TAFormer: A Transmission-Aware Transformer for Underwater Image Enhancement

Yuanyuan Li, Zetian Mi, Yulin Wang, Shuaiyong Jiang, and Xianping Fu

***Abstract*— The attenuation and scattering of different colors of light underwater are wavelength- and distance-dependent, leading to various degradation problems in underwater images. When enhancing underwater images, many deep learning-based methods rely solely on convolutional neural networks to learn a mapping from degraded images to clear images to achieve enhanced effects. However, such methods have limitations in capturing long-term dependencies, preventing them from accurately capturing the global information of images. Although Transformers can solve this problem, there is a lack of inductive bias in training due to the limited number of training datasets with certain degradation phenomena. To address this issue, a novel Swin Transformer based on physical perception is proposed for the first time. Swin Transformer is used to solve the long- and short-distance dependency problem. Additionally, the underwater image degradation process is considered in network design to solve the problem of poor inductive bias. Combining the advantages of physical imaging, convolutional neural networks and Transformer can effectively improve the visual quality of underwater images. Rich qualitative and quantitative experimental results show that our Transformer achieves competitive performance on 5 benchmark datasets.**

***Index Terms*— Transmission-aware transformer, transmission-guided multi-head self-attention, spatio-frequency domain interaction, underwater image enhancement.**

I. Introduction

UNDERWATER image enhancement technology [1] is crucial for obtaining high-quality underwater images and exploring the underwater environment. It is widely used in the fields of marine environmental investigation and underwater robots. However, the complexity of the underwater optical imaging process requires various considerations. For example, light of different colors attenuates and is absorbed at different rates during underwater propagation, and the presence of soluble substances and suspended particles in the water affects the light received by the camera. There are different types of scattering, and the complex imaging processes often result in issues with directly captured underwater images such as color anomalies, loss of detail, and low contrast. Therefore, it is highly important to develop innovative and practical underwater image enhancement methods.

Generally, current mainstream underwater image enhancement methods can be divided into three categories: methods based on pixel adjustment, methods based on imaging models and methods based on deep learning. Methods based on pixel adjustment [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] directly adjust pixel values by modifying aspects such as brightness, contrast and saturation, to improve the visual quality of the image. However, these methods ignore the causes and processes of image degradation under water, so the enhancement effects often have limitations, such as artifacts and excessive enhancement. In addition, imaging model-based methods [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] construct underwater imaging models by analyzing the underwater imaging process and solve the relevant model parameters based on specific assumptions, such as background light and medium transmission, to reverse engineer theoretically clear underwater images. However, for complex and dynamic underwater environments, specific assumptions and priors are not always established, and simultaneously estimating multiple parameters is difficult and computationally intensive. In recent years, underwater image enhancement technologies based on deep learning [2], [3], [4], [5], [6], [7], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55] have developed rapidly. They improve the visual perception of images by learning the mapping of source images to high-quality images from large-scale underwater images, and have achieved impressive results. However, the networks have limited receptive fields of convolutional operators and have limitations in capturing long-distance dependencies and nonlocal self-similarity. From the underwater imaging process, we must admit that capturing long-distance dependencies is very important for image enhancement. For example, the two parameters of background light and medium transmittance in the imaging model rely on the global information of an image. As shown in Fig. 1, compared with 5 advanced deep learning-based methods, our method achieves significant advantages in PSNR and SSIM on 3 underwater benchmark datasets.

In response to the abovementioned shortcomings, the Transformer [56] has emerged in the past two years. It benefits from self-focus and can capture the global dependencies between data. It is also increasingly used in the field of computer vision [57]. Subsequently, Swin Transformer [58]

Manuscript received 4 January 2024; revised 6 May 2024 and 14 June 2024; accepted 2 September 2024. Date of publication 6 September 2024; date of current version 30 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62176037 and Grant 62002043; in part by the Liaoning Revitalization Talents Program under Grant XLYC1908007; in part by the Foundation of Liaoning Key Research and Development Program under Grant 201801728; in part by the Liaoning Applied Basic Research Project under Grant 2022JH2/101300264; and in part by the Dalian Science and Technology Innovation Fund under Grant 2018J12GX037, Grant 2019J11CY001, Grant 2021JJ12GX028, and Grant 2022JJ12GX016. This article was recommended by Associate Editor S. T. Kim. *(Corresponding author: Xianping Fu.)*

The authors are with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: liyuanyuan@dlmu.edu.cn; mizetian@dlmu.edu.cn; wyulin@dlmu.edu.cn; lucky_star924@163.com; fxp@dlmu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3455353

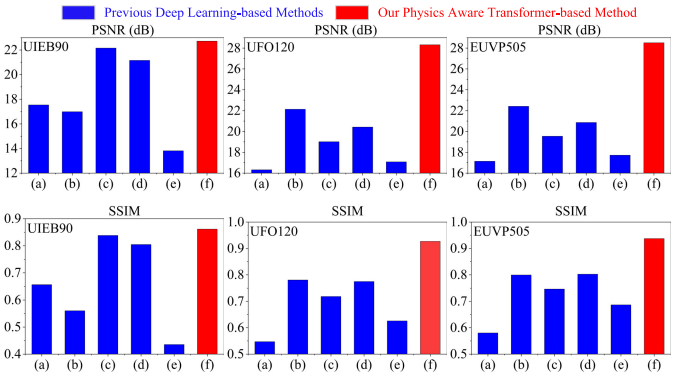

Fig. 1. Our TAFormer (f) is compared with five advanced deep learning based methods: (a) CWR [2], (b) FUnIE [3], (c) PUIE [4], (d) Ucolor [5], (e) UWCNN-II [6], on three benchmark underwater datasets: UIEB [7], UFO [8], and EUVP [3]. It can be seen that our method significantly outperformed other methods in terms of PSNR and SSIM evaluation metrics.

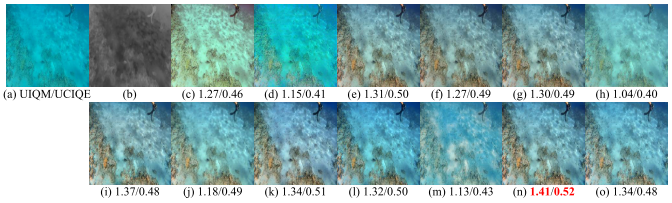

Fig. 2. Our TAFormer is compared with six advanced deep learning based methods ((c) CWR [2], (d) FUnIE [3], (e) PUIE-MC [4], (f) PUIE-MP [4], (g) Ucolor [5], (h) UWCNN-II [6]) and five Transformer based methods ((i) URSCT [59], (j) U-shape [60], (k) Uformer [61], (l) Restormer [62], (m) SwinIR [63]). (a), (b), and (n) are the input, transmission map, and our result, respectively. (o) is the result of our network without transmission guidance.

was proposed to use a moving window to limit the calculation of attention in a window, thereby improving the computational efficiency of the model. However, due to the lack of inductive bias, the generalization performance of Transformer may not be as good as that of convolutional networks. To achieve ideal results, a fairly large dataset is required for training. Although some researchers have combined Swin Transformer with convolutional neural networks for underwater image enhancement [59], due to the lack of correct inductive bias, this method will result in overenhancement or underenhancement for an underwater scene with a small dataset. This issue motivates us to consider the physical characteristics of underwater images. According to the underwater imaging process [27], [28], the transmission of the underwater image represents the percentage of scene light that reaches the camera after attenuation, which reveals the degree of quality degradation in different areas of the underwater image, as shown in Fig. 2(b). Therefore, with the guidance of medium transmission map, areas with severe fading can be given more attention and the advantages of imaging model-based methods can be integrated into the Transformer model, thereby utilizing the physical properties of the underwater images to accelerate network optimization and improve enhancement performance.

To solve the above problems, a new transmission-aware Swin Transformer is developed for underwater image enhancement. The key component of this network framework is the transmission-aware Swin Transformer block, which contains transmission-guided multihead self-attention, providing the most discriminative self-attention value for feature aggregation, and a transmission-guided spatio-frequency domain interactive feed-forward network to achieve better detail enhancement. First, to enable the model to capture local dependencies and global dependencies, the depthwise separable convolution block and Swin Transformer [59] are combined, and the depthwise separable convolution is used to replace the linear layer in the traditional Swin Transformer, capturing more local attention in each window enables spatial and channelwise enhancement. Then, transmission-guided window multihead self-attention replaces ordinary self-attention, using transmission as a guide for image values (V), thus facilitating better feature aggregation. Finally, the proposed spatio-frequency domain interactive feedforward network further explores spatial domain information and Fourier domain information to better enhance the aggregated features. As shown in Fig. 2, for underwater image with extremely severe fading, methods based on GANs and CNNs often have poor overall or local enhancement effects, while traditional Transformer-based methods tend to exhibit under-enhancement phenomena, as shown in Fig. 2(i), (j), (l), (m) and our network without transmission guidance. In contrast, the visual perception and indicator results show that our method produces higher contrast, richer local details, and the best overall reconstruction effect. The main innovations of this work are summarized as follows:

- To overcome the limitations of simply using imaging models, convolutional neural networks or Transformers for underwater image enhancement, a Swin Transformer based on physical perception is proposed. It integrates an image degradation mechanism and convolutional neural networks into Swin Transformer, which can effectively incorporate the advantages of the three technologies.
- Since the transmission of an underwater image can implicitly characterize the degree of different degraded areas in the image, transmission-guided window multihead self-attention is proposed instead of traditional window self-attention. This approach can utilize the attenuation information of underwater images as key clues to guide the modeling of long-range dependencies, thereby achieving better feature aggregation.
- Considering that the global background light and structural information of underwater images can be partially decomposed in the Fourier domain, abandoning the feed-forward neural network composed of traditional spatial domain convolution operations, a transmission-guided spatio-frequency domain interaction feedforward network is designed to achieve receptive field coverage of the entire image by incorporating feature information in the Fourier domain, thereby improving the perceptual quality of the network and enhancing the aggregated features well.
- Subjective and objective comparative experiments with multiple recent mainstream underwater image enhancement methods on various datasets show that our method can achieve good results in improving color shift, enhancing contrast, and enriching details, which proves that our method can solve the problem of traditional Transformer's lack of correct inductive bias.

The remainder of the paper is organized as follows: Section II describes the current research status of underwater image enhancement in detail. Section III provides a detailed description of the proposed method. The experiments are presented in Section IV, and Section V concludes the paper.

II. Related Work

Underwater image enhancement can be divided into pixel-based adjustment methods, imaging model-based methods and deep learning-based methods. In recent years, visual Transformer models have been increasingly used in the field of image processing.

