U-shape Transformer for Underwater Image Enhancement

Lintao Peng, Chunli Zhu, and Liheng Bian*

***Abstract*—The light absorption and scattering of underwater impurities lead to poor underwater imaging quality. The existing data-driven based underwater image enhancement (UIE) techniques suffer from the lack of a large-scale dataset containing various underwater scenes and high-fidelity reference images. Besides, the inconsistent attenuation in different color channels and space areas is not fully considered for boosted enhancement. In this work, we built a large scale underwater image (LSUI) dataset, which covers more abundant underwater scenes and better visual quality reference images than existing underwater datasets. The dataset contains 4279 real-world underwater image groups, in which each raw image's clear reference images, semantic segmentation map and medium transmission map are paired correspondingly. We also reported an U-shape Transformer network where the transformer model is for the first time introduced to the UIE task. The U-shape Transformer is integrated with a channel-wise multi-scale feature fusion transformer (CMSFFT) module and a spatial-wise global feature modeling transformer (SGFMT) module specially designed for UIE task, which reinforce the network's attention to the color channels and space areas with more serious attenuation. Meanwhile, in order to further improve the contrast and saturation, a novel loss function combining RGB, LAB and LCH color spaces is designed following the human vision principle. The extensive experiments on available datasets validate the state-of-the-art performance of the reported technique with more than 2dB superiority. The dataset and demo code are available on https://lintaopeng.github.io/_pages/UIE%20Project%20Page.html.**

***Index Terms*—Underwater image enhancement, Transformer, Multi-color space loss function, Underwater image dataset**

I. Introduction

UNDERWATER Image Enhancement (UIE) technology [1], [2] is essential for obtaining underwater images and investigating the underwater environment, which has wide applications in ocean exploration, biology, archaeology, underwater robots [3] and among other fields. However, underwater images frequently have problematic issues, such as color casts, color artifacts and blurred details [4]. Those issues could be explained by the strong absorption and scattering effects on light, which are caused by dissolved impurities and suspended matter in the medium (water). Therefore, UIE-related innovations are of great significance for improving the visual quality and merit of images in accurately understanding the underwater world.

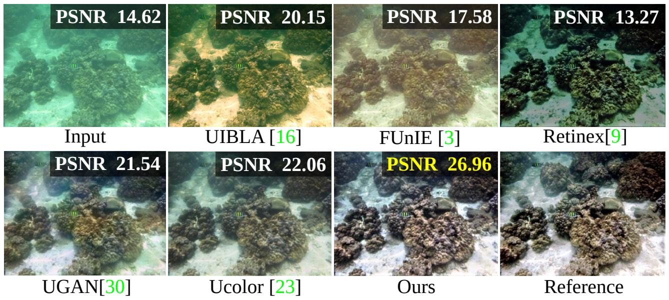


Fig. 1. Compared with the existing UIE methods, the image produced by our U-shape Transformer has the highest PSNR[5] score and best visual quality.

In general, the existing UIE methods could be categorized into three types, which are physical model-based, visual prior-based and data-driven methods, respectively. Among them, visual prior-based UIE methods [6], [7], [8], [9], [10], [11] mainly concentrated on improving the visual quality of underwater images by modifying pixel values from the perspectives of contrast, brightness and saturation. Nevertheless, the ignorance of the physical degradation process limits the improvement of enhancement quality. In addition, physical-model based UIE methods [12], [13], [14], [15], [16], [17], [18], [19], [20] mainly focus on the accurate estimation of medium transmission. With the estimated medium transmission and other key underwater imaging parameters such as the homogeneous background light, a clean image can be obtained by reversing a physical underwater imaging model. However, the performance of physical model-based UIE is restricted to complicated and diverse real-world underwater scenes. That is because, (1) *model hypothesis is not always plausible with complicated and dynamic underwater environment*; (2) *evaluating multiple parameters simultaneously is challenging.* More recently, as to the data-driven methods [21], [3], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [4], which could be regarded as deep learning technologies in UIE domain, exhibit impressive performance on UIE task. However, the existing underwater datasets more-or-less have the disadvantages, such as a small number of images, few underwater scenes, or even not real-world scenarios, which limits the performance of the data-driven UIE method. Besides, the inconsistent attenuation of the underwater images in different color channels and space areas have not been unified in one framework.

In this work, we first built a large scale underwater image (LSUI) dataset, which covers more abundant underwater scenes (water types, lighting conditions and target categories) and better visual quality reference images than existing underwater datasets [32], [28], [26], [22]. The dataset contains 4279 real-world underwater images, and the corresponding clear images are generated as comparison references. We also provide

L. Peng, C. Zhu and L. Bian are with the Advanced Research Institute of Multidisciplinary Science & School of Information and Electronics, Beijing Institute of Technology, Beijing, China. Correspondence to L. Bian: bian@bit.edu.cn.

the semantic segmentation map and medium transmission map for each image. Furthermore, with the prior knowledge that the attenuation of different color channels and space areas in underwater images is inconsistent, we designed a channel-wise multi-scale feature fusion transformer (CMSFFT) and a spatial-wise global feature modeling transformer (SGFMT) based on the attention mechanism, and embedded them in our U-shape Transformer which is designed based on [33]. Moreover, according to the color space selection experiment and [34], [23], we designed a multi-color space loss function including RGB, LAB and LCH color space. Fig. 1 shows the result of our UIE method and some comparison UIE methods, and the main contributions of this paper can be summarized as follows:

- We reported a novel U-shape Transformer dealing with the UIE task, in which the designed channel-wise and spatial-wise attention mechanism based on transformer enables to effectively remove color artifacts and casts.
- We designed a novel multi-color space loss function combing the RGB, LCH and LAB color-space features, which further improves the contrast and saturation of output images.
- We released a large-scale dataset containing 4279 real underwater images and the corresponding high-quality reference images, semantic segmentation maps, and medium transmission maps, which facilitates further development of UIE techniques.

II. Related work

A. UIE Methods

UIE is an indispensable step to improve the visual quality of recorded underwater images. A variety of methods have been proposed and can be categorized as visual prior, physical models, and data-driven methods.

UIE methods based on visual prior. This approach aims to restore a clear underwater image by modifying its pixel value. Typical methods involve: 1) *Modify the pixel value with single metric.* Such as contrast adjustment [35], histogram equalization [35], and white balance [11]. For instance, Hitam et al. [35] used contrast adjustment and adaptive histogram equalization methods in RGB color space and HSV color space to enhance the contrast of underwater images and reduce noise. 2) *Modify the pixel value with multiple metrics.* For example, fusion-based methods, which exhibit the final enhancement image via the weighted fusion of multiple traditional UIE methods. For example, Fang et al. [6] first applied white balance and global contrast adjustments to enhance underwater images, and then the two enhancement results are combined into one image by weighted addition to obtain the final enhanced underwater image. 3) *Retinex based UIE methods*. Fu et al. [9] proposed a retinex model-based UIE method including color correction, layer decomposition and enhancement. Furthermore, Zhang et al. [36] proposed a multi-scale UIE method based on the retinex model.

The way of modifying pixel value has the inherent advantage of improving the contrast and saturation of the raw underwater image. However, as visual prior neglected the inconsistent attenuation degree of underwater images in varied color channels and space areas, it performs not well on real underwater images with complex underwater environments.

UIE methods based on physical models. This approach regard UIE as a problem of inversion, and researchers usually enhance underwater images based on the following three steps, 1) establishing the prior conditions of the hypothetical physical imaging model; 2) estimating the key parameters; 3) reversing the degradation process of the underwater imaging process to obtain a clear image.

Prior is the basis of the physical model based UIE, in which existing work includes underwater dark channel priors [13], attenuation curve priors [15], fuzzy priors [18] and minimum information priors [20], etc. Early-stage research enhanced the underwater image by modifying the dark channel prior (DCP) [13] algorithm. Chiang et al.[18] restored the underwater image by combining the DCP with the wavelength compensation algorithm. Drews Jr et al. [13] proposed an underwater dark channel prior algorithm (UDCP) based on the priori that the red channel in the underwater image is more attenuated. Carlevaris Bianca et al. [37] used the attenuation difference prior between the three color channels in RGB color space to predict the transmission characteristics of the underwater scene, which feasibility is basically due to red light generally decays faster than green and blue light. In addition, Peng et al. [16] proposed a depth map estimate method for underwater scenes based on the intrinsic characteristics of underwater image blurriness and light absorption to effectively recover underwater images. Li et al. [7] integrate the minimum information loss and histogram distribution prior for depth estimation to recover underwater images.

This branch of UIE methods could achieve satisfactory results only when underwater scenes are in accordance with the selected physical imaging model. Therefore, the manually established priors restrain the model's robustness and scalability under the complicated and varied circumstances. Moreover, as the underwater physical imaging model does not taken human eye's perception characteristics into account, the visual quality of the restored images are of poor presentation effect. In recent years, underwater physical imaging models are gradually utilized in combination with data-driven methods[23].

Data-driven UIE methods. Current data-driven UIE methods can be divided into two main technical routes, (1) *designing an end-to-end module;* (2) *utilizing deep models directly to estimate physical parameters, and then restore the clean image based on the degradation model.* To alleviate the need for real-world underwater paired training data, Li et al. [22] proposed a WaterGAN to generate underwater-like images from in-air images and depth maps in an unsupervised manner, in which the generated dataset is further used to train the WaterGAN. Moreover, [24] exhibited a weakly supervised underwater color transmission model based on CycleGAN [38]. Benefiting from the adversarial network architecture and multiple loss functions, that network can be trained using unpaired underwater images, which refines the adaptability of the network model to underwater scenes. However, images in the training dataset used by the above methods are not matched real underwater images, which leads to limited enhancement effects of the above methods in diverse real-world underwater scenes. Recently, Li et al. [28] proposed a gated fusion network named WaterNet, which uses gamma-corrected images, contrast-improved images, and white-balanced images as the inputs to enhance underwater images. Yang et al. [39] proposed a conditional generative adversarial network (cGAN) to improve the perceptual quality of underwater images.

