Deep Color-Corrected Multi-scale Retinex Network for Underwater Image Enhancement

Hao Qi, Huiyu Zhou, Junyu Dong, *Member, IEEE*, and Xinghui Dong, *Member, IEEE*

***Abstract*—The acquisition of high-quality underwater images is of great importance to ocean exploration activities. However, images captured in the underwater environment often suffer from degradation due to complex imaging conditions, leading to various issues, such as color cast, low contrast and low visibility. Although many traditional methods have been used to address these issues, they usually lack robustness in diverse underwater scenes. On the other hand, deep learning techniques struggle to generalize to unseen images, due to the challenge of learning the complicated degradation process. Inspired by the success achieved by the Retinex-based methods, we decompose the Underwater Image Enhancement (UIE) task into two consecutive procedures, including color correction and visibility enhancement, and introduce a novel deep Color-Corrected Multi-scale Retinex Network (CCMSR-Net). With regard to the two procedures, this network comprises a Color Correction subnetwork (CC-Net) and a Multi-scale Retinex subnetwork (MSR-Net), which are built on top of the Hybrid Convolution-Axial Attention Block (HCAAB) that we design. Thanks to this block, the CCMSR-Net is able to efficiently capture local characteristics and the global context. Experimental results show that the CCMSR-Net outperforms, or at least performs comparably to, 11 baselines across five test sets. We believe that these promising results are due to the effective combination of color correction methods and the multi-scale Retinex model, achieved by jointly exploiting Convolutional Neural Networks (CNNs) and Transformers.**

***Index Terms*—Underwater image enhancement (UIE), underwater image processing, Retinex model, color correction, deep learning.**

I. Introduction

UNDERWATER image processing [1, 2, 3] is key to various ocean exploration activities, such as marine biology [4], underwater inspection [5] and navigation of Autonomous Underwater Vehicles (AUVs) [6]. However, images captured in the underwater environment often suffer from degradation problems, attributed to wavelength-dependent attenuation and forward and backward scattering caused by the water body [7]. Consequently, underwater images frequently exhibit apparent color cast, low contrast and blurring, which can interfere with downstream tasks, such as object detection [8] and segmentation [9]. Therefore, Underwater Image Enhancement (UIE) plays an important role in addressing that challenge.

Over the years, traditional approaches have been developed, to enhance the visual quality of underwater images, including restoration-based [10, 11, 12] and enhancement-based methods [13, 14, 15, 16]. Restoration-based methods normally aim to estimate the parameters associated with the degradation process. These parameters are then used to recover a clear underwater image. On the other hand, enhancement-based methods primarily focus on adjusting the pixel values contained in an underwater image for purposes of improving the visibility, colorfulness and contrast of the image. Among these methods, in particular, the Retinex-based approaches [16, 17] have produced impressive results.

The Retinex model, which was originally proposed by Land [18], simulates the color perception of the Human Visual System (HVS). Many UIE methods [19, 20, 21] were introduced on top of this model. According to the Retinex model, a degraded image can be perceived as the combination of the reflectance and the illumination. The reflectance represents the intrinsic property of a scene, while the illumination is the primary factor which results in image distortion [18]. Given that the effect of the illumination is estimated and removed, the desired reflectance can be then derived. An enhanced image with increased visibility and perceptual quality can be further obtained using the reflectance data.

However, the Retinex model [18, 20] cannot properly handle color cast alone (refer to “MSR” [20] in Fig. 1). Therefore, this model was often brought together with a color correction algorithm [19, 21]. Despite the effectiveness of this strategy has been demonstrated, two problems still remain with these methods. First, the predefined procedures of existing color correction algorithms struggle to adapt to unpredictable scenes. Second, it is challenging to design an appropriate *prior* for illumination estimation while the fixed prior falls short in accommodating diverse degraded phenomena. Hence, those methods usually encountered the challenge in adapting to the diverse characteristics manifested in underwater images, which might lead to under-enhancement or over-enhancement [22].

With the rapid development of deep learning techniques in the field of computer vision [23, 24, 25], they have been applying to the UIE task. Supervised learning approaches [26, 27, 28, 29] were used to train a Convolutional Neural Network (CNN) which aimed to map a degraded underwater image to its clear counterpart. The performance of these approaches severely relied on the availability of the annotated training data. However, it was hard to derive the ground-truth (clear) image for a degraded underwater image. Consequently, the limited training data failed to comprehensively capture the intricacies of the degradation process. Since the supervised

This study was in part supported by the National Natural Science Foundation of China (NSFC) (No. 42176196) and was in part supported by the Young Taishan Scholars Program (No. tsqn201909060) (Corresponding author: Xinghui Dong).