A. Underwater Image Enhancement

1) Pixel-Based Adjustment Methods: These approaches improve the visual perception of the image by adjusting the image pixel values, and include methods based on image decomposition and fusion [9], [10], [11], [12], [13], [14], [15], [16], methods based on histogram distribution prior [17], [18], staged processing methods [19], [20], [21], and methods based on retinex model decomposition [22], [23], [24], [25], [26]. For example, Ancuti et al. successively proposed different multiscale fusion methods for enhancing underwater images, such as the fusion-based method [10]. Similarly, Zhang et al. [12] used the color correction method to correct color and then calculated two versions of global and local contrast enhancement from color-corrected images. The weight and multiscale fusion of the two versions are calculated to obtain the enhanced image. However, this method weakens the colors in the image. Yuan et al. [13] proposed a new multiscale fusion method based on blurriness and color. However, the results produced by this method are not always ideal, and noise is amplified during enhancement. Zhou et al. proposed multi-feature prior fusion [15] and multi-view fusion mechanisms [16] to enhance underwater images, both of which achieved significant results. In addition, Zhou et al. [18] proposed the multi-interval sub-histogram equalization technique, which estimates the degree of fading of different features in an image based on the statistical characteristics of the image, thereby guiding feature enhancement and improving the visual effect of the degraded image. Zhang et al. [21] proposed a method based on minimum color loss and local adaptive contrast enhancement to enhance underwater images in stages. However, this method cannot produce satisfactory results under low light conditions. In addition, some scholars have been inspired by Retinex theory. For example, Zhuang et al. [25] proposed a Bayesian Retinex algorithm that utilizes reflection priors and multi-order gradients prior to illumination maps to enhance underwater images. This method first performed color correction on the input image, then established a maximum posterior model for the corrected image, and finally used the norm to enhance the spatial smoothness and spatial linear smoothness of the illumination map. The maximum posterior problem was transformed into an energy minimization problem and solved using the alternating direction multiplier method (ADMM) as well as threshold shrinkage and fast Fourier transform methods. Similarly, the team proposed a Retinex variational model inspired by a hyper-Laplacian reflection prior [26], which first corrected the color and then built a hyper-Laplacian on the first- and second-order gradients of reflectance. This method can produce better enhancement results.

Methods of this type have good enhancement results for underwater images in a variety of degraded scenes, but because they ignore the underwater imaging process, the results are either overenhanced or underenhanced and may even introduce new colors.

2) Imaging Model-Based Methods: Due to the establishment of underwater imaging models [27], [28], these methods are widely used to retrieve clear images from degraded underwater images. Since the dark channel prior [29] method achieves convincing results in haze removal and because of the similarity of imaging models between underwater images and haze images, some variants of DCP have been used to restore underwater images. Such methods rely mainly on prior information to estimate the parameters of the imaging model, such as the attenuation prior [30], red channel prior [31], underwater dark channel prior [32], blurriness prior [33], general dark channel prior [34], and illumination channel sparsity prior [36]. For example, Chiang and Chen advocated the use of dehazing algorithms to enhance underwater images [30] to compensate for attenuation differences along the propagation path. Galdran et al. asserted that light of different wavelengths has different attenuation rates, among which red wavelength light attenuates the most quickly, so they proposed a red channel prior method [31] to restore underwater images, that is, to restore the colors corresponding to short wavelengths and then restore the degraded details. Drews et al. contended that the red channel of underwater images was unreliable and proposed using only the information of the green channel and blue channel to estimate the transmission map [32]. Peng and Cosman [33] estimated underwater scene depth using image blurriness and light absorption. Subsequently, they proposed a generalized dark channel prior method [34] that can restore a variety of degraded images and integrate an adaptive color correction method into the imaging model. In addition, Xie et al. proposed a variational framework guided by the red channel prior based on the underwater imaging model [35]. Hou et al. [36] explored an illumination channel sparsity prior to address the issue of uneven lighting in underwater images. In such methods, the accuracy of the model directly affects the recovery results, and methods based on complex models [37], [38] are more suitable for scenarios with severe fading. For example, [37] combined a comprehensive imaging formation model with unsupervised learning to improve the robustness of the method under various lighting conditions and complex underwater environments. In recent years, some scholars have combined physical models with deep learning methods [39], [40], [41], using the advantages of both to achieve underwater image enhancement. Tang et al. [42] first developed a generalized imaging model using deep learning that can effectively solve the problems of color distortion and semantic information loss.

Although such methods can achieve good restoration effects, due to the limitations of the assumptions and priors themselves, complex and diverse water environments, and degraded

images that cannot be explained by imaging models (such as images with multiple artificial light sources), the enhancement results of such methods often suffer from low credibility and poor robustness.

3) Deep Learning-Based Methods: Deep learning is flourishing [64], [65], making it a very competitive technical method in the field of underwater image enhancement. These methods can be divided into GAN-based methods [2], [3], [43], [44], [45], [46], [47], [48] and CNN-based methods [4], [6], [7], [49], [50], [51], [52], [53], [54], [55]. GAN-based methods use the generator to produce a result that tends to be a real image and then use the discriminator to distinguish real images from fake images. For example, Li et al. [43] first used a generative adversarial network to generate extremely realistic underwater images from aerial images and their depth maps, and then used these data to train a color correction network. Fabbri et al. [44] first used the CycleGAN model to generate a dataset. Islam et al. [3] proposed a real-time underwater image enhancement network based on a fully convolutional generative adversarial network for image enhancement and introduced a large-scale underwater dataset EUVP containing pairwise and nonpairwise. However, this method does not work well for severely degraded images, and some enhancements may appear oversaturated. Han et al. [2] proposed a method based on contrastive learning and generative adversarial networks to maximize the mutual information between degraded images and enhanced images. To increase adaptability, Chen et al. [47] proposed a multi-scale feature fusion network to enhance underwater images, and Cong et al. [48] combined a physical model with a generative adversarial network; however, this method achieved poor performance in enhancing images with severe degradation and non-uniform illumination. CNN-based methods obtain enhanced results by learning the mapping from degraded images to clear images. For example, Li et al. [7] constructed an underwater image dataset called UIEB and designed a simple convolutional neural network to verify the usability of the proposed dataset. To reduce the complexity of the model, Li et al. [6] designed a lightweight CNN model based on underwater scene priors. Jiang et al. [49] proposed a lightweight cascade network based on a Laplacian pyramid. Jiang et al. [50] proposed a lightweight real-time underwater image enhancement network with only 9k parameters. Similarly, Liu et al. [51] proposed a 'Best of Both Worlds' super lightweight network for enhancing underwater images. In order to merge the advantages of physical models and CNNs, Li et al. [5] proposed using transmission maps to guide the decoder network for feature reconstruction. To solve the problem that real high-quality underwater images cannot be obtained, Fu et al. [4] designed a probabilistic network to learn the enhanced distribution of underwater images, and introduced Monte Carlo likelihood estimation (MC) and Maximum Probability estimation (MP), two consistency processes, to predict the final result. To reduce the dependence on paired datasets, Zhou et al. [53] proposed hybrid comparative learning regulation, which increased the model's generalization ability. In addition, they proposed using the luminance features of underwater images under mixed illumination to guide consistent enhancement of similar brightness regions, and using complementary colors to guide color correction [54]. To better reduce the color deviation of underwater images, Lin et al. [55] introduced knowledge transfer into the color correction network and achieved good results.

Deep learning-based methods rely on large-scale pairwise underwater datasets for training and fail to consider the complex underwater imaging process. Thus, models trained on one dataset show unsatisfactory performance on other datasets.

B. Transformer

A popular model in the field of natural language processing, the Transformer was first proposed in [56]. This model is based on the attention mechanism and has achieved superior performance on machine translation tasks. In recent years, the original model and its variants have been used by an increasing number of scholars to complete vision-related tasks, such as image classification [57], target detection [66], image synthesis [67] and image restoration [62], [63], [68], and have achieved surprising results. In addition, remarkable results have been achieved in underwater image enhancement. For example, Peng et al. [60] introduced the Transformer model into the image enhancement task for the first time. This model included a channel multi-scale feature fusion module and spatial global feature modeling module, forcing the network to focus on color channels and areas with severe spatial attenuation. To reduce manual traces, Tang et al. [69] used Neural Architecture Search to automatically find an automatic U-Net structure that is beneficial for underwater image enhancement, and proposed an optional Transformer structure, thus improving the learning ability of deep models. However, due to the lack of inductive bias in this model, achieving ideal augmentation results requires training on a fairly large dataset. To save computational effort, Liu et al. [58] proposed the Swin Transformer model, which calculates self-attention in non-overlapping sliding windows while also allowing cross-window connections. This type of hierarchical structure exhibits the flexibility to model at various scales and has linear computational complexity relative to image size, so it is used in many underwater imaging tasks. For example, Huang et al. [70] proposed an adaptive group attention mechanism and applied it to the Swin Transformer model, achieving good enhancement results; however, the model is less robust, and the underwater image processing effect in low-light scenes is poor. Ren et al. [59] combined a CNN with a Swin Transformer to improve the ability to capture global dependencies and local dependencies.

Due to the lack of prior knowledge of the underwater imaging process, the results produced by these methods sometimes lead to overexposure or under compensation of the red channel.

III. Proposed Method

In this section, the overall pipeline, TAFormer, and symmetric hierarchical network architecture for underwater image enhancement are first introduced. Then, the definition of transmission and its estimation method used in this paper are

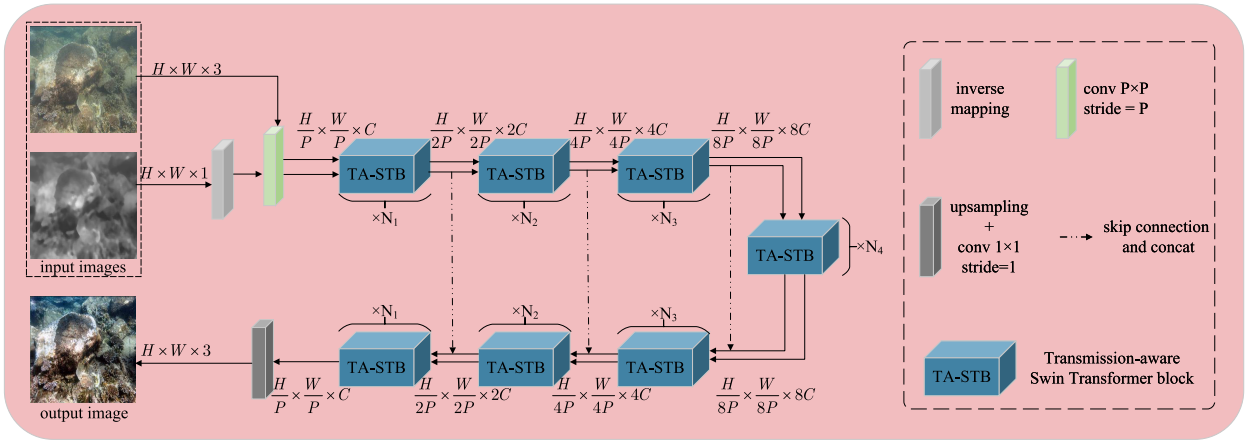

Fig. 3. The transmission-aware Swin Transformer architecture (TAFormer) for underwater image enhancement mainly includes 7 transmission-aware Swin Transformer blocks, with the top three are used to map inputs to deeper feature spaces, the middle one is used to learn useful features, and the bottom three are used to reconstruct images from feature spaces.

introduced briefly. Next, a transmission-aware Swin Transformer block (TA-STB), as the basic building block of this method, is described. It contains two key elements: transmission guided multihead self-attention (TG-MSA) and a transmission guided spatio-frequency domain interactive feed-forward network (TG-SFDI-FFN). Finally, a brief description of the loss function required for network training is provided.