The methods mentioned above usually use existing deep neural networks for general purposes directly on UIE tasks and neglect the unique characteristics of underwater imaging. For example, [24] directly used the CycleGAN [38] network structure, and [28] adopted a simple multi-scale convolutional network. Other models such as UGAN [30],WaterGAN [22] and cGAN [39], still inherited the disadvantage of GAN-based models, which produces unstable enhancement results. In addition, Ucolor [23] combined the underwater physical imaging model and designed a medium transmission guided model to reinforce the network's response to areas with more severe quality degradation, which could improve the visual quality of the network output to a certain extent. However, physical models sometimes failed with varied underwater environments.

From above, our proposed network aims at generating high visual quality underwater images by properly accounting the inconsistent attenuation characteristics of underwater images in different color channels and space areas.

B. Underwater Image Datasets

The sophisticated and dynamic underwater environment results in extreme difficulties in the collection of matched underwater image training data in real-world underwater scenes. Present datasets can be classified into two types, they are, (1) Non-reference datasets. Liu et al. [32] proposed the RUIE dataset, which encompasses varied underwater lighting, depth of field, blurriness and color cast scenes. Akkaynak et al. [26] published a non-reference underwater dataset with a standard color comparison chart. Those datasets, however, cannot be used for end-to-end training for lacking matched clear reference underwater images. (2) Full-reference datasets. Li et al. [22] presented an unsupervised network dubbed WaterGAN to produce underwater-like images using in-air images and depth maps. Similarly, Fabbri et al. [30] used CycleGAN to generate distorted images from clean underwater images based on weakly-supervised distribution transfer. However, these methods rely heavily on training samples, which is easy to produce artifacts that are out of reality and unnatural. Li et al. [28] constructed a real UIE benchmark UIEB, including 890 images pairs, in which reference images were hand-crafted using the existing optimal UIE methods. Although those images are authentic and reliable, the number, content and coverage of underwater scenes are limited. In contrast, our LSUI dataset contains 4279 real-world underwater images with more abundant underwater scenes (water types, lighting conditions and target categories) than existing underwater datasets [32], [28], [26], [22], and the corresponding clear images are generated as comparison references. We also provide the semantic segmentation map and medium transmission map for each raw underwater image.

C. Transformers

Although CNN-based UIE methods [28], [3], [30], [31], [23] achieved significant improvement compared with traditional UIE methods. There are still two aspects that limit its further promotion, (1) *uniform convolution kernel is not able to characterize the inconsistent attenuation of underwater images in different color channels and spatial regions;* (2) *the CNN architecture concerns more on local features, while ineffective for long-dependent and global feature modeling.*

Recently, transformer [40] has gained more and more attention, its content-based interactions between image content and attention weights can be interpreted as spatially varying convolution, and the self-attention mechanism is efficient at modeling long-distance dependencies and global features. Benefiting from these advantages, transformers have shown outstanding performance in several vision tasks [41], [42], [43], [44]. Compared with previous CNN-based UIE networks, our CMSFFT and SGFMT modules designed based on the transformer can guide the network to pay more attention to the more serious attenuated color channels and spatial areas. Moreover, by combining CNN with transformer, we achieve better performance with a relatively small amount of parameters.

III. Proposed dataset and method

A. LSUI Dataset

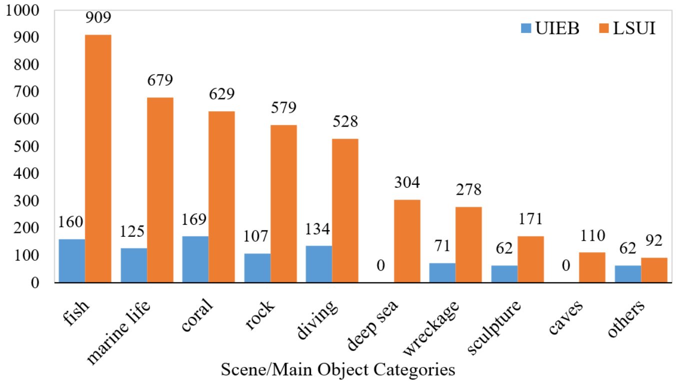


Fig. 2. Statistics of our LSUI dataset and the existing underwater dataset UIEB [28].

Data Collection. We have collected 8018 underwater images, which is composed of images collected by ourself and from other existing public datasets [32], [26], [22], [30] (All images have been licensed and used only for academic purposes). Real underwater images with rich water scenes, water types, lighting conditions and target categories, are selected to the extent possible, for further generating clear reference images.

Reference Image Generation. The reference images were selected with two round subjective and objective evaluations, to eliminating the potential bias to the extent possible. In the first round, inspired by ensemble learning [45] that multiple weak classifiers could form a strong one, we firstly use 18 existing optimal UIE methods [6], [9], [16], [13], [14], [17], [18], [19], [20], [3], [22], [24], [25], [27], [29], [31], [46], [47] to process the collected underwater images successively,

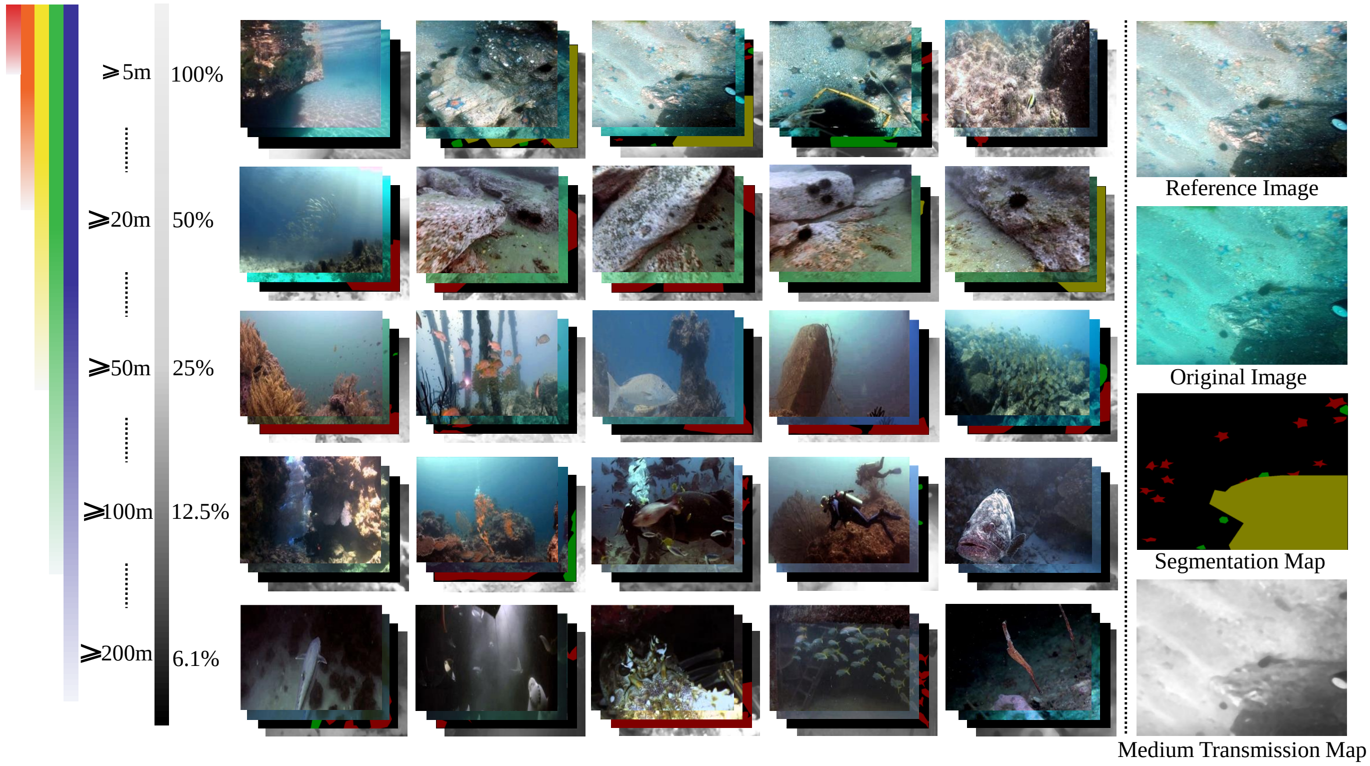


Fig. 3. Example images in the LSUI dataset. Our LSUI dataset contains 4279 real-world underwater images with more abundant underwater scenes (water types, lighting conditions and target categories) than existing underwater datasets [32], [28], [26], [22], and the corresponding clear images are generated as comparison references. We also provide the semantic segmentation map and medium transmission map for each raw underwater image. The top of each image group is the clear reference image, followed by the raw underwater image, semantic segmentation map, and medium transmission map.

and a set with $18 * 8018$ images is generated for the next-step optimal reference dataset selection. Unlike [28], to reducing the number of images that need to be selected manually, non-reference metrics UIQM [48] and UCIQE [49] are adopted to score all generated images with equal weights. Then, the top-three reference images of each original one form a set with the size $3 * 8018$. Considering individual differences, 20 volunteers with image processing experience were invited to rate images according to 5 most important judgments (contrast; saturation; color correction effects; artifacts degree; over or under-enhancement degree) of UIE tasks with a score from 0-10, where the higher score represents the more contentedness. And the total score of each reference picture is 100 $(5 * 20)$ after normalizing each score to 0-1. The top-one reference image of each raw underwater image was chosen with the highest summation value. In addition, images with the highest summation lower than 70 have been removed from the dataset.

