# USWformer: Efficient Sparse Wavelet Transformer for Underwater Image Enhancement

Priyanka Mishra[1] Nancy Mehta[2*] Santosh Kumar Vipparthi[1] Subrahmanyam Murala[3]
[1]Indian Institute of Technology Ropar, INDIA
[2]University of Würzburg, Germany
[3]Trinity College Dublin, Ireland
priyanka.20eez0010@iitrpr.ac.in

## Abstract

*Transformer-based methods have shown great promise in underwater image enhancement (UIE) tasks due to their capability to model long-range dependencies, which are vital for reconstructing clear images. While numerous effective attention mechanisms have been devised to handle the computational requirements of transformers, they frequently incorporate redundant information and noisy interactions from irrelevant regions. Additionally, the current methods focusing solely on the raw pixel space constrains the exploration of the underwater image frequency dynamics, thus hindering the models from fully leveraging their potential for producing high-quality images. To address these challenges, we propose USWformer, an efficient **UIE Sparse Wavelet** Transformer Network (**1.19 M** parameters) to eliminate the redundant features in both the spatial and frequency domains. The USWformer consists of two fundamental components: a Sparse Wavelet Self-Attention (SWSA) block and a Multi-scale Wavelet Feed-Forward Network (MWFN). The SWSA block selectively preserves essential attention scores from the keys corresponding to each query, adjusting the feature details. MWFN further diminishes the feature redundancy in the aggregated features thereby improving the enhancement of the underwater images. We assess the efficacy of our approach across benchmark datasets comprising synthetic and real-world underwater images, showcasing its superiority via thorough ablation studies and comparative analyses.*

## 1. Introduction

Over the past decades, underwater image analysis has emerged as a significant area of focus within the computer vision community, becoming even more critical as the pace of the human quest for sea exploration increases. This is particularly relevant in domains such as marine biology, ecology [38], autonomous underwater vehicles (AUV) [5, 30], underwater robotics [47], and archaeology [2, 9]. Unlike costly and complex, specialized underwater imaging devices, deep learning methods for underwater image enhancement provide a more efficient and practical solution. Despite significant advances in terrestrial image enhancement [13, 34, 37], underwater image enhancement remains challenging due to light attenuation, scattering, and water turbidity [1, 20], resulting in issues such as low contrast, color distortion, blur, and noise [3]. These challenges impede subsequent computer vision tasks [49], emphasizing the need for a robust underwater image enhancement method.

To handle these issues, traditional underwater image enhancement (UIE) methods based on the physical properties of underwater images have been proposed [11, 21, 32, 33]. These approaches examine the degradation mechanisms caused by color cast or scattering and compensate for them to enhance the images. However, these physics-based mod-

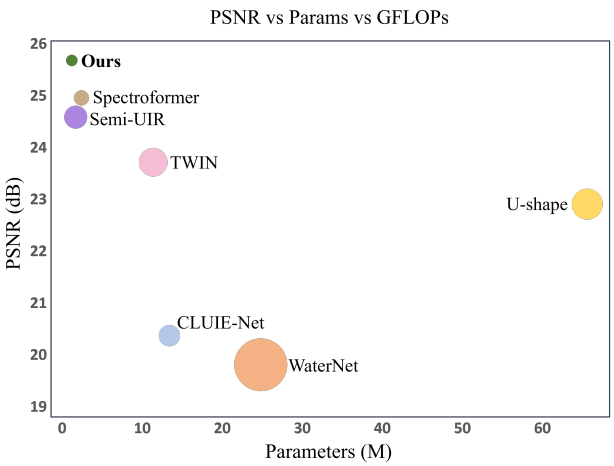


Figure 1. *Model complexity trade-off.* Visualization of PSNR, parameters, and GFLOPS on the challenging UIEB dataset. The proposed USWformer delivers state-of-the-art performance while maintaining the **lowest parameters** (1.19M) and **GFLOPs** (9.14).

*Corresponding Author

els, with their limited representation capacity, fail to address all the complex physical and optical factors in underwater scenes, leading to sub-optimal enhancement under highly complex and diverse conditions. Recently, learning-based methods [16, 19, 20, 35] have demonstrated superior results due to the powerful feature representation and nonlinear mapping capabilities of neural networks using substantial paired training data. Besides the inevitable success of the above-mentioned Convolutional neural network based UIE methods, their limited receptive field reduces effectiveness in capturing long-range dependencies, that are usually crucial for accurate color restoration and attenuation mitigation in degraded images. The more recent, Transformer-based networks [40, 45] are capable of dealing this limitation via incorporating self-attention mechanisms where they typically consider all query-key pair attention relations to aggregate the incoming features. However, this approach is inefficient for underwater image reconstruction, as not all query tokens are closely relevant to their corresponding key tokens. Thus, crafting an efficient mechanism that discerns the most valuable features within information flow, while maintaining reduced sensitivity to specific UIE tasks seems to be the probable solution. Intuitively, the development of a sparse Transformer that selectively identifies the most pertinent interactions among tokens could significantly enhance feature aggregation efficiently.

Few works on sparsity [39, 52] deploy squared-ReLU based activation function in the spatial domain as a probable solution to eradicate any negative irrelevant interaction for general image restoration tasks. Nevertheless, the properties of sparsity for underwater images remain insufficiently explored, thereby hindering the effective utilization of the representational capabilities of the proposed works. Since underwater image acquisition engages both frequency and spatial domains to extract valuable insights, integrating these domains enhances the overall color accuracy and contrast. Building on these insights and owing to the inherent sparsity of of Discrete Wavelet Transform (DWT), we explore the concept of sparsity for efficient UIE, fundamentally in the frequency domain.