H. Qi, J. Dong and X. Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao, 266100. (e-mail: qihao@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn, xinghui.dong@ouc.edu.cn). H. Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, LE1 7RH Leicester, U.K. (e-mail: hz143@leicester.ac.uk).

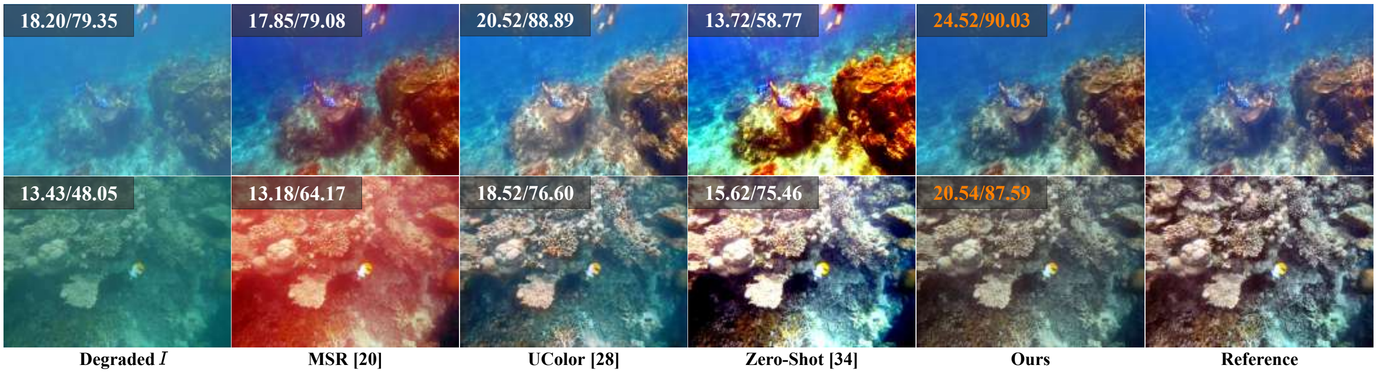

Fig. 1. Two sets of resultant images obtained from a real-world degraded underwater image in the *UIEB* [27] data set. Within each set, a degraded image I, the enhanced images produced by MSR [20], UColor [28], Zero-Shot [34] and the proposed CCMSR-Net, and the reference image are displayed from left to right. The PSNR and SSIM [39] values are shown at the top-left corner of the degraded image or a resultant image, which are computed between this image and the reference image.

learning approaches directly learned the non-linear mapping, they tended to solely focus on the limited training data. As a result, the model trained could not be popularized to unseen images (refer to “UColor” [28] in Fig. 1).

Many efforts were also made on performing the UIE task using semi-supervised learning methods [1, 30, 31, 32, 33] and unsupervised learning approaches [34, 35]. Although these methods avoided the requirement of the annotated underwater images, they were difficult to train and usually produced dissatisfying results (refer to “Zero-Shot” [34] in Fig. 1). In particular, the majority of existing deep UIE methods were designed on top of CNNs and rarely paid attention to modeling the global context. The limited size of effective receptive fields restricted their capability to capture long-range dependencies [36], which are important to image enhancement [37, 38].

Motivated by the success achieved by Retinex-based methods, we decompose the intricate UIE operation into two consecutive procedures, i.e., color correction and visibility enhancement, rather than directly learning the non-linear mapping from degraded images to their clear counterparts, performed by supervised learning methods. We propose a novel deep Color-Corrected Multi-scale Retinex Network (CCMSR-Net). This network comprises a Color Correction subnetwork (CC-Net) and a Multi-scale Retinex subnetwork (MSR-Net), in terms of the two procedures. Compared with traditional color correction algorithms [14, 15, 17, 40], which severely depend on predefined procedures and the meticulous parameter tuning, the CC-Net can be end-to-end trained and therefore is more robust to diverse degraded images.

For purposes of enhancing the visibility and visual quality of degraded images, the MSR-Net is introduced on top of the classical Multi-scale Retinex (MSR) model [20], to estimate and eliminate the illumination. Unlike some traditional Retinex-based methods [17, 22], which involve the intricate prior design, this network gets rid of the reliance on priors. The MSR-Net obtains the final enhanced image by fusing the results produced by the classical MSR model at multiple scales using an encoder-decoder network (refer to “Ours” in Fig. 1). As a result, the robustness of the MSR-Net is improved.