After the first round, some of the generated reference images still have problems such as blur, color cast and noise. So in the second round, we invited volunteers to vote on each reference picture again to select its existing problems and determine the corresponding optimization method, and then use appropriate image enhancement methods [43], [50], [51] to process it. Next, all volunteers were invited to conduct another round of voting to remove image pairs that more than half of the volunteers were dissatisfied with. To improve the utility of the LSUI dataset, we also hand-labeled a segmentation map and generated a medium transmission map for each image. Eventually, our LSUI dataset contains 4279 images and the corresponding high-quality reference images, semantic segmentation maps, and medium transmission maps for each image.

As shown in Fig .2, compared with UIEB [28], our LSUI dataset contains large number of images with richer underwater scenes and object categories. In particular, our LSUI dataset includes deep-sea scenes and underwater cave scenes that are not available in previous underwater datasets. We provide some examples of our LSUI dataset in Fig .3, which includes varied underwawter scenes, water types, lighting conditions and target categories. As to the authors' best knowledge, LSUI is the largest real underwater image dataset with high quality reference images at the present time, which could facilitate the further development of the UIE methods.

B. U-shape Transformer

1) Overall Architecture: The overall architecture of the U-shape Transformer is shown as Fig. 4, which includes a CMSFFT & SGFMT based generator and a discriminator.

In the generator, (1) Encoding: Except being directly input to the network, the original image will be downsampled three times respectively. Then after 1*1 convolution, the three scale feature maps are input into the corresponding scale convolution block. The outputs of four convolutional blocks are the inputs of the CMSFFT and SGFMT; (2) Decoding: After feature remapping, the SGFMT output is directly sent

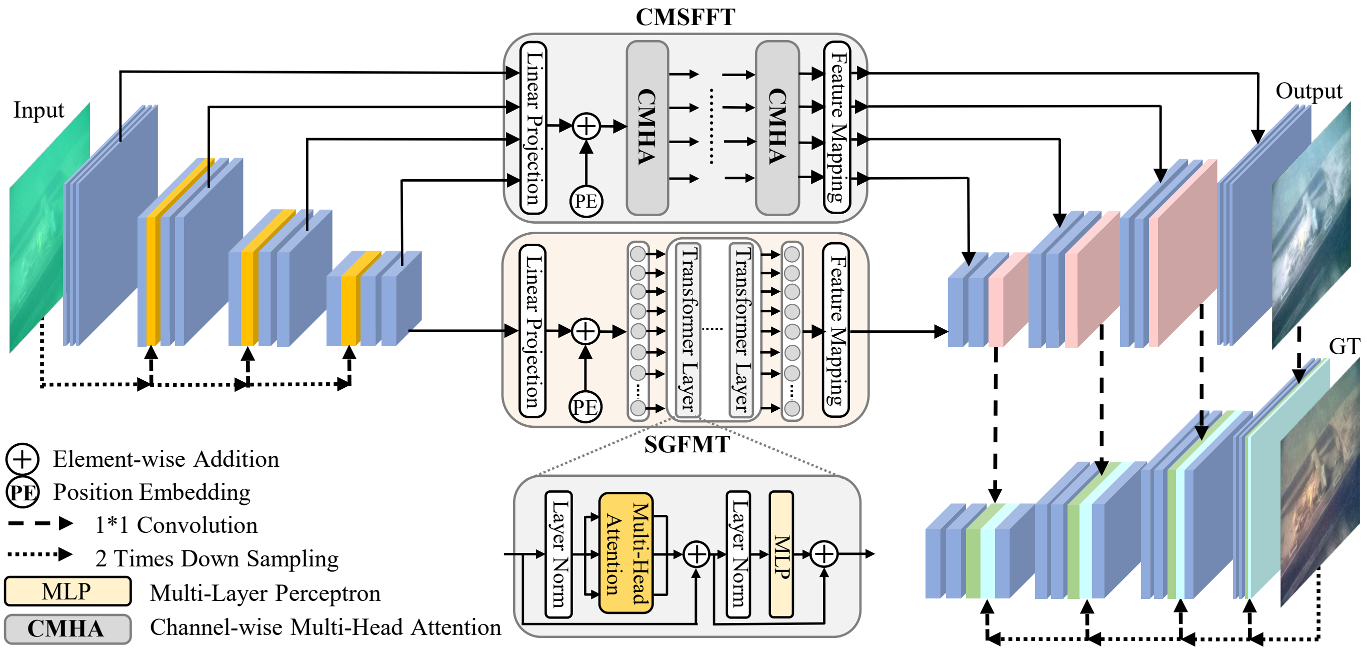


Fig. 4. The network structure of U-shape Transformer. CMSFFT and SGFMT modules specially designed for UIE tasks reinforce the network's attention to the more severely attenuated color channels and spatial regions. The multi-scale connections of the generator and the discriminator make the gradient flow freely between the generator and the discriminator, therefore making the training process more stable.

to the first convolutional block. Meanwhile, four convolutional blocks with varied scales will receive the four outputs from CMSFFT.

In the discriminator, the input of the four convolutional blocks includes: the feature map output by its own upper layer, the feature map of the corresponding size from the decoding part and the feature map generated by $1 * 1$ convolution after downsampling to the corresponding size using the reference image. With the described multi-scale connections, the gradient flow can flow freely on multiple scales between the generator and the discriminator, such that a stable training process could be obtained, details of the generated images could be enriched. The detailed structure of SGFMT and CMSFFT in the network will be described in the following two subsections.

2) SGFMT: The SGFMT (as shown in Fig. 5) is used to replace the original bottleneck layer of the generator, which can assist the network to model the global information and reinforce the network's attention on severely degraded parts. Assuming the size of the input feature map is $F_{in} \in \mathbb{R}^{\frac{H}{16}*\frac{W}{16}*C}$. For the expected one-dimensional sequence of the transformer, linear projection is used to stretch the two-dimensional feature map into a feature sequence $S_{in} \in \mathbb{R}^{\frac{HW}{256}*C}$. For preserving the valued position information of each region, learnable position embedding is merged directly, which can be expressed as,

$$S_{in} = W * F_{in} + \mathrm{PE}, \tag{1}$$

where $W * F_i$ represents a linear projection operation, PE represents a position embedding operation.

Then we input the feature sequence S_{in} to the transformer block, which contains 4 standard transformer layers [40].

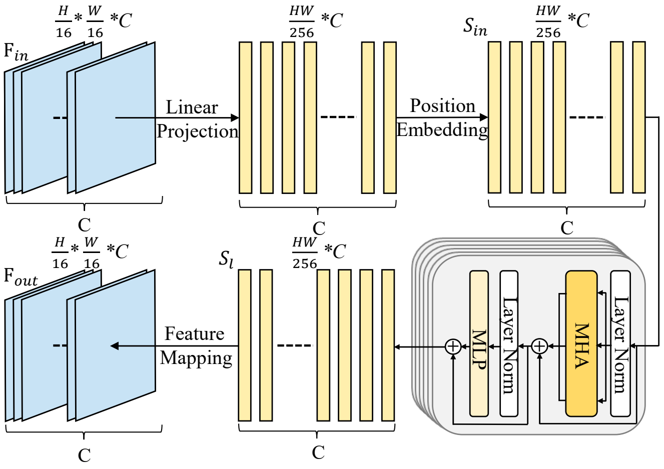


Fig. 5. Data flow diagram of the SGFMT module. Based on the prior that underwater images are not uniformly degraded in different spatial regions, we designed a novel **spatial-wise global feature modeling transformer (SGFMT)** based on the spatial self-attention mechanism to replace the original bottleneck layer of the generator. It can accurately model the global feature of underwater images and reinforce the network's attention to the space areas with more serious attenuation, thus achieving uniform UIE.

Each transformer layer contains a multi-head attention block (MHA) and a feed-forward network (FFN). The FFN includes a normalization layer and a fully connected layer. The output of the l-th($l \in [1, 2, \ldots, l]$) layer in the transformer block can be calculated by,

$$S_l^{'} = \mathrm{MHA}(\mathrm{LN}(S_{l-1})) + S_{l-1} \tag{2}$$

$$S_l = \mathrm{FFN}(\mathrm{LN}(S_l^{'})) + S_l^{'}, \tag{3}$$

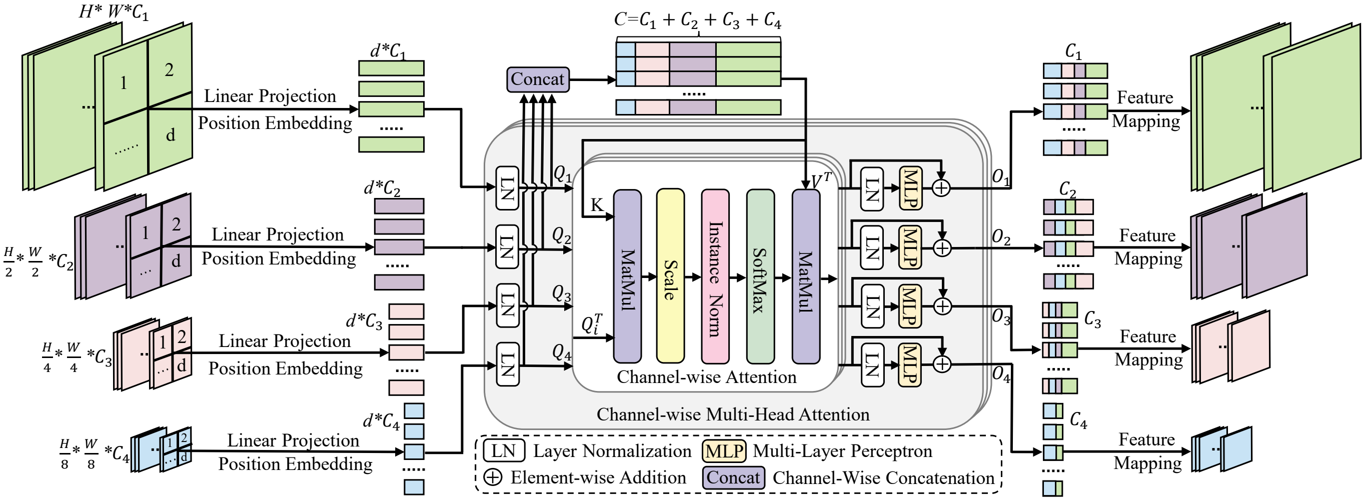


Fig. 6. Detailed structure of the CMSFFT module. According to the prior that underwater images are inconsistently attenuated in different color channels, we propose a novel **channel-wise multi-scale feature fusion transformer(CMSFFT)**. Specifically, CMSFFT replaces the skip connection of the generator, uses the channel-wise self-attention mechanism to perform channel-wise multi-scale feature fusion on the features output by the encoder of the generator, and transmits the fusion results to the decoder efficiently, so as to reinforce the network's attention to the color channels with more serious attenuation and realize accurate UIE.