In this work, we propose an efficient transformer-based network, USWformer for underwater image enhancement that leverages the most valuable features within information flow in the wavelet space to save the computational complexity. As shown in Figure 1, our model outperforms the SoTA UIE methods by a considerable margin, while utilising around half or even less GFLOPs. The central element of the proposed USWformer is the Sparse Wavelet Transformer Block (SWTB) which comprises a Sparse Wavelet Self-Attention (SWSA) to retain the most pertinent color features, and the Multi-scale Wavelet Feed-Forward Network (MWFN) that refines the aggregated multi-scale features at different resolutions. SWSA incorporates two branches: the upper branch ensures the essential information flow from the low-frequency components acquiring comprehensive features and complex details, and the lower branch filters out the irrelevant tokens from the high-frequency components to significantly augment the image visibility. To efficiently integrate the pixel-specific and globally consistent information while minimizing the computational complexity, we adaptively weigh the outputs of the two branches, following the approach in [52]. Additionally, our effective alternative to the regular feed-forward network [48], *i.e.,* MWFN, enhances the feature transformation by suppressing the redundant operations in the wavelet space. Specifically, the recursive application of DWT to the low-frequency sub-band at each resolution helps in enhancing the visual clarity and detail. To summarise, the main contributions of our work are:

- We propose USWformer, an efficient Transformer-based network for enhanced texture and detail recovery in underwater image enhancement.
- We propose an efficient learnable Sparse Wavelet Self-Attention mechanism that adaptively integrates the most pertinent self-attention values.
- We propose a Multi-scale Wavelet Feed-Forward Network leveraging multi-resolution analysis to further suppress any invaluable information.

The ablation study is conducted on various configurations of the proposed approach. Through a series of experiments on both synthetic and real-world settings, the effectiveness of the proposed method has been validated.

## 2. Related Work

### 2.1. Underwater Image Enhancement

Existing Underwater image enhancement (UIE) methods can be broadly divided into physical model-based, visual prior-based, and deep model-based approaches [11, 20, 31, 32, 40]. The majority of physical model-based UID methods leverage prior knowledge to construct models, such as attenuation curve priors [43], fuzzy priors [8], and water dark channel priors [33]. However, the scalability and robustness of the model are hindered by externally set priors when faced with complex and diverse conditions. Recently, deep learning methods have shown promising performance in underwater imaging. To address the scarcity of real-world underwater paired training data, numerous methods have adopted GAN-based frameworks for UIE, including UGAN [12], UIE-DAL [41], and WaterGAN [23]. Semi-UIR [15] introduced a mean teacher-based semi-supervised network that utilizes unlabeled data effectively. Lately, few research works have been done where the frequency domain properties have been utilized which showcases the tremendous potential that the frequency domain holds. Spectroformer [18] exploits the frequency domain characteristics

via its Hybrid Fourier-Spatial Upsampling for improving the feature resolution of degraded images. WF-Diff [50] utilises the frequency domain characteristics and diffusion models for image enhancement and adjustment. Recent work on wavelet-pixel domain fusion, such as WPFNet [28], has shown improved underwater image enhancement by combining wavelet and pixel domains, preserving details, and improving color and noise reduction compared to existing methods. However, the aforementioned frequency-based approaches, owing to their computational complexity potentially introduce unwanted interactions in the irrelevant areas. Unlike these approaches, we adopt a novel Transformer based approach in the wavelet space to enhance the most useful features and relieve the less informative ones.

### 2.2. Transformers in Vision

Inspired by the success of Transformers in NLP and high-level vision tasks, they have been applied to image restoration, outperforming previous CNN-based methods by effectively modelling non-local information [7]. However, the quadratic complexity of vanilla self-attention limits the application of Transformers to high-resolution images. To address this, the authors in [48] introduced an efficient transformer network designed for restoration tasks such as image deraining, deblurring, and denoising, which calculates attention along the channel dimension to reduce computational costs. Another solution is window-based attention, as seen in Uformer [45], which enhances locality within the Transformer architecture. SwinIR [25] also employs window-based attention, incorporating a shift mechanism for improved cross-window interactions. Few other approaches have investigated the novel use of Transformers through channel-wise and spatial-wise attention layers [31], or by employing customized transformer blocks that utilize both frequency and spatial domains as inputs for self-attention [18]. Unlike the aforementioned approaches, we introduce an adaptive sparse self-attention mechanism in the wavelet space to minimize the overall redundancy by selecting the most informative interactions.

### 2.3. Sparse Representation

While efficient attention mechanisms reduce computational costs but still suffer from redundancy and irrelevant features [7, 51]. DRSformer [7] addresses this with a top-k channel selection, while CODE [51] reduces redundancy in super-pixel space, though both face challenges. Sparse representations, inspired by neural activity, have proven effective for tasks like image deraining [44] and super-resolution [29]. Unlike previous methods that limit attention to local windows or sparse token interactions, we propose a simpler and more efficient approach by designing sparse wavelet self-attention and a novel wavelet feedforward network in our Transformer architecture.

## 3. Proposed Method

In this section, we first outline the overall pipeline of the proposed USWformer for the underwater image enhancement task as shown in Figure 2. Next, we delve into the Sparse Wavelet Transformer Block (SWTB), the cornerstone of our method, comprising two primary components: Sparse Wavelet Self-Attention (SWSA) and the Multi-scale Wavelet Feed-Forward Network (MWFN).

