Spectroformer: Multi-Domain Query Cascaded Transformer Network For Underwater Image Enhancement

MD Raqib Khan[1], Priyanka Mishra[2], Nancy Mehta[3], Shruti S. Phutke[4]
Santosh Kumar Vipparthi[2], Sukumar Nandi[5], Subrahmanyam Murala[1]
[1]CVPR Lab, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2]CVPR Lab, Indian Institute of Technology Ropar, India
[3]Computer Vision Lab, CAIDAS, IFI, University of Würzburg, Germany
[4]Institute for Integrated and Intelligent Systems, Griffith Univeriesty, Australia
[5]Indian Institute of Technology Guwahati, India
khanmd@tcd.ie

Abstract

Underwater images often suffer from color distortion, haze, and limited visibility due to light refraction and absorption in water. These challenges significantly impact autonomous underwater vehicle applications, necessitating efficient image enhancement techniques. To address these challenges, we propose a Multi-Domain Query Cascaded Transformer Network for underwater image enhancement. Our approach includes a novel Multi-Domain Query Cascaded Attention mechanism that integrates localized transmission features and global illumination features. To improve feature propagation from the encoder to the decoder, we propose a Spatio-Spectro Fusion-Based Attention Block. Additionally, we introduce a Hybrid Fourier-Spatial Upsampling Block, which uniquely combines Fourier and spatial upsampling techniques to enhance feature resolution effectively. We evaluate our method on benchmark synthetic and real-world underwater image datasets, demonstrating its superiority through extensive ablation studies and comparative analysis. The testing code is available at: https://github.com/Mdraqibkhan/Spectroformer.

1. Introduction

Underwater Image Enhancement (UIE) algorithms are vital for aquatic exploration with wide applications in Autonomous Underwater Vehicles (AUVs), underwater mine detection [52], submerged robots [17], and among other fields. However, the major challenges in underwater imaging include poor equipment quality [27], insufficient illumination, and light absorption/scattering [45]. These issues lead to quality problems like color shifts, haziness, and blurriness, reducing image interoperability and thus limiting its application in the underwater world [32].

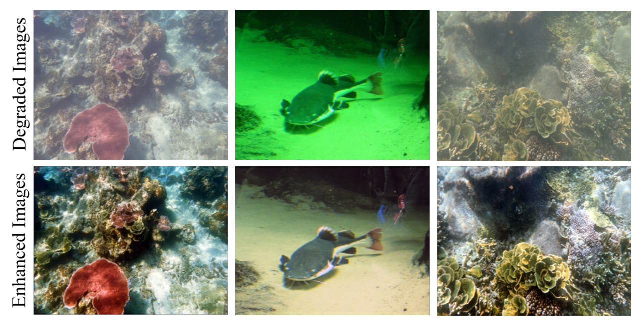

Figure 1. Sample visual results of the proposed network (Spectroformer) on real-world underwater scenarios.

Generally, the existing UIE methods fall into three categories. The first category employs a physical model-based approach [8, 16], centered on accurately estimating the transmission maps to generate enhanced images. However, the effectiveness of these model-based approaches is limited to less complex environments. Visual prior-based UIE approaches [1, 28] in the second category focus on refining the perceptual quality by adjusting the pixel values for contrast, and brightness. Nevertheless, they are constrained by the ignorance of the physical deterioration process.

On the other hand, deep learning methods in the third category [11, 12, 24, 30] exhibit remarkable performance in UIE task. Particularly, recent attempts [35, 41] have been made to tailor transformers [49] for this task on account of their ability to exploit long-range information. Though these aforementioned transformer approaches have shown promising results in underwater applications, they are mainly centered on spatial domains. However, the underwater image acquisition taps into both the frequency and spatial domains, extracting valuable insights. The former (frequency) domain analysis uncovers fine details [50] (high-frequency components) and overarching patterns (low-frequency components), while the latter domain focuses on pixel values and positions for scene understanding. Thus, integrating both domains enhances visibility, color accuracy, and contrast, enabling effective image enhancement in challenging aquatic conditions. Acknowledging this, we introduce a streamlined architecture in Multi-Domain for enhancing the underwater images.

In this work, we propose a novel transformer-based network, Spectroformer for underwater image enhancement that leverages the intrinsic underwater image degradation factors of transmission and atmospheric light. This includes localized transmission (pixel-specific) and globally consistent ambient light. The inclusion of frequency-domain characteristics further enables the pixel positions to encapsulate the overall image properties in the spatial domain. In order to further bridge the gap between spatial and frequency-domains, capturing complex details and comprehensive features, respectively, we design a Multi-Domain Query Cascaded Attention (MQCA) mechanism. Our Multi-Domain Query Cascaded Transformer Network, guided by the innovative MQCA mechanism, seamlessly combines spatial and frequency-domain information to significantly enhance the underwater image quality. *To the best of the authors's knowledge, this is the first effort that focuses on a multi-domain query cascaded attention technique in a transformer for underwater image enhancement.*

Additionally, in order to reinforce the proposed models's attention to the crucial color channels, we introduced a Spatio-Spectro Fusion-Based Attention Block by integrating both domains. It basically replaces the direct skip connections and transmits non-redundant attention-enhanced features from the encoder to the corresponding decoder. Further, we observe that the pixel shuffling technique rearranges pixels to increase spatial resolution [46], which improves visual clarity and detail. In contrast, frequency-domain upsampling draws out small features from various frequencies, to improve the overall quality of the image [57]. To boost the merits of upsampling from individual domains, we propose a Hybrid Fourier-Spatial Upsampling Block. It effectively mixes Fourier and spatial upsampling techniques to significantly enhance the feature clarity. In summary, the main contributions of our work are:

- We propose Spectroformer, a Multi-Domain Query Cascaded Transformer network for underwater image enhancement.
- We propose a Multi-Domain Query Cascaded Attention mechanism that integrates localized transmission features and global illumination features.
- A Spatio-Spectro Fusion-Based Attention Block is proposed to transmit attention-enhanced features from the encoder to the corresponding decoder, effectively boosting performance and feature enhancement.
- A Hybrid Fourier-Spatial Upsampling Block is introduced that uniquely combines Fourier and spatial upsampling techniques to effectively enhance feature resolution.