where LN represents layer normalization, and S_l represents the output sequence of the l-th layer in the transformer block. The output feature sequence of the last transformer block is $S_l \in \mathbb{R}^{\frac{HW}{256} * \ C}$, which is restored to the feature map of $F_{out} \in \mathbb{R}^{\frac{H}{16} * \ \frac{W}{16} \ *C}$ after feature remapping.

3) CMSFFT: To reinforce the network's attention on the more serious attenuation color channels, inspired by [52], we designed the CMSFFT block to replace the skip connection of the original generator's encoding-decoding architecture (Fig.6), which consists of the following three parts.

Multi-Scale Feature Encoding. The inputs of CMSFFT are the feature maps $F_i \in \mathbb{R}^{\frac{H}{2^i} * \frac{W}{2^i} * C_i} (i = 0, 1, 2, 3)$ with different scales. Differs from the linear projection in Vit [53] which is applied directly on the partitioned original image, we use convolution kernels with related filter size $\frac{P}{2^i} * \frac{P}{2^i} (i = 0, 1, 2, 3)$ and step size $\frac{P}{2^i} (i = 0, 1, 2, 3)$, to conduct linear projection on feature maps with varied scales. In this work, P is set as 32. After that, four feature sequence $S_i \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$ could be obtained, where $d \in \frac{HW}{P^2}$. Those four convolution kernels divide feature maps into the same number of blocks, while the number of channels $C_i (i = 1, 2, 3, 4)$ remains unchanged. Then, four query vectors $Q_i \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$, $K \in \mathbb{R}^{d* \ C}$ and $V \in \mathbb{R}^{d* \ C}$ can be obtained by Eq.(4).

$$Q_i = S_i W_{Q_i} \quad K = SW_K \quad V = SW_V, \tag{4}$$

where $W_{Q_i} \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$, $W_K \in \mathbb{R}^{d* \ C}$ and $W_V \in \mathbb{R}^{d* \ C}$ stands for learnable weight matrices; S is generated by concatenating $S_i \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$ via the channel dimension, where $C = C_1 + C_2 + C_3 + C_4$. In this work, C_1, C_2, C_3, and C_4 are set as 64, 128, 256, 512, respectively.

Channel-Wise Multi-Head Attention(CMHA). The CMHA block has six inputs, which are $K \in \mathbb{R}^{d* \ C}$, $V \in \mathbb{R}^{d* \ C}$ and $Q_i \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$. The output of channel-wise attention $\mathrm{CA_i} \in \mathbb{R}^{C_i* \ d} (i = 1, 2, 3, 4)$ could be obtained by,

$$\mathrm{CA}_i = \mathrm{SoftMax}(\mathrm{IN}(\frac{Q_i^T K}{\sqrt[2]{C}}))V^T, \tag{5}$$

where IN represents the instance normalization operation. This attention operation performs along the channel-axis instead of the classical patch-axis[53], which can guide the network to pay attention to channels with more severe image quality degradation. In addition, IN is used on the similarity maps to assist the gradient flow spreads smoothly.

The output of the i-th CMHA layer can be expressed as,

$$\mathrm{CMHA_i} = (\mathrm{CA_i}^1 + \mathrm{CA_i}^2 + ..., +\mathrm{CA_i}^N)/N + Q_i, \tag{6}$$

where N is the number of heads, which is set as 4 in our implementation.

Feed-Forward Network(FFN). Similar to the forward propagation of [53], the FFN output can be expressed as,

$$O_i = \mathrm{CMHA_i} + \mathrm{MLP}(\mathrm{LN}(\mathrm{CMHA_i})), \tag{7}$$

where $O_i \in \mathbb{R}^{d* \ C_i} (i = 1, 2, 3, 4)$; MLP stands for multi-layer perception. Here, The operation in Eq. (7) needs to be repeated l (l=4 in this work) times in sequence to build the l-layer transformer.

Finally, feature remappings are performed on the four different output feature sequences $O_i \in \mathbb{R}^{C_i* \ d} (i = 1, 2, 3, 4)$ to reorganize them into four feature maps $F_i \in \mathbb{R}^{\frac{H}{2^i} * \frac{W}{2^i} * C_i} (i = 0, 1, 2, 3)$, which are the input of convolutional block in the generator's decoding part.

C. Loss Function

To take advantage of the LAB and LCH color spaces' wider color gamut representation range and more accurate description of the color saturation and brightness, we designed a multi-color space loss function combining RGB, LAB and LCH color spaces to train our network. The image from RGB space is firstly converted to LAB and LCH space, and reads,

$$\begin{aligned} L^{G(x)}, A^{G(x)}, B^{G(x)} &= \mathrm{RGB2LAB(G(x))} \\ L^y, A^y, B^y &= \mathrm{RGB2LAB(y)}, \end{aligned} \tag{8}$$

$$\begin{aligned} &L^{G(x)}, C^{G(x)}, H^{G(x)} = \mathrm{RGB2LCH}(\mathrm{G(x)}), \\ &L^{y}, C^{y}, H^{y} = \mathrm{RGB2LCH}(\mathrm{y}), \end{aligned} \tag{9}$$

where x, y and $G(x)$ represents the original inputs, the reference image, and the clear image output by the generator, respectively.

Loss functions in the LAB and LCH space are written as Eq.(10) and Eq.(11).

$$\begin{aligned} &Loss_{LAB}(G(x), y) = E_{x,y}[(L^{y} - L^{G(x)})^{2} - \\ &\sum_{i=1}^{n} Q(A_i^{y}) log(Q(A_i^{G(x)})) - \sum_{i=1}^{n} Q(B_i^{y}) log(Q(B_i^{G(x)}))], \end{aligned} \tag{10}$$

$$\begin{aligned} &Loss_{LCH}(G(x), y) = E_{x,y}[-\sum_{i=1}^{n} Q(L_i^{y}) log(Q(L_i^{G(x)})) \\ &+ (C^{y} - C^{G(x)})^{2} + (H^{y} - H^{G(x)})^{2}], \end{aligned} \tag{11}$$

where Q stands for the quantization operator.

L_2 loss in the RGB color space $Loss_{RGB}$ and the perceptual loss $Loss_{per}$[54] , as well as $Loss_{LAB}$ and $Loss_{LCH}$ are the four loss functions for the generator.

Besides, standard GAN loss function is introduced for minimizing the loss between generated and reference pictures, and written as,

$$L_{GAN}(G, D) = E_{y}[logD(y)] + E_{x}[log(1 - D(G(x)))], \tag{12}$$

where D represents the discriminator. D aims at maximizing $L_{GAN}(G, D)$, to accurately distinguish the generated image from the reference image. And the goal of generator G is to minimize the loss between generated pictures and reference pictures.

Then, the final loss function is expressed as,

$$\begin{aligned} &G^{*} = arg \min_{G} \max_{D} L_{GAN}(G, D) + \alpha Loss_{LAB}(G(x), y) \\ &+ \beta Loss_{LCH}(G(x), y) + \gamma Loss_{RGB}(G(x), y) \\ &+ \mu Loss_{per}(G(x), y), \end{aligned} \tag{13}$$

where $\alpha, \beta, \gamma, \mu$ are hyperparameters, which are set as 0.001, 1, 0.1, 100, respectively, with numerous experiments.

IV. Experiments

In this section, we first introduce the training details of the U-shape Transformer and the detailed settings of the experiment. Next, we conduct experiments on the selection of color space. Then we retrain some network models we collected on the existing underwater datasets and the LSUI dataset to evaluate our proposed dataset. Moerover, we also compare our UIE method with state-of-the-arts on five datasets. Finally, series of ablation studies are conducted to demonstrate the effectiveness of each component in U-shape Transformer.

A. Implementation Details

The LSUI dataset was randomly divided as Train-L (3879 images) and Test-L400 (400 images) for training and testing, respectively. The training set was enhanced by cropping, rotating and flipping the existing images. All images were adjusted to a fixed size (256*256) when input to the network, and the pixel value will be normalized to [0,1].

We use python and pytorch framework via NVIDIA RTX3090 on Ubuntu20 to implement the U-shape Transformer. Adam optimization algorithm is utilized for the total of 800 epochs training with batchsize set as 6. The initial learning rate is set as 0.0005 and 0.0002 for the first 600 epochs and the last 200 epochs, respectively. Besides, the learning rate decreased 20% every 40 epochs. For $Loss_{RGB}$, L_2 loss is used for the first 600 epochs, and L_1 loss is used for the last 200 epochs.

