Li *et al.* [20] introduced a network that employs an embedding strategy spanning multiple color spaces, guided by transmission properties. Their approach utilizes an encoder that combines different color space representations and a decoder that enhances degraded regions based on transmission guidance. Fu *et al.* [9] models UIE into a distribution estimation problem. It first used a probabilistic network based on a conditional VAE and adaptive instance normalization that learns to approximate the posterior over meaningful appearance. Cong *et al.* [6] proposed a physical model-guided Generative Adversarial Network (GAN) for UIE. The network incorporates a Parameters Estimation subnetwork for learning physical model parameters and a Two-Stream Interaction Enhancement subnetwork with a Degradation Quantization module for key region enhancement, along with Dual-Discriminators for style-content adversarial constraints. Huang *et al.* [15] developed a Semi-supervised Underwater Image Restoration framework (Semi-UIR) based on the mean-teacher model. To address limitations with the naive approach, they introduced a reliable bank for pseudo ground truth and incorporated contrastive regularization to combat confirmation bias. However, CNNs [23, 36, 37] suffer from the inherent limitation of the local receptive field mechanism and insufficient to learn global representations.

2.2 Transformer-based Image Enhancement

Ren *et al.* [34] proposed a novel approach using the U-Net based Reinforced Swin-Convs Transformer. By embedding Swin Transformer into U-Net, they enhanced the model's ability to capture global dependencies while reintroducing convolutions to capture local attention. Zamir *et al.* [43] introduced an efficient Transformer model, designed to handle high-resolution image restoration tasks. It makes strategic modifications to the multi-head attention and feed-forward network modules, enabling it to capture long-range pixel interactions while being computationally manageable. Gu *et al.* [11] presented a hierarchical CNN and Transformer hybrid architecture. This architecture includes a residual-shaped hybrid stem combining convolutions with an Enhanced Deformable Transformer (DeTrans), capable of learning both local and global representations and exploiting multi-scale features effectively. Nervertheless, self-attention mechanism

in Transforms [5] scales massively for high-resolution images, which is impractical for real-world applications.

2.3 SSM-based Image Enhancement

Shi *et al*. [38] proposed the Residual State Space Block (RSSB), which demonstrated a significant breakthrough in Mamba-based image restoration. By processing both local and global information while maintaining linear complexity, it achieved high efficiency and commendable performance, highlighting the potential of local-global integration in image enhancement. Guan *et al*. [12] introduced a state space model (SSM) for underwater image enhancement (UIE) that aims to combine linear computational complexity with effective degradation handling. To address spatial and channel dependencies, their model incorporates spatial-channel omnidirectional selective scan blocks and multi-scale feedforward networks, promoting coordinated information flow and fine-tuning image details. By leveraging SSM, they have shown superior performance in overcoming the limitations of convolutional neural networks (CNNs) in generalizability and the computational inefficiency of Transformers.

3 Methods

3.1 Preliminaries

Structured State Space Models (S4) are a recent class of sequence models that build upon and extend principles from Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and classical state space models. Inspired by continuous systems, S4 fundamentally transforms an input sequence $x(t) \in \mathbb{R}$ into an output sequence $y(t) \in \mathbb{R}$ through the use of a hidden state $h(t) \in \mathbb{R}^N$. Continuous systems can be modeled by linear ordinary differential equations (ODEs) as follows:

$$h'(t) = Ah(t) + Bx(t), \tag{1}$$

$$y(t) = Ch(t) + Dx(t) \tag{2}$$

where $h(t) \in \mathbb{R}^N$ represents the hidden state, and $A \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^N, C \in \mathbb{R}^N$ are parameters associated with a state size of N, while $D \in \mathbb{R}$ accounts for the skip connection.

In practice, discretizing equations (1) and (2) is necessary. Using the zero-order hold (ZOH) method, we obtain the discrete form by converting A and B into their discrete equivalents via the time scale parameter Δ. The resulting discretization is defined as:

$$h'(t) = \overline{A}h_{t-1} + \overline{B}x_t, \tag{3}$$

$$y(t) = Ch_t + Dx_t, \tag{4}$$

$$\overline{A} = e^{\Delta A}, \tag{5}$$

$$\overline{B} = (\Delta A)^{-1}(e^{\Delta A} - I) \tag{6}$$

where $\Delta \in \mathbb{R}^D$ represents the time scaling parameter, and $B, C \in \mathbb{R}^{D\times N}$.

