U-ENHANCE: Underwater Image Enhancement Using Wavelet Triple Self-Attention

Priyanka Mishra[1], Santosh Kumar Vipparthi[1], and Subrahmanyam Murala[2]

[1] CVPR Lab, Indian Institute of Technology Ropar, India
[2] CVPR Lab, SCSS, Trinity College Dublin, Ireland
priyanka.20eez0010@iitrpr.ac.in

Abstract. Transformer-based methods have demonstrated remarkable performance in underwater image enhancement due to their ability to capture long-range dependencies, crucial for high-quality reconstruction of degraded images. However, existing Transformer-based techniques often treat all token similarities equally during self-attention, which can lead to the aggregation of irrelevant features, hampering clear image restoration. We propose **U-ENHANCE**, a novel **U**nderwater image **Enhance**ment framework that integrates wavelet-based frequency decomposition with spatial domain attention to address these challenges. In particular, we introduce a Wavelet Triple Self-Attention (WTSA) mechanism that performs self-attention across three dimensions—horizontal, vertical, and channel-wise, effectively capturing multi-scale features critical for restoring fine details and structural integrity. Additionally, we design a Self-Calibrated Feedforward Network (SCFN) that refines feature representation by dynamically adjusting the receptive field, further enhancing spatial and frequency domain integration. Extensive experiments on underwater image enhancement benchmarks demonstrate that U-ENHANCE outperforms state-of-the-art methods by providing superior restoration of color, clarity, and structural details. The code is available at: https://github.com/Priyanka01mishra/UENHANCE.

Keywords: Underwater image enhancement · Wavelet transform · Transformer

1 Introduction

Underwater imagery plays a critical role in various applications, including marine biology research [39], underwater archaeology [1, 9], environmental monitoring, offshore engineering, coastal border security. It also supports efficient fish farming [37] and the use of autonomous underwater vehicles (AUVs) [2, 30, 31] for ocean exploration and surveillance [52, 54]. High-quality underwater images are essential for tasks such as species identification, habitat mapping, and structural inspections of underwater infrastructure like pipelines and shipwrecks [30]. Moreover, in robotic and autonomous systems, the effectiveness of underwater navigation and object detection heavily relies on the clarity of captured images. However, capturing clear and accurate underwater images is a significant challenge

due to the unique properties of the aquatic environment. As light penetrates through water, it undergoes refraction, absorption, and scattering, resulting in several types of image degradation [51]. These effects lead to color distortion, where certain wavelengths of light (particularly red) are absorbed more quickly, causing a green or blue hue in images. Additionally, scattering caused by particles in the water results in reduced contrast, blurriness, and haziness, which obscure fine details and structural features. These factors make underwater images visually unappealing and can severely limit the performance of underwater robotic systems, which rely on accurate visual information for navigation and object detection. The restoration of underwater images is critical to overcoming these challenges, recovering lost details, increasing visibility, and improving the overall quality of underwater scenes [14, 17, 20]. Traditional image restoration techniques often rely on handcrafted priors and physical models to address specific degradations [16]. However, the diversity and complexity of underwater environments make it difficult to generalize these methods across different conditions. In contrast, deep learning-based approaches have gained popularity due to their ability to learn data-driven solutions that can adapt to various underwater conditions [6, 7, 13, 26]. These methods have shown promise in improving the quality of underwater images by effectively addressing color distortion, low contrast, and blurriness. CNN-based networks have demonstrated outstanding results in underwater image enhancement because of their ability to capture local features through convolutional operations. However, these methods inherently lack the ability to model long-range dependencies and global context [47], crucial for managing the complex degradations found in underwater environments. CNNs primarily rely on fixed-sized kernels [5], which restrict their ability to capture multi-scale features effectively, and they often struggle to generalize across diverse underwater conditions. As a result, exploring Transformer-based methods, which excel in modelling long-range dependencies and global feature interactions [44], presents a more promising direction for underwater image enhancement [32, 35]. While existing Transformer-based approaches have focused primarily on spatial domain attention, limited work has explored the integration of frequency domain information [19,45], which is crucial for fully addressing underwater degradations. Among the few methods that do incorporate frequency domain features, most rely on Fourier transforms [19,45]. However, Fourier-based methods, while useful for capturing global frequency components, lack the ability to localize features in both time (spatial) and frequency domains simultaneously. This limitation makes it challenging to accurately restore fine details and structural elements, which are essential in underwater scenes. In contrast, wavelet transforms offer a distinct advantage by providing multi-scale frequency representation with spatial localization [38,41]. Unlike Fourier transforms, which only capture global frequency information, wavelets allow for the decomposition of an image into both frequency and spatial components, making them highly effective in preserving localized details while also addressing global distortions. Given the clear benefits of wavelet transforms in providing both frequency domain and spatial domain localization, incorporating wavelet-based techniques into Trans-

former models for underwater image enhancement is a logical and advantageous step forward.

