WaterFormer

A Global–Local Transformer for Underwater Image Enhancement With Environment Adaptor

By Junjie Wen, Jinqiang Cui, Guidong Yang, Benyun Zhao, Yu Zhai, Zhi Gao, Lihua Dou, and Ben M. Chen

Underwater image enhancement (UIE) is crucial for high-level vision in underwater robotics. While convolutional neural networks (CNNs) have made significant achievements in UIE, the locality of convolution poses a challenge in capturing the global context. In contrast, transformer-based networks, adept at handling long-range dependencies, have shown promise in various vision tasks. Nonetheless, directly applying a transformer to UIE faces critical challenges: 1) it tends to produce results with coarse details due to the negligence of local texture and 2) the varicolored degraded images require the network to be adaptable to different underwater environments. In this article, we propose a novel transformer-based network that can effectively leverage both the global contextual and local detailed information with some key designs (a global–local transformer [GL-Trans] block and a detail-enhanced skip connector [DESC]) while being computationally efficient. Moreover, by introducing a simple but effective learnable environment adaptor, the proposed network is flexible to deal with different underwater environments. Extensive experiments have been conducted and have demonstrated the superiority of our proposed network compared with other state-of-the-art (SOTA) methods both qualitatively and quantitatively.

©SHUTTERSTOCK.COM/FRANZISKA REINHARDT

INTRODUCTION

Benefiting from the development of deep learning, UIE, which is vital for autonomous underwater vehicles (AUVs) to acquire clear underwater images and perform downstream vision-related tasks, has gained impressive success in recent years.

Existing learning-based UIE methods [1], [2] mainly rely on CNNs. The fixed geometric structure of a CNN module makes it outstanding at efficiently extracting local representations but also prevents it from extracting long-range dependencies. Moreover, each activation unit within the same CNN layer shares identical receptive fields with restricted regions, which is disadvantageous for high-level layers that encapsulate global semantic information.

Digital Object Identifier 10.1109/MRA.2024.3351487
Date of current version: 26 January 2024

The popularity of transformers [4] in the natural language processing (NLP) field has led to the success of transformer-based networks in a variety of discriminative vision tasks [5], [6]. Unlike the locality of convolution operations, the self-attention (SA) mechanism, which is a core component of a transformer-based network, helps the network better extract the global context of input images by calculating the feature responses with a weighted sum of all other features.

However, directly applying a transformer to generative vision tasks like UIE is still challenging. As has been observed before [7], [8], transformer-based networks tend to generate coarse images with unclear details. It might be caused by the following reasons: 1) traditional vision transformers divide the input content into a limited number of patches to draw global dependencies, which inevitably lose detailed local information, or 2) the recently proposed window-based transformers [6], [9] apply SA locally on windows of small spatial sizes to alleviate the quadratic computational cost but might introduce blocking artifacts due to its window-based SA manner [7]. Previous researchers have either replaced the network decoders [8] with CNNs or introduced a generative adversarial network (GAN) loss [7], [10] to help alleviate the problem.

Moreover, the property of degraded underwater images also poses special demands for UIE methods. Owing to the absorption of light that varies with wavelength, as well as backscattering, degraded underwater images could exhibit varicolored effects in different underwater environments, requiring the UIE methods to be adaptive to images in various conditions. Previous works have either applied complex domain adaptation (DA) methods [11], [12] or explicitly classified different types of water [3] to help the network adapt to different underwater environments.

In this article, we propose a novel transformer-based network, dubbed WaterFormer, for UIE. With the well-designed GL-Trans block and DESC as our key components, the network is effective in balancing the global semantic and local detailed features while being computationally efficient. Moreover, we also present a simple but effective environment adaptor to make the network flexible to various underwater environments.

Our main contributions are as follows:

- A novel transformer-based UIE network that effectively integrates both global and local features while maintaining computational efficiency has been proposed.
- A simple but effective environment adaptor is designed to enable the network's adaptation to various underwater environments.
- Extensive experiments demonstrate that our method outperforms other SOTA methods both qualitatively and quantitatively.

“WITH THE WELL-DESIGNED GL-TRANS BLOCK AND DESC AS OUR KEY COMPONENTS, THE NETWORK IS EFFECTIVE IN BALANCING THE GLOBAL SEMANTIC AND LOCAL DETAILED FEATURES WHILE BEING COMPUTATIONALLY EFFICIENT.”

RELATED WORK

UIE

UIE techniques could be broadly categorized into two groups: traditional and learning-based approaches. The majority of traditional strategies aim to produce clear images by estimating the direct transmission and backscattering with certain prior assumptions [13], [14], [15]. Nevertheless, these strategies might be ineffective in scenarios where the underlying assumptions are inapplicable.

Learning-based UIE methods, in contrast, are mostly developed with CNNs. Due to the difficulty of acquiring reference images, early researchers combined underwater image generation and enhancement with the guidance of the underwater image formation model [16]. For example, Li et al. [17] introduced WaterGAN for synthesizing underwater images using in-air images and depth data pairs and employed a two-stage CNN-based network for the task of monocular underwater image color correction. Li et al. [3] trained multiple UWCNN models with synthetic underwater images based on different Jerlov water types [18] for the enhancement of underwater images. As these methods only used simple CNNs, other works applied special designs to make the network more suitable for UIE. Fu and Cao [2] adopted a two-branch global–local network to compensate for the global color distortion and local contrast reduction and designed a compressed histogram equalization with fixed parameters to further enhance the quality. Wen et al. [11] designed a CNN-based network with wavelet transform and a complex DA method to make the network better adaptable to various underwater environments.

However, the locality nature of convolution limits the CNN-based network's ability to effectively extract the global context of the input content, which may result in suboptimal performance. In this article, we take advantage of the long-range understanding capability of the SA mechanism and design an efficient and effective transformer-based network to boost the enhancement performance of various underwater images.

VISION TRANSFORMERS

A transformer [4] is a network solely based on SA mechanisms, which was initially proposed for machine translation tasks. Since then, transformer-based networks [19], [20] have achieved SOTA performances on various NLP tasks. The SA mechanism employed by the transformer greatly improves the network's ability to capture long-range dependencies across input features and helps transformer-based networks adapt to various vision tasks. For instance, Dosovitskiy et al. [5] first proposed a vision transformer on image classification tasks with a pure transformer instead of

CNNs by directly treating the input image as sequences of image patches, achieving excellent results compared with SOTA CNN-based networks. Carion et al. [21] then proposed DETR with a transformer encoder–decoder architecture for object detection and significantly outperformed competitive CNN-based baselines.

Although transformer-based networks have achieved great success in various discriminative vision tasks, their capability in UIE, which requires the network to generate clear images with detailed texture, is still under study and exploration. However, the long-range capturing nature of the SA mechanism makes the local detailed information inevitably ignored by the transformer. Although some researchers limit SA on small size windows [9], [10], [22] to extract local attentions, it could generate blocking artifacts due to the window-based SA mechanism [7].