B. Experiment Settings

Benchmarks. Besides Train-L, the second training set Train-U contains 800 pairs of underwater images from UIEB [28] and 1,250 synthetic underwater images from [55]; the third training set Train-E contains the paired training images in the EUVP [3] dataset. Testing datasets are categorized into two types, (1) full-reference testing dataset: Test-L400 and Test-U90 (remaining 90 pairs in UIEB); (2) non-reference testing dataset: Test-U60 and SQUID. Here, Test-U60 includes 60 non-reference images in UIEB; 16 pictures from SQUID [26] forms the second non-reference testing dataset.

Compared Methods. We compare U-shape Transformer with 10 UIE methods to verify our performance superiority. It includes two physical-based models (UIBLA [16], UDCP [13]), three visual prior-based methods (Fusion [6], retinex based [9], RGHS [10]), and five data-driven methods (WaterNet [28], FUnIE [3], UGAN [30], UIE-DAL [31], Ucolor [23]).

Evaluation Metrics. For the testing dataset with reference images, we conducted full-reference evaluations using PSNR [5] and SSIM [56] metrics. Those two metrics reflect the proximity to the reference, where a higher PSNR value represents closer image content, and a higher SSIM value reflects a more similar structure and texture. For images in the non-reference testing dataset, non-reference evaluation metrics UCIQE [49] and UIQM [48] are employed, in which higher UCIQE or UIQM score suggests better human visual perception. For UCIQE and UIQM cannot accurately measure the performance

TABLE I
Statistical results of color space selection experiments. We test U-shape Transformers trained with different color space loss functions on Test-L400 and Test-U90 datasets, respectively, and the color spaces that obtain the top three PSNR scores are marked with red, green, and blue, respectively.

Color Space	RGB	HSV	HSI	XYZ	LAB	LUV	LCH	YUV
Tset-L400	**23.79**	23.32	23.37	22.63	**23.86**	22.81	**23.62**	23.43
Test-U90	**22.72**	22.01	22.17	21.69	**22.53**	21.77	**22.49**	22.23

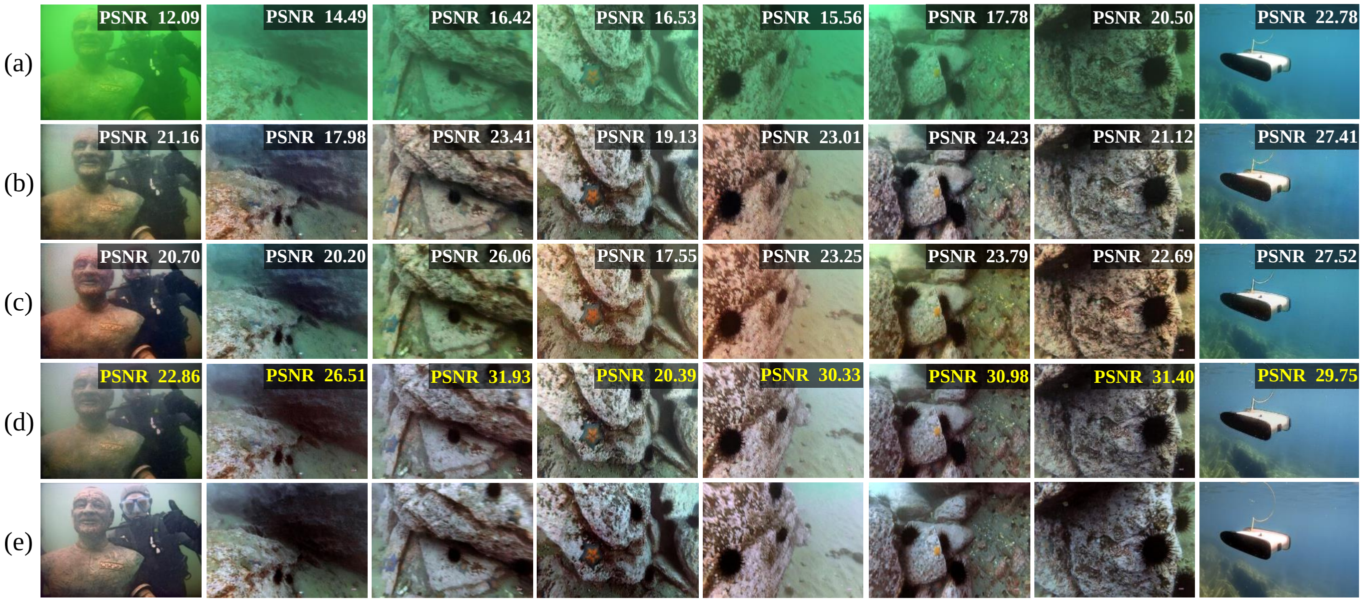


Fig. 7. Enhancement results of U-shape transformer trained on different underwater datasets. (a): Input images; (b): Enhanced results using the model trained on the Train-U; (c): Enhanced results using the model trained on the Train-E; (d): Enhanced results using the model trained by our proposed dataset Train-L; (e): Reference images(recognized as ground truth (GT)).

in some cases [28] [57], we also conducted a survey following [23], which results are stated as "perception score (PS)". PS ranges from 1-5, with higher scores indicating higher image quality. Moreover, NIQE [58], which lower value represents a higher visual quality, is also adopted as the metrics.

C. Color Space Selection

In order to select the appropriate color space to form the multi-color space loss function, we use the mixed loss function composed of the single color space loss function and other loss functions to train the U-shape transformer. We use Train-L to train the network, and then test and calculate PSNR on Test-L400 and Test-U90 data sets, respectively. The results are shown in Tab. I,

As in Tab. I, We note that the LAB, LCH, and RGB color spaces achieve the top-3 PSNR scores on both test datasets. In RGB color space, image is easy to store and display because of its strong color physical meaning, but these three components (R, G, and B) are highly correlated and easily affected by brightness, shadows, noise, and other factors. Compared with other color spaces, LAB color space is more consistent with the characteristics of human visual, can express all colors that human eyes can perceive, and the color distribution is more uniform. LCH color space can intuitively express brightness, saturation, and hue. Combined with the experimental results and the above analysis, we choose LAB, LCH, and RGB color space to form our multi-color space loss function.

D. Dataset Evaluation

The effectiveness of LSUI is evaluated by retraining the compared methods (U-net [59], UGAN [30] and U-shape Transformer) on Train-L, Train-U and Train-E. The trained network was tested on Test-L400 and Test-U90.

TABLE II
DATASET EVALUATION RESULTS. THE HIGHEST PSNR AND SSIM SCORES ARE MARKED IN RED.

Methods	Training Data	Test-U90		Test-L400	
		PSNR	SSIM	PSNR	SSIM
U-net[59]	Train-U	17.07	0.76	19.19	0.79
	Train-E	17.46	0.76	19.45	0.78
	Ours	**20.14**	**0.81**	**20.89**	**0.82**
UGAN[30]	Train-U	20.71	0.82	19.89	0.79
	Train-E	20.72	0.82	19.82	0.78
	Ours	**21.56**	**0.83**	**21.74**	**0.84**
Ours	Train-U	21.25	0.84	22.87	0.85
	Train-E	21.75	0.86	23.01	0.87
	Ours	**22.91**	**0.91**	**24.16**	**0.93**

As shown in Tab.II, the model trained on our dataset is the best of PSNR and SSIM. It could be explained that LSUI contains richer underwater scenes and better visual quality reference images than existing underwater image datasets, which could improve the enhancement and generalization ability of the tested network.

Fig. 7 is the sampled enhancement results of U-shape transformer trained on different underwater datasets, which is a supplement of the Data Evaluation part of the paper. Enhancement results training on Train-L (a portion of our LSUI dataset) demonstrates the highest PSNR value and preferable visual quality, while results training on other datasets show a certain degree of color cast. For the high-quality reference images and rich underwater scenes (lighting conditions, water types and target categories), our constructed LSUI dataset could improve the imaging quality and generalization performance of the UIE network.

TABLE III
QUANTITATIVE COMPARISON AMONG DIFFERENT UIE METHODS ON THE FULL-REFERENCE TESTING SET. THE HIGHEST SCORES OF PSNR AND SSIM ARE MARKED IN RED, AND ALL UIE METHODS ARE TESTED ON A PC WITH AN INTEL(R) I5-10500 CPU, 16.0GB RAM, A NVIDIA GEFORCE RTX 1660 SUPER.

Methods	Test-L400		Test-U90		FLOPs↓	#param.↓	time↓
	PSNR↑	SSIM↑	PSNR↑	SSIM↑			
UIBLA[16]	13.54	0.71	15.78	0.73	×	×	42.13s
UDCP[13]	11.89	0.59	13.81	0.69	×	×	30.82s
Fusion[6]	17.48	0.79	19.04	0.82	×	×	6.58s
Retinex based[9]	13.89	0.74	14.01	0.72	×	×	1.06s
RGHS[10]	14.21	0.78	14.57	0.79	×	×	8.92s
WaterNet[28]	17.73	0.82	19.81	0.86	193.7G	24.81M	0.61s
FUnIE[3]	19.37	0.84	19.45	0.85	10.23G	7.019M	0.09s
UGAN[30]	19.79	0.78	20.68	0.84	38.97G	57.17M	0.05s
UIE-DAL[31]	17.45	0.79	16.37	0.78	29.32G	18.82M	0.07s
Ucolor[23]	22.91	0.89	20.78	0.87	443.85G	157.4M	2.75s
Ours	**24.16**	**0.93**	**22.91**	**0.91**	66.2G	65.6M	0.07s

TABLE IV
QUANTITATIVE COMPARISON AMONG DIFFERENT UIE METHODS ON THE NON-REFERENCE TESTING SET. THE HIGHEST SCORES ARE MARKED IN RED.