A. Overall Pipeline

The overall structure of the proposed TAFormer, which is a hierarchical encoder-decoder framework, is shown in Fig. 3. The proposed network consists of three main parts: an encoder, a bottleneck and a decoder. The encoder includes 3 transmission-aware Swin Transformer Layers (TA-STLs). Each TA-STL contains a continuous even number of transmission-aware Swin Transformer Blocks (TA-STBs), which can map a given input to a deeper feature domain while reducing the spatial dimensions. The bottleneck is located at the bottom of the network model, allowing the network to learn more useful deep features while keeping the spatial dimensions unchanged. The decoder also includes 3 TA-STLs, which increase the spatial dimensions of the image while reducing the channel dimensions to reconstruct underwater images.

The specific expression is to produce an underwater image $X \in \mathbb{R}^{H*W*3}$ and its corresponding transmission $T \in \mathbb{R}^{H*W*1}$, where $H * W$ is the image resolution. First, reverse mapping is performed on the transmission to obtain the reverse transmission (RT), and then, non-overlapping image patch embedding is performed using the P $\times$ P convolution for the underwater images and RT, resulting in shallow underwater image features and RT features, respectively, where P is the size of the block. Then, the features are input into TA-STLs to extract rich spatial variation feature distributions. To tap the multi-scale representation of image degradation, each level of the encoder-decoder pipeline contains its own specific spatial resolution and channel resolution. For downsampling and upsampling, pixel-unshuffle and pixel-shuffle operations are used. Similar to [59] and [60], the characteristics of a jump-connection are added to the encoder and decoder to achieve stable training. Different from the standard attention from [58], TG-MSA is developed to achieve the embedding of degenerate image physical attributes, which aims to achieve more targeted execution feature aggregation. In addition, the TG-SFDI-FFN is proposed in TA-STB, which uses transmission as a feature selector to enrich the local-global information of the spatio-frequency domain to better improve the image quality.

B. Transmission Estimation

According to the underwater imaging model proposed by McGladery [27] and Jaffe [28], the underwater imaging process can be divided into three parts: direct radiation, forward scattering, and backward scattering. Due to the camera being too close to the underwater scene and ignoring forward scattering, the underwater imaging model can be simplified as the following formula:

$$X^c(x) = Y^c(x) \odot T(x) + A^c \odot (1 - T(x)), c \in \{r, g, b\}, \tag{1}$$

where x is the pixel index, $X(x)$ means the captured underwater image, $Y(x)$ means a real underwater scene, A is the global background light, $\odot$ denotes pixel multiplication, and $T(x)$ stands for a medium transmission map, representing the percentage of scene light that reaches the camera after passing through the medium transmission, and implicitly indicates the degree of quality degradation in different areas of the image.

Considering that the underwater imaging process is similar to the outdoor hazy imaging process, most researchers use the technique based on DCP [29] to estimate the transmission map. The global background light is usually determined by the brightest pixel in the captured image. According to DCP [29], the medium transmission map can be estimated as follows:

$$T(x) = 1 - \min_{y \in \Omega(x)} \left(\min_{c} \frac{X^c(y)}{A^c} \right), \tag{2}$$

where $\Omega(x)$ represents a local block centered on x, which is set to a size of 15 $\times$ 15.

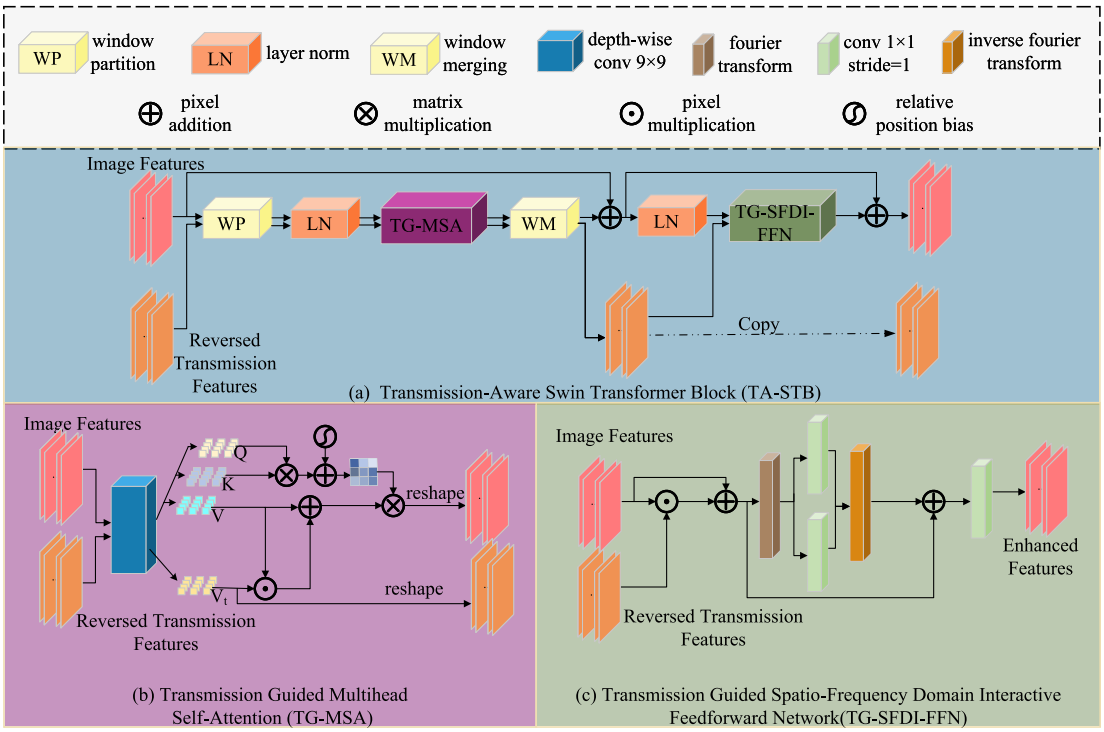

Fig. 4. Detailed structure of TAFormer. Among them, (a) is a basic unit of TAFormer, namely TA-STB, which mainly includes window partition (WP), two layer normalization (LN), transmission guided multi-head self-attention (TG-MSA), window merging (WM), and transmission guided spatio-frequency domain interactive feedforward network (TG-SFDI-FFN). (b) provides a detailed network structure for TG-MSA. (c) provides a detailed network structure for TG-SFDI-FFN.

According to the Beer-Lambert law, the attenuation process of the transmission can usually be expressed as $T(x) = e^{-\beta D(x)}$, where $D(x) \geq 0$ represents the distance from the camera to the target object and β is the spectral volume attenuation coefficient. According to [29], when $\min_c \left(\min_{y \in \Omega(x)} \frac{I^c(y)}{A^c} \right) \to 1$, $T(x)$ tends towards 0, which is not always accurate. Taking inspiration from [34], we estimate $T(x)$ using Eq.3, which considers all possible relationships between A^c and $X^c(x)$, making the estimated transmission map not only more accurate, but also more robust and universal. A^c is estimated from depth-dependent color channels. Due to limited space, it is recommended that readers consult the literature [34] for a more detailed process.

$$T(x) = \max_{c, y \in \Omega(x)} \left(\frac{|A^c - X^c(y)|}{\max(A^c, 1 - A^c)} \right). \tag{3}$$

C. Transmission-Aware Swin Transformer Block

Previous deep learning-based methods rely on CNNs and have limitations in capturing long-distance dependencies. The traditional Transformer [58], [59] relies on the Swin Transformer Block to obtain local attention, or combines convolution with the Swin Transformer to obtain more local attention. The former ignores local dependencies, while the latter merges global and local dependencies. However, using only data-driven methods to calculate the self-attention of degraded images can easily overlook the physical properties of underwater images during the imaging process, which is not conducive to image restoration because the attenuation degree of different regions in the same image varies. To overcome this limitation, inspired by [59], a transmission-aware Swin Transformer block is developed as a feature extraction unit by combining deep convolution and a Swin Transformer, while utilizing the physical properties of underwater imaging processes, as shown in Fig. 4(a).

More precisely, TG-MSA includes two attention calculation strategies, namely, transmission-guided window multihead self-attention and transmission-guided sliding window multihead self-attention, denoted as TG-MSA1 and TG-MSA2, respectively. TG-MSA2 can be understood as a combination of a sliding window and TG-MSA1 with a masking mechanism.

Specifically, given the image features and RT features of the i-1th block, $F_X^{i-1} \in \mathbb{R}^{H*W*C}$ and $F_T^{i-1} \in \mathbb{R}^{H*W*C}$, since each TA-STL contains a continuous even number of TA-STBs, the encoding process for two consecutive TA-STBs can be defined as:

$$\begin{aligned}
F_T^{i\,\prime} &= LN(WP(F_T^{i-1})), \\
F_X^{i\,\prime} &= WM(TG-MSA1(LN(WP(F_X^{i-1})), F_T^{i\,\prime})) + F_X^{i-1}, \\
F_T^{i} &= WM(DSC(F_T^{i\,\prime})), \\
F_X^{i} &= TG-SFDI-FFN(LN(F_X^{i\,\prime}), F_T^{i}) + F_X^{i\,\prime}, \\
F_T^{i+1\prime} &= LN(WP(F_T^{i})), \\
F_X^{i+1\prime} &= WM(TG-MSA2(LN(WP(F_X^{i})), F_T^{i+1\prime})) + F_X^{i}, \\
F_T^{i+1} &= WM(DSC(F_T^{i+1\prime})), \\
F_X^{i+1} &= TG-SFDI-FFN(LN(F_X^{i+1\prime}), F_T^{i+1}) + F_X^{i+1\prime},
\end{aligned} \tag{4}$$

(5)

where $WP(\cdot)$ represents the window partition operation, $LN(\cdot)$ represents the layer normalization operation, and $WM(\cdot)$ represents the window merging operation. $TG-SFDI-FFN(\cdot)$ is a spatio-frequency domain interactive feedforward network guided by transmission. Considering that the RT features are only subjected to depthwise separable convolution in TG-MSA, we use the $DSC(\cdot)$ to represent depthwise separable convolution for $F_T^{i\prime}$ and $F_T^{i+1\prime}$ in Eq. 4 and Eq. 5, instead of the $TG-MSA(\cdot)$, which can be more clearly seen in Fig. 4(b).

1) TG-MSA: Due to the fact that different regions of the same image may have different degrees of degradation, inspired by [5], which indicates that the quality degradation of underwater images can be reflected through the transmission, transmission features are introduced to guide the calculation of self-attention. As shown in Fig. 4(b), the RT features that have undergone window partitioning and layer normalization operations are transmitted to the corresponding TA-STB's TG-MSA to facilitate the calculation of self-attention. To enhance the local attention in each patch, deep separable convolution is first introduced to encode the spatial and channel context information of degraded images, rather than a linear layer. In addition, in order to force the network to focus on areas with significant degradation, deep separable convolution for RT is also implemented.