### 3.1. Overall Pipeline

Our objective is to train a network that removes color cast from the degraded underwater image and enhances image details in the generated output. Our proposed USWformer as illustrated in Figure 2, utilizes a hierarchical encoder-decoder framework. Each component of the encoder-decoder pipeline operates at different spatial resolutions and channel dimensions to obtain a multi-scale representation from the input image. For feature downsampling and upsampling, pixel-unshuffle and pixel-shuffle operations are employed, respectively. Following the approach in [45, 48], skip connections are incorporated to link consecutive intermediate features, ensuring stable training. Given a degraded underwater image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$, USWformer initially applies overlapped patch embedding using a $3 \times 3$ convolution to produce shallow features, denoted as $X_0 \in \mathbb{R}^{H \times W \times C}$. After that, a group of Sparse Wavelet Transformer Blocks (SWTB) process these shallow features. We stack $N_i \in \{1, 2, 3, 4\}$ SWTBs to capture detailed features denoted as $X_d \in \mathbb{R}^{H \times W \times C}$ for spatially-varying information within the network backbone. Each SWTB harnesses the wavelet space characteristics to enhance the Transformer's robust capabilities. Further, the standard Transformer self-attention [10] is switched out for Sparse Wavelet Self-Attention (SWSA) in each SWTB to get feature sparsity, which makes the process of aggregating relevant features more effective. Additionally, the proposed Multi-scale Wavelet Feed-Forward Network (MWFN) in the SWTB strives to improve the multi-scale local details for underwater image enhancement in an efficient way. Ultimately, a $3 \times 3$ convolutional layer is employed on the resultant deep features to procure the final output. This comprehensive procedure culminates in the generation of an output image $\boldsymbol{O}$.

### 3.2. Sparse Wavelet Transformer Block

Standard Transformers [10, 42, 48] compute self-attention globally by considering all tokens, which can lead to noisy interactions between irrelevant features, making them unsuitable for image enhancement. To address this issue, we introduce a Sparse Wavelet Transformer Block (SWTB), as shown in Figure 2, that harnesses wavelet-based sparsity and frequency domain processing. Our approach integrates frequency and spatial domain insights to

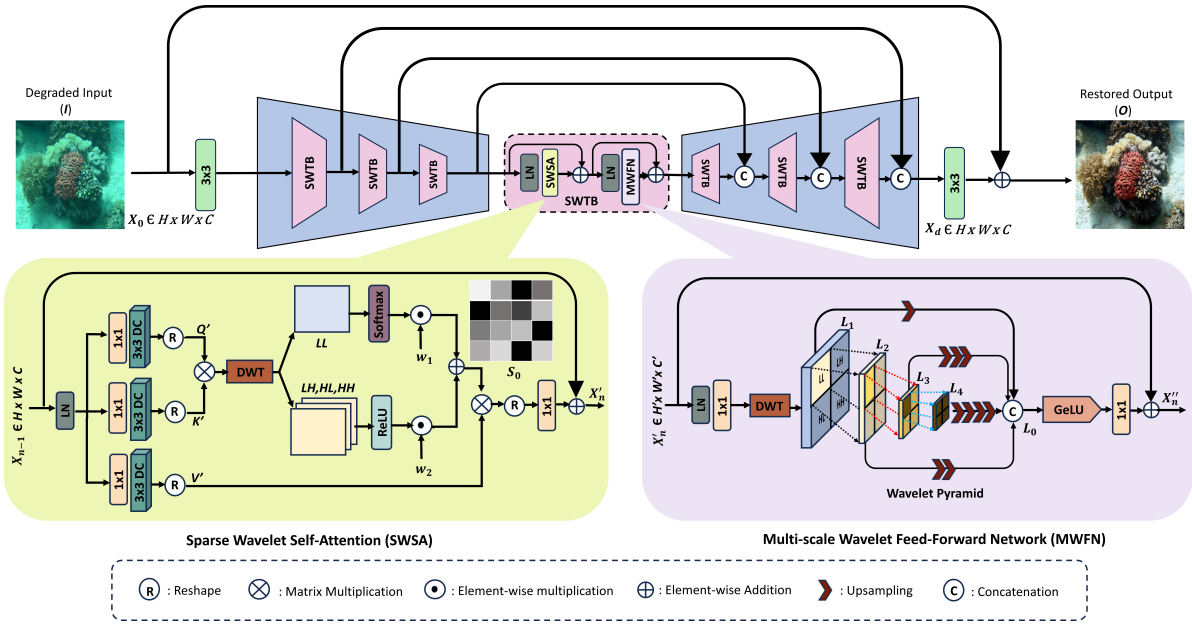


Figure 2. Architectural overview of the proposed USWformer. It primarily includes a Sparse Wavelet Transformer Block (SWTB) with a Sparse Wavelet Self-Attention (SWSA) and a Multi-scale Wavelet Feed-forward Network (MWFN). Here, DC denotes the depth-wise convolution, and LN represents the layer normalization.

uncover intricate details and patterns in degraded underwater images. This approach reduces irrelevant interactions and enhances feature extraction by facilitating information interaction between high and low-frequency features, thus boosting the overall enhancement performance. Specifically, given the input features at the $(n-1)^{th}$ block denoted as, $X_{n-1}$, the encoding procedure of the proposed SWTB is described as:

$$\mathbf{X'}_n = \mathbf{X}_{n-1} + \text{SWSA}(\text{LN}(\mathbf{X}_{n-1})) \quad (1)$$

$$\mathbf{X''}_n = \mathbf{X'}_n + \text{MWFN}(\text{LN}(\mathbf{X'}_n)) \quad (2)$$

where, LN is the layer normalization, $\mathbf{X'}_n$ and $\mathbf{X''}_n$ represents the outputs of SWSA and MWFN blocks. respectively, which are detailed in the subsequent subsection.