In addition, degraded underwater images may hold varicolored representations due to the light absorption and backscattering in different environments, requiring the UIE methods to be flexible to consistently enhance images from various environments, which has been ignored by previous transformer-based UIE methods [23], [24], [25], [26], [27], [28], [29], [30]. In this article, we propose a transformer-based UIE network that leverages the benefits of both global and local attention features to enhance images with detailed texture. A simple but effective environment adaptor to assist the network gain a better understanding of the underwater environment is also presented.

> “UNLIKE TRADITIONAL U-SHAPED NETWORKS, THE INPUTS FOR EACH DECODER STAGE ARE A COMBINATION OF UPSTREAM DECODER FEATURES AND FEATURES FROM OUR DESC.”

PROPOSED METHOD

In this section, the overall pipeline of WaterFormer is first introduced in the section “Overall Pipeline.” Then, the section “GL-Trans Block” details the proposed GL-Trans block. The DESC is discussed in the section “DESC.” Finally, the section “Environment Adaptor” describes the environment adaptor.

OVERALL PIPELINE

As shown in Figure 1, the overall design of our proposed WaterFormer is inspired by the U-shaped hierarchical networks [31] that have been widely applied in generative vision tasks. Specifically, the degraded underwater images $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ are first processed by a 3×3 convolutional layer for extracting the low-level feature embeddings $\mathbf{X_0} \in \mathbb{R}^{C\times H\times W}$, where C denotes the channels number and $H\times W$ is the spatial dimension. These features are then transformed through K encoder stages, resulting in deep features $\mathbf{X}_K \in \mathbb{R}^{2^K\times(H/2^K)\times(W/2^K)}$, with each stage comprising GL-Trans blocks that efficiently integrate both global and local information, where K is the number of encoder stages.

Following the K encoder stages, the bottleneck stage with L_b GL-Trans blocks is used to process the deep features $\mathbf{X}_K$. The GL-Trans block in the bottleneck stage differs from those in encoder/decoder stages as it contains an environment adaptor to enhance the network’s adaptability to diverse underwater environments.

The decoder stages mirror the encoder stages to progressively reconstruct clear underwater images. Unlike traditional U-shaped networks, the inputs for each decoder stage are a combination of upstream decoder features and features from our DESC. We observe that directly merging encoder and decoder features leads to a performance drop with coarse detail, possibly due to the patch-based SA mechanism’s limitations in detail extraction. Finally, a 1×1 convolutional layer generates a residual map $\mathbf{I}' \in \mathbb{R}^{3\times H\times W}$, resulting in the enhanced image $\tilde{\mathbf{I}} = \mathbf{I} + \mathbf{I}'$.

During training, our network’s total loss is calculated by combining the $\mathcal{L}_1$ loss and the structural similarity (SSIM) loss [32] $\mathcal{L}_{\text{ssim}}$ with their corresponding balanced weights λ_1 and λ_{ssim}, which is

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}}. \tag{1}$$

GL-TRANS BLOCK

Traditional vision transformers [5], [21] are adept at extracting long-range dependencies using full SA across all image patches. However, applying this approach to high-resolution images is computationally expensive, with computational cost rising quadratically $[O(H^2W^2)]$. Although larger image patches could alleviate the computational burden, it somehow results in diminished image detail. Some researchers [6], [10] have limited SA to smaller spatial windows, focusing on local features, but this approach compromises the ability to model long-range dependencies.

To tackle computational challenges and utilize both global and local SA advantages, we introduce the GL-Trans block. As shown in Figure 1(b), this block includes two parallel SA branches: a window-based local multihead SA (L-MSA) and a global MSA (G-MSA). As shown in Figure 1, after layer normalization, input features are passed to these branches.

The structure of L-MSA is detailed in Figure 2(a). The local features of size $\hat{C}\times\hat{M}\times\hat{M}$ ($\hat{C}$ for channel number and $\hat{M}$ for window size) are processed with 1×1 pointwise convolutions and 3×3 depthwise convolutions for both cross-channel context and spatial information extraction. They are transformed into three feature maps, Q_l, K_l, and V_l, which are then reshaped into $\hat{M}\hat{M}$ feature vectors with dimension $\hat{C}$. The SA mechanism is then applied to the $\hat{M}\hat{M}$ feature vectors, and the local attention matrix is generated to multiply to V_l with matrix multiplication.

Unlike L-MSA, G-MSA operates on global features of size $\hat{C}\times\hat{H}\times\hat{W}$. As depicted in Figure 2 (b), it uses 1×1 pointwise and 3×3 depthwise convolutions to accentuate local context, creating feature maps Q_g, K_g, and V_g. These