The ablation study is done on different configurations of the proposed approach. The effectiveness of the proposed method has been verified through various experiments conducted on both synthetic and real-world images for underwater image enhancement. Also, the applicability of the proposed method is verified for depth-estimation tasks.

2. Related Work

2.1. Underwater Image Enhancement

Underwater Image Enhancement (UIE) is an indispensable pre-processing step for high-level computer vision tasks such as object detection, recognition, and tracking. The existing UIE methods can be broadly categorized into four groups: hardware-dependent, physical model-dependent, non-physical model-dependent, and deep learning-dependent methods.

Hardware-dependent Methods: Prior underwater image enhancement efforts have utilized techniques like specialized hardware, stereo vision, and polarization filters [44, 47]. However, these methods have drawbacks: hardware-based ones are costly and complex, polarizers have moving parts causing image acquisition issues, and underwater conditions challenge stereo approaches. Methods relying on multiple images are unsuitable for real-time use [10]. In contrast, single-image enhancement stands out for challenging underwater scenes.

Physical Model-dependent Methods: Several studies have concentrated on enhancing underwater images using the image formation model. Yang *et al.* [53] introduced a modified dark channel prior algorithm, while Chiang *et al.* [7] combined it with a wavelength-dependent compensation method. Another approach, the Underwater Dark Channel Prior (UDCP) [10], addressed red channel unreliability. Liu and Chau [33] minimized costs to enhance contrast based on the dark channel, and Peng *et al.* [39] improved underwater images using light absorption insights. Additionally, Peng *et al.* [38] proposed a Generalized Dark Channel Prior (GDCP) incorporating adaptive color correction for image restoration.

Non-Physical Model-dependent Methods: These methods aim to enhance visual quality by adjusting the pixel values of an image. Iqbal *et al.* [19] expanded the pixel range in RGB and HSV color spaces to enhance contrast and saturation in underwater images. Ancuti *et al.* [3] introduced an enhancement technique blending contrast-enhanced and color-corrected images using a multi-scale fusion approach. Ghani and Isa [15], [14] refined the approach of Iqbal *et al.* [19] by shaping the stretching process following the Rayleigh distribution to mitigate over- and under-enhancement. Fu *et al.* [13] proposed a retinex-based method for underwater image enhancement involving color correction, layer decomposition, and enhancement.

Deep Learning-dependent Methods: The rapid progress in deep learning has significantly accelerated the development and performance of computer vision tasks. Li *et al.* [25] proposed UWCNN, an end-to-end deep network designed to tackle the underwater image enhancement problem across various underwater images. In [48], Pritish *et al.* improved underwater images by utilizing adversarial learning of their content features. In a recent development, Li *et al.* [26] introduced WaterNet, a gated fusion network that employs gamma-corrected, contrast-enhanced, and white-balanced images as inputs to enhance underwater images. Jiang *et al.* [20] introduced a target-oriented perceptual adversarial network featuring an adaptive fusion of latent features to counter the degradation of underwater images. Li *et al.* [30] introduced a WaterGAN that generates underwater-style images from images taken above water and depth maps through an unsupervised process to mitigate the requirement for paired underwater training data. The resulting dataset is then utilized to train the WaterGAN. Yang *et al.* [54] introduced a conditional generative adversarial network (cGAN) to enhance the visual quality of underwater images.

2.2. Transformers in Computer Vision Applications

Due to the Transformer's capacity to capture global contexts and its notable advancements in various high-level vision tasks such as image classification, semantic segmentation, and object detection, it has been extended to address image restoration tasks. Zamir *et al.* introduced an efficient transformer network, as outlined in [55], suitable for restoration tasks, including image deraining, denoising, and deblurring. Peng *et al.* [36] introduced a U-shaped transformer for enhancing underwater images, incorporating channel-wise and spatial-wise feature fusion modules within the network. In contrast to existing approaches, [23] introduced an efficient Transformer-based method for high-quality image deblurring that leverages frequency-domain characteristics to simplify scaled dot-product attention, avoiding complex matrix multiplication.

3. Proposed Method

Our main goal is to combine the insights from both frequency and spatial domains for revealing fine details [22,50] and patterns in the degraded underwater images. To alleviate the color distortion and contrast decline, we incorporate several key designs in our proposed network. We first present the holistic pipeline of Spectroformer as depicted in Figure 2. Thereafter, we provide a detailed overview of the proposed components: Multi-Domain Query Cascaded Transformer, Spatio-Spectro Fusion-Based Attention Block, and Hybrid Fourier-Spatial Upsampling.

Overall Pipeline: Given a degraded image **(I)**, Spectroformer perform first applies a convolution, resulting in shallow features denoted as $\mathbf{F}_o$ shown in Figure 2. Next, these shallow features are processed through a series of Multi-Domain Query Cascaded Transformer Blocks (MQCT), each incorporating the innovative Multi-Domain Query Cascaded Attention mechanism. The features obtained from the initial MQCT stage are further refined using the proposed Spatio-Spectro Fusion-Based Attention Block, which is strategically integrated into skip connections. On the decoder side, we employ a Hybrid Fourier-Spatial Upsampling Block to effectively enhance feature resolution. Finally, a convolution layer is applied to the resulting deep features, labeled as $\mathbf{F}_d$, to obtain the final output. This entire process culminates in the generation of an enhanced output image **(O)**.