### 3.2.1 Sparse Wavelet Self-Attention

Utilizing all the similarities between query and key tokens [42,45] for self-attention is inefficient for image reconstruction, as not all query tokens are closely relevant to their corresponding key tokens. Thus, to address the inherent issues in vanilla self-attention, such as the introduction of noisy interactions due to the consideration of all query-key token values, we proposed a sparse wavelet self-attention (SWSA) mechanism that selectively identifies and utilises the useful token predictions. Instead of computing the attention scores by applying softmax to the query-key dot product, we first apply Discrete Wavelet Transform (DWT) to decompose the input into multiple frequency sub-bands, providing a multi-resolution analysis of the attention map. This multi-resolution approach aids in understanding and preserving important features at various scales, while the inherent sparsity of DWT enables more efficient computations. The decomposed high $(LH, HL, HH)$ and the low-frequency $(LL)$ components from DWT are directed into two separate branches. A ReLU activation function processes the high-frequency bands in the lower branch to achieve sparsity. The ReLU activation function helps achieve sparsity by setting all negative values in the high-frequency $(HF)$ bands to zero. This highlights important features and ignores less significant details or noise in the incoming features. To circumvent any sort of potential oversparsity induced by the ReLU and wavelet-based self-attention, we input the low-frequency $(LF)$ band into the upper branch and follow it with a softmax operation. Since, this band contains the crucial information, employing it in combination with the Softmax guarantees the preservation of essential features.

Here, given a normalized tensor $\mathbf{X}_{n-1} \in \mathbb{R}^{H\times W\times C}$, the proposed SWSA, first outputs queries **Q'**, keys **K'** and values **V'** matrices:

$$\mathbf{Q}^{'} = \phi_3(\psi_1(\mathbf{X})); \mathbf{K}^{'} = \phi_3(\psi_1(\mathbf{X})); \mathbf{V}^{'} = \phi_3(\psi_1(\mathbf{X})) \quad (3)$$

where, $\phi_3(.)$, and $\psi_1(.)$ denotes the $3 \times 3$ depth-wise and $1 \times 1$ pointwise convolution, respectively. Motivated by [48], the self-attention is implemented across channels instead of the spatial dimension to reduce memory complexity and further deployment of DWT helps in identifying the useful tokens. The overall computation can be defined as:

$$(LL, LH, HL, HH) = \text{DWT}\left(\frac{\mathbf{Q}^{'} \cdot \mathbf{K}^{'\mathbf{T}}}{\alpha}\right) \quad (4)$$

This approach enhances the attention process by removing lower values from the query-key dot product, thereby focusing on more significant interactions and reducing noise. To highlight upon focusing the most informative component from the decomposed high-frequency sub-bands, the concatenated sub-bands are passed via a ReLU layer that removes the similarity scores with negative values, thus ensuring sparsity in the high-frequency domain:

$$SWSA_{HF} = \text{ReLU}([HH, LH, HL]) \tag{5}$$

where [.] denotes the concatenation operation. However, since both DWT and ReLU may trigger information loss owing to over-sparsity, hence the decomposed low-frequency component, carrying the important information is passed through a softmax layer. It considers all query-key pairs for attaining attention scores:

$$SWSA_{LF} = \text{Softmax}(LL) \tag{6}$$

However, the primary difficulty in the dual branch scheme lies in effectively leveraging the benefits of both paradigms. Therefore, inspired by [52], SWSA addresses this by adaptively fusing the output of the two branches, selectively incorporating features and thus contributing in controlling the sparsity of the tokens:

$$S_0 = (w_1 \times SWSA_{HF} + w_2 \times SWSA_{LF})\mathbf{V}^{\text{'}} \tag{7}$$

here, $w_1, w_2$ are the normalising weights defined as : $w_n = \frac{e^{\beta_n}}{\sum_{i=1}^{N} e^{\beta_i}}, \quad n \in \{1, 2\}$ and $\beta_n$ represent the learnable parameter. The final output is obtained as:

$$\mathbf{X}^{\text{'}}_n = \mathbf{X}_{n-1} + \psi_1(S_0) \tag{8}$$

#### 3.2.2 Multi-scale Wavelet Feed-Forward Network

Earlier works [45, 48] typically deploy single-scale depth-wise convolution in the feedforward network to learn local features. However, the limited receptive field of this approach hinders its ability to achieve high-quality image reconstruction, as it needs to capture both local and global representation. To address this limitation, here we design a multi-scale wavelet feedforward network (MWFN) to generate multi-resolution features that effectively capture local details and global context without relying on depth-wise convolution. Thus, by providing additional insights from both the frequency and spatial domains, MWFN enhances the model's capacity to identify patterns and textures that the spatial domain alone may not readily detect. This approach not only enhances both the local and global feature representations but also reduces computational complexity, minimizes redundancy, and better captures edge information by analyzing the data at multiple resolutions. Given an input tensor $\mathbf{X}'_n \in \mathbb{R}^{H' \times W' \times C'}$, we first apply layer normalization. Next, we use a $1 \times 1$ convolution to expand the channel dimension by a factor of $r$ = 2.66.

Following this, we apply the Discrete Wavelet Transform (DWT) to decompose the feature map into high-frequency $(LH, HL, HH)$ and low-frequency $LL$ bands at multiple scales. At every scale, we then take the LL band and recursively apply DWT to it until we reach a resolution of $H/16 \times W/16$. After obtaining the decomposed features, we concatenate the information from all the bands at different resolutions, using bilinear interpolation for upsampling to ensure consistent feature alignment. This interpolation not only helps maintain the spatial coherence of the features but also reduces artifacts and preserves the smoothness of the image, leading to more accurate reconstruction. The concatenated features $L_0$ are passed through the GeLU activation function, as the combination of layer normalization and GeLU improves performance by 0.1-0.2% [14]. Finally, we apply another $1 \times 1$ convolution to restore the original input dimension. In this way, the overall process of MWFN is formulated as below:

$$\mathbf{X}^{\text{''}}_n = \mathbf{X}^{\text{'}}_n + \psi_1(\text{GeLU}(L_0)) \tag{9}$$

$$L_0 = [L_1, L_2, L_3, L_4],\ L_1 = \text{DWT}(\psi_1(\text{LN}(\mathbf{X}^{\text{'}}_n))) \tag{10}$$

where, [.] denotes the concatenation, $L_2$, $L_3$, and $L_4$ are obtained by applying DWT on the $LL$ subbands of $L_1$, $L_2$, and $L_3$, respectively, $L_1 \rightarrow \frac{H'}{2} \times \frac{W'}{2} \times C'$; $L_2 \rightarrow \frac{H'}{4} \times \frac{W'}{4} \times C'$; $L_3 \rightarrow \frac{H'}{8} \times \frac{W'}{8} \times C'$; $L_4 \rightarrow \frac{H'}{16} \times \frac{W'}{16} \times C'$. These multi-resolution feature analyses make the model more robust to variations in scale and resolution, enabling it to handle complex underwater images more efficiently and improving its generalization capability across diverse datasets.