In this work, we propose an efficient transformer-based network, U-ENHANCE, for underwater image enhancement that effectively leverages wavelet-based frequency decomposition to capture both spatial and frequency domain features, significantly enhancing restoration quality. The key component of U-ENHANCE is the Wavelet Transformer Block (WTB), which integrates the Wavelet Triple Self-Attention (WTSA) mechanism and a Self-Calibrated Feedforward Network (SCFN). The WTSA captures essential multi-scale features by applying self-attention across three dimensions—horizontal, vertical, and channel-wise—thus effectively capturing long-range dependencies and fine details. Additionally, WTSA decomposes features into wavelet sub-bands to preserve both high-frequency details and low-frequency structures, ensuring comprehensive feature extraction for underwater image enhancement. Furthermore, SCGFN refines these features by dynamically adjusting the receptive field, resulting in improved feature representation in both spatial and frequency domains. By adaptively handling local and global features, SCFN enhances the model's ability to restore clear, sharp images from severely degraded underwater scenes. This novel combination of wavelet-based frequency analysis and transformer attention mechanisms enables U-ENHANCE to significantly outperform existing methods, achieving superior restoration quality across multiple underwater image benchmarks.

The main contributions of this paper are as follows:

- We propose a Wavelet Triple Self-Attention (WTSA) Module that decomposes input features into frequency sub-bands using Discrete Wavelet Transform (DWT) for better noise reduction and detail preservation. In this module, we further introduce a Triple Attention mechanism (horizontal, vertical, and channel self-attention) to reduce computational complexity and capture long-range dependencies essential for underwater image enhancement.
- We propose a Self-Calibrated Feedforward Network (SCFN) which dynamically adjusts the receptive field to capture richer spatial and inter-channel dependencies. This improves spatial and contextual information processing, leading to more discriminative feature representation and enhanced restoration performance.

Experimentation on synthetic and real-world datasets, along with ablation studies verify the effectiveness of the proposed method for underwater image enhancement.

2 Literature Survey

2.1 Underwater Image Enhancement

Underwater image enhancement (UIE) techniques are generally classified into three main categories: physical model-based methods, visual prior-based approaches, and deep learning-based strategies [10, 21, 32, 33, 42]. Physical model-based methods often employ prior knowledge to construct enhancement models,

utilizing concepts such as attenuation curve priors [46], fuzzy priors [8], and water dark channel priors [34]. While these methods can be effective, their reliance on externally defined priors can limit scalability and robustness in complex and diverse underwater environments.

Recent advancements in deep learning have shown significant promise in UIE. To address the challenge of limited real-world underwater paired training data, many researchers have turned to Generative Adversarial Network (GAN)-based frameworks. Notable examples include UGAN [12], FUnIE-GAN [17] UIE-DAL [43], and Watergan [23], which generate synthetic training data to improve enhancement performance. Additionally, Semi-UIR [16] introduced a mean teacher-based semi-supervised network that effectively leverages unlabeled data to enhance model training.

Emerging research has started to explore the utilization of frequency domain properties in UIE, highlighting the significant potential of frequency-based methods. Spectroformer [19], for instance, exploits frequency characteristics through a hybrid Fourier-spatial upsampling technique to enhance the resolution of degraded image features. Similarly, WF-Diff [50] combines frequency domain analysis with diffusion models for image enhancement and adjustment. Recent developments in wavelet-pixel domain fusion, such as WPFNet [28], have demonstrated improved UIE by integrating wavelet and pixel domains. This fusion preserves fine details and enhances color fidelity while reducing noise more effectively than previous methods. However, these frequency-based approaches may introduce unintended interactions in irrelevant areas due to their computational complexity and the potential for processing overhead.

2.2 Transformers in Vision

Building on the transformative impact of Transformers in NLP and high-level vision applications, recent advancements have extended their use to image restoration tasks, where they have outperformed traditional CNN-based models by effectively capturing long-range dependencies [4]. Nevertheless, the quadratic complexity of standard self-attention presents challenges for processing high-resolution images. To mitigate this issue, [49] proposed a transformer architecture optimized for restoration tasks like image deraining, deblurring, and denoising by computing attention across the channel dimension, thereby reducing computational burden. Another approach is the use of window-based attention, exemplified by Uformer [47], which enhances local interactions within the Transformer architecture. SwinIR [25] also leverages window-based attention but incorporates a shift mechanism to facilitate better cross-window communication. Additional strategies have explored the application of Transformers with channel-wise and spatial-wise attention layers [32], or through the integration of both frequency and spatial domains for self-attention as seen in [19]. In contrast to these methods, we propose an adaptive sparse self-attention mechanism in the wavelet domain, aimed at reducing redundancy by focusing on the most relevant interactions.

3 Proposed Method

3.1 Overall Pipeline

The proposed U-ENHANCE framework, as illustrated in Fig. 1, begins by embedding an input RGB image, $\boldsymbol{I}$ into a feature space using the Input Projection module, which consists of a 3×3 standard convolution. This step produces low-level features, $X_0 \in H \times W \times C$ where, H, W, C refers to the height, width and channel dimensions, respectively that are passed into a hierarchical encoder-decoder backbone network, designed with four stages. Between stages, down-sampling and up-sampling are achieved through pixel-unshuffle and pixel-shuffle operations, enabling multi-scale representation of underwater-degradation effects. At the core of each block is the use of Wavelet Triple Self-Attention (WTSA), which replaces the conventional Transformer self-attention [44]. This WTSA mechanism introduces a three-dimensional co-computation of horizontal, vertical, and channel-wise self-attention, making feature aggregation more efficient. Additionally, each U-ENHANCE Block incorporates a Self-Calibrated Feedforward Network (SCFN) to ensure more effective feature refinement. After encoding and decoding, the deep features, $X_d \in H \times W \times C$, are passed through an Output Projection module, another 3×3 standard convolution, which restores the feature map to its original dimensions of $H \times W \times 3$. The final restored image $\boldsymbol{O}$ is produced by combining this output with the original input via a residual connection, ensuring enhanced image restoration.