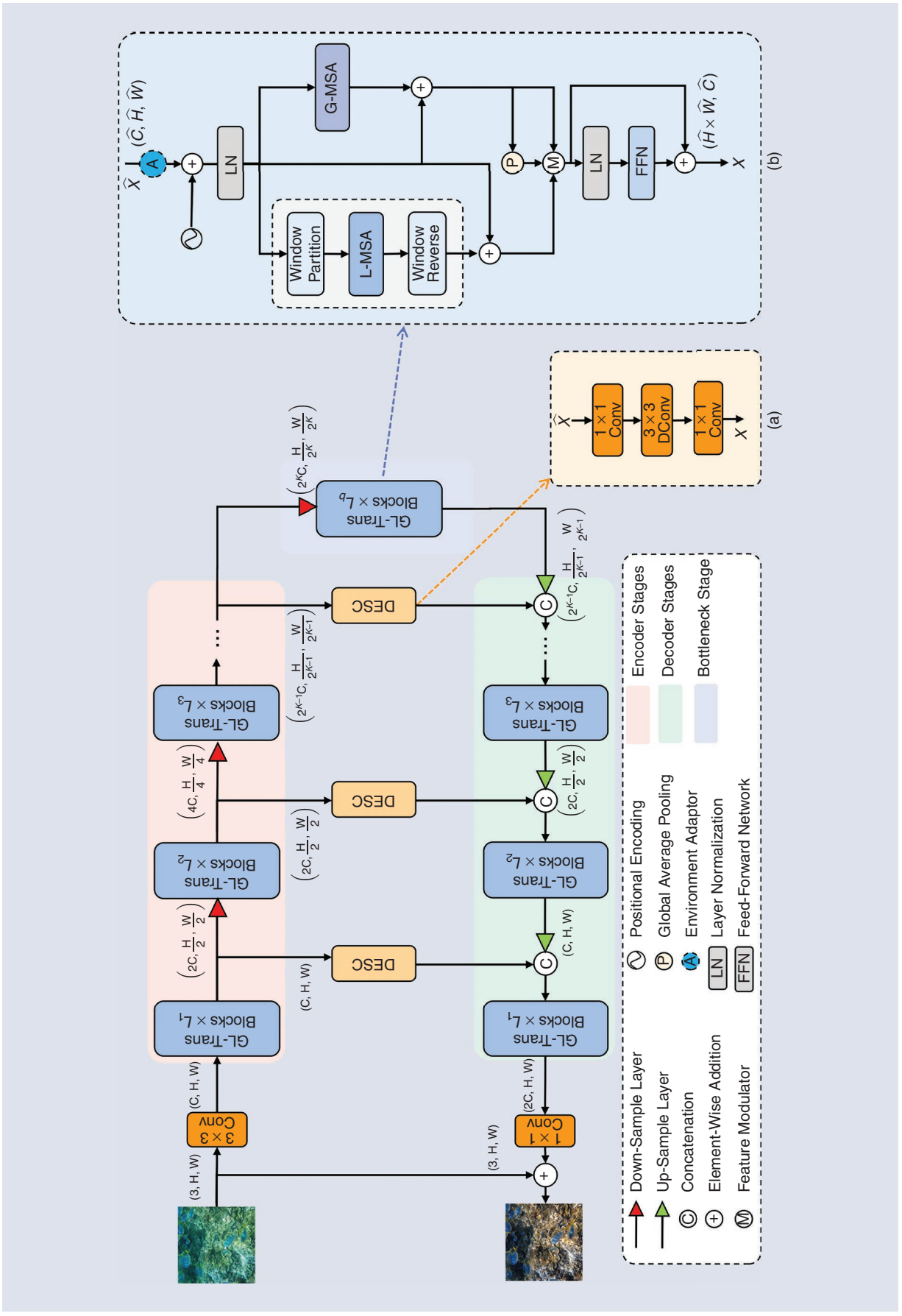

FIGURE 1. The overall architecture of our proposed WaterFormer for UIE. (a) The DESC for improving the expressivity of feature details. (b) The GL-Trans block with a parallel attention design. (The environment adaptor is only present in the bottleneck stage of the network.) L-MSA: local multihead SA; LN: layer normalization; G-MSA: global MSA.

maps are reconstructed into two-dimensional features of size $\hat{N} \times \hat{N}$, then reshaped into $\hat{C}$ feature vectors with dimension $\hat{N}\hat{N}$. Following this, the SA mechanism is applied to these vectors, forming a global attention matrix that is matrix-multiplied with V_g.

The computational costs for L-MSA and G-MSA can then be computed as follows:

$$\begin{aligned}\Omega(\text{L-MSA}) &= 3\hat{H}\hat{W}\hat{C}^2 + 27\hat{H}\hat{W}\hat{C} + 2\hat{H}\hat{W}\hat{M}^2\hat{C} \\ \Omega(\text{G-MSA}) &= 3\hat{H}\hat{W}\hat{C}^2 + 27\hat{H}\hat{W}\hat{C} + \hat{N}^2\hat{C}^2 + \hat{H}\hat{W}\hat{C}^2. \end{aligned} \quad (2)$$

The equations demonstrate that the computational costs for both MSA modules exhibit linear complexity $O(\hat{H}\hat{W})$.

In the proposed approach, as depicted in Figure 2(c), local X_l and global X_g features are combined using a feature modulator. This process involves averaging the global features and converting them into weighting parameters δ with a sigmoid function, which indicate the relative importance of X_g and X_l. The modulator then merges these features with weighting parameters.

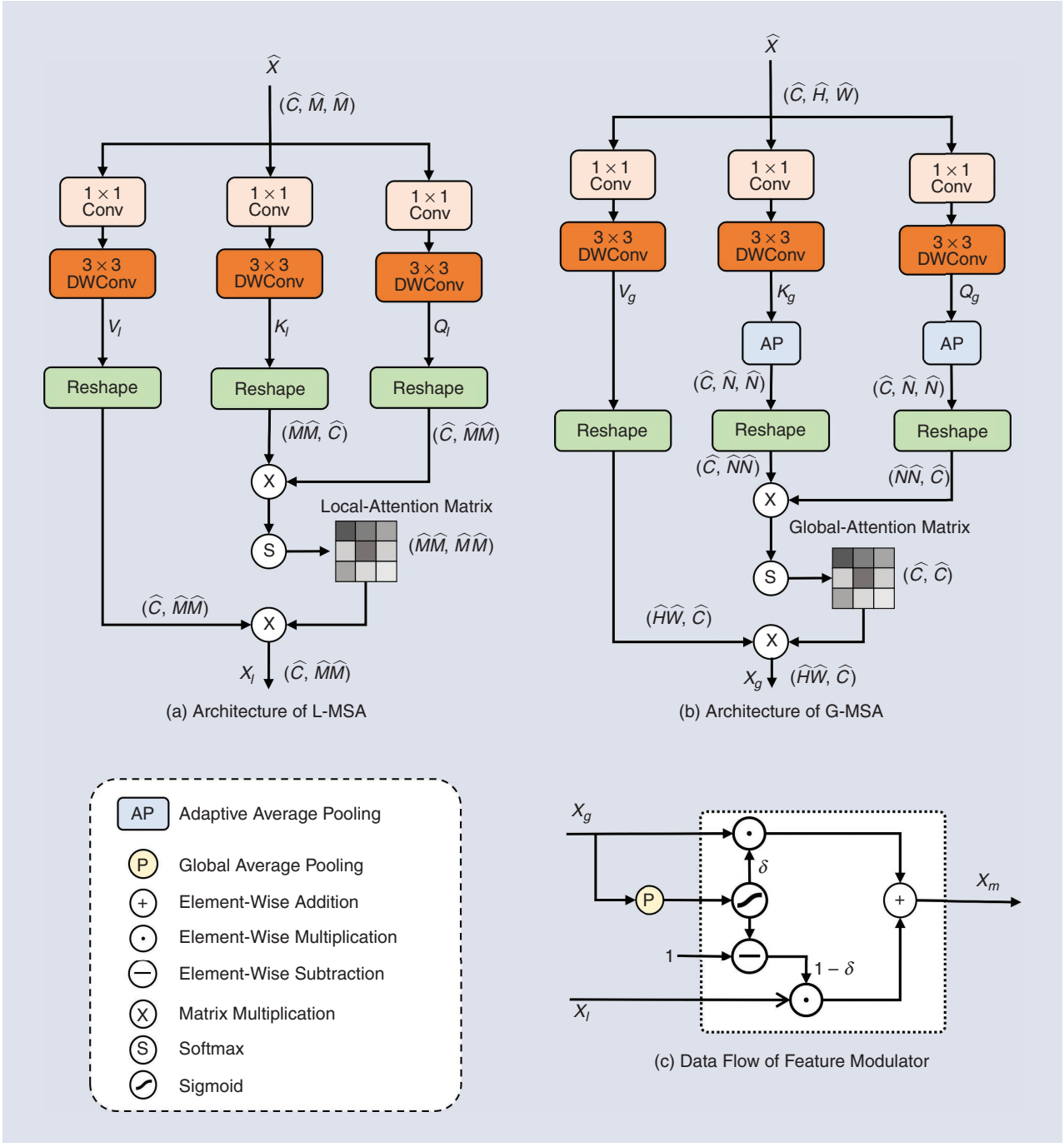

FIGURE 2. The details of our proposed GL-Trans block. (a) The L-MSA. (b) The G-MSA. (c) The feature modulator.

DESC

In generative vision tasks, maintaining detailed structural and textual context is crucial for producing high-quality images. While U-shaped networks based on CNNs effectively extract local context through direct connections among encoder and decoder features, transformer-based networks face challenges in capturing local information due to their inherent focus on longer contextual information through SA mechanisms. Consequently, directly linking encoder and decoder features in transformer-based architectures often leads to a loss of local detail.