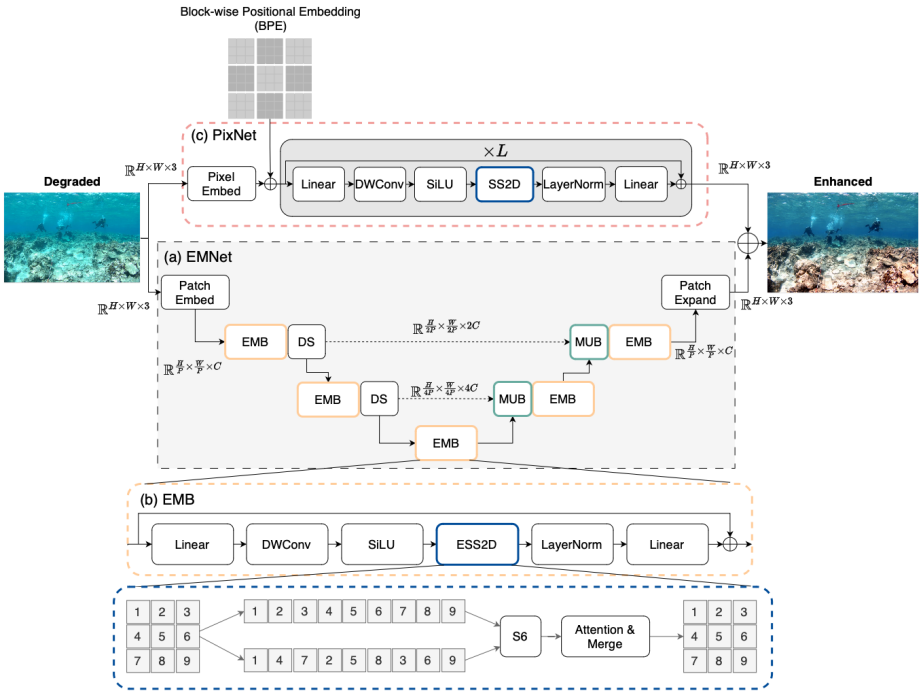


Fig. 1: Overall architecture of PixMamba. (a). EMNet: Efficient Mamba Net; (b). EMB: Efficient Mamba Block; (c). PixNet: PixMamba Net; MUB: Mamba Upsampling Block; DS: Downsampling Block; DWConv: Depth-wise Convolution Block; S6: Mamba SSM [10].

3.2 Overall Architecture

The architecture of our proposed framework, termed PixMamba, builds on an Efficient Mamba Network (EMNet) as the backbone while incorporating PixMamba Net (PixNet) in parallel, as depicted in Fig. 1. Given an underwater degraded image $I \in \mathbb{R}^{H\times W\times 3}$, the EMNet processes the input by first encoding it via PatchEmbed, which generates a patched image $I_E^0 \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$. This encoded image is subsequently passed through three stages of the Efficient Mamba Block (EMB), with a downsampling layer applied after each stage. The image features are progressively downsampled to dimensions $I_E^1 \in \mathbb{R}^{\frac{H}{2P}\times\frac{W}{2P}}$ and $I_E^2 \in \mathbb{R}^{\frac{H}{4P}\times\frac{W}{4P}}$. Following this, the features are decoded across three upsampling stages using the Mamba Upsampling Block (MUB) and the EMB, producing feature maps I_D^2, I_D^1, I_D^0 of sizes $\frac{H}{4P}\times\frac{W}{4P}$, $\frac{H}{2P}\times\frac{W}{2P}$, $\frac{H}{P}\times\frac{W}{P}$, respectively. Finally, the decoded image $I_D^0 \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$ is projected back to the original image resolution $I_{\text{FD}} \in \mathbb{R}^{H\times W\times 3}$.