Specifically, for the input degraded image feature $F_X^i \in \mathbb{R}^{H*W*C}$ and RT feature $F_T^i \in \mathbb{R}^{H*W*C}$, the WP operation first divides them into multiple windows. Next, we perform LN on them to obtain $F_{X-LN}^i \in \mathbb{R}^{Num*Win^2*C}$ and $F_{T-LN}^i \in \mathbb{R}^{Num*Win^2*C}$, where *Num* is the number of windows and *Win* is the window size. Then, deep separable convolution is implemented to obtain the query matrix $Q \in \mathbb{R}^{Num*Head*Win^2*(C/Head)}$, key matrix $K \in \mathbb{R}^{Num*Head*Win^2*(C/Head)}$, and value matrix $V \in \mathbb{R}^{Num*Head*Win^2*(C/Head)}$ of the image feature map. At the same time, the value matrix $V_T \in \mathbb{R}^{Num*Head*Win^2*(C/Head)}$ of the RT features is obtained, which can extract more feature information without increasing the number of parameters. Finally, self-attention is calculated.

$$TG-MSA1(Q,K,V,V_T) = soft\max\left(\frac{Q \otimes K^T}{\sqrt{C}} + pos\right) \otimes (V \oplus V \odot V_T), \tag{6}$$

where $pos \in \mathbb{R}^{Head*Win^2*Win^2}$ is the relative position deviation, K^T represents a transposed matrix of K, $\otimes$ is the matrix multiplication, $\oplus$ and $\odot$ represent pixel addition and pixel multiplication, respectively.

Due to the continuous and pairwise nature of TA-STB, $TG-MSA1(\cdot)$ is replaced by $TG-MSA2(\cdot)$ in another TA-STB, which can be considered a combination of cyclic shift and multi-head self-attention based on a window (TG-MSA1) with a masking mechanism.

2) TG-SFDI-FFN: Previous studies [60], [61], [70] have directly introduced single-scale or multi-scale convolution of the spatial domain into the feedforward neural network to enhance locality, often ignoring the large amount of global information existing in the Fourier domain, as recent researches [71], [72], [73] have shown that the global background light and structural information of underwater images can be partially decomposed in the Fourier domain. As shown in Fig. 5, first, a Fourier transform is performed on the underwater image, and then an inverse Fourier transform is performed on the obtained amplitude and phase. The phase contains the structural information of the image, and the amplitude contains the background light of the image.

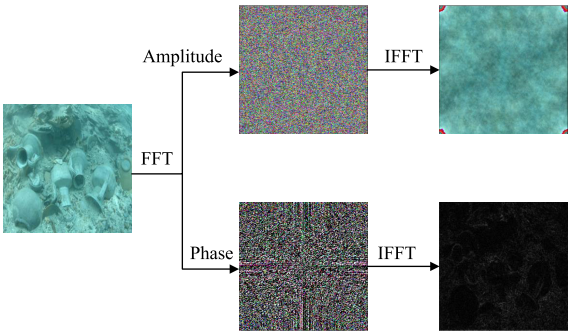

Fig. 5. The representation of underwater images in the Fourier domain. The result of applying inverse Fourier transform to the phase alone reveals the structural information, while the result of applying inverse Fourier transform to the amplitude alone reveals the global background light information.

In this paper, a transmission-guided spatio-frequency domain interactive feedforward network is designed, as shown in Fig. 4(c). During the transmission process, the RT is used for pixel level attention, and larger attention weights are assigned to pixels with higher quality degradation. The weighted image feature map is further decomposed by Fourier transform, and the decomposed amplitude and phase components are extracted separately. Finally, the inverse transform is used to transfer the data to the spatial domain.

Specifically, given an input image tensor $F_X^i \in \mathbb{R}^{H*W*C}$ and an RT tensor $F_T^i \in \mathbb{R}^{H*W*C}$, the RT is used as a feature selector to weight the importance of different spatial positions in the image feature map, resulting in a weighted image feature map. The formula is as follows:

$$F_{TG-X}^i = F_X^i \oplus F_X^i \odot F_T^i. \tag{7}$$

The weighted image features are transferred to the frequency domain through Fourier transform to obtain the amplitude and phase components of the image features. A 1×1 convolution is performed on the amplitude and phase components for feature extraction, and then the features are transferred to the spatial domain through an inverse Fourier transform, as shown in the following formula:

$$\begin{aligned} F_{mag-X}^i, F_{pha-X}^i &= fft(F_{TG-X}^i) \\ F_{S-X}^i &= ifft(conv(F_{mag-X}^i), conv(F_{pha-X}^i)). \end{aligned} \tag{8}$$

where $fft(\cdot)$ and $ifft(\cdot)$ represent the fast Fourier transform and the inverse fast Fourier transform, respectively.

Residual connection is performed on the features before and after Fourier transform, and then 1×1 convolution is performed to obtain the output image features of the TG-SFDI-FFN. The specific formula is as follows:

$$F_X^{i+1} = conv(\tau \odot F_{S-X}^i \oplus (1-\tau) \odot F_{TG-X}^i), \tag{9}$$

where τ is the superparameter used to control the spatial frequency domain information exchange ratio, which is set to 0.1. The RT weight is regarded as an identical connection, which not only avoids gradient vanishing but also tolerates errors caused by inaccurate transmission estimation. In addition, aggregating features through spatial and frequency domain interactions can not only effectively improve the local details of the image, but also facilitate further color compensation.

D. Loss Function

Following [59] and [69], to balance the high- and low-frequency information and structure of images more robustly, we construct the following loss function.

$$\ell_{total} = \ell_{char} + \kappa_1 \ell_{edge} + \kappa_2 \ell_{ms-ssim}, \tag{10}$$

where κ_1 and κ_2 are used to balance the size between loss values, so they are set to 1 and 2 respectively. ℓ_{char} is the Charbonnier loss, which can be expressed as:

$$\ell_{char} = \sqrt{\| Y - G\|^2 + \varepsilon^2}, \tag{11}$$

where Y and G represent the reconstructed image and the real reference image, respectively, and ε is set to 0.001 based on experience. In addition, the edge loss ℓ_{edge} is designed to constrain the consistency of edge textures, which can be represented as:

$$\ell_{edge} = \| \mathbb{C}(Y) - \mathbb{C}(G)\|_1, \tag{12}$$

where $\| \cdot \|_1$ represents the L1 loss, following [74], and $\mathbb{C}(\cdot)$ represents an edge extraction operation, which is used to extract edge information from reconstructed images and real reference images. Finally, because the MS-SSIM measures the detailed information of images at different scales, which is closer to the subjective quality evaluation results, we design the MS-SSIM loss as follows:

$$\ell_{ms-ssim} = 1 - MS - SSIM(Y, G). \tag{13}$$

IV. Experiments

In this section, the training details and experimental settings of TAFormer are first specified. Then, the proposed method is compared with 15 representative algorithms subjectively and objectively, and a series of ablation studies are performed to verify each component of TAFormer. Finally, the expansion experiments, application experiments and limitations analysis are presented at the end of this section.

A. Training Details

In our framework, $\{N_1, N_2, N_3, N_4\}$ is set to $\{8, 8, 8, 8\}$, and the number of attention heads for each TA-STB is set to 8. The window size is set to 8, and the patch size P is 2. The initial number of channels C is 32, and the expansion ratio in the TG-SFDI-FFN is set to 2. The dropping ratio in the skip path is 0.1. During the training process, the Adam optimizer with a batch size of 2 for 800 epochs is used. The initial learning rate is set to 5e-4. In addition, the cosine annealing learning rate decay strategy is synergized for training with warm-up epoch 3. The entire model is run on Python 3.7 and PyTorch 1.12 with NVIDIA GeForce GTX 1080Ti and 12 GB GPU memory servers, which do not require pretraining.

TABLE I

The Average MSE, PSNR, SSIM, UIQM, and UCIQE of Different Methods on the UIEB90 Test Set. The Red Color Represents the Optimal Value of the Indicator

Dataset		UIEB90				
Methods		MSE↓	PSNR↑	SSIM↑	UIQM↑	UCIQE↑
I	ACDC	1.30	17.91	0.69	1.48	0.56
	BR	1.15	18.57	0.71	1.51	0.59
II	RDCP	1.30	18.79	0.75	1.37	0.58
	IBLA	2.00	16.47	0.64	1.40	0.60
III	CWR	1.36	17.55	0.66	1.27	0.52
	FUnIE	1.69	16.99	0.56	1.37	0.56
	PUIE-MC	0.48	22.15	0.84	1.33	0.58
	PUIE-MP	0.56	21.38	0.81	1.28	0.56
	Ucolor	0.67	21.16	0.80	1.32	0.57
	UWCNN-II	3.51	13.81	0.44	1.21	0.49
IV	URSCT	0.61	21.85	0.85	1.42	0.60
	U-shape	0.73	21.02	0.79	1.34	0.58
	Uformer	0.49	22.31	0.86	1.41	0.61
	Restormer	0.50	23.37	0.86	1.39	0.61
	SwinIR	1.05	19.27	0.75	1.37	0.57
	Ours	0.48	22.72	0.86	1.42	0.61

B. Experimental Settings

1) Datasets: Five benchmark datasets are used to evaluate the performance of our method compared with other methods: (1) EUVP [3], is an underwater image dataset containing a large number of paired and unpaired images. The paired underwater scene dataset contains 2185 pairs of underwater images, 1680 pairs of images as training sets (EUVP1680) and 505 pairs of images as test sets (EUVP505). (2) UIEB [7], which includes 890 pairs of real underwater images with references and 60 challenging real underwater images without reference; 800 pairs of images are used as training sets (UIEB800), 90 pairs of images are used as test sets (UIEB90), and 60 challenging images are used as test sets (UIEB60). (3) UFO-120 [8] is a large-scale SESR training dataset, that includes more than 1500 training samples and 120 test samples (UFO120). (4) SQUID [75] contains 57 stereo pairs from 4 different locations in Israel. Due to the similar attenuation phenomenon of these images, we randomly selected 30 images as the test set (SQUID30). (5) Color-Check7 [9], which contains seven underwater Color-Checker images taken by different cameras as test sets CC7, is used to evaluate the accuracy of underwater image color correction.

2) Comparison Methods: TAFormer is compared with 15 representative underwater image enhancement methods. These methods include two methods based on pixel adjustment (type I: ACDC [12], BR [25]), two methods based on imaging models (type II: RDCP [31], IBLA [33]), six methods based on deep learning (type III: CWR [2], FUnIE [3], PUIE-MC [4], PUIE-MP [4], Ucolor [5], UWCNN-II [6]), and five methods based on the Transformer model (type IV: URSCT [59], U-shape [60], Uformer [61], Restormer [62], SwinIR [63]).