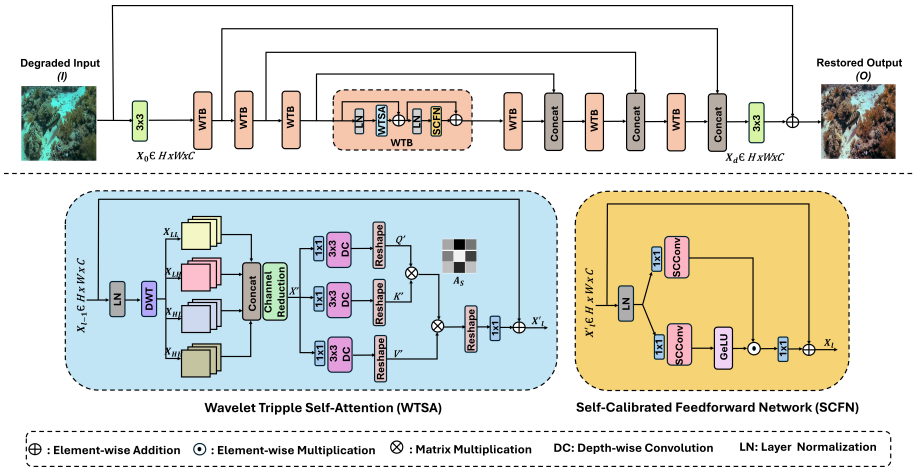

Fig. 1: Illustration of the proposed U-ENHANCE. It consists of a Wavelet Transformer Block (WTB) with a Wavelet Tripple Self-Attention (WTSA) and a Self-Calibrated Feed-forward Network (SCFN).

3.2 Wavelet Transformer Block

Existing underwater image enhancement methods based on Transformers emphasize spatial domain attention, often neglecting the critical role of frequency information [19, 45]. While some techniques incorporate frequency domain features, they frequently struggle to effectively capture both frequency and spatial characteristics due to a lack of a unified integration strategy. To overcome this limitation, we introduce the Wavelet Triple Self-Attention (WTSA) mechanism, which replaces the traditional Multi-head Self-Attention (MSA) in conventional transformer blocks. Our approach effectively fuses wavelet and spatial domain information, enabling superior feature extraction for restoring underwater-degraded effects like color distortion and blurriness while preserving structural details and edges. Wavelets offer a key advantage by preserving localized spatial information even in the frequency domain, providing a multiscale representation that captures both fine-grained frequency details and spatial localization, crucial for maintaining structural integrity across different scales [38, 41]. Additionally, we introduce a Self-Calibrated Feedforward Network (SCFN) to ensure efficient feature refinement. This network optimizes the processing of both spatial and frequency domain features, further enhancing the model's ability to accurately restore underwater images. The computational workflow of the Wavelet Transformer Block (WTB) for underwater image enhancement can be described as follows:

$$X'_\ell = X_{\ell-1} + \text{WTSA}(\text{LN}(X_{\ell-1})) \tag{1}$$

$$X_\ell = X'_\ell + \text{SCFN}(\text{LN}(X'_\ell)) \tag{2}$$

where $\text{LN}(\cdot)$ is layer normalization, X'_l and X_l denote the outputs of the WTSA and SCFN blocks, respectively, which are explained in the following subsection.

3.3 Wavelet Tripple Self-Attention

In the proposed Wavelet Triple Self-Attention (WTSA) module, the input features $X_{l-1} \in H \times W \times C$ first undergo layer normalization (LN) to stabilize the input distribution. Following normalization, the features are transformed into the frequency domain using the Discrete Wavelet Transform (DWT), which decomposes them into four sub-bands: $(X_{LL}, X_{LH}, X_{HL}, X_{HH})$ as shown below in eqn 3 :

$$(X_{LL}, X_{LH}, X_{HL}, X_{HH}) = \text{DWT}(\text{LN}(X_{\ell-1})) \tag{3}$$

Decomposing input features into these four sub-bands offers several advantages. Firstly, it allows for the preservation of both low-frequency and high-frequency components of the signal, which is essential for accurately capturing the overall structure and intricate details of the input data. Secondly, working in the frequency domain enables better noise reduction, as certain frequency components can be attenuated while retaining important features. These sub-bands retain

the same number of channels as the input features but have spatial dimensions downsampled by a factor of 2. After the DWT, the four sub-bands are concatenated, and a channel reduction operation is applied to reduce the number of channels to match the input feature dimensions as described in eqn 4 :