To capture finer, pixel-level details, PixNet is introduced. PixNet enhances the deep features of the input image in a sequential manner, processing it

through L stages using Mamba Blocks. At each stage l, the image features $I_P^l \in \mathbb{R}^{HW \times D}$ are progressively enriched, where $l \in \{1, 2, \ldots, L\}$. PixNet begins by embedding the image I pixel-wise into a feature map $I_P^0 \in \mathbb{R}^{HW \times D}$ and then augments local information through Block-wise Positional Embedding (BPE), $\text{BPE} \in \mathbb{R}^{\frac{HW}{B^2} \times D}$, which is bilinearly sampled. After processing the image through all L stages, PixNet projects the refined features into the final pixel-level output $I_{\text{FP}} \in \mathbb{R}^{H \times W \times 3}$.

By combining the patch-level information from EMNet (I_{FD}) and the pixel-level information from PixNet (I_{FP}), we generate a final enhanced image $I_F = I_{\text{FD}} + I_{\text{FP}}$, where $I_F \in \mathbb{R}^{H \times W \times 3}$, B is the block size, P is the patch size, and H, W, D are the height, width, and hidden dimension of the image, respectively.

3.3 EMNet

Our proposed EMNet, as illustrated in Fig. 1(a), integrates the SSM, which effectively captures both global and local feature dependencies, into a U-Net-inspired architecture for image restoration [35]. A direct integration of SSM into the U-Net architecture, however, significantly increases computational complexity by doubling the hidden dimension at each stage. To mitigate this, EMNet reduces the number of stages in the U-Net by one, thereby optimizing memory usage and computational efficiency. Additionally, the patch size in the initial stage is doubled to maintain the original U-Net's receptive field, which enhances restoration performance. EMNet further incorporates two key components: the Efficient Mamba Block (EMB) and the Mamba Upsampling Block (MUB).

Efficient Mamba Block As shown in Fig. 1(b), the Efficient Mamba Block (EMB) processes image patches using the Efficient SS2D (ESS2D) operation, a more computationally efficient variant of the 2D selective scan operation proposed in VMamba [25]. While the original SS2D models feature dependencies using four-directional scans, ESS2D simplifies this by reducing computational overhead with minimal impact on performance. Following feature extraction, a spatial and channel attention module [16] refines the representation by eliminating channel redundancy and adjusting attention weights. The module consists of two branches: one for channel attention, which captures global feature representations, and another for spatial attention, which assesses the significance of individual tokens. This attention-based filtering improves the overall feature quality.

Mamba Upsampling Block Traditional U-Net architectures employ a symmetric encoder-decoder structure, with upsampling used to restore the original image's spatial dimensions. However, this upsampling process often leads to a loss of fine details and the introduction of noise, which may degrade performance. To address these limitations, we propose the Mamba Upsampling Block, which incorporates SSM mechanisms before upsampling. By leveraging SSM to selectively maintain important feature dependencies, the Mamba Upsampling Block

preserves details during the upsampling process, leading to higher-quality image restoration. This process is formalized as:

$$I_D^{s-1} = \text{Norm}(\text{TransposeConv2D}(\text{EMB}(I_D^s W))) \tag{7}$$

where W is a learnable projection matrix, and I_D^s represents the decoded feature at stage s.

3.4 PixMamba Net

We introduce PixMamba Net (PixNet), illustrated in Fig. 1(c), as a complementary method that operates at the pixel level to capture finer details. While EMNet processes image patches (e.g., 2x2 or 4x4 pixels), PixNet performs pixel-wise operations, enabling it to extract more granular features and improve noise reduction. The core of PixNet is the Mamba Block, which leverages SSM to exploit each pixel's full potential, resulting in enhanced global consistency and finer feature extraction. To provide spatial information for SSM, we introduce Block-wise Positional Embedding (BPE), which splits the positional embedding into blocks and resizes it to the input sequence. The process is defined as follows:

$$\text{PE} = \textbf{Upsample}(\text{BPE}) \tag{8}$$

$$I_P^0 = [I^0 W; I^1 W; \ldots; I^{HW} W] + \text{PE} \tag{9}$$

$$I_P^l = \textbf{MambaBlock}(I_P^{l-1}) + I_P^{l-1} \tag{10}$$

$$I_{\text{FP}} = \textbf{Project}(I_P^L) \tag{11}$$

where $W \in \mathbb{R}^{HW \times D}$ is the learnable projection matrix, and I^i denotes the i-th pixel of the input image.