3.1. Multi-Domain Query Cascaded Transformer

Transformers are adept in modelling the global contexts by computing the scaled dot product attention between queries and keys. However, we observe that as the degraded underwater images usually contain blur, color, and contrast distortions, evaluating the scaled dot-product attention only in the spatial domain does not effectively exploit the global contents, resulting in unwanted artifacts. In light of this, we propose a novel Multi-Domain Query Cascaded Transformer Block (see Figure 2) where the queries are processed in the frequency-domain and keys in the spatial domain to generate a detailed and informative attention map.

Within the transformer block, the process initiates with the normalized tensor $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$, which is directed to the proposed Multi-Domain Query Cascaded Attention mechanism (MQCA) as depicted in Figure 2. In the MQCA mechanism, the generation of the final attentive feature map occurs through two stages. In the first stage, key ($\mathbf{K}_1$), query ($\mathbf{Q}_1$), and value ($\mathbf{V}_1$) are derived by applying 1×1 convolutions followed by 3×3 depth-wise convolutions. In a similar fashion, for the second stage, the key ($\mathbf{K}_2$) and value ($\mathbf{V}_2$) are obtained from the attentive feature of the first stage. However, the query ($\mathbf{Q}_2$) is a frequency-domain processed query ($\mathbf{Q}_2$) generated through the frequency-domain Feature Processor (FDFP) as shown in Figure 2. To generate the attentive feature at each stage, we follow the approach introduced in the Restormer model [56].

$$\mathbf{Q}_1 = \Phi_3(\psi_1(\mathbf{X}));\ \mathbf{K}_1 = \Phi_3(\psi_1(\mathbf{X}));\ \mathbf{V}_1 = \Phi_3(\psi_1(\mathbf{X})) \tag{1}$$

$$\mathbf{X'} = \psi_1\left(\text{Attention}\left(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1\right)\right) \tag{2}$$

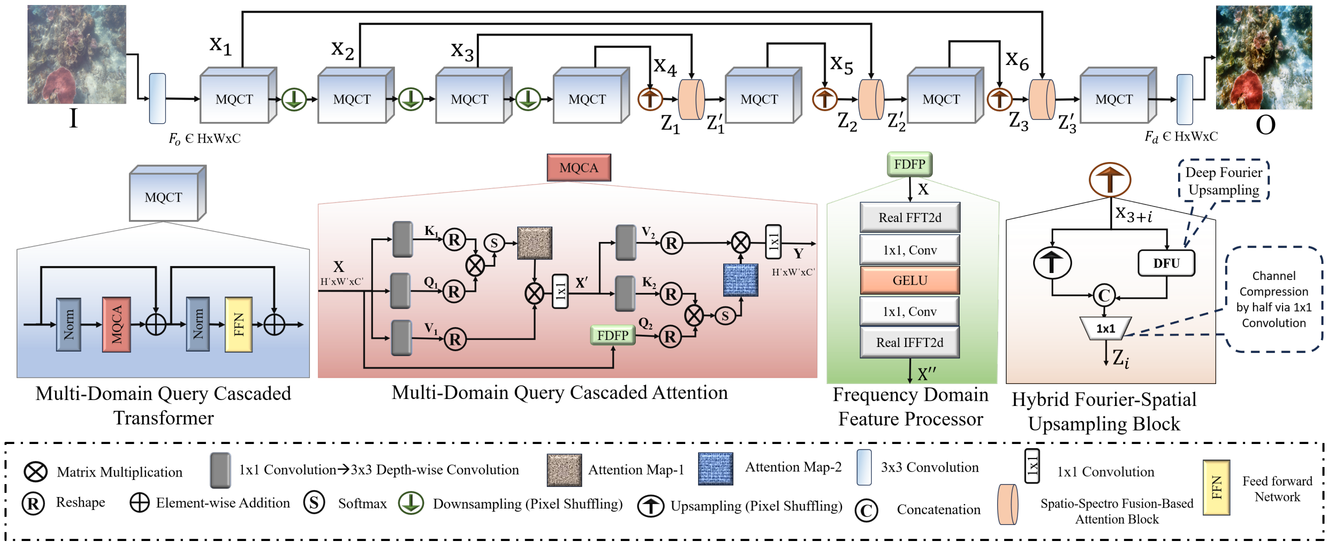

Figure 2. Overview of the proposed network (**Spectroformer**) for underwater image enhancement. The network consists of **Multi-Domain Query Cascaded Transformer, Spatio-Spectro Fusion-Based Attention Block, and Hybrid Fourier-Spatial Upsampling Block**. Multi-Domain Query Cascaded Transformer is proposed to tackle issues with color distortion and contrast reduction and seamlessly combines spatial and frequency-domain information. Spatio-Spectro Fusion-Based Attention Block is proposed to transmit attention-enhanced features from the encoder to the corresponding decoder. The Hybrid Fourier-Spatial Upsampling Block is proposed to uniquely combine Fourier and spatial upsampling techniques to effectively enhance feature resolution.

$$\mathbf{Q}_2 = FDFP(\mathbf{X})); \ \mathbf{K}_2 = \Phi_3(\psi_1(\mathbf{X'})); \ \mathbf{V}_2 = \Phi_3(\psi_1(\mathbf{X'})) \quad (3)$$

$$\mathbf{Y} = \psi_1\left(\text{Attention}\left(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2\right)\right) \quad (4)$$

$$\text{Attention}\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right) = \mathbf{V}_i \cdot \text{Softmax}\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i}{\alpha}\right) \quad (5)$$