$$X' = \text{CR}(\text{Concat}(LL, LH, HL, HH)) \tag{4}$$

where, CR is channel reduction convolution.
To restore the spatial resolution, the concatenated features are upsampled using bilinear interpolation to return them to their original spatial dimensions. The use of bilinear interpolation is advantageous because it provides a smooth upsampling process, preserving the visual quality of the images by minimizing artifacts. Additionally, bilinear interpolation is computationally efficient, allowing for quick reconstruction of spatial dimensions without significantly increasing processing time. This is followed by a 1×1 pointwise convolution to mix the channels effectively and a 3×3 depth-wise convolution to refine the spatial features and capture local dependencies. Subsequently, the features are processed through the Triple Attention mechanism. Here, the self-attention computation is decomposed into three separate components: horizontal self-attention, vertical self-attention, and channel self-attention, based on the concept of Triple Multi-Dconv Head Transposed Attention (TMDTA) [53]. This approach divides the feature correlation into three distinct directions as shown in Fig. 2, with the query (Q'), key (K'), and value (V') tensors being reshaped accordingly. Specifically, the query, key and value tensors are projected in three dimensions: horizontally (Q_H, K_H, V_H), vertically (Q_W, K_W, V_W), and along the channel dimension (Q_C, K_C, V_C), resulting in three attention matrices: $A_H \in R^{H \times H}$, $A_W \in R^{W \times W}$, and $A_C \in R^{C \times C}$, instead of the conventional full-pixel attention matrix $R^{HW \times HW}$. Eventually these three different attention matrices are concatenated to generate A_s matrix as defined by the eqn 5 and 6 where, A_x represents the self-attention along the horizontal, vertical and channel dimesions.

$$A_S = \text{Concat}(A_C, A_H, A_W) \tag{5}$$

$$A_x = \text{Softmax}\left(\frac{Q_x K_x^T}{\alpha}\right) \times V_x \tag{6}$$

where, $x \in \{C, H, W\}$.
By performing matrix multiplications in each of these three directions separately, the WTSA module preservs essential spatial and frequency information. The result is a more efficient feature representation that captures long-range dependencies across spatial and channel dimensions, leading to improved restoration of underwater-degraded images. This combination of wavelet-based frequency decomposition, spatial refinement, and directional attention ensures that the WTSA module efficiently captures both local and global features. The final output of WTSA block is X_l' which is shown in eqn 1.

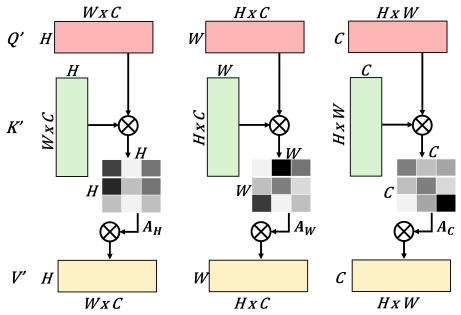

Fig. 2: Illustration of the Triple attention concept [53]. Characteristic pixel attention is decomposed into horizontal, vertical, and channel self-attention.

3.4 Self-Calibrated Feedforward Network

Previous studies on transformer architectures primarily employ standard feedforward networks (FFNs) in the transformer block to handle the token-wise transformations. However, these implementations typically overlook long-range spatial dependencies and inter-channel correlations, which are crucial for more discriminative feature representation. To address this, we propose a novel Self-Calibrated Feedforward Network (SCFN) that integrates self-calibrated convolutions [27] within the feedforward layers, enhancing both spatial and inter-channel information processing.

In particular, the SCFN builds upon the idea of self-calibrated convolution [27], which adapts the receptive field dynamically to consider richer contextual information. The result is a feedforward network that not only encodes local information but also expands its scope to adaptively capture long-range dependencies around each spatial position, thus improving the overall representation capability. The full process of the SCFN is formulated as follows (see eqn 7):

$$X_l = X'_l + \psi_1\Big[\text{GELU}\left(\text{SCConv}\left(\psi_1\left(\text{LN}(X'_l)\right)\right)\right) \odot \text{SCConv}\left(\psi_1\left(\text{LN}(X'_l)\right)\right)\Big] \tag{7}$$

where, $\psi_1(.)$ and SCConv denotes the 1×1 convolution and self-calibrated convolution, respectively.
This novel SCFN introduces an adaptive mechanism that dynamically adjusts the spatial and channel-wise dependencies, resulting in more robust and discriminative feature extraction. Moreover, by incorporating self-calibration, the SCFN avoids unnecessary complexity or parameter overhead while significantly improving the model's ability to capture multi-scale information, essential for underwater image enhancement.

4 Experimental Details

4.1 Training Losses

In training our proposed architecture, we utilized a comprehensive total loss function, denoted as L_T, which combines several distinct loss components. These components include perceptual loss (L_1) [18], Charbonnier loss (L_2) [3], multi-scale structural similarity index (MS-SSIM) loss (L_3) [48], and gradient loss (L_4) [36]. Together, these losses are integrated to ensure a robust and effective training process, addressing both perceptual quality and structural accuracy. The total loss function is formulated as follows:

$$L_T = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4 \tag{8}$$

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. The losses are explained individually as follows:

Perceptual Loss (L_1): Perceptual loss measures the perceptual similarity between generated and target images by utilizing feature representations from a pre-trained neural network. This approach has been shown to improve the quality of generated images across various image-generation tasks. Let O represent the target image and G_t represent the generated image. Using a pre-trained VGG19 [40] network (ϕ_i) we extract feature maps at different layers. The perceptual loss, L_1, is calculated as the difference between the feature maps of the target and generated images:

$$L_1 = \sum_{i=1}^{N=4} \|\phi_i(O) - \phi_i(G_t)\|_2^2 \tag{9}$$

Here, ϕ_i represents the feature extraction function at layer i of the CNN, and $(N = 4)$ is the total number of layers considered for perceptual loss calculation.