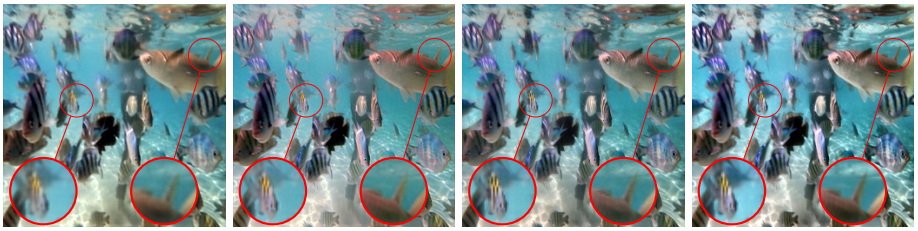
4 Experiments

4.1 Implementation Details

The proposed PixMamba was built using the PyTorch 2.1.0 and MMagic [29] toolkits. We used an NVIDIA RTX 3090 GPU for all training and testing experiments. The model was trained end-to-end using the Charbonnier Loss [19] and the AdamW [28] optimizer. The learning rate was set to $4e^{-4}$, with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. All images were resized to 256×256 pixels. Training batch size was set to 16, and PixMamba network was trained for 800 epochs. Learning rate was adjusted using a 20-epoch warm-up, followed by a cosine annealing scheduler [27].

Table 1: Quantitative comparisons across **C60** and **UCCS** datasets, model parameters, and FLOPs. Best highlighted in **bold** and second in underline.

Method	Venue	C60		UCCS		Params ↓	FLOPs ↓
		UIQM ↑	UCIQE ↑	UIQM ↑	UCIQE ↑		
Ucolor [20]	TIP 21	2.482	0.553	3.019	0.550	157.4M	34.68G
PUIE-Net [9]	ECCV 22	2.521	0.558	3.003	0.536	**1.41M**	30.09G
URSCT [34]	TGRS 22	2.642	0.543	2.947	0.544	11.41M	18.11G
Restormer [43]	CVPR 22	2.688	0.572	2.981	0.542	26.10M	140.99G
PUGAN [6]	TIP 23	2.652	0.566	2.977	0.536	95.66M	72.05G
MFEF [44]	EAAI 23	2.652	0.566	2.977	0.556	61.86M	26.52G
Semi-UIR [15]	CVPR 23	2.667	0.574	**3.079**	0.554	1.65M	36.44G
Convformer [11]	TETCI 24	2.684	0.572	2.946	0.555	25.9M	36.9G
X-CAUNET [33]	ICASSP 24	2.683	0.564	2.922	0.541	31.78M	261.48G
WaterMamba [12]	arXiv 24	2.853	0.582	3.057	0.55	3.69M	**7.53G**
PixMamba (Ours)	-	**2.868**	**0.586**	3.053	**0.561**	8.68M	7.60G



(a) WaterMamba [12] **(b)** Semi-UIR [15] **(c)** Ours **(d)** reference

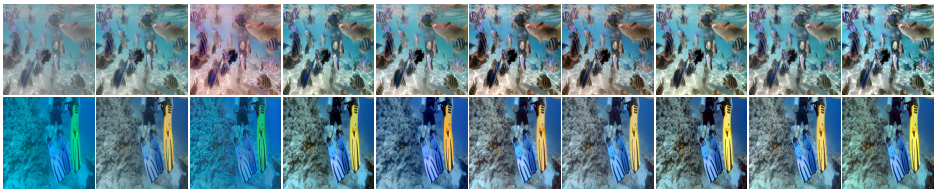
Fig. 2: Enhanced image detail visualization. Our method improves the detail features of the degraded image compared to WaterMamba [12] and Semi-UIR [15] As highlighted in the red circle, our approach shows superior result on the detail features over WaterMamba [12] and Semi-UIR [15], demonstrating the advantage of our proposed MUB and PixNet techniques.