### 3.3. Training Losses

In training our proposed architecture, we employed a total loss function $L_T$, which integrates multiple individual loss components. These components include perceptual loss ($L_1$) [17], Charbonnier loss ($L_2$) [6], multi-scale structural similarity index (MS-SSIM) loss ($L_3$) [46], and gradient loss ($L_4$) [36]. The total loss function is formulated as follows:

$$L_T = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4 \tag{11}$$

where, $\lambda_{1,2,3,4} \in \{2, 3, 1, 2.5\}$ are empirically determined weighting factors. This combination of loss functions is crucial for optimizing our model, enabling it to capture various aspects of intrinsic image attributes and produce visually appealing, high-quality output images. *Detailed explanations of these loss functions are provided in the supplementary material.*

## 4. Experimental Discussion

### 4.1. Datasets

For our comparative analysis, we used the synthetic Underwater Image Enhancement Benchmark (UIEB) [20]


Figure 3. Qualitative analysis on full-reference UIEB dataset [20] of our proposed USWformer and the existing SoTA methods. USWformer (Ours) generates sharper and visually-faithful results without any artifacts *(Noticeable differences in quality are highlighted with boxes)*.

Table 1. Results on the Same and Cross Dataset evaluation, (**Train-UIEB** ) [20] (↑: higher is better, ↓: lower is better, **blue** and purple indicate **best** and second best values respectively).

| Method | Test- UIEB | Test-LSUI | Params↓ | GFLOPs↓ |
|---|---|---|---|---|
| | PSNR↑/SSIM↑ | PSNR↑/SSIM↑ | | |
| Semi-UIR [15] | 24.59/0.90 | 24.67/0.86 | 1.67 | 36.43 |
| WaterNet [20] | 19.81/0.86 | 17.73/0.82 | 24.81 | 193.70 |
| U-shape [31] | 22.91/0.91 | 22.87/0.85 | 65.60 | 66.20 |
| TWIN [27] | 23.72/0.83 | 23.54/0.83 | 11.37 | 56.80 |
| CLUIE-Net [24] | 20.37/0.89 | 20.71/0.81 | 13.39 | 31.00 |
| Spectroformer [18] | 24.96/0.917 | 25.08/0.87 | 2.40 | 15.75 |
| **Ours** | **25.68/0.946** | **25.34/0.912** | **1.19** | **9.14** |

along with three real-world underwater datasets: U45 [22], UCCS [26], and SQUID [4]. The training set includes 800 randomly selected image pairs, while 90 images are set aside for testing. The U45 dataset [22] contains 45 real-world images with features like color casts, low contrast, and haze-like degradation commonly found in underwater scenes. The UCCS dataset [26] consists of 300 authentic underwater images showcasing a variety of marine life and environments. The SQUID dataset [4] includes 57 stereo image pairs taken at different locations in Israel.

### 4.2. Training Details

To address the limited number of images in the UIEB dataset for training purposes, we employed data augmentation techniques to expand the dataset. These techniques included horizontal and vertical flipping, noise injection, and contrast variation. By applying these methods, we effectively increased the variability and robustness of the training data, thereby enhancing the model's performance. Specifically, 4800 image pairs from the UIEB dataset were utilized for training purposes. The testing phase involved 90 images from the UIEB dataset. To ensure consistency, all input images were resized to $256 \times 256$ pixels. During training, we employed the ADAM optimizer with an initial learning rate of $3 \times 10^{-4}$, which was modulated using the cosine annealing strategy. The network was implemented using PyTorch and trained on an NVIDIA GeForce RTX 2080 GPU.

### 4.3. Comparison Methods

We perform a comparative analysis between USWformer and state-of-the-art (SOTA) UIE methods: Semi-UIR [15], WaterNet [20], U-shape [31], TWIN [27], CLUIE-Net [24], and Spectroformer [18].

### 4.4. Analysis on Synthetic Dataset

The proposed method is quantitatively evaluated against existing SoTA techniques using key metrics such as PSNR, and SSIM. The quantitative results for the same dataset evaluation (Train-UIEB, Test-UIEB) are presented in Table 1. Additionally, we have provided an analysis of parameters (in Millions) and GFLOPs, demonstrating the computational efficiency of our model. To further showcase the generalization capability of our model, we also conducted a cross-dataset evaluation (Train-UIEB, Test-LSUI [31]). This approach allowed us to verify that the model could generalize well beyond the training data, underscoring its ability to perform effectively across different datasets. The qualitative results are illustrated in Figure 3. These results clearly illustrate that USWformer surpasses existing approaches in enhancing the underwater images. These visual comparisons with significantly enhanced portions of the images highlighted in boxes showcase the superior performance of the proposed USWformer in enhancing underwater image quality, particularly in terms of color correction, contrast enhancement, and detail preservation.