Charbonnier loss (L_2): Training the network with MSE loss often results in blurry reconstructions because it maximizes the log-likelihood of a Gaussian distribution. To address this issue, we chose the Charbonnier loss, a differentiable version of the L_1 norm. The Charbonnier loss is computed between the restored images (O) and their corresponding ground-truth images (G_t), and it is defined as follows:

$$L_2 = \mathbb{E}_{O\sim Q(O), G_t\sim Q(G_t)} \sqrt{(O - G_t)^2 + \epsilon} \tag{10}$$

where, $Q(O)$ and $Q(G_t)$ are the distributions of the restored image (O) and the ground-truth image (G_t), respectively. Additionally, the value of ϵ is empirically set to 1×10^{-3}.

MS-SSIM loss (L_3): The Structural Similarity (SSIM) loss primarily addresses a single input resolution. In contrast, the Multi-Scale SSIM (MS-SSIM) loss provides greater flexibility by taking into account different input resolutions.

$$L_3 = 1 - (MSSSIM(O, G_t)) \tag{11}$$

Gradient loss (L_4): Generally, the Charbonnier loss prioritizes low-frequency components. However, when training the network to incorporate high-frequency details, the gradient loss becomes crucial. This second-order loss function enhances the sharpness of edges in the output [29]. Here, $\hat{G_O}$ and $\hat{G_{G_t}}$ represent the distributions of $Q(O)$ and $Q(G_t)$ respectively.

$$L_4 = \mathbb{E}_{\hat{G_O} \sim Q(O), \hat{G_O} \sim Q(G_t)} \left\| \hat{G}_{G_t} - \hat{G}_O \right\|_1 \tag{12}$$

4.2 Datasets

To perform a comparative analysis, we utilized the synthetic Underwater Image Enhancement Benchmark (UIEB) [21] along with the real-world underwater datasets U45 [22] and C60 [21]. The Underwater Image Enhancement Benchmark (UIEB) dataset contains 890 pairs of underwater images, including both degraded and clean versions, representing various scenes to capture the diversity of underwater environments. The dataset is divided into 800 image pairs for training, selected at random, while the remaining 90 pairs are designated for testing. Additionally, the dataset provides 60 real-world degraded underwater images for evaluation purposes. U45 consists of 45 real-world images that exhibit features like color casts, low contrast, and degradation effects similar to haze in underwater environments.

4.3 Training Details

To mitigate the limited size of the UIEB dataset for training, we employed several data augmentation strategies to expand the dataset's diversity. These augmentations included horizontal and vertical flips, noise addition, and contrast adjustments. This approach significantly increased the variability in the training data, improving the robustness and performance of the model. In total, 4800 augmented image pairs from the UIEB dataset were used for training, while 90 images were reserved for testing. All input images were uniformly resized to 256×256 pixels for consistency. The model was trained using the ADAM optimizer with an initial learning rate of 3×10^{-4}, which was progressively adjusted through a cosine annealing schedule. The implementation was done in PyTorch, and the training was conducted on an NVIDIA GeForce RTX 2080 GPU.

Table 1: Quantitative comparison of different UIE methods on the synthetic UIEB dataset (↑: higher is better, ↓: lower is better, **red** and blue indicate **best** and second best values, respectively).

Method	PSNR ↑	SSIM ↑	UIQM ↑	Parameters (M) ↓
UDCP [11]	13.81	0.692	1.825	-
UIBLA [34]	15.78	0.731	2.014	-
RGHS [15]	14.57	0.791	2.410	-
WaterNet [21]	19.81	0.864	2.818	193.70
FUnIE-GAN [17]	21.03	0.775	**3.092**	7.02
CLUIE-Net [24]	20.37	0.890	2.674	31.00
Ours	**22.07**	**0.911**	2.701	**5.52**

4.4 Results on Synthetic Dataset

The effectiveness of the proposed approach is evaluated through a quantitative comparison with current state-of-the-art methods, using metrics like PSNR and SSIM. We also compared the parameters (in millions, M) of learning-based methods. Table 1 presents the quantitative outcomes on the popular UIEB dataset, while Fig. 3 illustrates qualitative results. The proposed method exhibits superior performance in comparison to the state-of-the-art techniques.