To tackle this challenge, we introduce the DESC, consisting of a sequence of convolutional layers. As illustrated in Figure 1(a), the process begins with a 1×1 convolutional layer applied to the encoder features, expanding their dimensionality. This is followed by a 3×3 depthwise convolution for gathering information around each pixel. Subsequently, a 1×1 convolution is employed to gather cross-channel context and to resize features back to their original dimension.

> “WATERFORMER RANKS SECOND IN INFERENCE SPEED AND MAINTAINS FEWER PARAMETERS COMPARED WITH MOST OTHER NETWORKS, STRIKING AN EFFECTIVE BALANCE BETWEEN PERFORMANCE AND COMPUTATIONAL EFFICIENCY.”

ENVIRONMENT ADAPTOR

Given the challenges posed by wavelength-dependent light absorption and backscattering, effective UIE methods should handle degraded underwater images in various underwater environments. Instead of using intricate DA techniques [11], [12], we propose a straightforward yet effective environment adaptor that makes the network flexible to different underwater environments.

As illustrated in Figure 1(b), we incorporate the environment adaptor at the bottleneck stage, preceding the L-MSA/G-MSA modules. The insight is that the deep features in the bottleneck stage preserve more global information pertinent to the underwater environment. As shown in Figure 3, the core of the adaptor lies in the learnable environment embeddings with shape $L\times C$, where L is the possible environment type number and C is the deep feature dimension. The deep features first undergo a selection process from the learnable environment embeddings, aligning with the environment after Linear and Softmax processing. Subsequently, the selected environment embedding is channelwise multiplied with the deep features and integrated as a bias term. This approach improves the flexibility of the feature maps and contributes to improved enhancement performance.

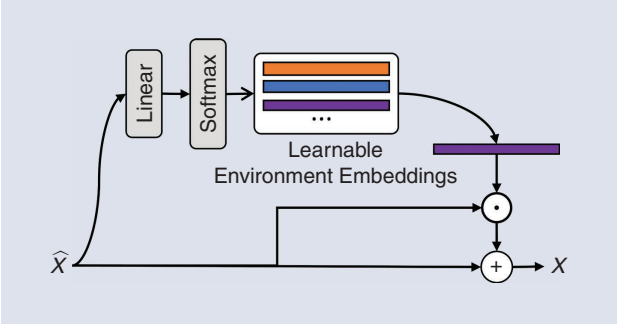

FIGURE 3. An illustration of the environment adaptor.

EXPERIMENTS AND ANALYSIS

EXPERIMENTAL SETTINGS

NETWORK SETUP

In the proposed WaterFormer, the count of encoder/decoder stages K is set to four. Across stages 1 to 4, each contains two GL-Trans blocks, with attention heads and channel dimensions configured as $\{1, 2, 4, 8\}$ and $\{32, 64, 128, 256\}$, respectively. The bottleneck stage is also designed with two GL-Trans blocks, each having eight attention heads. It is observed that increasing the number of network parameters beyond $K \geq 4$ does not lead to notable improvements in performance.

TRAINING DETAILS

Our network is trained on a single NVIDIA Tesla V100 GPU using PyTorch, with an ADAM optimizer initiated at learning rate of 0.0001. We employ cosine annealing for learning rate adjustment until convergence. The balanced weights, λ_1 and λ_{ssim}, are set to 1. Training involves initially cropping input images to 128×128 pixels at a batch size of 16 and progressively scaling up to full size with a smaller batch size. The training extends for 100 epochs, incorporating an early stop technique [33] applied to mitigate overfitting.

DATASETS

We first evaluate the performance of our network on synthetic underwater images. The synthetic training datasets are selected from UWCNN [3], LNRUD [34], and SYREA [11]. Their testing datasets are denoted as Test-600U, Test-1000L, and Test-600S, respectively. For evaluating the network performance on real-world images, two real-world underwater datasets are used, including UIEB [35] and EUVP [36]. The distribution of training, validation, and testing images across these datasets is detailed in Table 1.

EVALUATION METRICS

The commonly used peak signal-to-noise ratio (PSNR) and SSIM [32] are adopted for both the synthetic and real-world UIE evaluations. Moreover, for real-world underwater images, we also compute the nonreference metrics underwater image quality measure (UIQM) [37] and underwater color

image quality evaluation (UCIQE) [38] that are commonly used for evaluating underwater image qualities. A human subjective user study is also conducted to further evaluate the network performance on real underwater images.

EXPERIMENTS ON SYNTHETIC UNDERWATER IMAGES

Experiments are first conducted using synthetic underwater image datasets. The performances of various image enhancement methods have been compared, including traditional prior-based methods like UDCP [13] and SOTA CNN-based methods such as UNet [31], UColor [1], GLNet [2], and SyreaNet [11], along with transformer-based methods like SwinIR [10], URSCT [24], UshapeTrans [23], UDAFormer [27], Uformer [9], and Restormer [39]. For a fair comparison, the released codes of these methods are used if they are publicly available; otherwise, they are retrained with the same synthetic training data.

> “OUR METHOD, HOWEVER, GENERATES IMAGES THAT CLOSELY RESEMBLE THE GROUND TRUTH WITH SUPERIOR PSNR/SSIM SCORES.”

TABLE 1. The counts of training, validation, and testing images across each dataset.

DATASET	TRAINING	VALIDATION	TESTING
UWCNN [3]	2,342	552	596
LNRUD [34]	26,146	1,700	1,154
SYREA [11]	20,675	513	638
UIEB [35]	703	95	92
EUVP [36]	4,050	500	550

Figure 4 illustrates the visual comparison of synthetic images under different underwater conditions from Test-600U. The results indicate that the traditional prior-based method, UDCP [13], inadequately enhances image quality, leaving water effects visible. Other learning-based methods exhibit color deviations or obvious artifacts. Our method, however, generates images that closely resemble the ground truth with superior PSNR/SSIM scores.

Table 2 details the quantitative results on testing datasets, where our WaterFormer achieves the highest PSNR and SSIM scores across all synthetic datasets. It surpasses other SOTA methods substantially, averaging 2.27 dB higher in PSNR and 0.017 in SSIM. Furthermore, WaterFormer not only excels in image enhancement but also demonstrates computational efficiency. This efficiency is evident in Figure 5, where it attains the optimal PSNR in Test-600U without compromising computational efficiency. Table 3 presents a comprehensive analysis of transformer-based networks in terms of GFLops, parameter count, and inference time. WaterFormer ranks second in inference speed and maintains fewer parameters compared with most other networks, striking an effective balance between performance and computational efficiency. Figure 6 displays training loss curves for various UIE methods on the SYREA training set, showcasing