### 4.5. Analysis on Real-world Dataset

To assess the robustness of our proposed approach in real-world applications, we present results derived from the U45, UCCS, and SQUID datasets. Our quantitative analysis includes key metrics such as UIQM (Underwater Image Quality Measure), UISM (Underwater Image Sharpness Measure), NIQE (Natural Image Quality Evaluator), and BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator). Table 2 provides a summary of these results, showing that our method either outperforms or is on par with state-of-the-art techniques. Additionally, we offer qualitative insights into the U45, UCCS, and SQUID datasets, as shown in Figure 4, highlighting notable improvements in color balance, and enhanced visibility in the

Table 2. Quantitative comparison of different UIE methods on the real-world U45, UCCS, and SQUID datasets (↑: higher is better, ↓: lower is better, **blue** and purple indicate **best** and second best values, respectively).

| Dataset | Method | Semi-UIR [15] | WaterNet [20] | U-shape [31] | TWIN [27] | CLUIE-Net [24] | Spectroformer [18] | **Ours** |
|---|---|---|---|---|---|---|---|---|
| U45 | UIQM ↑ | **4.301** | 3.091 | 2.923 | 3.135 | 2.890 | 3.243 | 4.247 |
| | UISM ↑ | 7.142 | 6.187 | 5.567 | 6.698 | 5.988 | **7.354** | 7.250 |
| | NIQE ↓ | **3.767** | 4.596 | 4.309 | 3.992 | 3.874 | 3.842 | 3.804 |
| | BRISQUE ↓ | 23.020 | 21.156 | 21.565 | 20.089 | 20.612 | 19.957 | **17.398** |
| UCCS | UIQM ↑ | 3.773 | 3.134 | 2.874 | 3.119 | 3.066 | 3.209 | **4.279** |
| | UISM ↑ | **7.119** | 6.187 | 5.391 | 6.732 | 6.715 | 6.563 | 7.033 |
| | NIQE ↓ | 4.710 | 6.104 | 4.401 | 4.370 | 4.420 | **3.982** | 4.220 |
| | BRISQUE ↓ | **20.852** | 24.275 | 23.549 | 25.755 | 29.524 | 23.258 | 24.703 |
| SQUID | UIQM ↑ | 2.449 | **3.379** | 2.422 | 3.066 | 1.414 | 3.088 | 2.598 |
| | UISM ↑ | 7.373 | 7.071 | 7.004 | 7.148 | 7.208 | 7.368 | **7.447** |
| | NIQE ↓ | **3.439** | 3.752 | 4.621 | 4.377 | 3.710 | 3.613 | 4.011 |
| | BRISQUE ↓ | 20.189 | 23.364 | 29.164 | **13.874** | 25.981 | 21.477 | 22.419 |

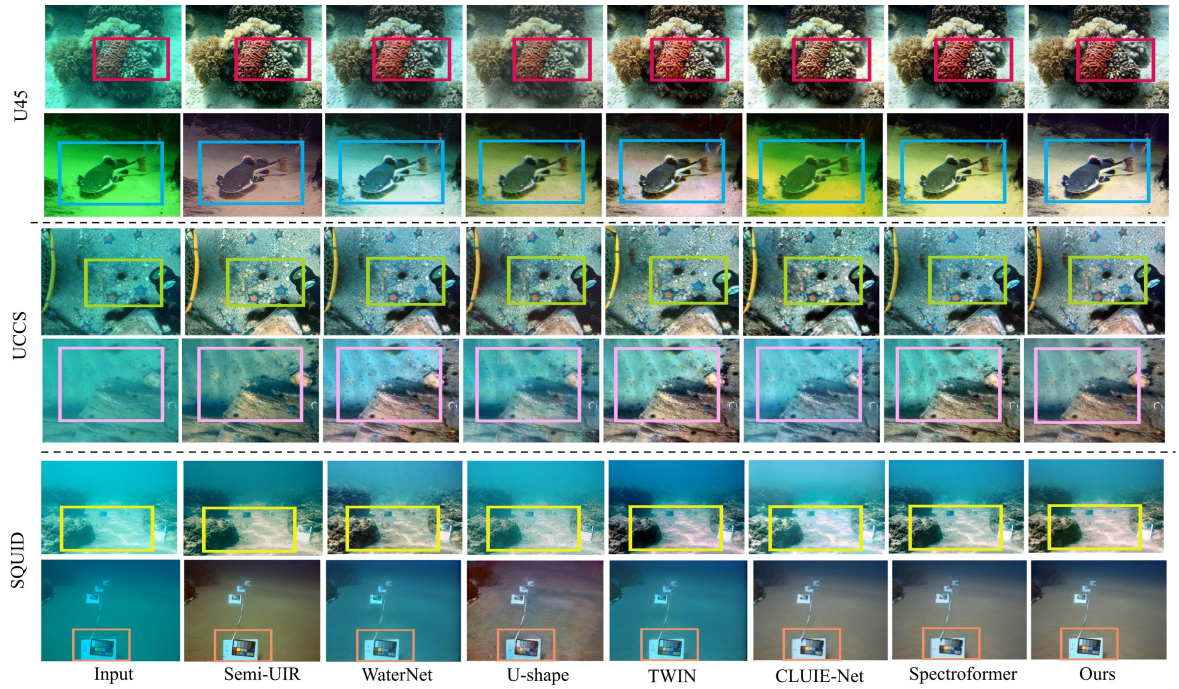


Figure 4. Qualitative analysis on non-reference benchmarks U45 [22], UCCS [26], and SQUID [4]. Unlike other approaches, for all the compared real-world datasets, USWformer (Ours) efficiently restore natural colors and preserve the fine details *(Noticeable differences in quality are highlighted with boxes)*.

reconstructed images. These improvements are attributed to the novel modules introduced in our approach.

## 5. Ablation Study

To examine the impact of each architectural module in our proposed network, we performed a comprehensive ablation study using the UIEB dataset [20].

### 5.1. Influence of SWSA

The Sparse Wavelet Self-Attention (SWSA) block enhances the extraction and utilization of relevant features by leveraging multi-resolution analysis and selective sparsity. To evaluate its effectiveness, we conducted two sets of experiments. In the first set, we integrated the SWSA block into our proposed network and compared its performance with the baseline network. The inclusion of SWSA resulted in a performance gain of around 0.38 dB as clear from Table 3. The visual results in Figure 5 further substantiate the quantitative findings, demonstrate that incorporating the SWSA block leads to superior performance for underwater image enhancement.