In addition, we design a Hybrid Convolution-Axial Attention Block (HCAAB), which captures not only local characteristics but also the global context through bringing together the convolution and self-attention mechanism [41], to address the limitations of the existing CNN-based methods [27, 28, 29]. In particular, we design a lightweight axial self-attention mechanism for the sake of overcoming the quadratic complexity associated with the original self-attention mechanism. Both the CC-Net and MSR-Net are built based on a set of HCAABs.

To our knowledge, either the CC-Net or the MSR-Net has not been explored for the UIE task before. To summarize, the contributions of this research can be identified as threefold.

- We explicitly divide the complicated UIE task into the color correction and visibility enhancement procedures and propose a deep Color-Corrected Multi-scale Retinex Network (CCMSR-Net), which contains two subnetworks, i.e., the CC-Net and the MSR-Net, with regard to the two procedures. In contrast to their traditional counterparts, the CC-Net is more robust to various degraded images while the MSR-Net does not rely on a prior. Compared with other deep learning methods, we avoid directly learning the complicated non-linear mapping from limited training data. As a result, our CCMSR-Net is able to adapt to diverse degraded underwater images.
- We design a Hybrid Convolution-Axial Attention Block (HCAAB), which combines convolutions with the lightweight axial self-attention mechanism that we introduce. This combination allows our CCMSR-Net, built on top of HCAABs, to capture both local characteristics and the global context.
- To evaluate the effectiveness of the proposed CCMSR-Net, we conduct extensive UIE experiments on five testing sets. The results provide the community with a series of benchmarks.

The remainder of this paper are organized as follows. In Section II, we review the literature related to this research. We introduce the proposed CCMSR-Net in Section III. Experimental setup and results are reported in Section IV. We examine the application of the enhanced images produced by our method to different vision tasks in Section V. Finally, we draw our conclusion in Section VI.

II. Related Work

A. Traditional Methods

In terms of whether or not the underwater Image Formation Model (IFM) is used, traditional UIE methods can be divided into restoration-based and enhancement-based methods.

1) Restoration-Based Methods: The restoration-based UIE methods are normally adopted on top of the IFM. These methods aim to estimate the physical parameters of the IFM, which describe the degradation of underwater images, for the purpose of restoring underlying clean images. Due to the similar degradation mechanisms of the images captured in the air under the foggy weather and the images acquired in the underwater environment, many methods were proposed based on the Dark Channel Prior (DCP) [42], which was originally used for image dehazing.

To adjust the color distortion of underwater images, Chiang and Chen [43] used the wavelength compensation technique. Then the DCP was used to remove the haze in the images. Drews et al. [10] proposed the Underwater Dark Channel Prior (UDCP) in order to restore the color of underwater images. The DCP was modified with respect to the characteristics of underwater scenes. In [44], the depth information was estimated based on the degree of blurriness and the DCP was adopted in order to estimate the background light.

The Generalized Dark Channel Prior (GDCP) was introduced by Peng et al. [11] through combing an adaptive color correction approach and the DCP for the sake of restoring the images with color cast and hazy degradation.

2) Enhancement-Based Methods: Basically, enhancement-based methods are focused on adjusting the pixel values of underwater images in order to enhance the visual quality without considering the IFM. For example, classic image enhancement methods, such as Histogram Equalization [45] and Gamma Correction [27], can be applied to some simple underwater scenes in which the degradation is nearly globally uniform. In [46], a multi-scale fusion strategy is used to enhance the details based on the enhanced results in RGB color space and CIELab color space. Recently, Zhang et al. [15] proposed an enhancement method, namely, MMLE, by adaptively balancing the color in the RGB color space and enhancing the contrast in the HSI color space. Wang et al. [47] performed underwater image enhancement by aligning the color distributions of underwater images in the RGB and CIELab color spaces with those of natural images.

The Retinex theory, which was proposed to simulate the human perception of lightness and color, was also used to enhance the underwater images. Fu et al. [17] first transformed the color-corrected underwater images into the CIE-Lab color space and then performed the Retinex decomposition in order to derive the illumination and reflectance compoments. Each component was further enhanced using different strategies. In [16], a multi-scale Retinex model, which was designed on top of both the bilateral and trilateral filters, was used to enhance underwater images and suppress the artifact in the resultant images. Recently, Zhuang et al. [22] proposed a variational Retinex model, i.e., HLRP, which leveraged the hyper-laplacian prior to enhance underwater images.