3) Evaluation Metrics: In this paper, three full reference evaluation indicators, MSE, PSNR and SSIM, are used, which reflect the relationship between the reconstructed image and the real reference image, as well as three non-reference

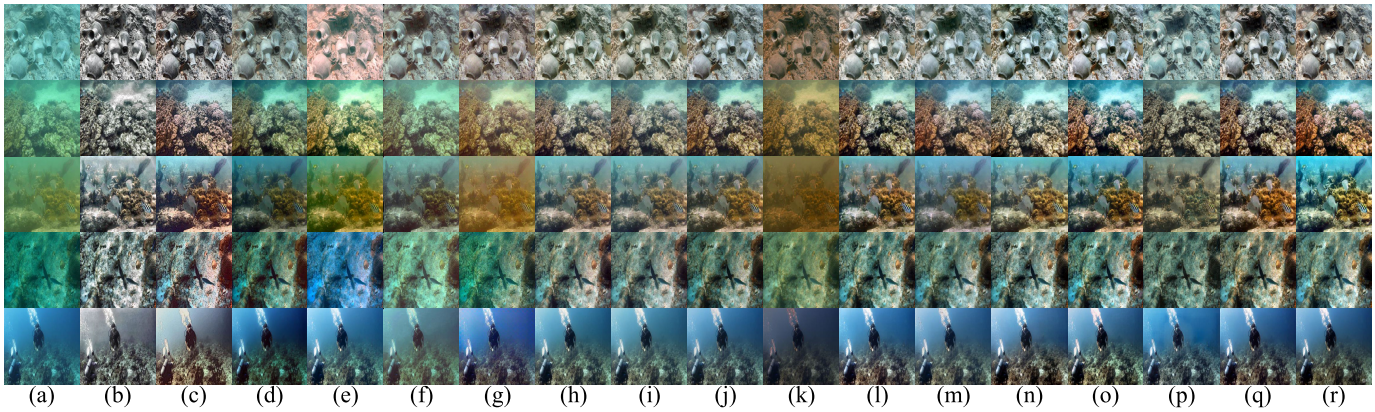

Fig. 6. Reconstruction results from all comparison methods on UIEB90 test set. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours. (r) References.

TABLE II

THE AVERAGE MSE, PSNR, SSIM, UIQM, AND UCIQE OF DIFFERENT METHODS ON THE UFO120 AND EUVP505 TEST SETS. THE RED COLOR REPRESENTS THE OPTIMAL VALUE OF THE INDICATOR

Datasets		UFO120					EUVP505				
Methods		MSE↓	PSNR↑	SSIM↑	UIQM↑	UCIQE↑	MSE↓	PSNR↑	SSIM↑	UIQM↑	UCIQE↑
I	ACDC	2.60	14.70	0.52	1.47	0.57	2.46	14.90	0.54	1.47	0.56
	BR	2.39	15.33	0.53	1.47	0.59	2.17	15.70	0.55	1.46	0.58
II	RDCP	0.68	20.43	0.70	1.37	0.59	0.66	20.80	0.73	1.39	0.59
	IBLA	2.03	17.64	0.65	1.43	0.63	1.41	18.67	0.71	1.43	0.62
III	CWR	1.87	16.32	0.55	1.28	0.51	1.64	17.14	0.58	1.33	0.53
	FUnIE	0.50	22.14	0.78	1.42	0.60	0.44	22.42	0.80	1.42	0.59
	PUIE-MC	1.01	19.02	0.72	1.36	0.60	0.86	19.55	0.75	1.37	0.60
	PUIE-MP	1.04	18.92	0.71	1.32	0.58	0.89	19.45	0.73	1.32	0.57
	Ucolor	0.70	20.43	0.77	1.34	0.59	0.62	20.87	0.80	1.35	0.59
	UWCNN-II	1.75	17.08	0.63	1.27	0.52	1.43	17.72	0.69	1.28	0.52
IV	URSCT	0.14	27.42	0.92	1.42	0.58	0.12	28.30	0.94	1.38	0.57
	U-shape	0.34	23.95	0.84	1.34	0.58	0.36	23.79	0.85	1.36	0.57
	Uformer	0.13	27.60	0.92	1.37	0.58	0.13	27.87	0.93	1.37	0.57
	Restormer	0.13	28.09	0.92	1.50	0.58	0.12	28.28	0.93	1.46	0.58
	SwinIR	0.19	26.69	0.89	1.37	0.59	0.17	26.90	0.91	1.38	0.58
	Ours	0.12	27.98	0.92	1.41	0.59	0.11	28.51	0.94	1.39	0.59

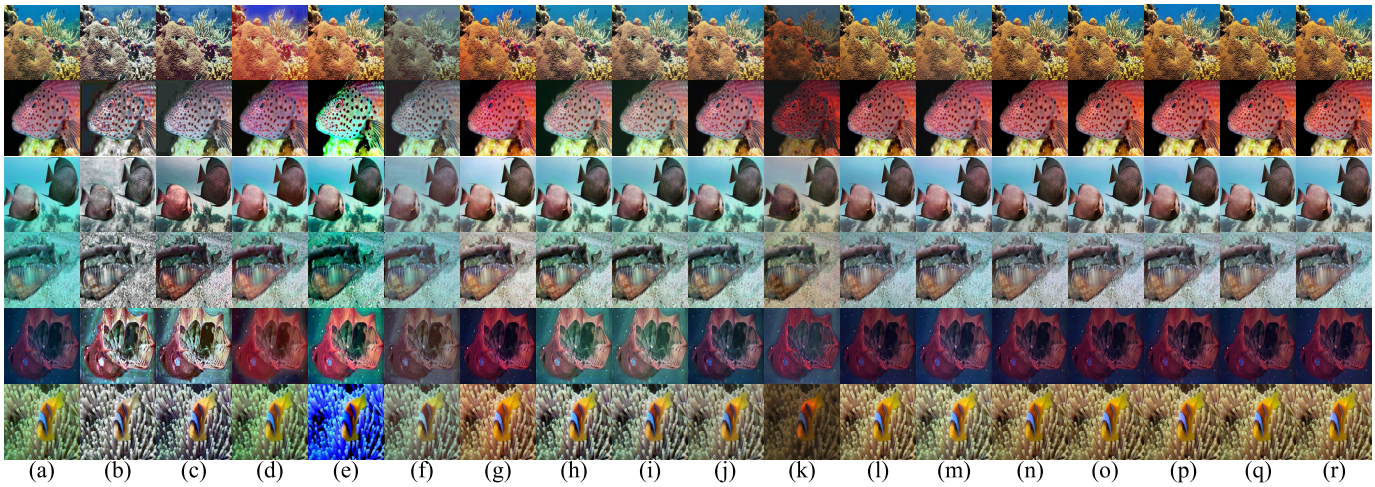

Fig. 7. Reconstruction results from all comparison methods on UFO120 and EUVP505 test sets. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours. (r) References.

evaluation indicators, UCIQE [76], UIQM [77], and PS (perception score), which reflect the quality of the reconstructed image. The lower the MSE is, the greater the similarity between the reconstructed image and the reference image

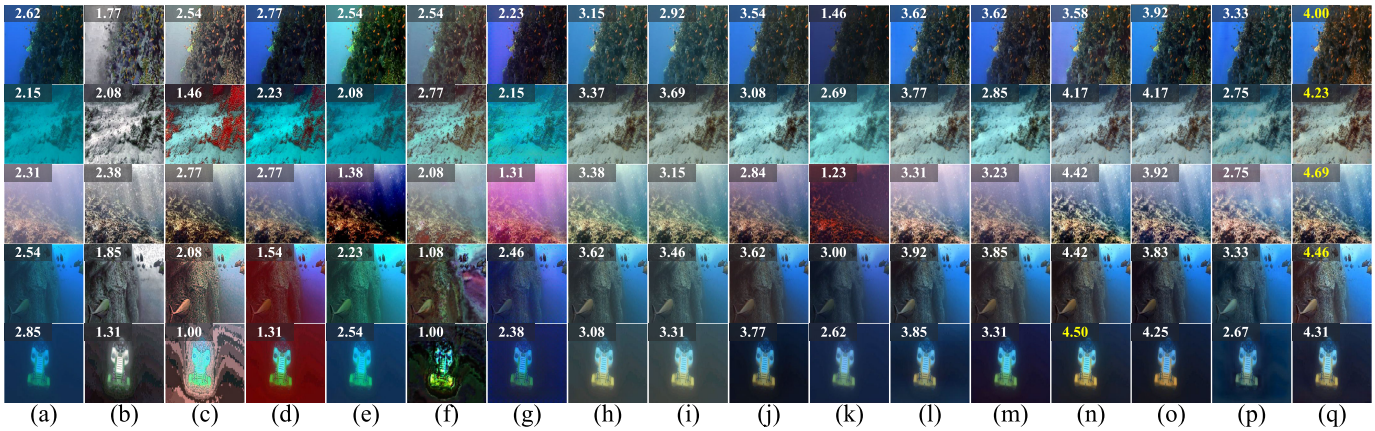

Fig. 8. Reconstruction results from all comparison methods UIEB60 test set. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours.

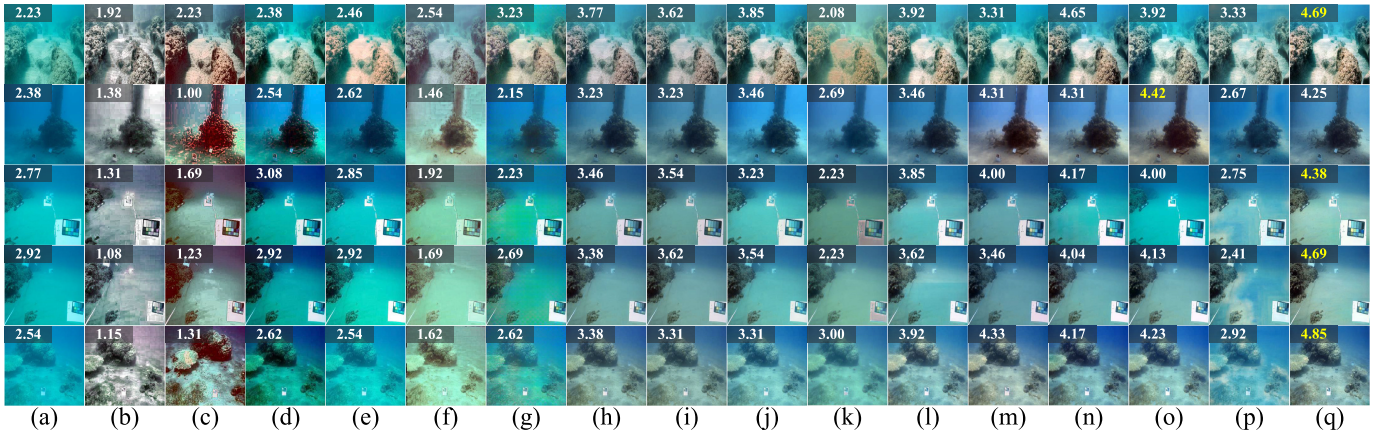

Fig. 9. Reconstruction results from all comparison methods on SQUID30 test set. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours.

is, and the higher the PSNR is, the more information the reconstructed image contains. The higher the SSIM value is, the greater the structural and textural similarity of the reconstructed image. The higher the UCIQE is, the better the color concentration, saturation, and contrast of the image are. The higher the UIQM value is, the better the color richness, clarity, and contrast of the image are. However, because UIQM and UCIQE cannot accurately measure the quality of reconstructed images in some cases, we invite 15 volunteers with experience in underwater image processing to rate the image quality enhanced by each method, with a PS ranging from 1-5, where higher scores represent higher quality images.

For test sets with reference images, both full reference evaluation indicators and non reference evaluation indicators are used. For test sets without reference images, only three non reference evaluation indicators are used. Notably, for CC7, CIEDE2000 [78] is used to measure the degree of color restoration in images. A smaller score indicates a more natural recovery outcome.

C. Comparisons on Full Reference Underwater Images

Full reference and non reference evaluations on the UIEB90, UFO120, and EUVP505 datasets are conducted, and the subjective perception results are shown in Fig. 6 and Fig. 7. The objective evaluation indicators are shown in Table I and II.