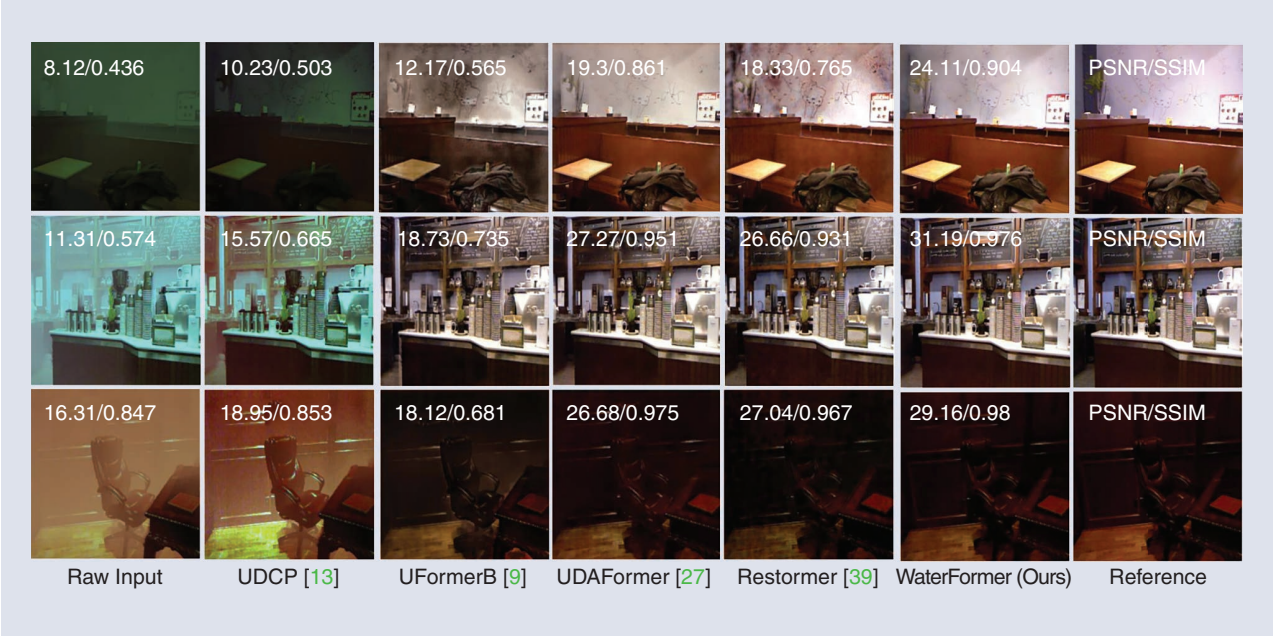

FIGURE 4. Visual comparisons of the performances of WaterFormer and other SOTA methods on synthetic underwater images sampled from Test-600U [3]. Zoom in for a better view.

TABLE 2. A quantitative comparison of our proposed method and other SOTA methods on various synthesis underwater dataset.

	TEST-600U [3]		TEST-1000L [34]		TEST-600S [11]		AVERAGE	
METHOD	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**
UDCP [13]	18.81	0.858	21.17	0.886	19.25	0.865	19.74	0.869
GLNet [2]	26.07	0.926	27.81	0.951	25.73	0.922	26.54	0.933
UColor [1]	25.63	0.901	27.31	0.949	23.44	0.909	25.46	0.919
UNet [31]	25.53	0.893	27.93	0.953	23.51	0.915	25.66	0.920
UFormerB [9]	22.59	0.857	27.49	0.947	22.41	0.897	24.16	0.900
SwinIR [10]	25.64	0.913	28.95	0.958	23.62	0.933	26.07	0.935
Restormer [39]	28.08	0.934	30.07	0.961	25.57	0.932	27.91	0.942
UDAFormer [27]	26.26	0.906	26.49	0.934	25.38	0.941	26.04	0.927
SyreaNet [11]	26.73	0.918	28.75	0.964	26.21	0.952	27.23	0.945
UShapeTrans [40]	24.35	0.863	27.31	0.941	23.27	0.903	24.98	0.902
URSCT [24]	24.86	0.891	27.52	0.950	23.39	0.902	25.26	0.914
WaterFormer	**30.09**	**0.943**	**32.61**	**0.978**	**27.84**	**0.965**	**30.18**	**0.962**

The bold values denote the best, while the underlined values indicate the second best.

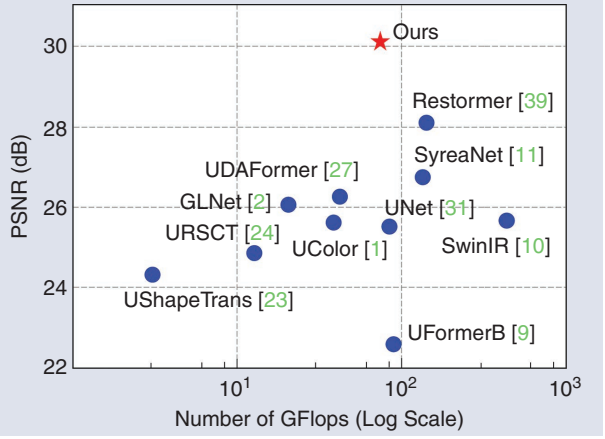

FIGURE 5. Our proposed WaterFormer achieves SOTA performance on the synthetic underwater image dataset (Test-600U) [3] while maintaining computational efficiency.

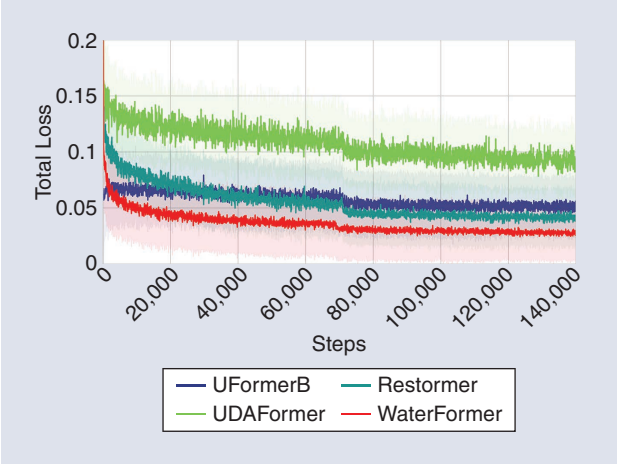

FIGURE 6. The training loss curves of various UIE methods on the SYREA [11] training set.

TABLE 3. Quantitative comparisons of GFLops, parameter counting, and inference time of transformer-based networks.

METHOD	GFLOPS	PARAMETER (M)	TIME (MS)
UFormerB	87.6	69.5	53.1
SwinIR	202.1	31.4	53.2
Restormer	87.2	16.9	59.1
UDAFormer	41.6	**9.6**	48.4
UShapeTrans	**3.03**	31.6	**28.5**
URSCT	14.2	11.4	49.8
WaterFormer	49.8	27.1	44.3

The bold values denote the best, while the underlined values indicate the second best.

WaterFormer's superiority by achieving the lowest training loss among SOTA methods.

EXPERIMENTS ON REAL UNDERWATER IMAGES

The efficacy of our WaterFormer has been further assessed on real underwater images from the UIEB [35] and EUVP [36] testing sets, comparing it with other SOTA methods. Figure 7 illustrates visual comparisons in greenish and bluish underwater conditions. Real underwater images pose greater challenges than synthetic ones due to their complex environments. In these scenarios, UDCP [13] inadequately addresses the color variations, often leading to darker tones. Methods like SwinIR [10], SyreaNet [11], UshapeTrans [23], and UDAFormer [27] tend to generate noticeable artifacts. GLNet [2] and Restormer [39] are less effective in eliminating greenish/bluish effects.