In the second set of experiments, we replaced the SWSA block with several efficient attention mechanisms: (1) Multi-DconvHead Transposed Attention (MDTA) [48], (2) Top-k Self-Attention (TKSA) [7], and (3) Multi-Domain Query Cascaded Attention (MQCA) [18]. The quantitative results on the UIEB dataset, shown in Table 4, indicate that SWSA provides favorable gains of around 2.95 dB as compared to the popular MDTA, further confirming its effectiveness in removing the redundant information.

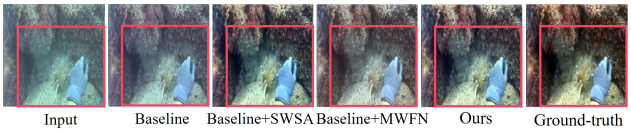


Figure 5. Ablation visual analysis for different network settings of the proposed architecture.

Table 3. Ablation studies conducted on various network configurations using the UIEB benchmark.

| Network Setting | PSNR ↑ | SSIM ↑ |
|---|---|---|
| Baseline | 22.51 | 0.862 |
| Baseline + SWSA | 22.89 | 0.896 |
| Baseline + MWFN | 22.73 | 0.911 |
| Ours (Baseline + SWSA + MWFN) | **25.68** | **0.946** |

Table 4. Ablation Study of Self-Attention Mechanisms.

| Models | MDTA [48] | TKSA [7] | MQCA [18] | SWSA (Ours) |
|---|---|---|---|---|
| PSNR ↑ | 22.73 | 23.28 | 22.45 | **25.68** |
| SSIM ↑ | 0.911 | 0.919 | 0.880 | **0.946** |

Table 5. Ablation study of Feed-Forward networks.

| Models | GDFN [48] | MSFN [7] | MWFN (Ours) |
|---|---|---|---|
| PSNR ↑ | 22.89 | 24.34 | **25.68** |
| SSIM ↑ | 0.896 | 0.917 | **0.946** |

### 5.2. Influence of MWFN

To validate the efficacy of the proposed Multi-Scale Wavelet Feed-Forward Network (MWFN), we conducted the following ablations. First, we evaluated the performance of our model with and without the MWFN block. The results, summarized in Table 3, highlight the improvement of around 0.22 dB achieved by integrating the MWFN block into the baseline. In addition to the quantitative analysis, a qualitative evaluation, shown in Figure 5, visualizes the enhanced performance. The images processed with the MWFN block exhibit noticeably better color balance, contrast, and detail preservation, clearly indicating that the inclusion of the MWFN block improves the model's ability to capture both local and global features compared to the baseline Furthermore, to assess the effectiveness of the MWFN block, we compared it against three other feed-forward network variants: (1) Gated-Dconv Feed-Forward Network (GDFN) [48] and (2) Mixed-scale Feed-forward Network (MSFN) [7]. The quantitative results of this comparison on the UIEB dataset, presented in Table 5, demonstrate that while GDFN, which introduces a gating mechanism in two same-scale depth-wise convolution streams, enhances performance, it overlooks the multi-scale knowledge crucial for image restoration tasks. In contrast, the MWFN block, with its ability to capture multi-scale features, provides superior performance in both PSNR and SSIM, further solidifying its efficacy for image restoration tasks.

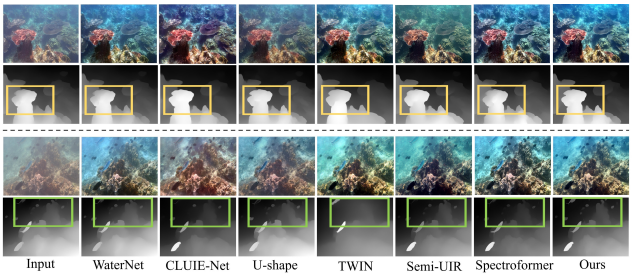


Figure 6. Application of proposed method (Ours) and existing methods (WaterNet [20], CLUIE-Net [24], U-shape [31], TWIN [27], Semi-UIR [15], and Spectroformer [18]) as a pre-processing step for depth-estimation on underwater U45 dataset [22].

## 6. Applicability to Downstream Computer Vision Task

In underwater scenarios, due to diminished visibility, the effective functioning of downstream computer vision tasks may be hindered. Hence, underwater image enhancement can act as a preprocessing step for these downstream applications to function with more accuracy. To demonstrate this, we conducted an experiment on underwater depth estimation. We first applied the depth-estimation algorithm to the degraded image, followed by enhanced images using our method and existing approaches. As shown in Figure 6, our enhanced output led to the most accurate depth estimation compared to other methods, highlighting its potential for underwater saliency detection.

## 7. Conclusion

In this paper, we proposed an efficient sparse transformer model, USWformer for underwater image enhancement. The network comprises several components, including Sparse Wavelet Transformer blocks, which integrate comprehensive features and intricate details effectively. Recognizing that vanilla self-attention in Transformers can be hindered by global interactions with irrelevant information, we implement a novel Sparse Wavelet Self-Attention to retain the most pertinent self-attention values and prevent any sort of information loss. For further enhancing the aggregation of relevant features, we develop a Multi-scale Wavelet Feed-Forward Network that effectively explores multi-scale representations and aims at eradicating redundant computations. An extensive analysis incorporating both synthetic and real-world datasets, along with thorough ablation studies is conducted to validate the effectiveness of USWformer.

## Acknowledgement

This work was supported by Project MoES/PAMC/DOM/04/2022 (E-12710), Project TIHIITG202204 and Project CRG/2022/006876. Also, I would like to thank all the CVPR Lab members for their support.