For the UIEB90 test set, we select five scenes with hazy, green, yellow-green, blue-green, and blue tones, and the enhancement results are shown in Fig. 6. The enhancement results of the RDCP, IBLA, CWR, FUnIE, and UWCNN-II methods are unstable, resulting in varying degrees of color deviation in the reconstruction results, further reducing image quality. The ACDC and BR methods can effectively remove color deviation by adjusting the pixel values, but they also reduce the richness of the images' colors. Although PUIE, Ucolor, U-shape and SwinIR effectively improve image quality, they still exhibit low saturation. URSCT, Uformer, Restormer and our method produce good enhancement results, but our method produces high contrast in local areas of the image. The objective evaluation indicators are shown in Table I. Our method's reconstruction results achieve the best scores on three evaluation indicators: MSE, SSIM, and UCIQE. Our method also outperforms all methods in type III. Although most of the full-reference metrics of the Restormer are also the highest, it can be seen from Fig.6 that its contrast is lower than that of our method. This demonstrates the effectiveness of incorporating underwater imaging processes and convolutions into the Transformer.

TABLE III
THE AVERAGE UIQM, AND UCIQE OF DIFFERENT METHODS ON THE UIEB60 AND SQUID30 TEST SETS. THE RED COLOR REPRESENTS THE OPTIMAL VALUE OF THE INDICATOR

	Datasets	UIEB60			SQUID30		
	Methods	UIQM↑	UCIQE↑	PS↑	UIQM↑	UCIQE↑	PS↑
	Input	1.06	0.48	2.54	0.79	0.40	2.58
I	ACDC	1.40	0.55	1.75	1.23	0.52	1.42
	BR	1.42	0.56	1.77	1.34	0.60	1.54
II	RDCP	1.27	0.56	1.96	1.15	0.57	2.73
	IBLA	1.32	0.57	2.35	0.95	0.48	2.71
III	CWR	1.27	0.52	1.85	0.98	0.48	1.91
	FUnIE	1.29	0.53	2.31	0.88	0.48	2.58
	PUIE-MC	1.16	0.55	3.31	0.88	0.51	3.46
	PUIE-MP	1.11	0.53	3.35	0.85	0.50	3.51
	Ucolor	1.17	0.54	3.51	0.87	0.50	3.51
	UWCNN-II	1.07	0.48	2.44	0.76	0.44	2.31
IV	URSCT	1.23	0.56	3.79	0.87	0.51	3.71
	U-shape	1.20	0.54	3.41	0.88	0.52	3.77
	Uformer	1.25	0.57	4.17	0.96	0.55	4.29
	Restormer	1.23	0.56	4.04	0.92	0.52	4.12
	SwinIR	1.24	0.53	3.02	0.85	0.47	2.79
	Ours	1.30	0.57	4.25	0.99	0.54	4.52

For the test sets UFO120 and EUVP505, due to the overlap between them, we select three images from each dataset and compare their subjective perception effects, as shown in Fig. 7. Similarly, the reconstruction results of the UWCNN-II method are generally darker and have lower contrast. The ACDC and BR methods can improve image quality by adjusting image pixel values, but their results produce a single color and distortion authenticity. The enhancement results of RDCP and IBLA are unstable, and prone to artifacts and the new color shifts. Although the CWR, PUIE-MC, PUIE-MP, and Ucolor methods can alleviate degradation to some extent, there is still a problem of low saturation. The FUnIE, and Transformer-based methods achieved good enhancement results, but the details of the U-shape and SwinIR results are poor, and the FUnIE results are slightly reddish. Table II reveals the objective evaluation indicators of all underwater image enhancement methods on these two datasets. Our method achieves the best results in terms of three indicators: MSE, PSNR, and SSIM. For the non reference evaluation metrics UIQM and UCIQE, most of the type I or II methods achieve the best results. Based on the subjective and objective evaluation results, it can be concluded that UIQM and UCIQE are not sensitive to color deviation or artifacts. Therefore, although they have value, they cannot be used as an absolute basis for evaluating the enhancement of underwater images.

D. Comparisons on Non-Reference Underwater Images

Non-reference evaluations on the UIEB60 and SQUID30 datasets are conducted, and the subjective perception results are shown in Fig. 8 and Fig. 9. The objective non reference indicators are shown in Table III.

Fig. 8 and Fig. 9 both show that most of our method's reconstruction results have the highest perception scores, which reflects that our method produces the most consistent representation of human visual perception. Because the two test sets are taken in extremely harsh environments, image degradation is severe, and the reconstruction results of some comparison methods are not always satisfactory. For example, the results of the BR, RDCP, IBLA, CWR, FUnIE, and UWCNN-II methods are extremely unstable, and prone to severe artifacts, excessive red channel compensation, and the introduction of new color biases. Although ACDC can remove color deviation, it also eliminates the richness of the colors in the image. The enhancement results of the two PUIE methods have a high degree of similarity. Although they can eliminate color deviation to some extent, they still have problems of low contrast and low saturation, similar to SwinIR. Although Ucolor also introduces transmission guidance in the network, it still ignores the long-distance dependencies of the network, leading to the problem of overall color deviation in the image. The Transformer-based methods URSCT, U-shape, Uformer and Restormer, have better reconstruction results, but there is still a problem of low local area contrast. Overall, the reconstruction results of our method show more natural colors and higher contrast.

From Table III, it is clear that our method achieved the best score in PS, which proves that considering the underwater imaging process and integrating it into the Transformer model has a better generalization ability for various real underwater scenes. Specifically, for some comparison methods, the PS score of their enhancement results is lower than that of the original image, which proves that the enhancement results of some methods are contrary to human perception. In addition, on the UIEB60 test set, our method, along with the IBLA method, achieve the highest UCIQE score, and on the SQUID30 test set, our method has the third highest in UCIQE score. It is worth noting that the UIQM and UCIQE metrics of no data-driven methods are generally higher than those of data-driven methods. Apart from the four methods of types I and II, our UIQM and UCIQE evaluation metrics are the highest among the eight methods of types III and IV. This reult further proves that our method outperforms deep learning-based methods and traditional Transformer methods.

E. Comparisons on Color Checker Images

To further verify the accuracy of the proposed network in color correction of underwater images, the color correction levels of 13 methods are compared on the CC7 dataset. Fig. 10 reveals that the professional Pentax W80 camera introduces color deviation when shooting underwater scenes. Among the compared methods, our method obtains the lowest CIEDE2000 score for correction, which means that our method has the most accurate color correction ability. The ACDC, BR, IBLA, CWR, and UWCNN-II methods exhibit color distortion. The FUnIE, PUIE-MP, Ucolor, and U-shape methods suffer from excessive red compensation, leading to image redness and low saturation. Although the RDCP, PUIE-MC, and URSCT methods can eliminate the influence of color shift to some extent, there is still a problem of low saturation. From Table IV, it can also be seen that the Transformer-based methods have a significant effect on color correction, but our method achieved

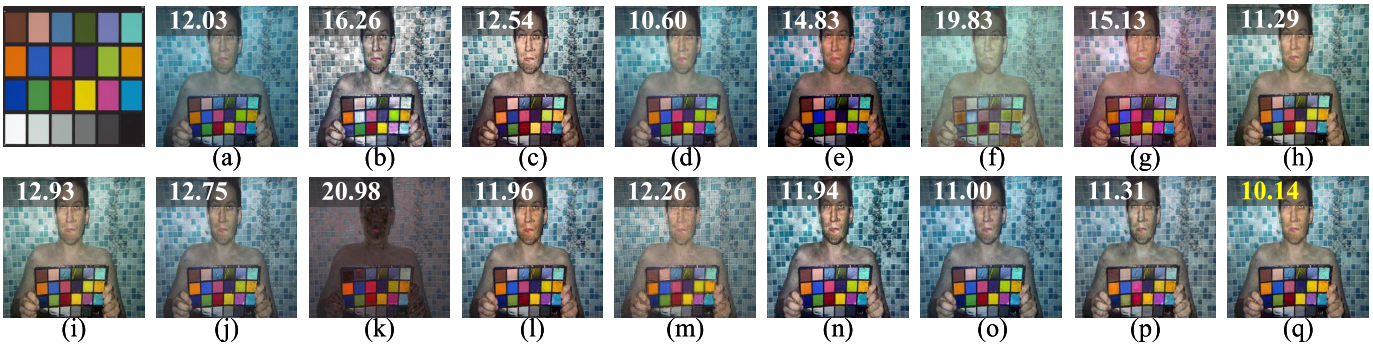

Fig. 10. Verify the color correction ability of all comparison methods on CC7. The first image is a real Macbeth ColorChecker, which is used to calculate the CIEDE2000 scores with the Macbeth ColorChecker in the enhanced results. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours.

TABLE IV
THE CIEDE2000 SCORES AND THEIR AVERAGE VALUES FOR ALL COMPARISON METHODS ON THE CC7 DATASET. RED REPRESENTS THE OPTIMAL VALUE OF THE INDICATOR

	Methods	Can D10	Fuj Z33	Oly T6000	Oly T8000	Pan TS1	Pen W60	Pen W80	Average
Original images		12.38	18.35	15.61	17.54	13.54	11.82	12.03	14.47
I	ACDC	15.27	20.70	17.13	13.03	11.48	14.32	16.26	15.45
	BR	11.82	16.71	12.02	10.11	10.89	11.01	12.54	12.16
II	RDCP	13.25	12.87	11.51	14.17	13.84	12.69	10.60	12.70
	IBLA	13.14	13.34	12.67	11.35	23.81	11.89	14.83	14.43
III	CWR	16.88	19.48	20.23	14.32	13.75	17.65	19.83	17.45
	FUnIE	15.43	22.78	13.35	14.24	10.73	15.17	15.13	15.26
	PUIE-MC	10.56	13.90	11.19	10.41	10.45	10.60	11.29	11.20
	PUIE-MP	9.97	14.63	9.51	11.34	11.25	10.50	12.93	11.45
	Ucolor	9.43	13.16	9.61	10.35	13.97	9.57	12.75	11.26
	UWCNN-II	19.47	22.95	23.02	17.75	32.00	19.41	20.98	22.23
IV	URSCT	10.97	14.31	13.24	13.11	12.16	10.59	11.96	12.34
	U-shape	8.35	13.40	9.57	13.69	12.06	9.26	12.26	11.22
	Uformer	11.53	11.49	11.12	14.06	10.57	11.75	11.94	11.78
	Restormer	12.04	12.27	11.94	12.86	11.37	9.52	11.00	11.57
	SwinIR	12.83	15.97	13.77	12.97	12.42	12.65	11.31	13.13
	Ours	11.34	12.05	11.35	11.13	10.41	11.71	10.14	11.16

the lowest score in most cases and achieved the lowest average score for all images.

F. Ablation Studies

To verify the necessity of various components and loss functions in the network, four ablation experiments are designed. First, we replace the designed TG-SFDI-FFN with an MLP network, which we refer to as the w/o TG-SFDI-FFN network. Second, we remove the multi-head self-attention component guided by the transmission, which we refer to as the w/o TG-MSA network here. Once again, to verify the necessity of edge loss, we remove the loss function and refer to it as the w/o edge loss network. Finally, in order to verify the effectiveness of the loss function MS-SSIM, we remove the loss and refer to it as the w/o ms-ssim loss network. It should be noted that based on past experience, removing the Charbonire loss is meaningless, so we avoid ablation experiments on this loss function.