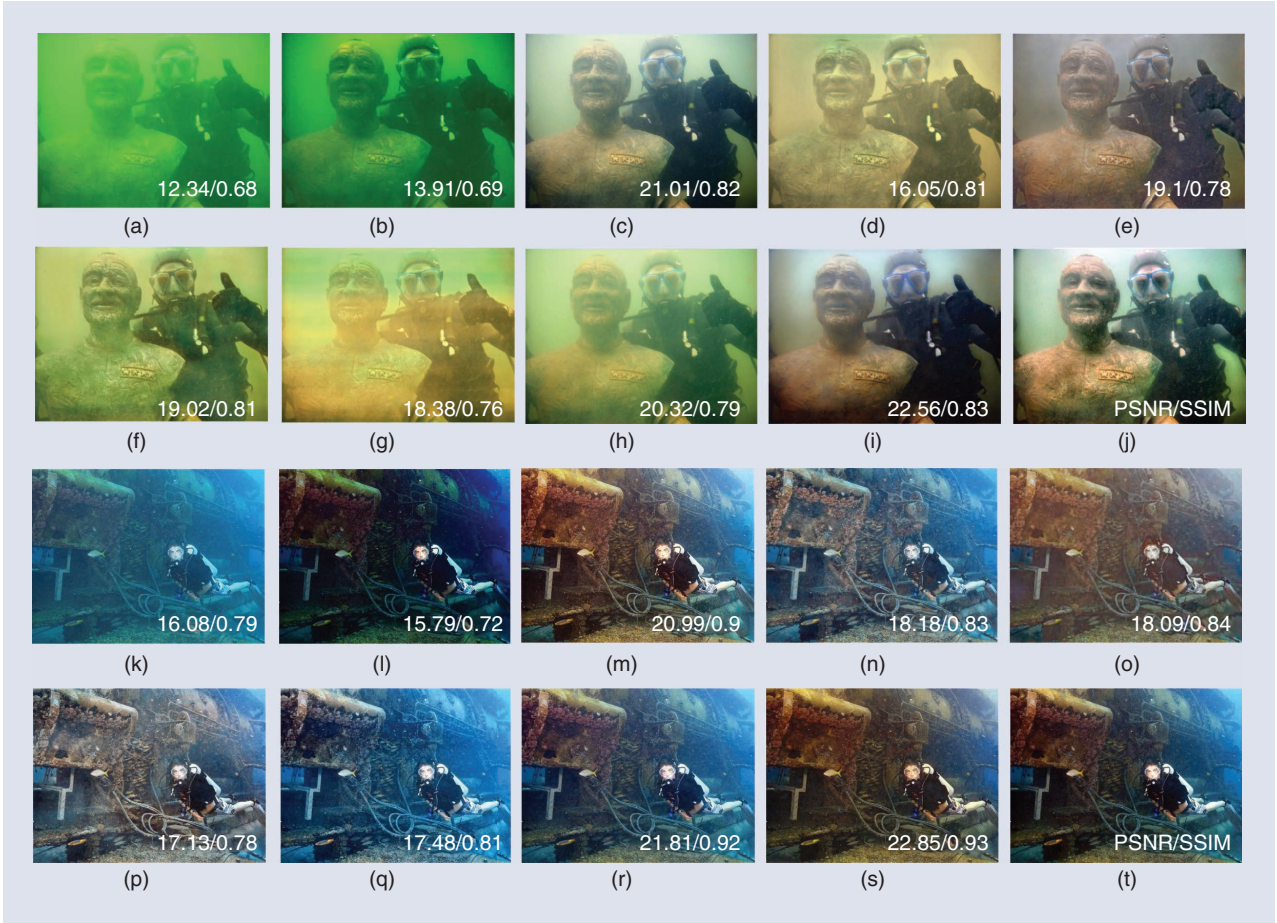

FIGURE 7. Visual comparisons of the performances of WaterFormer and other SOTA methods on real underwater images sampled from UIEB [35]. Zoom in for a better view. The enhancement results in a greenish underwater environment: (a) Raw input. (b) UDCP [13]. (c) GLNet [2]. (d) SwinIR [10]. (e) SyreaNet [11]. (f) UshapeTrans [40]. (g) UDAFormer [27]. (h) Restormer [39]. (i) WaterFormer. (j) Reference. The enhancement results in a bluish underwater environment: (k) Raw input. (l) UDCP [13]. (m) GLNet [2]. (n) SwinIR [10]. (o) SyreaNet [11]. (p) UshapeTrans [40]. (q) UDAFormer [27]. (r) Restormer [39]. (s) WaterFormer. (t) Reference.

Conversely, WaterFormer consistently improves image quality across various underwater conditions and achieves higher PSNR/SSIM scores with reference images. Table 4 presents the quantitative comparison of PSNR and SSIM scores for our network on the UIEB [35] and EUVP [36] datasets. Our network outperforms other SOTA methods in both datasets, demonstrating its capability on real underwater images.

To better validate the performances of various UIE methods on real-world images, we also compute the nonreference metrics UIQM [37] and UCIQE [38] that are commonly used for underwater image quality evaluation. Moreover, we conduct a user study by randomly choosing 10 enhanced underwater images from each method and invite 30 volunteers to score the results in a range of 0–10, with a higher score indicating better enhancement quality. As shown in Table 5, our proposed WaterFormer achieves comparable performances with other SOTA methods with respect to the metrics of UIQM/UCIQE and gets the highest score in the user study, indicating its effectiveness in generating visually pleasing enhanced images.

TABLE 4. Quantitative comparisons of real-world underwater image datasets.

	UIEB [35]		EUVP [36]	
METHOD	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**
UDCP [13]	15.18	0.763	16.38	0.772
GLNet [2]	22.85	0.913	23.08	0.889
UColor [1]	22.51	0.914	23.11	0.892
UFormerB [9]	22.28	0.918	23.88	0.887
SwinIR [10]	22.21	0.912	23.97	0.901
Restormer [39]	<u>23.38</u>	0.919	<u>24.12</u>	0.903
UDAFormer [27]	23.18	0.916	23.85	0.897
SyreaNet [11]	23.26	<u>0.926</u>	24.08	<u>0.908</u>
UShapeTrans [40]	22.08	0.909	23.18	0.885
URSCT [24]	22.55	0.913	23.98	0.891
WaterFormer	**23.79**	**0.932**	**24.58**	**0.915**

The bold values denote the best, while the underlined values indicate the second best.

TABLE 5. Nonreference evaluations of real-world images.

METHODS	UIQM↑	UCIQE↑	USER STUDY↑
UDCP [13]	1.290	0.556	4.11
GLNet [2]	1.541	0.619	6.18
UColor [1]	1.371	0.553	5.92
UFormerB [9]	1.259	0.552	3.56
SwinIR [10]	1.511	0.621	4.68
Restormer [39]	**1.689**	0.614	5.42
UDAFormer [27]	1.523	0.506	5.18
SyreaNet [11]	1.656	0.582	4.27
UShapeTrans [40]	1.404	0.577	4.52
URSCT [24]	1.310	0.528	4.35
WaterFormer	1.659	**0.625**	**6.53**

The bold values denote the best, while the underlined values indicate the second best.