B. Deep Learning Based Methods

Deep learning based methods can be divided into supervised, semi-supervised and unsupervised methods according to different training strategies.

1) Supervised Methods: A large number of labeled data is normally required in order to train deep learning models using supervised methods. Two different strategies can be used to construct a training data set, including synthesizing fake underwater images using the terrestrial data and selecting the most visually pleasant image from a set of resultant underwater images enhanced using different UIE methods as the pseudo ground-truth (i.e., the reference image). Li et al. [26] constructed a synthetic underwater image data set using the indoor RGB images and the associated depth data. A UIE network was trained using this data set.

In [27], a real-world data set, namely, UIEB, was built using the second strategy. Inspired by the success of the fusion-based method [40], Li et al. [27] proposed the WaterNet, which was used to learn fusion weights for fusing different enhancement results. Jiang et al. [48] designed the LCNet which learned high-quality residuals and reduced the number of parameters by introducing Laplacian pyramids into the network. For the purpose of performing the UIE task in the coarse-to-fine manner, Cai et al. [29] proposed a cascaded deep UIE network.

2) Semi-Supervised Methods: To reduce the requirement of the labeled data, semi-supervised learning techniques were also applied to the UIE task. Guo et al. [32] employed the CycleGAN [30], which was originally proposed for unpaired image translation, to enhance underwater images. Li et al. [33] further used the Structural Similarity Index (SSIM) loss function to improve the visual quality of the images enhanced using the CycleGAN. Fabbri et al. [31] first collected both severely degraded underwater images and relatively clear images. Then the CycleGAN was used to transfer the clear images to their degraded counterparts in order to construct a labeled training set. The Underwater GAN (UGAN) was proposed, which was trained using the data set, to enhance underwater images. Guo et al. [32] introduced the UWGAN, in which the multi-scale dense block was designed in order to improve the generator.

In [49], an image translation framework and the contrastive learning technique were brought together for the purpose of deriving the more visual-pleasant results. Han et al. [1] proposed a lightweight encoder-decoder network for UIE task and leveraged channel attention and spatial attention to learn the useful representations. Recently, Qi and Dong [50] proposed a physics-aware semi-supervised UIE method which leveraged both the labeled and unlabeled data to train a network. The parameters of the IFM, estimated using the network, were used to enhance degraded underwater images.

3) Unsupervised Methods: Unlike the supervised and semi-supervised methods, which require the ground-truth data for training a network, unsupervised methods seek to exploit the general characteristics inherent in the unlabeled training data, to obtain their counterparts directly. Recently, Kar et al. [34] revealed the relationship between the input image and the corresponding controlled-perturbation image in the associated IFM parameters. In [34], a zero-shot image restoration