Methods	Test-U60				SQUID			
	PS↑	UIQM↑	UCIQE↑	NIQE↓	PS↑	UIQM↑	UCIQE↑	NIQE↓
input	1.46	0.82	0.45	7.16	1.23	0.81	0.43	4.93
UIBLA[16]	2.18	1.21	0.60	6.13	2.45	0.96	0.52	4.43
UDCP[13]	2.01	1.03	0.57	5.94	2.57	1.13	0.51	4.47
Fusion[6]	2.12	**1.23**	0.61	4.96	2.89	**1.29**	0.61	5.01
Retinex based[9]	2.04	0.94	0.69	4.95	2.33	1.01	0.66	4.86
RGHS[10]	2.45	0.66	0.71	4.82	2.67	0.82	**0.73**	4.54
WaterNet[28]	3.23	0.92	0.51	6.03	2.72	0.98	0.51	4.75
FUnIE[3]	3.12	1.03	0.54	6.12	2.65	0.98	0.51	4.67
UGAN[30]	3.64	0.86	0.57	6.74	2.79	0.90	0.58	4.56
UIE-DAL[31]	2.03	0.72	0.54	4.99	2.21	0.79	0.57	4.88
Ucolor[23]	3.71	0.84	0.53	6.21	2.82	0.82	0.51	4.32
Ours	**3.91**	0.85	**0.73**	**4.74**	**3.23**	0.89	0.67	**4.24**

TABLE V
THE COLOR DISSIMILARITY COMPARISONS OF DIFFERENT METHODS ON COLOR-CHECK7 IN TERMS OF THE CIEDE2000. THE BEST SCORES ARE MARKED IN RED.

Methods	Pen W60	Pen W80	Can D10	Fuj Z33	Oly T6000	Oly T8000	Pan TS1	Avg
input	14.21	16.92	17.14	16.03	15.02	22.43	18.65	17.2
UIBLA[16]	13.45	16.31	14.48	14.29	12.46	14.91	20.13	15,15
UDCP[13]	15.32	24.12	16.53	13.21	12.65	16.78	12.85	15.92
Fusion[6]	12.65	13.54	14.43	12.31	11.78	**10.97**	11.15	12.41
Retinex based[9]	13.08	19.25	17.13	18.85	17.18	19.45	20.62	17.94
RGHS[10]	11.07	12.73	15.92	13.47	14.26	18.73	12.06	14.03
WaterNet[28]	12.54	19.82	15.71	12.73	17.75	21.87	18.91	17.05
FUnIE[3]	12.81	11.81	12.39	12.76	12.46	16.74	19.28	14.04
UGAN[30]	20.49	21.75	22.63	26.49	21.63	22.05	20.73	22.25
UIE-DAL[31]	12.94	16.73	14.64	12.93	16.78	17.21	18.34	15.65
Ucolor[23]	9.12	11.14	12.43	10.02	8.31	14.18	13.41	11.23
Ours	**7.87**	**9.70**	**9.96**	**8.23**	**7.71**	11.14	**9.81**	**9.20**

E. Network Architecture Evaluation

Full-Reference Evaluation. The Test-L400 and Test-U90 datasets were used for evaluation. The statistical results and visual comparisons are summarized in Tab. III and Fig. 8. We also provide the running time (image size is 256*256) of all UIE methods in Tab. III, as well as the FLOPs and parameter amount of each data-driven UIE method. And we retrianed the 5 open-sourced deep learning-based UIE methods on our dataset.

As in Tab.III, our U-shape Transformer demonstrates the best performance on both PSNR and SSIM metrics with relatively few parameters, FLOPs, and running time. The potential limitations of the performance of the 5 data-driven methods are analyzed as follows. The strength of FUnIE [3] lies in achieving fast, lightweight,and fewer parameter models, while naturally limits its scalability on complex and distorted testing samples. UGAN [30] and UIE-DAL [31] did not consider the inconsistent characteristics of the underwater images. Ucolor's media transmission map prior can not effectively represent the attenuation of each area, and simply introducing the concept of multi-color space into the network's encoder part cannot effectively take advantage of it, which causes unsatisfactory results in terms of contrast, brightness, and detailed textures.

The visual comparisons shown in Fig. 8 reveal that enhancement results of our method are the closest to the reference image, which has fewer color artifacts and high-fidelity object

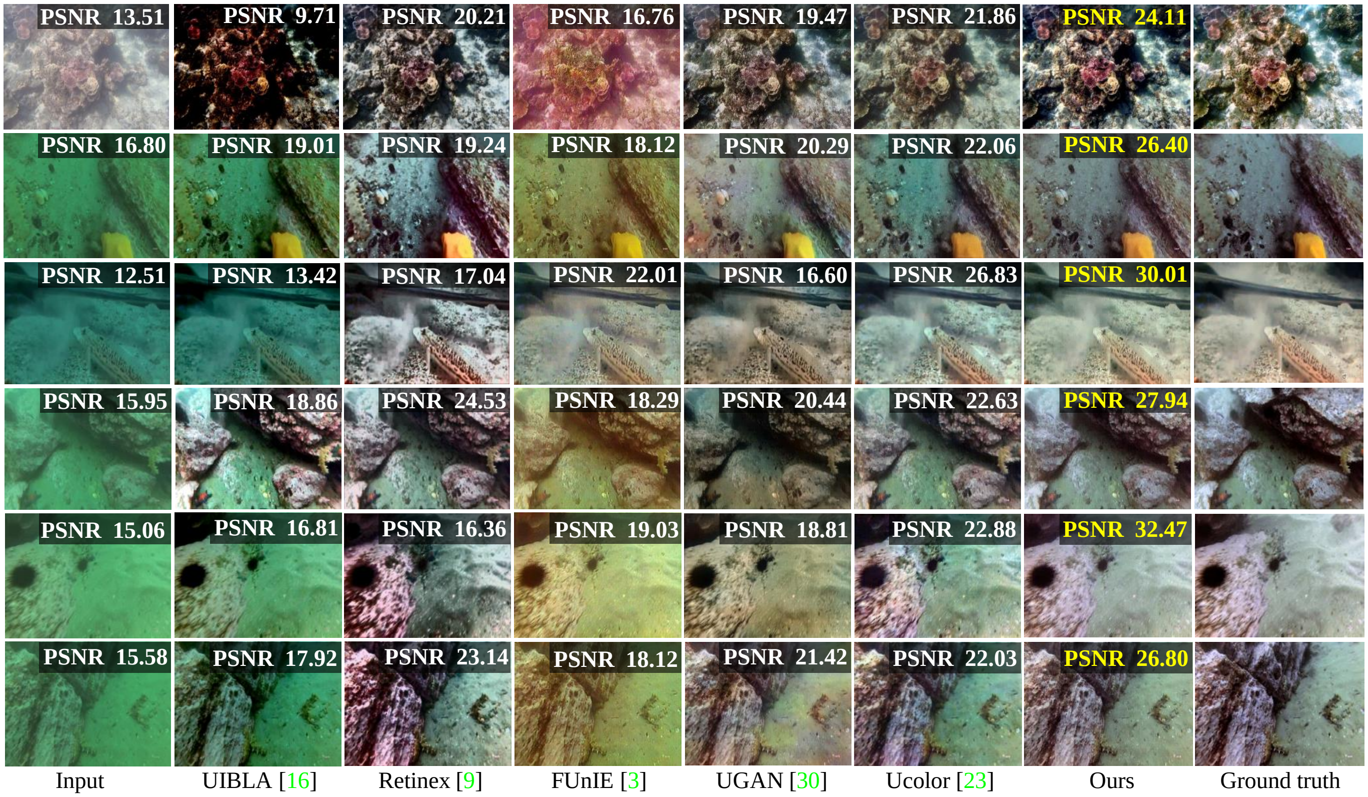


Fig. 8. Visual comparison of enhancement results sampled from the Test-L400(LSUI) and Test-U90(UIEB[28]) dataset. From left to right are raw underwater images, results of UIBLA[16], Retinex based[9], FUnIE[3], UGAN[30], Ucolor[23], our U-shape Transformer and the reference image (recognized as ground truth (GT)). The highest PSNR value of each raw is marked in yellow.

areas. Five selected methods tend to produce color artifacts that deviated from the original color of the object. Among the methods, UIBLA [16] exhibits severe color casts. Retinex based[9] could improve the image contrast to a certain extent, but cannot remove the color casts and color artifacts effectively. The enhancement result of FUnLE [3] is yellowish and reddish overall. Although UGAN [30] and Ucolor [28] could provide relatively good color appearance, they are often affected by local over-enhancement, and there are still some color casts in the result.

Non-reference Evaluation. The Test-U60 and SQUID datasets were utilized for the non-reference evaluation, in which statistical results and visual comparisons are shown in Tab. IV and Fig. 9.

As in Tab. IV, our method achieved the highest scores on PS and NIQE metrics, which confirmed the initial idea to contemplate the human eye's color perception and better generalization ability to varied real-world underwater scenes. Note that UCIQE and UIQM of all deep learning-based UIE methods are weaker than physical model-based or visual prior-based, also reported in [23]. Those two metrics are of valuable reference, but cannot as absolute justifications [28][57], for they are non-sensitive to color artifacts & casts and biased to some features.

As in Fig. 9, enhancement results of our method have the highest PS value, which index reflects the visual quality. Generally, compared methods are unsatisfactory, which includes undesirable color artifacts, over-saturation and unnatural color casts. Among the methods, results of the UIBLA [16] and FUnIE [3] have a certain degree of color cast. Retinex based [9] method introduces artifacts and unnatural colors. UGAN [30] and UIE-DAL [31] have the issue of local over-enhancement and color artifacts, which main reason is they ignore the inconsistent attenuation characteristics of the underwater images in the different space areas and the color channels. Although Ucolor [23] introduces the transmission medium prior to reinforcing the network's attention on the spatial area with severe attenuation, it still ignores the inconsistent attenuation characteristics of the underwater image in different color channels, which results in the problem of overall color cast. In our method, the reported CMSFFT and SGFMT modules could reinforce the network's attention to the color channels and spatial regions with serious attenuation, therefore obtaining high visual quality enhancement results without artifacts and color casts.