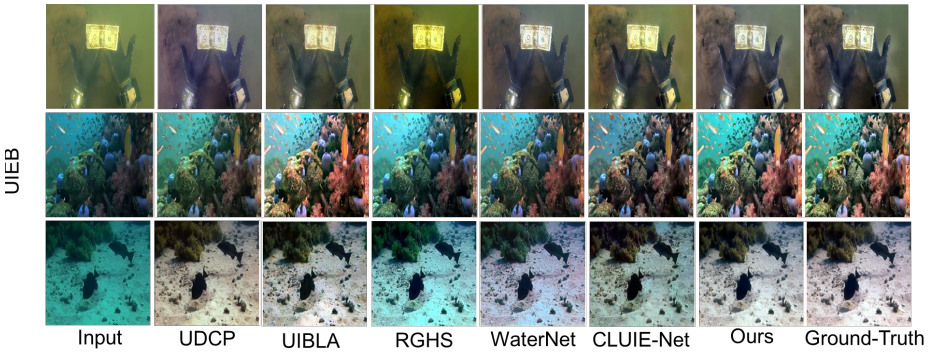

Fig. 3: Visual comparison of our method and state-of-the-art techniques on the full-reference UIEB dataset [21].

4.5 Results on Real-world Dataset

To evaluate the effectiveness of our proposed method in real-world conditions, we presented quantitative results on the C60 dataset [21]. Our analysis includes a comprehensive evaluation using metrics such as the Underwater Image Quality Measure (UIQM), Underwater Image Sharpness Measure (UISM), Naturalness Image Quality Evaluator (NIQE), and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). The quantitative outcomes are summarized in Table

Table 2: Quantitative comparison of different UIE methods on the real-world C60 dataset.

Method	UIQM ↑	UISM ↑	NIQE ↓	BRISQUE ↓
UDCP [11]	4.371	6.849	6.246	29.412
UIBLA [34]	4.654	7.084	6.230	27.968
RGHS [15]	3.942	7.347	6.141	25.485
WaterNet [21]	4.348	7.322	5.310	**23.115**
CLUIE-Net [24]	4.194	7.418	5.848	26.369
Ours	**4.713**	**7.531**	**5.269**	25.870

2. In addition, the qualitative comparison of the C60 and U45 datasets is illustrated in Figure 4. The results highlight a substantial improvement in color correction and overall visibility in the enhanced images, which can be attributed to the novel modules incorporated in our proposed approach.

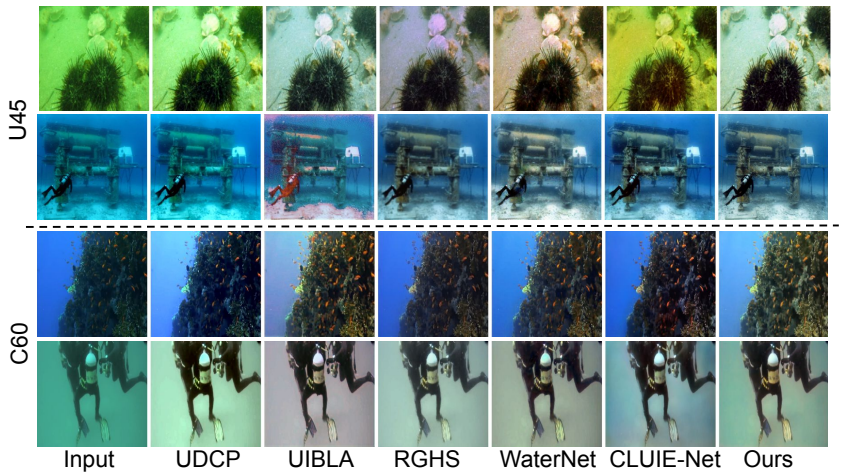

Fig. 4: Visual comparison of our method and state-of-the-art techniques on real-world U45 [22] and C60 [21] datasets.

5 Ablation Studies

To validate the effectiveness of the proposed components, we conduct a series of ablation experiments using the UIEB dataset.

5.1 Impact of the Proposed Wavelet Triple Self-Attention

To evaluate the effectiveness of the proposed Wavelet Triple Self-Attention (WTSA) module, we performed ablation experiments by comparing the model's perfor-

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

Network Setting	PSNR ↑	SSIM ↑
Baseline	20.82	0.651
Baseline + WTSA	21.53	0.787
Baseline + SCFN	21.74	0.834
Ours (Baseline + WTSA + SCFN)	**22.07**	**0.911**

mance with and without this module, provided in Table 3. Ablation experiments revealed significant improvements in underwater image enhancement when the Wavelet Triple Self-Attention (WTSA) module was included. Without WTSA, the model struggled with noise reduction and detail preservation, resulting in lower image quality. Incorporating WTSA led to higher performance across all metrics, including PSNR and SSIM, due to the wavelet-based frequency decomposition and triple attention mechanism, which efficiently captured long-range dependencies. These results underscore the importance of WTSA in enhancing image clarity while maintaining computational efficiency.

5.2 Impact of the Proposed Self-Calibrated Feedforward Network

To validate the effectiveness of the proposed Self-Calibrated Feedforward Network (SCFN) in U-ENHANCE, we conducted ablation experiments comparing performance with and without this module, highlighted in Table 3. The baseline model without SCFN showed a noticeable drop in image restoration quality, particularly in color accuracy and structural clarity. In contrast, incorporating SCFN led to superior results across all metrics. SCFN improved feature refinement by dynamically adjusting the receptive field, enhancing the integration of spatial and frequency domain features. These findings confirm SCFN's critical role in boosting U-ENHANCE's overall performance.