TABLE 6. Quantitative comparisons of ablated models on both synthetic and real-world underwater image testing sets.

	TEST-600U [3]		UIEB [35]	
MODELS	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Full model	**30.09**	**0.943**	**23.79**	**0.932**
Without G-MSA	26.82	0.923	22.27	0.910
Without L-MSA	27.04	0.926	22.38	0.912
Without feature modulator	28.56	0.931	23.28	0.926
Without DESC	27.81	0.929	22.50	0.917
Without environment adaptor	28.31	0.933	23.02	0.921

The bold values denote the best, while the underlined values indicate the second best.

ABLATION STUDY

This section details ablation studies conducted to evaluate the effects of our proposed modules on synthetic and real-world underwater images.

1) *The effectiveness of the GL-Trans block*: We evaluate the effectiveness of our GL-Trans block by removing the G-MSA/L-MSA branch in each transformer block. Table 6 reveals that removing G-MSA/L-MSA results in a PSNR decrease of approximately 3.0 on synthetic images and 1.5 on real images. Figure 8 shows that the model lacking L-MSA can produce noticeable artifacts and struggles to uniformly enhance image quality. Conversely, the model without G-MSA tends to create blocking artifacts in the background due to the window-based SA mechanism. Additionally, we explore the impact of the global–local feature modulator by replacing it with pointwise addition in each GL-Trans block. The results in Table 6 indicate that the feature modulator improves network performance, emphasizing the necessity of effectively integrating global and local transformer features.
2) *The effectiveness of DESC*: The efficacy of our proposed DESC is assessed by eliminating all DESCs and directly concatenating encoder with decoder features. As indicated in Table 6, this removal results in diminished performance, indicating the significance of utilizing detailed local information. As illustrated in Figure 8, visual comparisons show that the model without DESC struggles with handling the edges of foreground objects.
3) *The effectiveness of the environment adaptor*: The importance of the environment adaptor has been examined by removing it from the bottleneck GL-Trans block. Table 6 shows PSNR reductions of 1.78 and 0.77 for synthetic and real underwater images upon its removal. As seen in Figure 8, the network lacking the environment adaptor retains a greenish tone, indicating its limited ability to handle varied underwater environments. Further, to assess its effectiveness, input images are classified using the environment adaptor based on their underwater scenarios. The insight is that the environment adaptor can autonomously learn environmental features in an unsupervised manner during training. With the number of possible environment types set to $L=10$, Figure 9 validates the adaptor's capability in distinguishing

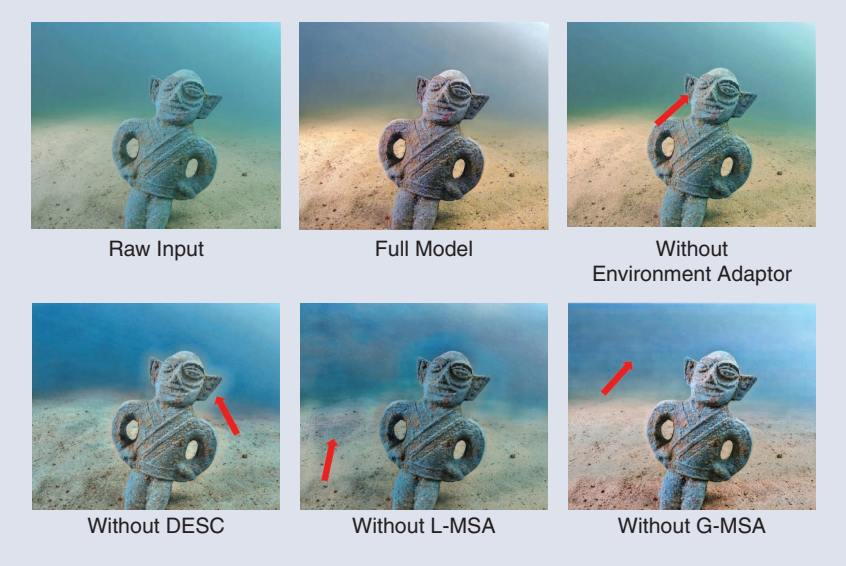

FIGURE 8. Visual comparisons of ablated models. The red arrows indicate the major differences between the ablated model and the full model. Zoom in for a better view.

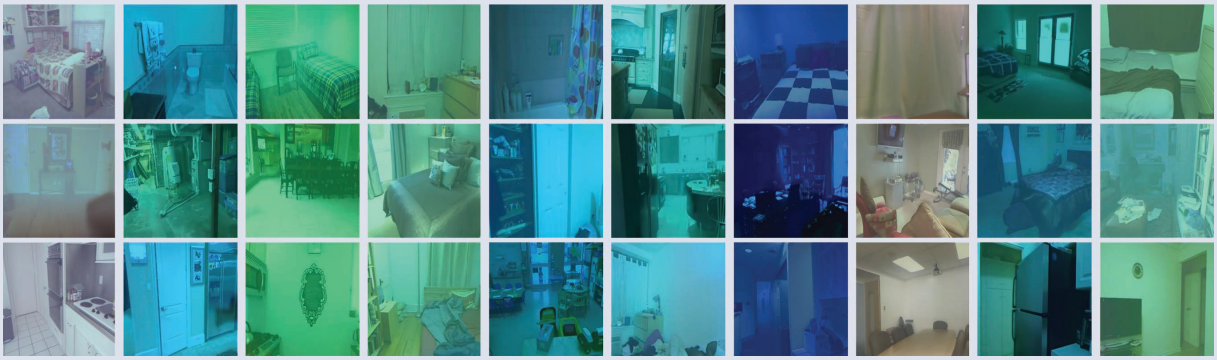

FIGURE 9. The visualization results of images sampled from Test-600S [11]. The images in each column belong to the same environment type.

FIGURE 10. Enhanced results are suboptimal for our proposed WaterFormer and other SOTA methods in environments with intricate lighting conditions.

different underwater environments. The images in each column correlate with a specific environment type, corresponding to distinct environment embeddings within the adaptor.

DISCUSSION

While our WaterFormer shows SOTA results on various UIE datasets, it faces challenges in certain conditions. The first row in Figure 10 illustrates that in shallow waters with sunlight reflections, our method may generate unwanted artifacts due to nonuniform illustration. Additionally, in highly turbid waters, as shown in the second row, WaterFormer's effectiveness is reduced. This issue is not unique to our method; other SOTA methods like Restormer [39], as demonstrated in the second column in Figure 10, also struggle under these complex illumination conditions. It can be attributed to the lack of comprehensive training data covering these specific underwater environments, limiting the generalizability of current learning-based approaches. Future efforts will thus be directed toward enhancing UIE performance in underwater scenarios with intricate lighting.

CONCLUSION

In this article, we propose a novel transformer-based network for UIE. Benefiting from our well-designed GL-Trans block, the network is superior in leveraging the global semantic and local detailed features while being computationally efficient. In addition, the proposed DESC module could boost the network's capability to generate fine details. Moreover, we introduce a simple but effective environment adaptor that helps the network better adapt to various underwater environments and further improves the network's performance. Extensive experiments have demonstrated that our proposed network surpasses other SOTA methods in both qualitative and quantitative evaluations on synthetic and real underwater images, which could establish a solid foundation for future AUV tasks with respect to vision-related applications.