F. Color Restoration Performance Evaluation

To demonstrate the robustness and accuracy of our UIE method for color correction, we compare the color correction ability of 10 UIE methods on the Color-Checker7 dataset. The Color-Checker7 dataset contains 7 underwater images taken from a shallow swimming pool with different cameras. Color checker is also photographed in each image. It provides a good path to demonstrate the robustness of our method to different imaging devices and the accuracy of color restoration.

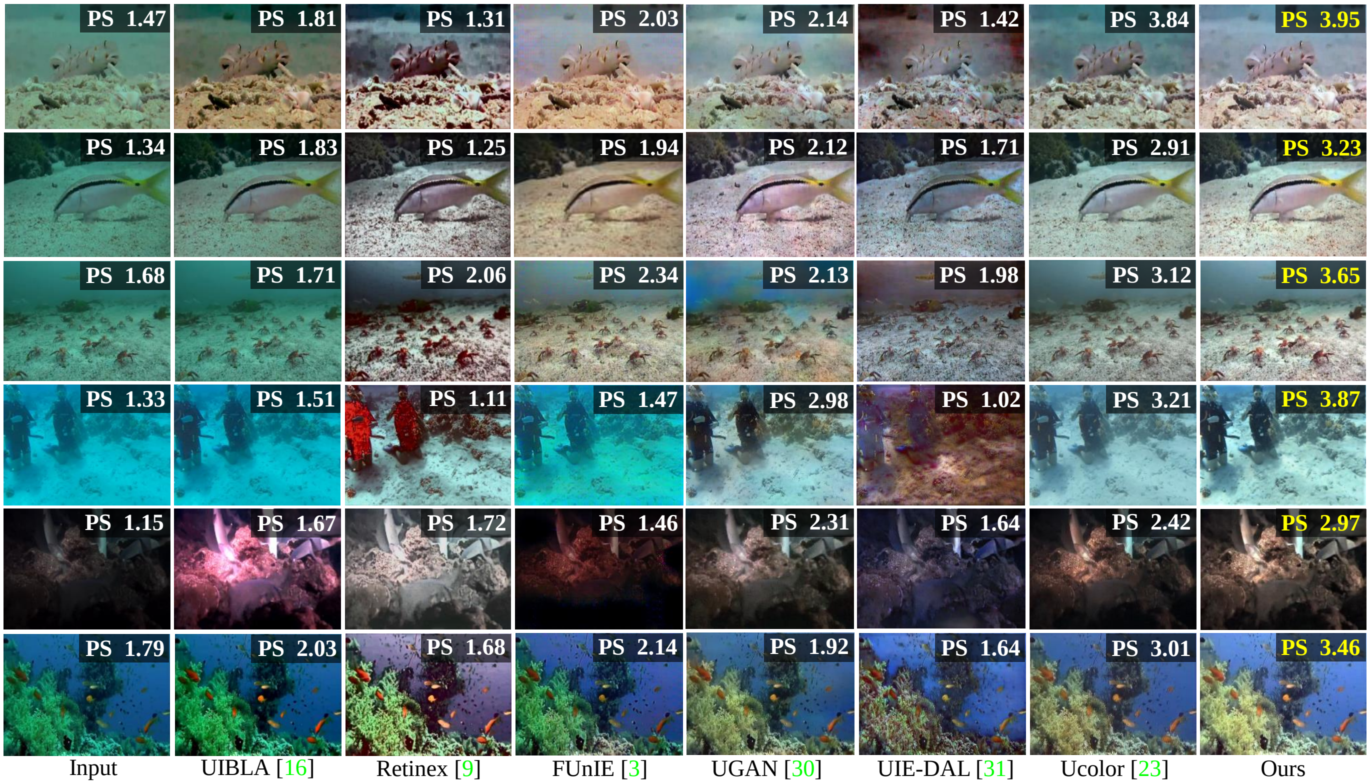


Fig. 9. Visual comparison of the non-reference evaluation sampled from the Test-U60(UIEB [28]) dataset. From left to right are raw underwater images, results of UIBLA [16], Retinex based [9], FUnIE [3], UGAN [30], UIE-DAL [31], Ucolor [23] and our U-shape Transformer. The score in the upper right corner of each image is the perception score(PS), and the highest PS value of each raw is marked in yellow.

We follow Ancuti et al. [60] to employ CIEDE2000 [61] to measure the relative differences between the corresponding color patches of ground-truth Macbeth Color Checker and the enhancement results of these comparison methods. The experimental results are shown in Tab. V and Fig .10.

As in Tab. V, for the cameras of Pentax W60, Pentax W80, Cannon D10, Fuji Z33, Panasonic TS1 and Olympus T6000, our U-shape Transformer obtains the lowest color dissimilarity. Moreover, our U-shape Transformer achieves the best average score. Such results demonstrate the superiority of our method for underwater color correction. It is worth mentioning that some comparable methods acquired lower score than that of the raw image, which reflected that those methods are incapable of recovering the real color and even break the inherent color.

As shown in Fig. 11, the professional underwater camera (Fuji Z33) also inevitably introduces various color casts. Among all the UIE methods involved in the comparison, our U-shape Transformer achieves the highest CIEDE 2000 score, which means our UIE method has the best color correction ability. The results of UDCP and UIBLA are bluish, and Retinex has the problem of color distortion. UGAN and UIE-DAL suffer from low saturation and excessive reddish compensation. Although FUnIE and Ucolor could remove the color cast to a certain extent, there are still problems of low contrast and saturation.

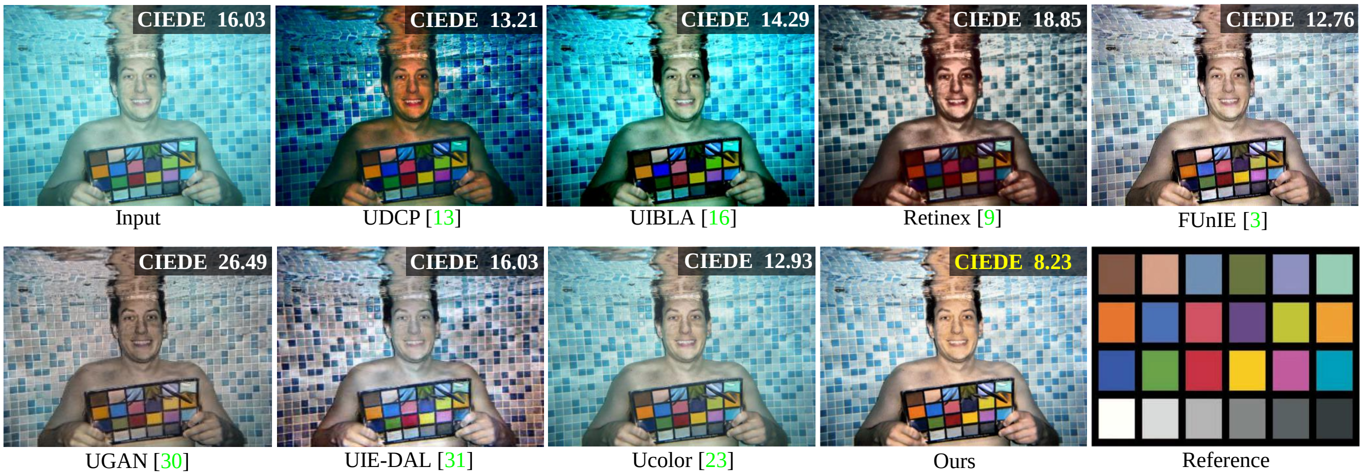
G. Ablation Study

To prove the effectiveness of each component, we conduct a series of ablation studies on the Test-L400 and Test-U90. Four factors are considered including the CMSFFT, the SGFMT, the multi-scale gradient flow mechanism (MSG), and the multi-color space loss function (MCSL).

TABLE VI
STATISTICAL RESULTS OF ABLATION STUDY ON THE TEST-L400 AND THE TEST-U90. THE HIGHEST SCORES ARE MARKED IN RED.

Models	Test-L400		Test-U90	
	PSNR	SSIM	PSNR	SSIM
BL	19.34	0.79	19.36	0.81
BL+CMSFFT	22.47	0.88	21.72	0.86
BL+SGFMT	21.78	0.86	21.36	0.87
BL+MSG	20.11	0.82	21.24	0.85
BL+MCSL	21.51	0.82	20.16	0.81
Full Model	**24.16**	**0.93**	**22.91**	**0.91**

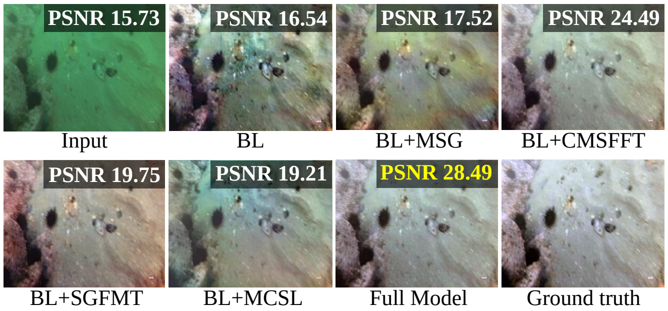
Experiments are all trained by Train-L. Statistical results are shown in Tab. VI, in which baseline model (BL) refers to [33], full models is the complete U-shape Transformer. In Tab. VI, our full model achieves the best quantitative performance on the two testing dataset, which reflects the effectiveness of the combination of CMSFFT, SGFMT, MSG, and MCSL modules. As in Fig .11, the enhancement result of the full model has the highest PSNR and best visual quality. The results of BL+MSG have less noise and artifacts than the BL module because the MSG mechanism helps to reconstruct local details. Thanks to the multi-color space loss function, the overall



Input UDCP [13] UIBLA [16] Retinex [9] FUnIE [3]

UGAN [30] UIE-DAL [31] Ucolor [23] Ours Reference

Fig. 10. Visual comparison of the color restoration performance evaluation. The input image is sampled from color-check7 dataset and it's taken by Fuji Z33. The values of CIEDE2000 metric for the regions of Color Checker are reported on the top-left corner of the images (the smaller, the better).