6 Applicability to Higher Level Computer Vision Task

In underwater environments, reduced visibility often impairs the performance of computer vision tasks. To address this, enhancing underwater images can serve as a crucial pre-processing step, improving the accuracy of downstream applications. To validate this, we performed an experiment centred on underwater depth estimation. We first applied a depth estimation algorithm to the degraded images and then compared the results with those obtained from images enhanced through our method as well as other existing techniques. As illustrated in Fig. 5, our enhanced images outperformed all other methods in depth estimation, demonstrating superior accuracy. This outcome underscores the effectiveness of our approach for depth estimation. The enhanced image quality ensures more reliable data for subsequent computational tasks, which is particularly important in the challenging underwater domain, where clarity and detail can significantly impact the success of vision-based algorithms.

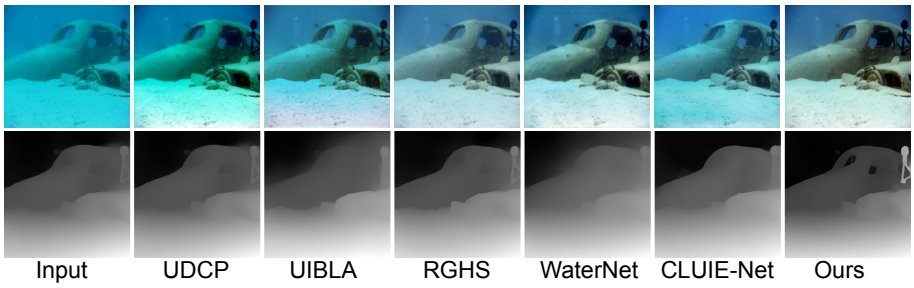

Fig. 5: Application of our proposed method and existing techniques as a pre-processing step for depth estimation on the underwater U45 dataset [22].

7 Conclusion

In this paper, we proposed U-ENHANCE, a novel framework for underwater image enhancement that effectively integrates wavelet-based frequency decomposition with spatial domain attention to address the complex challenges of underwater-degraded images. By introducing the Wavelet Triple Self-Attention (WTSA) mechanism, we demonstrated that self-attention across three dimensions—horizontal, vertical, and channel-wise—enables superior capture of multi-scale features, essential for preserving fine details and structural integrity. Additionally, the Self-Calibrated Feedforward Network (SCFN) refined feature representation by dynamically adjusting the receptive field, enhancing the model's ability to process spatial and frequency domain information. Through extensive experiments on benchmark datasets, U-ENHANCE consistently outperformed state-of-the-art methods, delivering superior restoration of color accuracy, clarity, and structural details. These results confirm the effectiveness of our approach and its potential to advance the field of underwater image enhancement.

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. Bailey, G.N., Flemming, N.C.: Archaeology of the continental shelf: marine resources, submerged landscapes and underwater archaeology. Quaternary Science Reviews **27**(23-24), 2153–2165 (2008) 1
2. Blidberg, D.R.: The development of autonomous underwater vehicles (auv); a brief summary. In: Ieee Icra. vol. 4, pp. 122–129 (2001) 1
3. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. International journal of computer vision **61**, 211–231 (2005) 9

4. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023) 4
5. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al.: Cross aggregation transformer for image restoration. Advances in Neural Information Processing Systems **35**, 25478–25490 (2022) 2
6. Chi, Z., Shu, X., Wu, X.: Joint demosaicking and blind deblurring using deep convolutional neural network. In: 2019 IEEE International conference on image processing (ICIP). pp. 2169–2173. IEEE (2019) 2
7. Chi, Z., Wang, Y., Yu, Y., Tang, J.: Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9137–9146 (2021) 2
8. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE transactions on image processing **21**(4), 1756–1769 (2011) 4
9. Coleman, D.F., Newman, J.B., Ballard, R.D.: 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). vol. 1, pp. 661–665. IEEE (2000) 1
10. Drews, P., Nascimento, E., Moraes, F., Botelho, S., Campos, M.: Transmission estimation in underwater single images. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 825–830 (2013) 3
11. Drews, P.L., Nascimento, E.R., Botelho, S.S., Campos, M.F.M.: Underwater depth estimation and image restoration based on single images. IEEE computer graphics and applications **36**(2), 24–35 (2016) 11, 12
12. Fabbri, C., Islam, M.J., Sattar, J.: Enhancing underwater imagery using generative adversarial networks. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 7159–7165. IEEE (2018) 4
13. Fu, M., Liu, H., Yu, Y., Chen, J., Wang, K.: Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–212 (2021) 2
14. Han, J., Shoeiby, M., Malthus, T., Botha, E., Anstee, J., Anwar, S., Wei, R., Armin, M.A., Li, H., Petersson, L.: Underwater image restoration via contrastive learning and a real-world dataset. Remote Sensing **14**(17), 4297 (2022) 2
15. Huang, D., Wang, Y., Song, W., Sequeira, J., Mavromatis, S.: Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24. pp. 453–465. Springer (2018) 11, 12
16. Huang, S., Wang, K., Liu, H., Chen, J., Li, Y.: Contrastive semi-supervised learning for underwater image restoration via reliable bank. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18145–18155 (2023) 2, 4
17. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters **5**(2), 3227–3234 (2020) 2, 4, 11
18. Johnson, J., Alahi, A., Fei-Fei, L.: 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. pp. 694–711. Springer (2016) 9