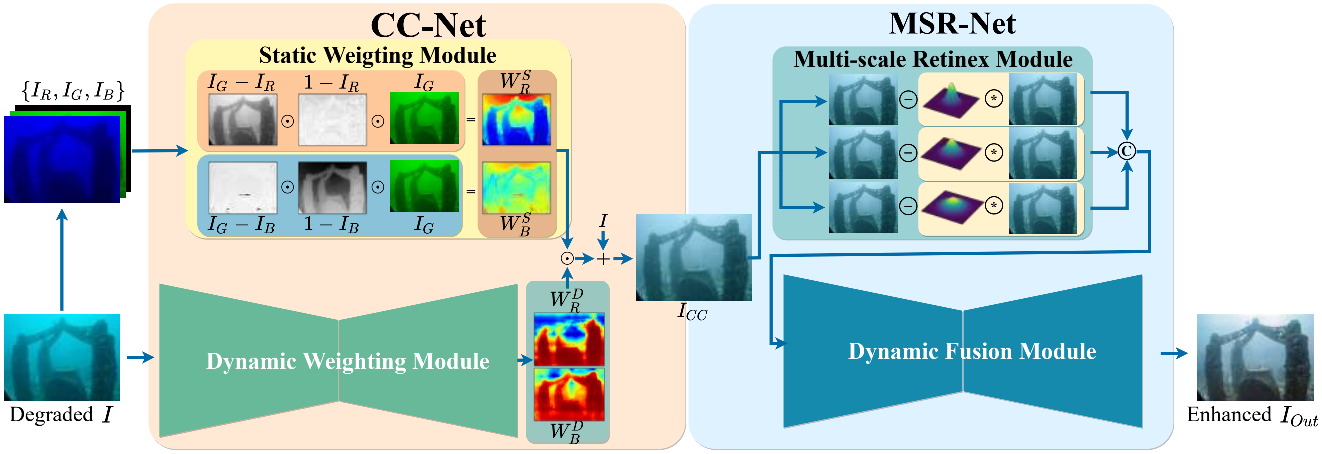

Fig. 2. The proposed Color-Corrected Multi-scale Retinex Network (CCMSR-Net), which comprises a Color Correction subnetwork (CC-Net) and a Multi-scale Retinex subnetwork (MSR-Net). The CC-Net uses both the static and dynamic weighting modules to correct color distortion. Given that the corrected image is sent to the Multi-scale Retinex (MSR) module, the MSR-Net further enhances the visibility by dynamically fusing the output of this module.

approach was proposed for the purpose of restoring clear underwater scenes on top of this relationship and the multiple regularizations on saturation, color and local consistency. In addition, Fu et al. [35] leveraged the relationship between the raw image and its re-degraded counterpart for unsupervised underwater image restoration.

In addition, deep reinforcement learning techniques [51, 52] have been employed for the UIE task, in which a reinforcement learning model was trained in order to leverage a set of image enhancement algorithms to produce the optimal result.

III. Methodology

Inspired by the achievements of Retinex-based UIE methods, we decompose the UIE task into two stages, including color correction and visibility enhancement. Regarding the color correction stage, our emphasis is put on adjusting color distributions, to rectify color distortions. In terms of the visibility enhancement stage, we leverage the classical Multi-scale Retinex (MSR) model [20] to improve the visibility of color-corrected images.

To address those issues, we propose a novel deep Color-Corrected Multi-scale Retinex Network (CCMSR-Net). As shown in Fig. 2, this network contains a Color Correction subnetwork (CC-Net) and a Multi-scale Retinex subnetwork (MSR-Net). For the purpose of performing color correction, the CC-Net uses a Static Weighting Module (SWM) and a Dynamic Weighting Module (DWM) to adjust color distributions. To further enhance the visibility, the MSR-Net is adopted based on the classic Multi-scale Retinex (MSR) model [20], which is used to estimate and eliminate the illumination. Since multiple results are produced by this model, a Dynamic Fusion Module (DFM) is used to fuse them into the final enhanced image. Both the CC-Net and MSR-Net are built on top of an encoder-decoder network. As depicted in Fig. 3, this network is developed using a set of Hybrid Convolution-Axial Attention Blocks (HCAABs) that we design. In contrast to the existing CNN-based methods [27, 28, 29], these blocks capture both local characteristics and the global context by jointly exploiting the convolution and self-attention mechanism [41].

A. Color Correction Network

According to the underwater Image Formation Model (IFM) [7], the red channel suffers from the more severe attenuation than the green and blue channels. It has also been demonstrated that the blue channel will significantly degrade in the turbid water [53]. In contrast, the green channel retains the richer information. Therefore, it is important to compensate the red and blue channels for the purpose of correcting the distorted color. We are motivated to propose a Color Correction Network (CC-Net), in order that both the channels can be adaptively compensated with the green channel. Compared with the traditional color correction methods [14, 17, 40], which rely on predefined procedures and parameters, this network can be end-to-end trained and hence is more robust to diverse degraded images. As shown in Fig. 2, the CC-Net contains two parallel modules, i.e., a Static Weighting Module (SWM) and a Dynamic Weighting Module (DWM).

1) Static Weighting Module: Inspired by traditional UIE methods [17, 40], the SWM is designed based on the differences between the red or blue channel and the green channel, to dominate the extent of information compensation. To be specific, this module generates two weight maps, i.e., W_R^S and W_B^S, by exploiting the rich information contained in the green channel. The two maps are used for the red and blue channels respectively. Given a degraded image, the computation process of the SWM can be expressed as:

$$W_R^S = (I_G - I_R) \odot (1 - I_R) \odot I_G, \quad (1)$$

$$W_B^S = (I_G - I_B) \odot (1 - I_B) \odot I_G, \quad (2)$$

where I_R, I_G and I_B denote the red, green and blue channels respectively, and $\odot$ stands for the Hadamard product. Traditional UIE methods often follow a similar procedure based on the region-wise information extraction. Although this procedure can be advantageous, the region-based computation is less efficient and the region size chosen may affect the result. In this case, the adaptability of those methods is limited. To address this problem, we employ a pixel-wise approach in the