Here, $\mathbf{X}$ and $\mathbf{Y}$ represent the input and output feature maps of the MQCA. $\Phi_m(\cdot)$ denotes a depth-wise convolution operator with a kernel size of $(m \times m)$ for channel-wise spatial context, and $\psi_m(\cdot)$ denotes a convolution operator with a kernel size of $(m \times m)$ to capture pixel-wise cross-channel context, where m can take values from the set $\{1, 2, 3\}$. Notably, the convolution layers within the network do not have biases. Matrices $\mathbf{Q}_{1,2} \in \mathbb{R}^{H'W' \times C'}$, $\mathbf{K}_{1,2} \in \mathbb{R}^{C' \times H'W'}$, and $\mathbf{V}_{1,2} \in \mathbb{R}^{H'W' \times C'}$ are acquired after reshaping tensors from their original dimensions $\mathbb{R}^{H' \times W' \times C'}$. The parameter α can be learned to modulate the dot product of $\mathbf{Q}_{1,2}$ and $\mathbf{V}_{1,2}$ before the application of the softmax function. This scaling factor enables control over the magnitude of the dot product, influencing the attention strength.

3.2. Spatio-Spectro Fusion-Based Attention Block

Typically, skip connections are employed to facilitate the reconstruction process by transferring encoder features to the corresponding decoder features [43]. However, the direct propagation of these features can sometimes lead to the transmission of redundant information. By merging insights from both the frequency and spatial domains, where

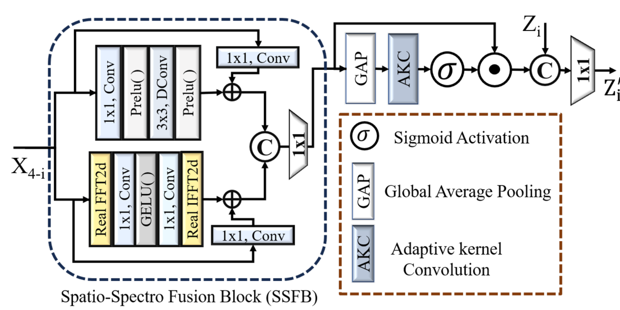

Figure 3. Overview of the proposed Spatio-Spectro Fusion-Based Attention Block. The encoder feature ($\mathbf{X}_{4-i}$) is first passed through the block to generate an attentive feature that captures the relevant information. It is then concatenated with the corresponding decoder feature ($\mathbf{Z}_i$). Lastly, a 1×1 convolution is applied to compress the channel dimension by half, which helps to refine and consolidate the combined information before further processing.

the former delves into revealing fine details, and the latter focuses on the interpretation of the pixel values, underwater image enhancement can benefit from the exploitation of the non-redundant features. Hence, to address the shortcomings of direct skip connections, we introduce the concept of "Spatio-Spectro Fusion-Based Attention Block" as vividly depicted in Figure 3. This novel block bridges the gap between the encoder and decoder by transmitting attentive features enhanced with spatio-spectral fusion mechanisms. It serves as an alternative to the traditional direct connections, contributing to improved performance by generating more enhanced features, $\mathbf{Z'}_i$ as:

$$\mathbf{Z'}_i = \psi_1(\langle \Omega(\mathbf{X}_{4-i}) \otimes \sigma(\omega_m(GAP(\Omega(\mathbf{X}_{4-i})))), \mathbf{Z}_i \rangle) \tag{6}$$

where, $\mathbf{X}_i$ are the input features of dimension $\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}} \times 2^{i-1}C$, $i \in (1, 2, 3)$, $\psi_1(\cdot)$ denotes a convolution operator with a kernel size of 1 × 1, $< \cdot >$ represents a concatenation operator, and Ω represents the function of spatio-spectro fusion block (see SSFB in Figure 3). Here, ω_m is a 1D convolution operator with adaptive kernel size (see AKC in Figure 3). The proposed SSFB block concurrently processes the spatial and spectral information for each encoder layer. To do this, the input features $\mathbf{X}_i$ are processed as:

$$\psi_1\left(\left\langle \begin{array}{c} \psi_1^P\left(DC_3^P(X_i)\right) + \psi_1(X_i) \\ IFFT\left(\psi_1^G(\psi_1(FFT(X_i)) + \psi_1(X_i)\right) \end{array} \right\rangle\right) \tag{7}$$

where, ψ_1^P and DC_3^P are 1×1 convolution and 3×3 depthwise separable convolution $\rightarrow$ PReLu activation, respectively. ψ_1^G is 1×1 convolution $\rightarrow$ GeLu activation (see SSFB Figure 3).

Further, in traditional CNNs, the kernel size is fixed and does not change during the training process. This means that some features may be over-smoothed (due to large kernel size) or under-smoothed (due to small kernel size) by the fixed kernel size [2], resulting in loss of important information and hence reduced performance. To circumvent this issue, SSFB attention block adaptively selects the kernel size based on the number of input feature channels. It does this by applying a learnable 1D convolution layer to the encoder features, which is then used to weigh the features at each channel. This allows the network to learn which kernel size is best suited to capture the features in each channel of the input. The adaptive kernel size k is determined by:

$$k = \alpha\left(C'\right) = \left| \frac{log_2(C')}{b} + \frac{a}{b} \right|_{odd} \tag{8}$$

where, $C' = 2^{i-1}C$ is the number of channels after GAP, $|x|_{\text{odd}}$ indicates the nearest odd number of x. In this work, we set a and b to 1 and 2, respectively.