## References


[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1682–1691, 2019. 1

[2] Geoffrey N Bailey and Nicholas C Flemming. Archaeology of the continental shelf: marine resources, submerged landscapes and underwater archaeology. *Quaternary Science Reviews*, 27(23-24):2153–2165, 2008. 1

[3] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. 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, 2020. 1

[4] Dana Berman, Tali Treibitz, and Shai Avidan. Single image dehazing using haze-lines. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):720–734, 2018. 6, 7

[5] D Richard Blidberg. The development of autonomous underwater vehicles (auv); a brief summary. In *Ieee Icra*, volume 4, pages 122–129, 2001. 1

[6] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61:211–231, 2005. 5

[7] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023. 3, 7, 8

[8] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE transactions on image processing*, 21(4):1756–1769, 2011. 2

[9] Dwight F Coleman, James B Newman, and Robert D Ballard. Design and implementation of advanced underwater imaging systems for deep sea marine archaeological surveys. In *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158)*, volume 1, pages 661–665. IEEE, 2000. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Paul Drews, Erickson Nascimento, Filipe Moraes, Silvia Botelho, and Mario Campos. Transmission estimation in underwater single images. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 825–830, 2013. 1, 2

[12] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7159–7165. IEEE, 2018. 2

[13] Barak Fishbain, Leonid P Yaroslavsky, Ianir Ideses, et al. Spatial, temporal, and interchannel image data fusion for long-distance terrestrial observation systems. *Advances in Optical Technologies*, 2008:546808, 2008. 1

[14] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5

[15] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18155, 2023. 6, 7, 8

[16] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 2

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[18] Raqib Khan, Priyanka Mishra, Nancy Mehta, Shruti S Phutke, Santosh Kumar Vipparthi, Sukumar Nandi, and Subrahmanyam Murala. Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1454–1463, 2024. 2, 3, 6, 7, 8

[19] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021. 2

[20] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019. 1, 2, 5, 6, 7, 8

[21] Chong-Yi Li, Ji-Chang Guo, Run-Min Cong, Yan-Wei Pang, and Bo Wang. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Transactions on Image Processing*, 25(12):5664–5677, 2016. 1

[22] Hanyu Li, Jingjing Li, and Wei Wang. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*, 2019. 6, 7, 8

[23] Jie Li, Katherine A Skinner, Ryan M Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation letters*, 3(1):387–394, 2017. 2

[24] Kunqian Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. Beyond single reference for training: Underwater image enhancement via comparative learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2561–2576, 2022. 6, 7, 8

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration us-

ing swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[26] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. 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. 6, 7

[27] Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31:4922–4936, 2022. 6, 7, 8

[28] Shiben Liu, Huijie Fan, Qiang Wang, Zhi Han, Yu Guan, and Yandong Tang. Wavelet-pixel domain progressive fusion network for underwater image enhancement. *Knowledge-Based Systems*, page 112049, 2024. 3

[29] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2021. 3

[30] Liam Paull, Sajad Saeedi, Mae Seto, and Howard Li. Auv navigation and localization: A review. *IEEE Journal of oceanic engineering*, 39(1):131–149, 2013. 1

[31] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. 2, 3, 6, 7, 8

[32] Yan-Tsung Peng, Keming Cao, and Pamela C Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. 1, 2

[33] Yan-Tsung Peng and Pamela C Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE transactions on image processing*, 26(4):1579–1594, 2017. 1, 2

[34] Airton Marco Polidorio, Franklin César Flores, Clélia Franco, Nilton Nobuhiro Imai, and Antonio MG Tommaselli. Enhancement of terrestrial surface features on high spatial resolution multispectral aerial images. In *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 295–300. IEEE, 2010. 1

[35] Qi Qi, Yongchang Zhang, Fei Tian, QM Jonathan Wu, Kunqian Li, Xin Luan, and Dalei Song. Underwater image co-enhancement with correlation feature matching and joint learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1133–1147, 2021. 2

[36] JL Ribeiro and EA Elsayed. A case study on process optimization using the gradient loss function. *International Journal of Production Research*, 33(12):3233–3248, 1995. 5

[37] Catur Aries Rokhmana and Hanif Muhammad Fauzi. Some enhancement of aerial and terrestrial photo for 3d modeling of texture-less object surface. In *International Conference on Unmanned Aerial System in Geomatics*, pages 289–299. Springer, 2021. 1

[38] Mark Shortis and Euan Harvey& Dave Abdo. A review of underwater stereo-image measurement for marine biology and ecology applications. *Oceanography and marine biology*, pages 269–304, 2016. 1

[39] DR So, W Manke, H Liu, Z Dai, N Shazeer, and QV Le. Primer: Searching for efficient transformers for language modeling. arxiv 2021. *arXiv preprint arXiv:2109.08668*. 2

[40] Yi Tang, Hiroshi Kawasaki, and Takafumi Iwaguchi. 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*, pages 5419–5427, 2023. 2

[41] Pritish M Uplavikar, Zhenyu Wu, and Zhangyang Wang. All-in-one underwater image enhancement using domain-adversarial learning. In *CVPR workshops*, pages 1–8, 2019. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[43] Yi Wang, Hui Liu, and Lap-Pui Chau. Single underwater image restoration using adaptive attenuation-curve prior. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(3):992–1002, 2017. 2

[44] Yinglong Wang, Chao Ma, and Bing Zeng. Multi-decoding deraining network and quasi-sparsity based training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13375–13384, 2021. 3

[45] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2, 3, 4, 5

[46] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 5

[47] Junku Yuh and Michael West. Underwater robotics. *Advanced Robotics*, 15(5):609–639, 2001. 1

[48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2, 3, 4, 5, 7, 8

[49] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11852–11861, 2024. 1

[50] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024. 3

[51] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14122–14132, 2023. 3

[52] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2952–2963, 2024. 2, 5