4.2 Datasets

The experiments used two publicly available underwater image datasets: UIEB [21] and UCCS [24]. UIEB dataset has total of 950 images, and was split into a train set of 800 samples (U800), a validation set of 90 samples (T90), and a challenge set of 60 samples (C60). Each sample in U800 and T90 includes a raw degraded underwater image and its corresponding human-curated reference image, while C60 has only degraded image [21]. The UCCS dataset includes three different underwater color settings: bluish, greenish and blue-green tones, each setting contains 100 images, totaling 300 images [24].

Table 2: Quantitative comparisons on **T90** dataset, model parameters, and FLOPs. Best highlighted in **bold** and second in underline.

Method	Venue	T90				
		PSNR ↑	SSIM ↑	MSE ↓	UIQM ↑	UCIQE ↑
Ucolor [20]	TIP 21	21.093	0.872	0.096	**3.049**	0.555
Shallow-uwnet [30]	AAAI 21	18.278	0.855	0.131	2.942	0.544
UIEC^2-Net [40]	SPIC 21	22.958	0.907	0.078	2.999	0.599
PUIE-Net [9]	ECCV 22	21.382	0.882	0.093	3.021	0.566
NU^{2}Net [13]	AAAI 23	23.061	0.923	0.086	2.936	0.587
FiveA+ [17]	BMVC 23	23.061	0.911	0.076	2.828	0.616
WaterMamba [12]	arXiv 24	**24.715**	**0.931**	-	-	-
PixMamba (Ours)	-	23.587	0.921	**0.061**	3.048	**0.617**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 3: The qualitative comparisons. T90 [21] samples are presented in each row from top to bottom. (a) raw; (b) Ucolor [20]; (c) PUGAN [6]; (d) MFEF [44]; (e) Semi-UIR [15]; (f) Convformer [11]; (g) X-CAUNET [33]; (h) WaterMamba [12]; (i) PixMamba; (j) reference.

4.3 Evaluataion Metrics

We evaluated our PixMamba method using five criteria. First, there's the Mean Squared Error (MSE) calculates average squared per-pixel error. Then, Peak Signal-to-Noise Ratio (PSNR) [18] gauges the ratio between the image's signal to its noise, offering a measure of the overall image quality. The Structural Similarity Index (SSIM) [41] measures how similar the image structure is, which aligns closely with human vision. Underwater Image Quality Measure (UIQM) [31] comprises of three underwater image attributed measures: image colorfulness, sharpness, and contrast. Lastly, the Underwater Color Image Quality Evaluation (UCIQE) [42] metric assesses the overall image smoothness, clarity, and contrast. It is worth noting that both UIQM and UCIQE are no-reference evaluation methods, designed to assess the quality of images without the need for a reference.

4.4 Qualitative Comparison

The visual qualitative comparison of our proposed PixMamba and other state-of-the-art models is depicted in Fig. 3 and Fig. 4. We reported most representa-

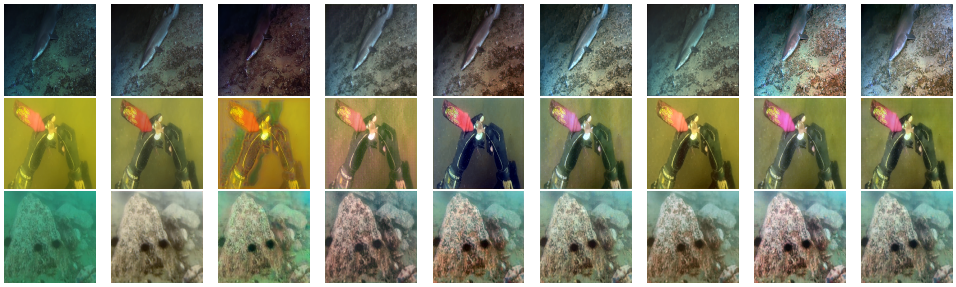


Fig. 4: The qualitative comparisons. First and second row are C60 [21] samples. Third row is UCCS [24] samples. (a) raw; (b) Ucolor [20]; (c) PUGAN [6]; (d) MFEF [44]; (e) Semi-UIR [15]; (f) Convformer [11]; (g) X-CAUNET [33]; (h) WaterMamba [12]; (i) PixMamba.

tive samples from the datasets. Additional, we illustrated the detail comparison in Fig. 2. By zooming in on fine-grained details, PixMamba enhances the entire degraded underwater image while preserving the quality of image. This advancement enables UIE to be further applied in high-resolution scenarios.