Input BL BL+MSG BL+CMSFFT

BL+SGFMT BL+MCSL Full Model Ground truth

Fig. 11. Visual comparison of the ablation study sampled from the Test-L400 dataset.

color of BL+MCSL's result is close to the reference image. The unevenly distributed visualization and artifacts in local areas of BL+MCSL are due to the lack of efficient attention guidance. Although the enhanced results of BL+CMSFFT and BL+SGFMT are evenly distributed, the overall color is not accurate. The investigated four modules have their particular functionality in the enhancement process, which integration could improve the overall performance of our network.

V. Conclusions

In this work, we released a large scale underwater image (LSUI) dataset, which contains 4279 real-world underwater images with more abundant underwater scenes (water types, lighting conditions and target categories) than existing underwater datasets [32], [28], [26], [22], and the corresponding clear images are generated as comparison references. We also provide the semantic segmentation map and medium transmission map for each raw underwater image. Besides, we reported an U-shape Transformer network for state-of-the-art UIE performance. The network's CMSFFT and SGFMT modules could solve the inconsistent attenuation issue of underwater images in different color channels and space regions, which has not been considered among existing methods. Extensive experiments validate the superior ability of the network to remove color artifacts and casts. Combined with the multi-color space loss function, the contrast and saturation of output images are further improved. Nevertheless, it is impossible to collect images of all the complicated scenes such as deep-ocean low-light scenarios. Therefore, we will introduce other general enhancement techniques such as low-light boosting [62] for future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61991451, 61971045, 62131003), National Key Research and Development Program of China (2020YFB0505601).

References

[1] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du, and K. Hu, "An in-depth survey of underwater image enhancement and restoration," *IEEE Access.*, vol. 7, pp. 123 638–123 657, 2019. 1

[2] P. Sahu, N. Gupta, and N. Sharma, "A survey on underwater image enhancement techniques," *IJCA*, vol. 87, no. 13, 2014. 1

[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, 2020. 1, 3, 7, 9, 10, 11

[4] R. Schettini and S. Corchs, "Underwater image processing: State of the art of restoration and image enhancement methods," *EURASIP. J. Adv. Signal Process.*, vol. 2010, pp. 1–14, 2010. 1

[5] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *QoMEX*. IEEE, 2012, pp. 37–38. 1, 7

[6] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *CVPR*, 2012, pp. 81–88. 1, 2, 3, 7, 9

[7] 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 T. Image Process.*, vol. 25, no. 12, pp. 5664–5677, 2016. 1, 2

[8] A. S. A. Ghani and N. A. M. Isa, "Underwater image quality enhancement through composition of dual-intensity images and rayleigh-stretching," in *ICCE*, 2014, pp. 219–220. 1

[9] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *ICIP*, 2014, pp. 4572–4576. 1, 2, 3, 7, 9, 10, 11

[10] D. Huang, Y. Wang, W. Song, J. Sequeira, and S. Mavromatis, "Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition," in *MMM*. Springer, 2018, pp. 453–465. 1, 7, 9

[11] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib, "Enhancing the low quality images using unsupervised colour correction method," in *IEEE Int. Conf. Syst. Man. Cybern.*, 2010, pp. 1703–1709. 1, 2

[12] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *CVPR*, 2009, pp. 1956–1963. 1

[13] P. Drews Jr, E. do Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *ICCV workshops*, 2013, pp. 825–830. 1, 2, 3, 7, 9

[14] P. L. Drews, E. R. Nascimento, S. S. 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, 2016. 1, 3

[15] Y. Wang, H. Liu, and L.-P. Chau, "Single underwater image restoration using adaptive attenuation-curve prior," *IEEE Trans. Circuits. Syst. I. Regul. Pap.*, vol. 65, no. 3, pp. 992–1002, 2018. 1, 2

[16] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE T. Image Process.*, vol. 26, no. 4, pp. 1579–1594, 2017. 1, 2, 3, 7, 9, 10, 11

[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 T. Image Process.*, vol. 25, no. 12, pp. 5664–5677, 2016. 1, 3

[18] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE T. Image Process.*, vol. 21, no. 4, pp. 1756–1769, 2012. 1, 2, 3

[19] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *JVCIR*, vol. 26, pp. 132–145, 2015. 1, 3

[20] C. Li, J. Guo, S. Chen, Y. Tang, Y. Pang, and J. Wang, "Underwater image restoration based on minimum information loss principle and optical properties of underwater imaging," in *ICIP*, 2016, pp. 1993–1997. 1, 2, 3

[21] Y. Guo, H. Li, and P. Zhuang, "Underwater image enhancement using a multiscale dense generative adversarial network," *IEEE J. OCEANIC. ENG.*, vol. 45, no. 3, pp. 862–870, 2019. 1

[22] 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, 2017. 1, 2, 3, 4, 12

[23] 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 T. Image Process.*, vol. 30, pp. 4985–5000, 2021. 1, 2, 3, 7, 8, 9, 10, 11

[24] C. Li, J. Guo, and C. Guo, "Emerging from water: Underwater image color correction based on weakly supervised color transfer," *IEEE Signal. Process. Lett.*, vol. 25, no. 3, pp. 323–327, 2018. 1, 2, 3

[25] H.-Y. Yang, P.-Y. Chen, C.-C. Huang, Y.-Z. Zhuang, and Y.-H. Shiau, "Low complexity underwater image enhancement based on dark channel prior," in *IBICA*, 2011, pp. 17–20. 1, 3

[26] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *CVPR*, 2019, pp. 1682–1691. 1, 3, 4, 7, 12

[27] X. Fu, Z. Fan, M. Ling, Y. Huang, and X. Ding, "Two-step approach for single underwater image enhancement," in *ISPACS*, 2017, pp. 789–794. 1, 3

[28] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE T. Image Process.*, vol. 29, pp. 4376–4389, 2020. 1, 3, 4, 7, 8, 9, 10, 11, 12

[29] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *PCM*. Springer, 2018, pp. 678–688. 1, 3

[30] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," *ICRA*, pp. 7159–7165, 2018. 1, 3, 7, 8, 9, 10, 11

[31] P. M. Uplavikar, Z. Wu, and Z. Wang, "All-in-one underwater image enhancement using domain-adversarial learning." in *CVPR Workshops*, 2019, pp. 1–8. 1, 3, 7, 9, 10, 11

[32] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 30, pp. 4861–4875, 2020. 1, 3, 4, 12

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134. 2, 11

[34] X. Li and A. Li, "An improved image enhancement method based on lab color space retinex algorithm," in *ICGIP*, C. Li, H. Yu, Z. Pan, and Y. Pu, Eds., vol. 11069. SPIE, 2019, pp. 756 – 765. 2

[35] M. S. Hitam, E. A. Awalludin, W. N. Jawahir Hj Wan Yussof, and Z. Bachok, "Mixture contrast limited adaptive histogram equalization for underwater image enhancement," in *ICCAT*, 2013, pp. 1–5. 2

[36] S. Zhang, T. Wang, J. Dong, and H. Yu, "Underwater image enhancement via extended multi-scale retinex," *Neurocomputing*, vol. 245, pp. 1–9, 2017. 2

[37] N. Carlevaris, Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *OCEANS*, 2010, pp. 1–8. 2

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2242–2251. 2, 3

[39] M. Yang, K. Hu, Y. Du, Z. Wei, Z. Sheng, and J. Hu, "Underwater image enhancement based on conditional generative adversarial network," *Signal Process., Image Commun.*, vol. 81, p. 115723, 2020. 3

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008. 3, 5

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 3

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. 3

[43] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," 2021. 3, 4

[44] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, October 2021, pp. 16 259–16 268. 3

[45] R. Polikar, "Ensemble learning," in *Ensemble machine learning*. Springer, 2012, pp. 1–34. 3

[46] 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," *arXiv preprint arXiv:2201.02832*, 2022. 3

[47] Z. Ma and C. Oh, "A wavelet-based dual-stream network for underwater image enhancement," *arXiv preprint arXiv:2202.08758*, 2022. 3

[48] 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, 2016. 4, 7

[49] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE T. Image Process.*, vol. 24, no. 12, pp. 6062–6071, 2015. 4, 7

[50] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCV Workshops*, October 2021, pp. 1833–1844. 4

[51] T. Ye, M. Jiang, Y. Zhang, L. Chen, E. Chen, P. Chen, and Z. Lu, "Perceiving and modeling density is all you need for image dehazing," 2021. 4

[52] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," 2021. 6

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2021. 6

[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016. 7

[55] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020. 7

[56] A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," in *ICPR*, 2010, pp. 2366–2369. 7

[57] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE T PATTERN ANAL*, vol. 43, no. 8, pp. 2822–2837, 2021. 8, 10

[58] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013. 8

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241. 8

[60] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, “Color balance and fusion for underwater image enhancement,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 379–393, 2018. 11

[61] 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, 2005. 11

[62] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in *CVPR*, 2018, pp. 3291–3300. 12