Fig. 11 shows the references and original images of some images in the UIEB90 test set, as well as the reconstructed

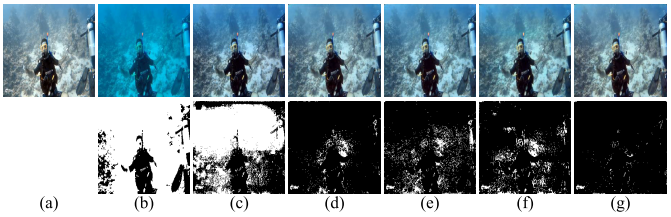

Fig. 11. Residual images between reconstruction results of different ablation networks and real reference images. (a) References. (b) Inputs and their residual images. (c) The reconstruction results of w/o TG-SFDI-FFN network and their residual images. (d) The reconstruction results of w/o TG-MSA network and their residual images. (e) The reconstruction results of w/o edge loss network and their residual images. (f) The reconstruction results of w/o ms-ssim loss network and their residual images. (g) The reconstruction results of our network and their residual images.

results and corresponding residual images of all ablation networks. Notably, the fewer white areas there are in the residual image, the closer it is to the real reference image. The results of the w/o TG-SFDI-FFN network have a more severe blue color deviation in the background area, so its residual figures have a larger area of white, indicating that removing this component produces the most different results from the

TABLE V

ANALYSIS OF ABLATION EXPERIMENTS ON DIFFERENT COMPONENTS OF TAFORMER, WITH RED INDICATING THE BEST VALUE OF THE INDICATOR

Networks	UIEB90					Params (M)	FLOPS
	MSE↓	PSNR↑	SSIM↑	UIQM↑	UCIQE↑		
w/o TG-SPDI-FFN	0.66	21.20	0.83	1.41	0.59	11.262	466.683
w/o TG-MSA	0.53	22.40	0.86	1.40	0.60	12.141	295.951
w/o edge loss	0.55	22.37	0.85	1.43	0.60	—	—
w/o ms-ssim loss	0.52	22.57	0.85	1.41	0.60	—	—
Ours	0.49	22.72	0.87	1.41	0.61	12.141	402.221

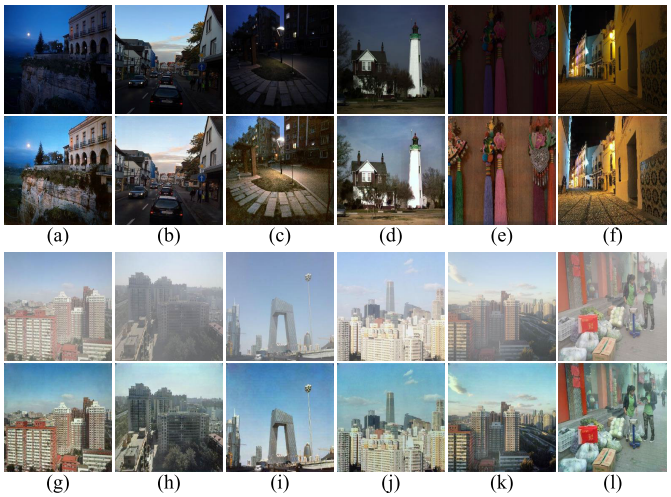

Fig. 12. Expand the proposed method to low illumination and hazy scenes. (a)-(f) are low illumination images, and (g)-(l) are hazy images. The first and third rows are the original images, while the second and fourth rows are our enhanced results.

real reference images. The reconstruction results of the w/o TG-MSA network indicate that due to the lack of guidance from the transmission representing the physical properties of the image, some defect areas may appear in the results. The reconstruction results of the w/o ms-ssim loss network and w/o edge loss network demonstrate the necessity of introducing two loss functions during the training process. In summary, the reconstruction results of our complete network exhibit the highest degree of similarity to the real reference image. The evaluation metrics of the different networks on the UIEB90 test set are shown in Table V, which shows that our network basically achieves the optimal values among all the indicators.

In addition, to demonstrate the effectiveness and necessity of the two key modules, we separately calculate the model parameters (Params) and floating point operations per second (FLOPS) of the w/o TG-SFDI-FFN, w/o TG-MSA and our networks, as shown in Table V. Although the design of the TG-SPDI-FFN slightly increases the number of model parameters, it results in a PSNR gain of 1.52 dB, and other indicators are also improved. The design of TG-MSA does not add additional parameters, and the gains of multiple evaluation metrics it brings are evident.

G. Extended Experiments

Fig. 12 validates the effectiveness of the proposed method in low illumination and hazy scenes. The first row shows the images captured in a low light scene, and the second row shows the corresponding reconstruction results. It can be seen that our method markedly improves the brightness of the image. The third row is the images captured in a hazy scene, and the fourth row is their corresponding restoration results. Our method not only effectively eliminates the influence of haze, but also significantly improves the image contrast. This result indicates that our network can also improve image quality in other challenging scenarios. It is worth noting that the networks used to enhance these two scenarios are obtained using the UIEB800 training set. when low illumination and hazy datasets are used for training, better restoration results may be obtained.

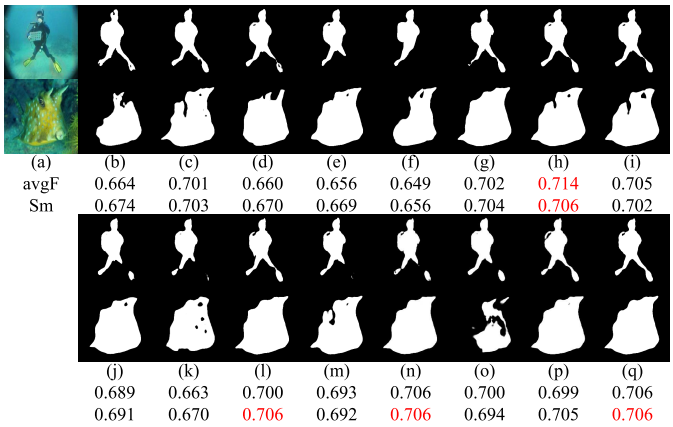

Fig. 13. Results of saliency detection. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours.

H. Application Experiments

To further illustrate the significance of the proposed method, we apply its enhanced results with all the comparison methods to the following two downstream tasks.

1) Salient Object Detection: The saliency detection method [79] is used to detect the salient areas in images. It is evident from Fig. 13 that the proposed method achieves better detection results. We use the average F-measure and Sm for quantitative evaluation, both of which indicate that the higher the score is, the better the detection effect is. Therefore, the superiority of the proposed method can be further demonstrated.

2) Keypoint Detection: The SIFT method [80] is used to detect the number of key points in all thr enhancement results, and it is evident from Fig. 14 that except for U-shape, all the Transformer based methods, including our method, detect a considerable number of key points.

I. Cross Validation and Limitation Analysis

This article uses three training sets: UIEB800, EUVP1680, and their superposition UE2480. Cross testing is conducted on the networks of the three training sets, as shown in Fig. 15. The first two rows show the enhancement results using different training sets in the EUVP505 test set, while the last two rows show the enhancement results using different training sets in

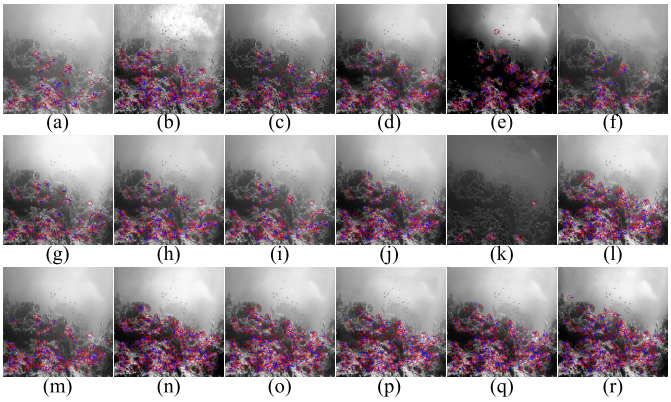

Fig. 14. Results of keypoint detection. (a) Inputs. (b) ACDC [12]. (c) BR [25]. (d) RDCP [31]. (e) IBLA [33]. (f) CWR [2]. (g) FUnIE [3]. (h) PUIE-MC [4]. (i) PUIE-MP [4]. (j) Ucolor [5]. (k) UWCNN-II [6]. (l) URSCT [59]. (m) U-shape [60]. (n) Uformer [61]. (o) Restormer [62]. (p) SwinIR [63]. (q) Ours. (r) References.

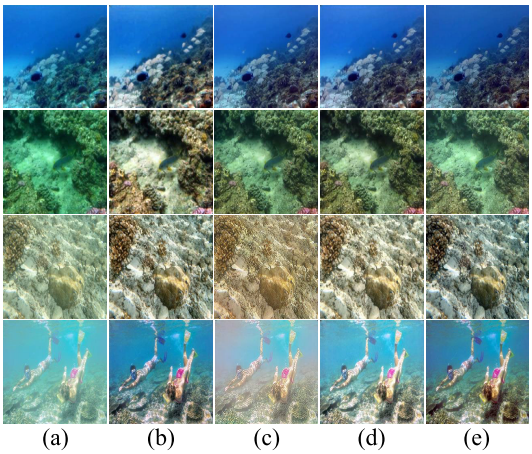

Fig. 15. Perform cross validation on networks trained on different datasets. (a) Input. (b) Test results using UIEB800 for training. (c) Test results using EUVP1680 for training. (d) Test results using UE2480 for training. (e) References.

the UIEB90 test set. The figure shows that the reconstruction results of the network obtained using the UIEB800 training set have clearer contrast because the labels in UIEB800 have high contrast. However, the network's enhancement results on images such as EUVP, where the original images are relatively blurry, still show blurry details. The reconstruction results of the network obtained using the EUVP1680 training set have insufficient contrast and are applicable to this type of image only. For other degraded underwater images such as UIEB90, the enhancement effect is not ideal, resulting in a reddish tint. The reason is that the contrast of the labels in the EUVP1680 training set is low. Using the superposition of two datasets for training, due to the diversity of underwater degradation types in the training set, the contrast of the test results is improved, but the details are not as good.

In summary, networks trained on different datasets have limitations in terms of generalizability, and accurate labels are crucial for training end-to-end networks. However, due to the difficulty of obtaining high-quality underwater images, most of the labels used in training are either the results of existing enhancement methods or synthetic results, which produces labels either lacking details or having low contrast. Inaccurate labels result in trained networks not always being very suitable for various underwater scenarios.

V. Conclusion

An underwater image enhancement network based on a transmission-aware Swin Transformer is proposed to achieve superior underwater image enhancement performance. This network uses the transmission of underwater images as weight guidance to achieve more accurate self-attention and more robust enhanced features, which have not been considered in previous methods. More specifically, the TG-MSA module of the network is used to integrate the transmission that characterizes the degree of image degradation into the multi-head self-attention process, solving the problem of inaccurate attention calculation. The TG-SFDI-FFN module uses the transmission as the attenuation weight of underwater images, which can enhance and fuse features in the spatial and frequency domains, thereby enhancing image details. Rich subjective and objective experiments verify the excellent ability to enhance details and correct colors.