3.3. Hybrid Fourier-Spatial Upsampling

The essence of upsampling is to retrieve the high-frequency channel information in the image. The existing popular upsampling operations ($e.g.$, transposed convolutions, un-pooling, interpolation) typically operate in the spatial domain and the current works [6,34] seldom exploits the potency of up-sampling in the frequency-domain. Since these spatial upsamplers are highly reliant on local pixel interactions [57], they may be unsuitable for exploring global dependency for the task of UIE. Nevertheless, frequency-domain features may help in the reconstruction of missing global details in the degraded image, and can substantially improve the reconstruction performance. Taking this into consideration, we design a "Hybrid Fourier-Spatial Upsampling Block" as shown in Figure 2 that intelligently combines Fourier (Deep Fourier Upsampling) and spatial up-sampling (Pixel-shuffle) techniques to significantly enhance the feature clarity.

3.4. Training Losses

To train our proposed architecture, we have incorporated the following losses as depicted in the equation below:

$$L_T = \lambda_1 L_C + \lambda_2 L_G + \lambda_3 L_M + \lambda_4 L_P \tag{9}$$

where, $\lambda_{1,2,3,4} \in \{0.03, 0.02, 0.01, 0.025\}$ weighting factors. The training involved a total loss function L_T comprising, Charbonnier loss (L_C) [5], Gradient loss (L_G) [42], Multiscale Structural Similarity Index (MS-SSIM) loss (L_M) [51], and Perceptual loss (L_P) [21]. This combined loss function effectively optimized our model, capturing diverse image attributes and producing high-quality output images. *More details about loss functions are given in the supplementary material.*

4. Experimental Discussion

This section covers datasets, training specifics, comparative analysis, and an ablation study of the proposed network.

4.1. Datasets

To conduct a comparative analysis, we have considered synthetic Underwater Image Enhancement Benchmark (UIEB) [26] and real-world underwater U45 [29], UCCS [32], SQUID [4] datasets. The training set is composed of randomly selected 800 image pairs, while the remaining 90 images are considered for testing purposes. U45 comprises 45 real-world images that showcase characteristics such as color casts, low contrast, and the degradation effects resembling haze in underwater scenarios. The UCCS dataset [32] comprises 300 genuine underwater images, providing a diverse range of marine organisms and environments for analysis. The SQUID dataset comprises 57 sets of stereo pairs captured at various locations within Israel.

4.2. Training Details

For the generation of images in our training set, we employed data augmentation methods including horizontal and vertical flipping, noise addition, and contrast variation. Specifically, we used 4800 image pairs from the UIEB dataset for training. Testing was performed using 90 images from UIEB. All input images were resized to dimensions of 256×256 pixels for consistency. During training, we utilized the ADAM optimizer with an initial learning rate of 3×10^{-4}, adjusting it via the cosine annealing strategy. Our network was implemented using PyTorch and trained on an NVIDIA GeForce RTX 2080 GPU.

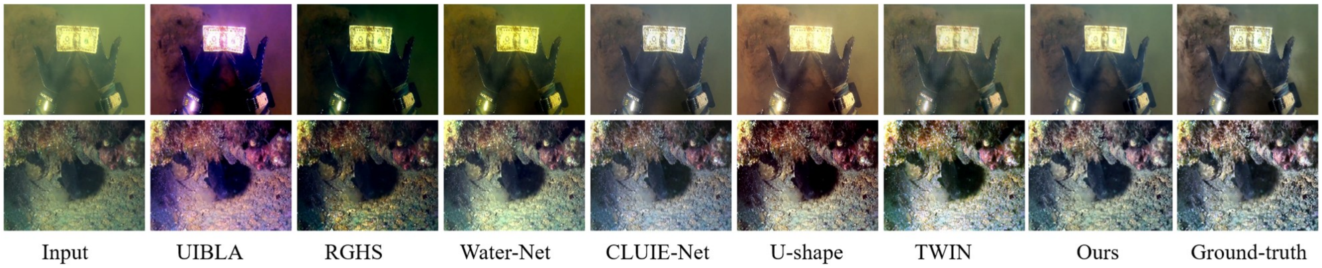

Figure 4. Qualitative comparison of the proposed method (Ours) with existing state-of-the-art methods (UIBLA [39], RGHS [18], Water-Net [26], CLUIE-Net [31], U-shape [37], TWIN [33]) for underwater image enhancement on UIEB dataset.