19. Khan, R., Mishra, P., Mehta, N., Phutke, S.S., Vipparthi, S.K., Nandi, S., Murala, S.: Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1454–1463 (2024) 2, 4, 6
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**, 4985–5000 (2021) 2
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 (2019) 3, 10, 11, 12
22. Li, H., Li, J., Wang, W.: A fusion adversarial underwater image enhancement network with a public test dataset. arXiv preprint arXiv:1906.06819 (2019) 10, 12, 14
23. Li, J., Skinner, K.A., Eustice, R.M., Johnson-Roberson, M.: Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. IEEE Robotics and Automation letters **3**(1), 387–394 (2017) 4
24. Li, K., Wu, L., Qi, Q., Liu, W., Gao, X., Zhou, L., Song, D.: 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) 11, 12
25. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021) 4
26. Liu, H., Wu, Z., Li, L., Salehkalaibar, S., Chen, J., Wang, K.: Towards multi-domain single image dehazing via test-time training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5831–5840 (2022) 2
27. Liu, J.J., Hou, Q., Cheng, M.M., Wang, C., Feng, J.: Improving convolutional networks with self-calibrated convolutions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10096–10105 (2020) 8
28. Liu, S., Fan, H., Wang, Q., Han, Z., Guan, Y., Tang, Y.: Wavelet-pixel domain progressive fusion network for underwater image enhancement. Knowledge-Based Systems p. 112049 (2024) 4
29. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015) 10
30. Naveen, P.: Advancements in underwater imaging through machine learning: Techniques, challenges, and applications. Multimedia Tools and Applications pp. 1–20 (2024) 1
31. Paull, L., Saeedi, S., Seto, M., Li, H.: Auv navigation and localization: A review. IEEE Journal of oceanic engineering **39**(1), 131–149 (2013) 1
32. Peng, L., Zhu, C., Bian, L.: U-shape transformer for underwater image enhancement. IEEE Transactions on Image Processing **32**, 3066–3079 (2023) 2, 3, 4
33. Peng, Y.T., Cao, K., Cosman, P.C.: Generalization of the dark channel prior for single image restoration. IEEE Transactions on Image Processing **27**(6), 2856–2868 (2018) 3
34. 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) 4, 11, 12
35. Ren, T., Xu, H., Jiang, G., Yu, M., Luo, T.: Reinforced swin-convs transformer for underwater image enhancement. arXiv preprint arXiv:2205.00434 (2022) 2

36. Ribeiro, J., Elsayed, E.: A case study on process optimization using the gradient loss function. International Journal of Production Research **33**(12), 3233–3248 (1995) 9
37. Rout, D.K., Kapoor, M., Subudhi, B.N., Thangaraj, V., Jakhetiya, V., Bansal, A.: Underwater visual surveillance: A comprehensive survey. Ocean Engineering **309**, 118367 (2024) 1
38. Scholl, S.: Fourier, gabor, morlet or wigner: comparison of time-frequency transforms. arXiv preprint arXiv:2101.06707 (2021) 2, 6
39. Shortis, M., Abdo, E.H.D.: A review of underwater stereo-image measurement for marine biology and ecology applications. Oceanography and marine biology pp. 269–304 (2016) 1
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
41. Sun, Q., Ren, Y., Jiao, L., Li, X., Shang, F., Liu, F.: Mwq: Multiscale wavelet quantized neural networks. arXiv preprint arXiv:2103.05363 (2021) 2, 6
42. Tang, Y., Kawasaki, H., Iwaguchi, T.: 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. pp. 5419–5427 (2023) 3
43. Uplavikar, P.M., Wu, Z., Wang, Z.: All-in-one underwater image enhancement using domain-adversarial learning. In: CVPR workshops. pp. 1–8 (2019) 4
44. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017) 2, 5
45. Wang, D., Sun, Z.: Frequency domain based learning with transformer for underwater image restoration. In: Pacific Rim International Conference on Artificial Intelligence. pp. 218–232. Springer (2022) 2, 6
46. Wang, Y., Liu, H., Chau, L.P.: Single underwater image restoration using adaptive attenuation-curve prior. IEEE Transactions on Circuits and Systems I: Regular Papers **65**(3), 992–1002 (2017) 4
47. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022) 2, 4
48. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003) 9
49. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022) 4
50. Zhao, C., Cai, W., Dong, C., Hu, C.: 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. pp. 8281–8291 (2024) 4
51. Zhou, J., Li, B., Zhang, D., Yuan, J., Zhang, W., Cai, Z., Shi, J.: Ugif-net: An efficient fully guided information flow network for underwater image enhancement. IEEE Transactions on Geoscience and Remote Sensing (2023) 2
52. Zhou, J., Zhang, D., Ren, W., Zhang, W.: Auto color correction of underwater images utilizing depth information. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) 1

53. Zhou, Y., Lin, J., Ye, F., Qu, Y., Xie, Y.: Efficient lightweight image denoising with triple attention transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7704–7712 (2024) 7, 8
54. Zhuang, P., Ding, X.: Underwater image enhancement using an edge-preserving filtering retinex algorithm. Multimedia Tools and Applications **79**(25), 17257–17277 (2020) 1