4.5 Quantitative Comparisons

As shown in Tab. 1, we compare our PixMamba with several state-of-the-art models included UColor [20], UIEC^2-Net [40], Shallow-UWNet [30], PUIE-Net [9], URSCT [34], Restormer [43], PUGAN [6], MFEF [44], Semi-UIR [15], Convformer [11], X-CAUNET [33], Five A^+ [17], NU^2Net [13], and WaterMamba [12]. The proposed PixMamba outperforms other state-of-the-art models across various datasets. Compared to Semi-UIR [15], our method has improved UIQM and UCIQE by 0.201 and 0.012 on C60 dataset, and improved UCIQE by 0.007 on UCCS datasets. On T90 dataset, compared to NU^2Net [13], the PSNR, UIQM, and UCIQE were improved by 0.526, 0.112, and 0.03, respectively.

4.6 Ablation Studies

Table 3: Ablation study. Trained on **U800** dataset and validated on **T90** dataset.

Method	EMNet	MUB	PixNet	BPE	T90 PSNR ↑	T90 SSIM ↑	Params ↓	FLOPs ↓
U-Net (ResBlock) [35]					18.102	0.822	3.35M	17.42G
PixMamba	✓				22.857	0.913	7.05M	7.15G
PixMamba	✓	✓			22.969	0.919	8.66M	5.99G
PixMamba	✓	✓	✓		23.295	0.920	8.68M	7.60G
PixMamba	✓	✓	✓	✓	23.587	0.921	8.68M	7.60G

To evaluate the contributions of each component in our proposed PixMamba model, we conduct an ablation study, summarized in Tab. 3. U-Net (ResBlock) [35], reaches 18.102 PSNR and 0.822 SSIM with 3.35M parameters and 17.42G FLOPs. Introducing the Efficient Mamba Net (EMNet) module and replaced the Mamba Upsampling Block (MUB) with vanilla patch expand upsampling [3] to form the initial PixMamba architecture boosts the PSNR to 22.857 and the SSIM to 0.913, albeit with a higher parameter count of 7.05M and reduced FLOPs of 7.15G. Adding the Mamba Upsampling Block (MUB) to PixMamba further improves the PSNR to 22.969 and SSIM to 0.919, though it increases the parameter count to 8.66M while reducing the FLOPs to 5.99G. Adding the PixMamba Net (PixNet) without integrates our proposed Block-wise Positional Embedding (BPE) improves the PSNR to 23.29 and SSIM to 0.920, with a slight increase in parameters to 8.68M and FLOPs to 7.60G. Finally, incorporating the PixMamba Net (PixNet) into the architecture enhances the performance further, achieving a PSNR of 23.587 and an SSIM of 0.921. This comprehensive analysis demonstrates that each module in the PixMamba architecture incrementally contributes to the overall performance.

Table 4: Comparison of PixMamba and Restormer [43] performance across different image resolutions. Speed is measured in seconds per image, and GPU memory is measured in MB.