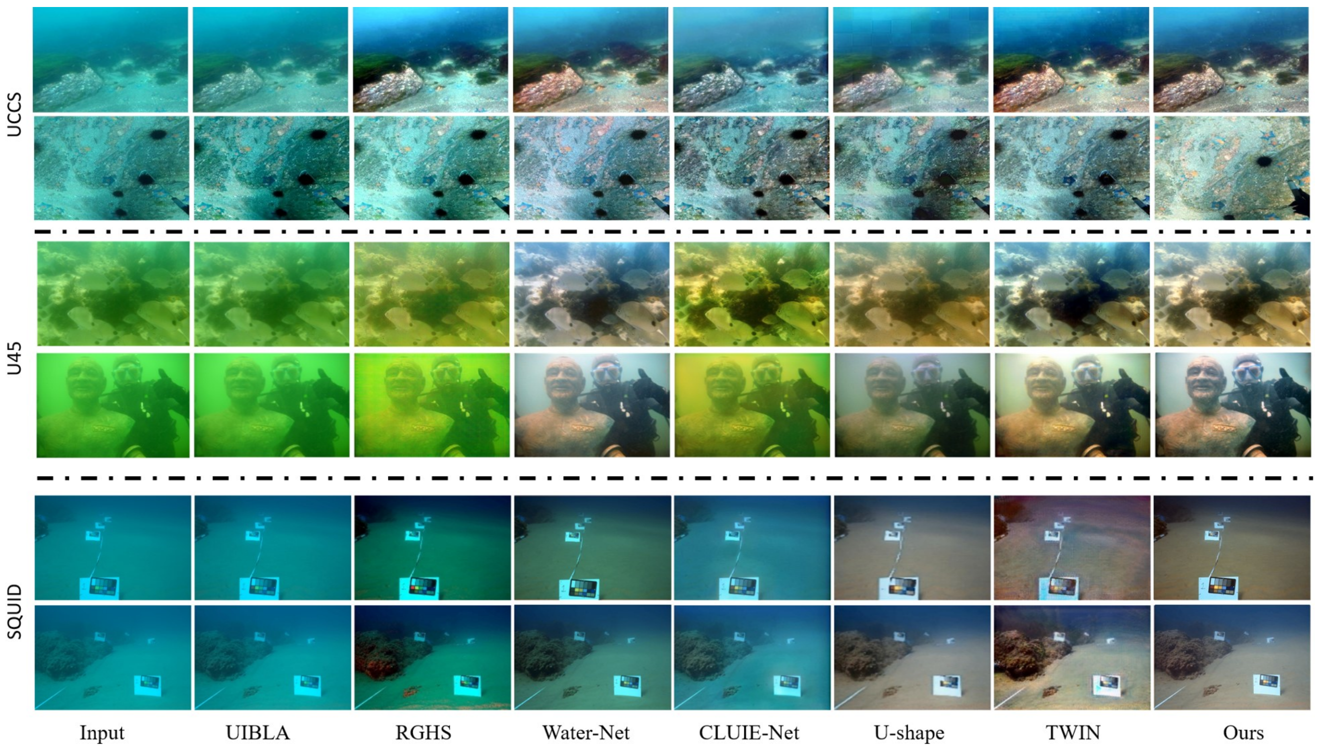

Figure 5. Qualitative comparison of the proposed method (Ours) with existing state-of-the-art methods (UIBLA [39], RGHS [18], Water-Net [26], CLUIE-Net [31], U-shape [37], TWIN [33]) for underwater image enhancement on real-world UCCS, U45, and SQUID datasets.

4.3. Analysis on Synthetic Datasets

The proposed method is quantitatively compared against existing state-of-the-art techniques, using metrics such as PSNR, SSIM, and UIQM for evaluation. Quantitative results for the most widely used UIEB dataset are in Table 1. Qualitative results for UIEB are shown in Figure 4 The proposed method demonstrates competitive performance compared to the state-of-the-art methods.

4.4. Analysis on Real-world Dataset

To assess the effectiveness of our proposed approach in real-world scenarios, we present results derived from the U45 dataset. Our quantitative analysis covers various metrics, including UIQM (Underwater Image Quality Measure), UISM (Underwater Image Sharpness Measure), NIQE (Naturalness Image Quality Evaluator), and BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator). Summarized results are available in Table 3. Furthermore, we provide qualitative insights into the U45, UCCS, and SQUID datasets via Figure 5. These findings underscore the significant enhancement in color balance and visibility within the enhanced images, attributed to the innovative modules introduced in our proposed method. *Additional qualitative outcomes are provided in the supplementary material.*

Table 1. Quantitative comparison of the proposed method (Ours) and existing state-of-the-art methods on the UIEB dataset for underwater image enhancement (↑: higher is better, **bold** and underline indicate **best** and second best values respectively).

Method	PSNR ↑	SSIM ↑	UIQM ↑
UDCP [9]	13.81	0.692	1.825
UIBLA [39]	15.78	0.731	2.014
RGHS [18]	14.57	0.791	2.410
WaterNet [26]	19.81	0.864	2.818
CLUIE-Net [31]	20.37	0.890	2.674
U-shape [37]	22.91	0.910	2.725
TWIN [33]	23.72	0.830	3.024
Ours	**24.96**	**0.917**	**3.075**

Table 2. Quantitative comparison of the proposed method and existing state-of-the-art methods on the real-world U45 dataset for underwater image enhancement (↑ - higher is better, ↓ - lower is better).

Method	UIQM ↑	UISM ↑	NIQE ↓	BRISQUE ↓
UIBLA [39]	1.710	4.012	4.2263	20.6737
RGHS [18]	2.506	5.558	3.8727	**18.5190**
WaterNet [26]	3.091	6.187	4.5966	21.1563
CLUIE-Net [31]	2.890	5.988	3.8743	20.6126
U-shape [37]	2.923	5.567	4.3098	21.5656
TWIN [33]	3.135	6.698	3.9929	20.0891
Ours	**3.243**	**7.354**	**3.8420**	19.9573

Table 3. Quantitative comparison of the proposed method and existing state-of-the-art methods on the real-world UCCS dataset [29] for underwater image enhancement (↑ - higher is better, ↓ - lower is better).

Method	UIQM ↑	UISM ↑	NIQE ↓	BRISQUE ↓
UIBLA [39]	2.555	5.939	**3.927**	25.455
RGHS [18]	2.506	5.558	4.209	26.360
Water-Net [26]	3.134	6.187	6.104	24.275
CLUIE-Net [31]	3.066	6.715	4.420	29.524
U-shape [37]	2.874	5.391	4.401	23.549
TWIN [33]	3.119	**6.732**	4.370	25.755
Ours	**3.209**	6.563	3.982	**23.258**

5. Ablation Study

To demonstrate the efficacy of the proposed components, we undertake the subsequent ablation studies on the UIEB dataset [26].