Designing an unsupervised or semi-supervised underwater image enhancement network will be the focus of future research.

References

[1] S. Anwar and C. Li, "Diving deeper into underwater image enhancement: A survey," *Signal Process., Image Commun.*, vol. 89, Nov. 2020, Art. no. 115978.
[2] J. Han et al., "Underwater image restoration via contrastive learning and a real-world dataset," *Remote Sens.*, vol. 14, no. 17, p. 4297, Aug. 2022.
[3] 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.
[4] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *Proc. IEEE Conf. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 465–482.
[5] 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, 2021.
[6] 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.
[7] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2019.
[8] M. Jahidul Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," 2020, *arXiv:2002.01155*.
[9] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 81–88.
[10] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2018.
[11] Z. Mi, Y. Li, J. Jin, Z. Liang, and X. Fu, "A generalized enhancement framework for hazy images with complex illumination," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[12] W. Zhang, Y. Wang, and C. Li, "Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement," *IEEE J. Ocean. Eng.*, vol. 47, no. 3, pp. 718–735, Jul. 2022.
[13] J. Yuan, Z. Cai, and W. Cao, "TEBCF: Real-world underwater image texture enhancement model based on blurriness and color fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
[14] Y. Kang, Q. Jiang, C. Li, W. Ren, H. Liu, and P. Wang, "A perception-aware decomposition and fusion framework for underwater image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 988–1002, Mar. 2022.

[15] J. Zhou, D. Zhang, and W. Zhang, "Underwater image enhancement method via multi-feature prior fusion," *Appl. Intell.*, vol. 52, no. 14, pp. 16435–16457, Mar. 2022.

[16] J. Zhou, J. Sun, W. Zhang, and Z. Lin, "Multi-view underwater image enhancement method via embedded fusion mechanism," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 105946.

[17] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.

[18] J. Zhou, L. Pang, D. Zhang, and W. Zhang, "Underwater image enhancement method via multi-interval subhistogram perspective equalization," *IEEE J. Ocean. Eng.*, vol. 48, no. 2, pp. 474–488, Apr. 2023.

[19] X. Fu, Z. Fan, M. Ling, Y. Huang, and X. Ding, "Two-step approach for single underwater image enhancement," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 789–794.

[20] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and M. Sbert, "Color channel compensation (3C): A fundamental pre-processing step for image enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 2653–2665, 2020.

[21] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Trans. Image Process.*, vol. 31, pp. 3997–4010, 2022.

[22] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 4572–4576.

[23] M. Li, K. Wang, L. Shen, Y. Lin, Z. Wang, and Q. Zhao, "UIALN: Enhancement for underwater image with artificial light," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3622–3637, Aug. 2023.

[24] S. Gao, M. Zhang, Q. Zhao, X. Zhang, and Y. Li, "Underwater image enhancement using adaptive retinal mechanisms," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5580–5595, Nov. 2019.

[25] P. Zhuang, C. Li, and J. Wu, "Bayesian retinex underwater image enhancement," *Eng. Appl. Artif. Intell.*, vol. 101, May 2021, Art. no. 104171.

[26] P. Zhuang, J. Wu, F. Porikli, and C. Li, "Underwater image enhancement with hyper-Laplacian reflectance priors," *IEEE Trans. Image Process.*, vol. 31, pp. 5442–5455, 2022.

[27] B. McGlamery, "A computer model for underwater camera systems," *Proc. SPIE*, vol. 208, pp. 221–231, Mar. 1980.

[28] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE J. Ocean. Eng.*, vol. 15, no. 2, pp. 101–111, Apr. 1990.

[29] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2010.

[30] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1756–1769, Apr. 2012.

[31] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 132–145, Jan. 2015.

[32] P. L. J. Drews, E. R. Nascimento, S. S. C. Botelho, and M. F. Montenegro Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Comput. Graph. Appl.*, vol. 36, no. 2, pp. 24–35, Mar. 2016.

[33] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1579–1594, Apr. 2017.

[34] Y. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.

[35] J. Xie, G. Hou, G. Wang, and Z. Pan, "A variational framework for underwater image dehazing and deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3514–3526, Jun. 2022.

[36] G. Hou, N. Li, P. Zhuang, K. Li, H. Sun, and C. Li, "Non-uniform illumination underwater image restoration via illumination channel sparsity prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 799–814, Feb. 2024.

[37] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *Int. J. Comput. Vis.*, vol. 2023, pp. 1–19, Aug. 2023.

[38] J. Zhou, T. Yang, W. Chu, and W. Zhang, "Underwater image restoration via backscatter pixel prior and color compensation," *Eng. Appl. Artif. Intell.*, vol. 111, May 2022, Art. no. 104785.

[39] L. Chen et al., "Perceptual underwater image enhancement with deep learning and physical priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3078–3092, Aug. 2021.

[40] S. Gao, W. Wu, H. Li, L. Zhu, and X. Wang, "Atmospheric scattering model induced statistical characteristics estimation for underwater image restoration," *IEEE Signal Process. Lett.*, vol. 30, pp. 658–662, 2023.

[41] S. Yan, X. Chen, Z. Wu, M. Tan, and J. Yu, "HybrUR: A hybrid physical-neural solution for unsupervised underwater image restoration," *IEEE Trans. Image Process.*, vol. 32, pp. 5004–5016, 2023.

[42] Y. Tang, X. Liu, Z. Zhang, and S. Lin, "Adaptive underwater image enhancement guided by generalized imaging components," *IEEE Signal Process. Lett.*, vol. 30, pp. 1772–1776, 2023.

[43] 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.

[44] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7159–7165.

[45] P. Lin, Y. Wang, G. Wang, X. Yan, G. Jiang, and X. Fu, "Conditional generative adversarial network with dual-branch progressive generator for underwater image enhancement," *Signal Process., Image Commun.*, vol. 108, Oct. 2022, Art. no. 116805.

[46] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 4922–4936, 2022.

[47] R. Chen, Z. Cai, and W. Cao, "MFFN: An underwater sensing scene image enhancement method based on multiscale feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[48] R. Cong et al., "PUGAN: Physical model-guided underwater image enhancement using GAN with dual-discriminators," *IEEE Trans. Image Process.*, vol. 32, pp. 4472–4485, 2023.

[49] N. Jiang, W. Chen, Y. Lin, T. Zhao, and C.-W. Lin, "Underwater image enhancement with lightweight cascaded network," *IEEE Trans. Multimedia*, vol. 24, pp. 4301–4313, 2022.

[50] J. Jiang et al., "Five A^+ network: You only need 9K parameters for underwater image enhancement," 2023, *arXiv:2305.08824*.

[51] X. Liu, S. Lin, K. Chi, Z. Tao, and Y. Zhao, "Boths: Super lightweight network-enabled underwater image enhancement," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2022.

[52] Q. Qi, K. Li, H. Zheng, X. Gao, G. Hou, and K. Sun, "SGUIE-Net: Semantic attention guided underwater image enhancement with multi-scale perception," *IEEE Trans. Image Process.*, vol. 31, pp. 6816–6830, 2022.

[53] J. Zhou et al., "HCLR-net: Hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement," *Int. J. Comput. Vis.*, pp. 1–25, Feb. 2024.

[54] J. Zhou, Q. Gai, D. Zhang, K.-M. Lam, W. Zhang, and X. Fu, "IACC: Cross-illumination awareness and color correction for underwater images under mixed natural and artificial lighting," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4201115.

[55] P. Lin, Y. Wang, Y. Li, Z. Fan, and X. Fu, "Underwater color correction network with knowledge transfer," *IEEE Trans. Multimedia*, vol. 26, pp. 8088–8103, 2024.

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

[57] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 558–567.

[58] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[59] 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, 2022, Art. no. 4209616.

[60] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 32, pp. 3066–3079, 2023.

[61] 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. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[62] 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. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.
[63] 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. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[64] H. Wang, M. Yao, G. Jiang, Z. Mi, and X. Fu, "Graph-collaborated auto-encoder hashing for multiview binary clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 10121–10133, Jul. 2024.
[65] J. Zhou et al., "UGIF-net: An efficient fully guided information flow network for underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4206117.
[66] G. Li, Z. Bai, Z. Liu, X. Zhang, and H. Ling, "Salient object detection in optical remote sensing images driven by transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 5257–5269, 2023.
[67] B. Zhang et al., "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11294–11304.
[68] M. Shao, Y. Qiao, D. Meng, and W. Zuo, "Uncertainty-guided hierarchical frequency domain transformer for image restoration," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110306.
[69] Y. Tang, T. Iwaguchi, H. Kawasaki, R. Sagawa, and R. Furukawa, "AutoEnhancer: Transformer on U-Net architecture search for underwater image enhancement," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2022, pp. 1403–1420.
[70] 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, 2022.
[71] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4479–4488.
[72] J. Huang et al., "Deep Fourier-based exposure correction network with spatial-frequency interaction," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 163–180.
[73] C. Li et al., "Embedding Fourier for ultra-high-definition low-light image enhancement," 2023, *arXiv:2302.11831*.
[74] B. Kamgar-Parsi, B. Kamgar-Parsi, and A. Rosenfeld, "Optimally isotropic Laplacian operator," *IEEE Trans. Image Process.*, vol. 8, no. 10, pp. 1467–1472, Oct. 1999.
[75] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2822–2837, Aug. 2020.
[76] 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.
[77] 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.
[78] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005.
[79] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9060–9069.
[80] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

Yuanyuan Li received the B.S. degree in computer science and technology from Qufu Normal University, Qufu, China, in 2019, and the M.S. degree in computer science and technology from Dalian Maritime University, Dalian, China, in 2022, where she is currently pursuing the Ph.D. degree. Her research interests include computer vision, deep learning, and image processing.

Zetian Mi received the B.S. and Ph.D. degrees from the College of Computer Science and Technology, Sichuan University, Chengdu, China, in 2012 and 2017, respectively. She is currently an Associate Professor with the College of Information Science and Technology, Dalian Maritime University. Her main research interests include image processing and machine learning.

Yulin Wang received the M.S. degree in mathematics from Dalian Maritime University, Dalian, China, in 2022, where she is currently pursuing the Ph.D. degree in computer science and technology with the School of Information Science and Technology. Her research interests include underwater image processing, low-rank matrix recovery, and artificial intelligence.

Shuaiyong Jiang received the B.S. degree from the School of Mathematics and Information Science, Yantai University, Yantai, China, in 2021. He is currently pursuing the M.S. degree with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. His research interests include underwater saliency detection and deep learning.

Xianping Fu received the Ph.D. degree from Dalian Maritime University (DMU), Dalian, China, in 2005. He was a Post-Doctoral Researcher with Tsinghua University, Beijing, China, in 2008, and a Senior Research Fellow with Harvard University, Cambridge, MA, USA, in 2009. His group was included with Liaoning Revitalization Talents Program. He is currently a Full Professor and the Dean of the College of Information Science and Technology, DMU, and the Director of Liaoning Underwater Robot Engineering Research Center. His major research interests include image processing for content recognition, multimedia technology, and underwater robot vision. He has authored more than 100 journal articles and conference papers in these fields, which have been published in IJCAI, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, ICME, ICMR, and OCEANS.