ACKNOWLEDGMENT

This study is supported by the Major Key Project of the PCL Department of Broadband Communication and the Talent Program of Guangdong Province (2021QN02Z107). Jinqiang Cui is the corresponding author. This article has supplementary downloadable material available at https://doi.org/10.1109/MRA.2024.3351477, provided by the authors.

AUTHORS

Junjie Wen, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T. 99907, Hong Kong, and Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen 518055, China. E-mail: jjwen@mae.cuhk.edu.hk.

Jinqiang Cui, Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen 518055, China. E-mail: cuijq@pcl.ac.cn.

Guidong Yang, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T. 99907, Hong Kong. E-mail: gdyang@mae.cuhk.edu.hk.

Benyun Zhao, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T. 99907, Hong Kong. E-mail: 1155145791@link.cuhk.edu.hk.

Yu Zhai, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T. 99907, Hong Kong. E-mail: zhaiyu@link.cuhk.edu.hk.

Zhi Gao, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China. E-mail: gaozhinus@gmail.com.

Lihua Dou. School of Automation, Beijing Institute of Technology, Beijing 100081, China. E-mail: doulihua@bit.edu.cn.

Ben M. Chen, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T. 99907, Hong Kong. E-mail: bmchen@cuhk.edu.hk.

REFERENCES

[1] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, May 2021, doi: 10.1109/TIP.2021.3076367.

[2] X. Fu and X. Cao, "Underwater image enhancement with global–local networks and compressed-histogram equalization," *Signal Process., Image Commun.*, vol. 86, Aug. 2020, Art. no. 115892, doi: 10.1016/j.image.2020.115892.

[3] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107038, doi: 10.1016/j.patcog.2019.107038.

[4] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929.*

[6] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 9992–10,002, doi: 10.1109/ICCV48922.2021.00986.

[7] B. Zhang et al., "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 11,294–11,304, 2022, doi: 10.1109/CVPR52688.2022.01102.

[8] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3D position embedding," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 5802–5810, 2022, doi: 10.1109/CVPR52688.2022.00572.

[9] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2022, pp. 17,662–17,672, doi: 10.1109/CVPR52688.2022.01716.

[10] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 1833–1844, doi: 10.1109/ICCVW54120.2021.00210.

[11] J. Wen et al., "SyreaNet: A physically guided underwater image enhancement framework integrating synthetic and real images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 5177–5183, doi: 10.1109/ICRA48891.2023.10161531.

[12] Q. Jiang, Y. Zhang, F. Bao, X. Zhao, C. Zhang, and P. Liu, "Two-step domain adaptation for underwater image enhancement," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108324, doi: 10.1016/j.patcog.2021.108324.

[13] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2013, pp. 825–830, doi: 10.1109/ICCVW.2013.113.

[14] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *J. Vis. Commun. Image Representation*, vol. 26, pp. 132–145, Jan. 2015, doi: 10.1016/j.jvcir.2014.11.006.

[15] D. Berman, T. Treibitz, and S. Avidan, "Diving into haze-lines: Color restoration of underwater images," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2017, vol. 1, p. 2.

[16] B. McGlamery, "A computer model for underwater camera systems," in *Proc. Ocean Opt. VI*, 1980, vol. 208, pp. 221–231, doi: 10.1117/12.958279.

[17] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robot. Autom. Lett.*, vol. 3, no. 1, pp. 387–394, Jan. 2018, doi: 10.1109/LRA.2017.2730363.

[18] N. G. Jerlov, "Classification of sea water in terms of quanta irradiance," *ICES J. Mar. Sci.*, vol. 37, no. 3, pp. 281–287, 1977, doi: 10.1093/icesjms/37.3.281.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805.*

[20] A. Radford et al., "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vision (ECCV),* Glasgow, U.K. Cham, Switzerland: Springer-Verlag, Aug. 23–28, 2020, pp. 213–229, 2020.

[22] R. Wang, Y. Zhang, and J. Zhang, "An efficient swin transformer-based method for underwater image enhancement," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18,691–18,708, 2023, doi: 10.1007/s11042-022-14228-6.

[23] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," 2021, *arXiv:2111.11843.*

[24] T. Ren et al., "Reinforced swin-convs transformer for simultaneous underwater sensing scene image enhancement and super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Sep. 2022, doi: 10.1109/TGRS.2022.3205061.

[25] A. Boudiaf et al., "Underwater image enhancement using pre-trained transformer," in *Proc. Int. Conf. Image Anal. Process.*, Cham, Switzerland: Springer-Verlag, 2022, pp. 480–488, doi: 10.1007/978-3-031-06433-3_41.

[26] Y. Tang, T. Iwaguchi, H. Kawasaki, R. Sagawa, and R. Furukawa, "AutoEnhancer: Transformer on U-Net architecture search for underwater image enhancement," in *Proc. Asian Conf. Comput. Vision*, 2022, pp. 1403–1420.

[27] Z. Shen, H. Xu, T. Luo, Y. Song, and Z. He, "UDAformer: Underwater image enhancement based on dual attention transformer," *Comput. Graph.*, vol. 111, pp. 77–88, Apr. 2023, doi: 10.1016/j.cag.2023.01.009.

[28] Z. Huang, J. Li, Z. Hua, and L. Fan, "Underwater image enhancement via adaptive group attention-based multiscale cascade transformer," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–18, Jul. 2022, doi: 10.1109/TIM.2022.3189630.

[29] Y. Zhang, D. Chen, Y. Zhang, M. Shen, and W. Zhao, "A two-stage network based on transformer and physical model for single underwater image enhancement," *J. Mar. Sci. Eng.*, vol. 11, no. 4, 2023, Art. no. 787, doi: 10.3390/jmse11040787.

[30] N. Cheng, Z. Sun, X. Zhu, and H. Wang, "A transformer-based network for perceptual contrastive underwater image enhancement," *Signal Process., Image Commun.*, vol. 118, Oct. 2023, Art. no. 117032, doi: 10.1016/j.image.2023.117032.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer-Verlag, Oct. 5–9, 2015, vol. 18, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[33] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Constructive Approximation*, vol. 26, pp. 289–315, Apr. 2007, doi: 10.1007/s00365-006-0663-2.

[34] T. Ye, S. Chen, Y. Liu, Y. Ye, E. Chen, and Y. Li, "Underwater light field retention: Neural rendering for underwater imaging," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2022, pp. 487–496, doi: 10.1109/CVPRW56347.2022.00064.

[35] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, Nov. 2019, doi: 10.1109/TIP.2019.2955241.

[36] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, Apr. 2020, doi: 10.1109/LRA.2020.2974710.

[37] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, Jul. 2016, doi: 10.1109/JOE.2015.2469915.

[38] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, Dec. 2015, doi: 10.1109/TIP.2015.2491020.

[39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2022, pp. 5718–5729, doi: 10.1109/CVPR52688.2022.00564.

[40] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," in *Proc. Comput. Vision (ECCV) Workshops,* Tel Aviv, Israel. Cham, Switzerland: Springer-Verlag, 2023, pp. 290–307, doi: 10.1007/978-3-031-25063-7_18.