5.1. Effectiveness of the Multi-Domain Query Cascaded Attention in Transformer

Our Multi-Domain Query Cascaded Transformer Network," guided by the innovative "Multi-Domain Query Cascaded Attention" mechanism, adeptly merges information from spatial and frequency-domains, leading to substantial enhancements in underwater image quality. To substantiate this claim, we conducted experiments with and without the Multi-Domain Query Cascaded Attention mechanism within the transformer. Quantitative validation from Table 4 and qualitative validation in Figure 6 reinforces our assertion that the proposed Multi-Domain Query Cascaded Attention mechanism effectively addresses challenges related to color distortion and contrast reduction, resulting in improved quality of underwater images.

Table 4. Quantitative results comparison of various network settings and losses optimization. *Note: B- Baseline, C- Multi-Domain Query Cascaded Attention, D- Spatio-Spectro Fusion Based Attention, E- Hybrid Fourier-Spatial Upsampling.*

Network Setting	PSNR	SSIM
B	22.51	0.862
B+C	24.24	0.891
B+C+D	24.46	0.901
Ours (B+C+D+E)	**24.96**	**0.917**

5.2. Effectiveness of the Spatio-Spectro Fusion-Based Attention Block in feature propagation

The Spatio-Spectro Fusion-Based Attention Block facilitates the transmission of attention-enhanced features from the encoder to the corresponding decoder, thereby enhancing performance and feature augmentation. To assess this, we conducted experiments both with and without the Spatio-Spectro Fusion-Based Attention Block in the proposed network. Observing the results presented in Table 4 and Figure 6, we can validate that the inclusion of the Spatio-Spectro Fusion-Based Attention Block leads to superior performance.

5.3. Effectiveness of the Hybrid Fourier-Spatial Upsampling

Pixel shuffling enhances spatial resolution for clearer visuals and details [46]. Frequency-domain upsampling improves overall image quality by extracting fine features across frequencies [57]. However, when used alone, they may miss subtle fluctuations. Our "Hybrid Fourier-Spatial Upsampling Block" combines both methods. Table 4 and Figure 6 demonstrate that this hybrid approach results in quality improvement. *More ablation studies are given in the supplementary material.*

6. Application of the Proposed Method for Depth-Estimation

We have seamlessly incorporated our approach with the method proposed by [40], positioning it as a pre-processing

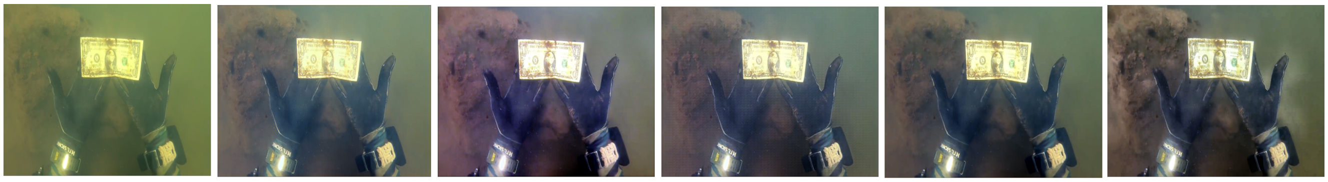

Figure 6. Qualitative results comparison of various network settings and losses optimization. *Note: I-Degraded, B- Baseline, C- Multi-Domain Query Cascaded Attention, D- Spatio-Spectro Fusion Based Attention, E- Hybrid Fourier-Spatial Upsampling.*

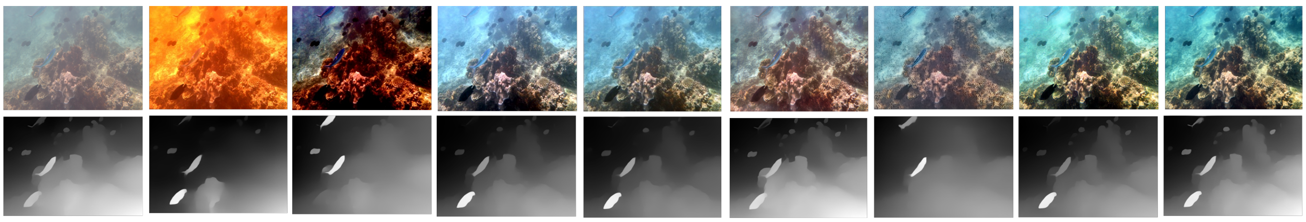

Figure 7. Applicability of the proposed and the existing underwater image restoration approaches (UDCP [9], UIBLA [39], RGHS [18], Water-Net [26], CLUIE-Net [31], U-shape [37], TWIN [33]) for depth-estimation task (top row: degraded input and restored output by respective methods; bottom row: the corresponding depth-map).

step to augment the accuracy of depth estimation. This integration has resulted in significant enhancements in precision, as illustrated in Figure 7. This adaptation to intricate challenges in advanced computer vision validates the versatility of our approach and its capacity to elevate different aspects of the field. The amalgamation of restoration and depth estimation effectively corroborates the potential of our approach to driving advancements in computational visual analysis.

7. Conclusion

In this paper, we proposed an underwater image enhancement model, Spectroformer. The network encompasses multiple components, including the Multi-Domain Query Cascaded Transformer that integrates localized transmission and global illumination features. Additionally, a Spatio-Spectro Fusion-Based Attention Block is proposed to transmit attention-enhanced features from the encoder to the decoder. Moreover, a Hybrid Fourier-Spatial Upsampling Block is introduced, combining Fourier and spatial upsampling techniques to enhance feature resolution effectively. Extensive analysis is conducted on both synthetic and real-world datasets, supplemented by comprehensive ablation studies, to validate the efficacy of the proposed method for underwater image enhancement. Furthermore, the versatility of the proposed approach is demonstrated through its applicability to other widely used application, depth estimation.