Image Resolution	**PixMamba**		**Restormer [43]**	
	Speed	GPU Mem.	Speed	GPU Mem.
256^2	0.0166	1578	0.0478	1296
512^2	0.0303	4183	0.2059	3650
1024^2	0.1335	14880	0.8637	12934

4.7 Efficiency Comparisons

The computational efficiency analaysis of our proposed SSM-based PixMamba and the Transformer-based Restormer [43], as shown in Tab. 4. The evaluation focuses on their performance in terms of processing time per image and GPU memory consumption during the inference phase. PixMamba demonstrates a linear growth in inference time across varying image sizes (256x256, 512x512, and 1024x1024). In contrast, Restormer [43] exhibits a non-linear increase in processing time as image dimensions expand, suggesting potential scalability limitations for high-resolution imagery. Regarding GPU memory utilization, both models show comparable consumption patterns across different image sizes, with PixMamba slightly exceeding Restormer [43] memory usage. This marginal difference can be attributed to the relatively small size of the PixMamba model, where the overhead associated with the State Space Model (SSM) architecture may surpass that of the Transformer architecture. The efficiency evaluations are conducted on a single NVIDIA RTX 4090 GPU.

5 Conclusion

In this paper, we introduced a novel architecture for underwater image enhancement (UIE) task: PixMamba, which leverages State Space Models (SSM) for linear complexity and effective feature modeling. PixMamba employs a dual-level processing approach, which contains Efficient Mamba Net (EMNet) for patch-level modeling and PixMamba Net (PixNet) for pixel-level modeling to improve overall image quality and model efficiency. Specially, PixNet contains Block-wise Positional Embedding (BPE) while modeling at pixel-level patch, it allows PixNet to have both spatial information and global fine-grained features seamlessly. EMNet utilizes an SSM-based U-Net architecture at the patch level. It incorporates two key components: Efficient Mamba Block (EMB) for lower memory computational cost and Mamba Upsampling Block (MUB) for more detail-preserving restoration. Comprehensive experiments demonstrate that PixMamba performs advantageously against existing methods, substantiating its efficiency and effectiveness.

References

1. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Bekaert, P.: Color balance and fusion for underwater image enhancement. IEEE Transactions on Image Processing **27**(1), 379–393 (2018)
2. Berman, D., Levy, D., Avidan, S., Treibitz, T.: 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 (2021)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision Workshops (ECCVW). pp. 205–218 (2022)
4. Cao, X., Ren, L., Sun, C.: Dynamic target tracking control of autonomous underwater vehicle based on trajectory prediction. IEEE Transactions on Cybernetics **53**(3), 1968–1981 (2023)
5. Chen, S.F., Wen, C.X., Cheng, W.H., Hua, K.L.: Representation and boundary enhancement for action segmentation using transformer. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5965–5969 (2024)
6. Cong, R., Yang, W., Zhang, W., Li, c., Guo, C.L., Huang, Q., Kwong, S.: PUGAN: Physical model-guided underwater image enhancement using GAN with dual-discriminators. IEEE Transactions on Image Processing **32**, 4472–4485 (2023)
7. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: International Conference on Machine Learning (ICML) (2024)
8. Fayaz, S., Parah, S.A., Qureshi, G.J., Lloret, J., Ser, J.D., Muhammad, K.: Intelligent underwater object detection and image restoration for autonomous underwater vehicles. IEEE Transactions on Vehicular Technology **73**(2), 1726–1735 (2024)
9. Fu, Z., Wang, W., Huang, Y., Ding, X., Ma, K.K.: Uncertainty inspired underwater image enhancement. In: European Conference on Computer Vision (ECCV). pp. 465–482 (2022)

10. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
11. Gu, P., Zhang, Y., Wang, C., Chen, D.Z.: Convformer: Combining cnn and transformer for medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 642–651 (2023)
12. Guan, M., Xu, H., Jiang, G., Yu, M., Chen, Y., Luo, T., Song, Y.: WaterMamba: Visual state space model for underwater image enhancement. arXiv preprint arXiv:2405.08419 (2024)
13. Guo, C., Wu, R., Jin, X., Han, L., Chai, Z., Zhang, W., Li, C.: Underwater Ranker: Learn which is better and how to be better. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 702–709 (2023)
14. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: MambaIR: A simple baseline for image restoration with state-space model. arXiv preprint arXiv:2402.15648 (2024)
15. Huang, S., Wang, K., Liu, H., Chen, J., Li, Y.: Contrastive semi-supervised learning for underwater image restoration via reliable bank. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18145–18155 (2023)
16. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: LocalMamba: Visual state space model with windowed selective scan. arXiv preprint arXiv:2403.09338 (2024)
17. Jiang, J., Ye, T., Bai, J., Chen, S., Chai, W., Jun, S., Liu, Y., Chen, E.: Five A^+ Network: You only need 9k parameters for underwater image enhancement. In: British Machine Vision Conference (BMVC) (2023)
18. Korhonen, J., You, J.: Peak signal-to-noise ratio revisited: Is simple beautiful? In: International Workshop on Quality of Multimedia Experience Workshop (QoMEX). pp. 37–38 (2012)
19. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE/CVF Conferene on Computer Vision and Pattern Recognition (CVPR). pp. 624–632 (2017)
20. Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., Ren, W.: Underwater image enhancement via medium transmission-guided multi-color space embedding. IEEE Transactions on Image Processing **30** (2021)
21. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing **29**, 4376–4389 (2020)
22. Li, C., Quo, J., Pang, Y., Chen, S., Wang, J.: Single underwater image restoration by blue-green channels dehazing and red channel correction. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1731–1735 (2016)
23. Lin, Y.X., Tan, D.S., Cheng, W.H., Chen, Y.Y., Hua, K.L.: Spatially-aware domain adaptation for semantic segmentation of urban scenes. In: IEEE International Conference on Image Processing (ICIP). pp. 1870–1874 (2019)
24. Liu, R., Fan, X., Zhu, M., Hou, M., Luo, Z.: Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4861–4875 (2020)
25. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: VMamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)

27. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
29. MMagic Contributors: MMagic: OpenMMLab multimodal advanced, generative, and intelligent creation toolbox. `https://github.com/open-mmlab/mmagic` (2023)
30. Naik, A., Swarnakar, A., Mittal, K.: Shallow-UWnet: Compressed model for underwater image enhancement. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 15853–15854 (2021)
31. Panetta, K., Gao, C., Agaian, S.: Human-visual-system-inspired underwater image quality measures. IEEE Journal of Oceanic Engineering **41**(3), 541–551 (2016)
32. Peng, Y.T., Cosman, P.C.: Underwater image restoration based on image blurriness and light absorption. IEEE Transactions on Image Processing **26**(4), 1579–1594 (2017)
33. Pramanick, A., Sarma, S., Sur, A.: X-caunet: Cross-color channel attention with underwater image-enhancing transformer. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3550–3554 (2024)
34. Ren, T., Xu, H., Jiang, G., Yu, M., Zhang, X., Wang, B., Luo, T.: Reinforced swinconvs transformer for simultaneous underwater sensing scene image enhancement and super-resolution. IEEE Transactions on Geoscience and Remote Sensing (2022)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
36. Shahid, M., Chien, I.F., Sarapugdi, W., Miao, L., Hua, K.L.: Deep spatial-temporal networks for flame detection. Multimedia Tools and Applications **80**, 1–22 (11 2021)
37. Shahid, M., Virtusio, J., Wu, Y.H., Chen, Y.Y., Tanveer, M., Muhammad, K., Hua, K.L.: Spatio-temporal self-attention network for fire detection and segmentation in video surveillance. IEEE Access **PP**, 1–1 (12 2021)
38. Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., Yang, W.: VmambaIR: Visual state space model for image restoration. arXiv preprint arXiv:2403.11423 (2024)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)
40. Wang, Y., Guo, J., Gao, H., Yue, H.: UIEC^2-Net: Cnn-based underwater image enhancement using two color space. Signal Processing: Image Communication **96**, 116250 (2021)
41. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
42. Yang, M., Sowmya, A.: An underwater color image quality evaluation metric. IEEE Transactions on Image Processing **24**(12), 6062–6071 (2015)
43. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5728–5739 (2022)
44. Zhou, J., Sun, J., Zhang, W., Lin, Z.: Multi-view underwater image enhancement method via embedded fusion mechanism. Engineering Applications of Artificial Intelligence **121**, 105946 (2023)

45. Zhuang, P., Wu, J., Porikli, F., Li, C.: Underwater image enhancement with hyper-laplacian reflectance priors. IEEE Transactions on Image Processing **31**, 5442–5455 (2022)