Acknowledgement

This work was made possible through the generous support of the projects MoES/PAMC/DOM/04/2022 (E-12710) and TIHIITG202204. The authors would also like to thank all the members of CVPR Lab for their support on this work.

References

[1] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through composition of dual-intensity images and rayleigh-stretching. *SpringerPlus*, 3(1):1–14, 2014. 1

[2] Abhinav Agrawal and Namita Mittal. Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *The Visual Computer*, 36(2):405–412, 2020. 5

[3] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–88, 2012. 2

[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. 5

[5] 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(3):211–231, 2005. 5

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution

transformer. arxiv 2022. *arXiv preprint arXiv:2205.04437*, 1, 2022. 5

[7] 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

[8] 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

[9] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016. 7, 8

[10] Paulo L.J. Drews, Erickson R. Nascimento, Silvia S.C. Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE Computer Graphics and Applications*, 36(2):24–35, 2016. 2

[11] Akshay Dudhane, Praful Hambarde, Prashant Patil, and Subrahmanyam Murala. Deep underwater image restoration and beyond. *IEEE Signal Processing Letters*, 27:675–679, 2020. 1

[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. 1

[13] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding. A retinex-based enhancing approach for single underwater image. In *2014 IEEE international conference on image processing (ICIP)*, pages 4572–4576. IEEE, 2014. 3

[14] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Enhancement of low quality underwater image through integrated global and local contrast correction. *Applied Soft Computing*, 37:332–344, 2015. 3

[15] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through integrated color model with rayleigh distribution. *Applied soft computing*, 27:219–230, 2015. 3

[16] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1

[17] Jon Henderson, Oscar Pizarro, Matthew Johnson-Roberson, and Ian Mahon. Mapping submerged archaeological sites using stereo-vision photogrammetry. *International Journal of Nautical Archaeology*, 42(2):243–256, 2013. 1

[18] Dongmei Huang, Yan Wang, Wei Song, Jean Sequeira, and Sébastien Mavromatis. 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*, pages 453–465. Springer, 2018. 6, 7, 8

[19] Kashif Iqbal, Michael O. Odetayo, Anne E. James, Rosalina Abdul Salam, and Abdullah Zawawi Talib. Enhancing the low quality images using unsupervised colour correction method. *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 1703–1709, 2010. 2, 3

[20] Zhiying Jiang, Zhuoxiao Li, Shuzhou Yang, Xin Fan, and Risheng Liu. Target oriented perceptual adversarial fusion network for underwater image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6584–6598, 2022. 3

[21] 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

[22] MD Raqib Khan, Ashutosh Kulkarni, Shruti S Phutke, and Subrahmanyam Murala. Underwater image enhancement with phase transfer and attention. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023. 3

[23] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 3

[24] 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. 1

[25] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020. 3

[26] 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. 3, 5, 6, 7, 8

[27] Chongyi Li, Jichang Guo, and Chunle Guo. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal processing letters*, 25(3):323–327, 2018. 1

[28] 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

[29] 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. 5, 7

[30] 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. 1, 3

[31] Kunqian Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. Beyond single reference for train-

ing: underwater image enhancement via comparative learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 6, 7, 8

[32] 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. 1, 5

[33] 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. 2, 6, 7, 8

[34] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Gated multi-resolution transfer network for burst restoration and enhancement. *arXiv preprint arXiv:2304.06703*, 2023. 5

[35] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *arXiv preprint arXiv:2111.11843*, 2021. 1

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

[37] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 290–307. Springer, 2023. 6, 7, 8

[38] 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. 2

[39] 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. 2, 6, 7, 8

[40] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 7

[41] Tingdi Ren, Haiyong Xu, Gangyi Jiang, Mei Yu, and Ting Luo. Reinforced swin-convs transformer for underwater image enhancement. *arXiv preprint arXiv:2205.00434*, 2022. 1

[42] 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

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[44] Yoav Y Schechner and Nir Karpel. Clear underwater vision. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[45] Raimondo Schettini and Silvia Corchs. Underwater image processing: state of the art of restoration and image enhancement methods. *EURASIP journal on advances in signal processing*, 2010:1–14, 2010. 1

[46] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2, 7

[47] Tali Treibitz and Yoav Y Schechner. Active polarization descattering. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):385–399, 2008. 2

[48] 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. 3

[49] 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. 1

[50] Danxu Wang and Zhonglin Sun. Frequency domain based learning with transformer for underwater image restoration. In *Pacific Rim International Conference on Artificial Intelligence*, pages 218–232. Springer, 2022. 2, 3

[51] 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

[52] David P Williams. On optimal auv track-spacing for underwater mine detection. In *2010 IEEE International Conference on Robotics and Automation*, pages 4755–4762. IEEE, 2010. 1

[53] Hung-Yu Yang, Pei-Yin Chen, Chien-Chuan Huang, Ya-Zhu Zhuang, and Yeu-Horng Shiau. Low complexity underwater image enhancement based on dark channel prior. In *2011 Second International Conference on Innovations in Bio-inspired Computing and Applications*, pages 17–20. IEEE, 2011. 2

[54] Miao Yang, Ke Hu, Yixiang Du, Zhiqiang Wei, Zhibin Sheng, and Jintong Hu. Underwater image enhancement based on conditional generative adversarial network. *Signal Processing: Image Communication*, 81:115723, 2020. 3

[55] 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 *CVPR*, 2022. 3

[56] 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. 3

[57] Man Zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. *arXiv preprint arXiv:2210.05171*, 2022. 2, 5, 7