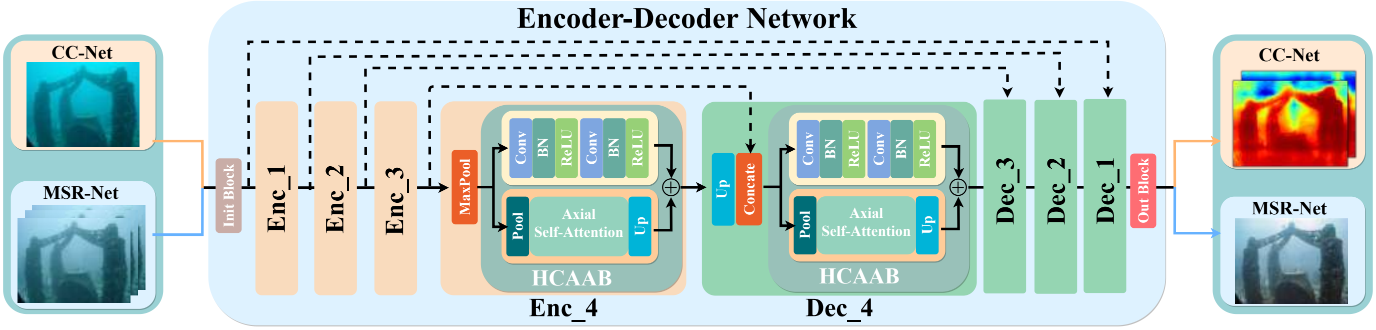

Fig. 3. The encoder-decoder network shared by both the dynamic weighting module in the CC-Net and the dynamic fusion module in the MSR-Net.

SWM while aggregating the information contained in different regions using the DWM.

2) Dynamic Weighting Module: To increase the robustness of the color correction operation, we introduce the DWM, which implicitly learns a second set of weight maps. The DWM is appended to the SWM and is implemented as an encoder-decoder network (see Fig. 3). This module is able to capture the context information across different regions through the encoding process. Then the decoding process uses this information to allocate weights to individual pixels. This strategy enables the adaptable color correction operation across multiple regions. The output of the DWM is two additional weight maps: W_R^D and W_B^D.

3) Color Correction: The compensation operation for the red and blue channels is performed as:

$$I_R^{'} = I_R + W_R^S \odot W_R^D, \tag{3}$$

$$I_B^{'} = I_B + W_B^S \odot W_B^D. \tag{4}$$

Finally, $I_R^{'}$, I_G and $I_B^{'}$ are comprised of a color image, i.e., the color-corrected image I_{CC}. Due to both the SWM and DWM, the CC-Net is able to adaptively compensate the attenuated red and blue channels using the information contained in the green channel. The color-corrected image I_{CC} presents the more balanced color distribution than the degraded image I.

B. Multi-scale Retinex Network

Although the CC-Net can reduce the distinct color distortion, other degradation phenomena still remain. To further increase the visibility, we introduce a Multi-scale Retinex Network (MSR-Net) on top of the Multi-scale Retinex (MSR) model [20]. This network aims to integrate the multi-scale results produced by the MSR model into an enhanced image. Despite some methods [17, 22] were adopted based on the Retinex model [18], they normally put their emphasis on formulating intricate priors for illumination estimation, while rarely considering the resilience of varied underwater images. In contrast, we employ the efficient MSR model for illumination estimation and increase the robustness of our method through the utilization of deep learning techniques. In this subsection, we will first introduce the MSR model and then describe the MSR-Net in detail.

1) Multi-scale Retinex Model: In terms of the Retinex [18] model, an image I can be treated as the Hadamard product of the reflectance R and the illumination L:

$$I = R \odot L. \tag{5}$$

The reflectance R represents the underlying clear image while the illumination L is the factor which causes the decrease in image quality. To avoid the numerical error, the computation of illumination elimination is usually conducted in logarithmic space [19]. According to the Single-Scale Retinex (SSR) [19] model, the estimation of L is derived by applying a Gaussian filter to I. Subsequently, the reflectance R, i.e., the enhanced image, can be obtained by eliminating the illumination. The computation can be expressed as:

$$R = log(I) - log(I * G), \tag{6}$$

where G is the Gaussian function and $*$ denotes the convolution operation.

In contrast, the Multi-scale Retinex (MSR) [20] model uses multiple SSR models to achieve the better performance. In this case, the reflectance R can be computed as:

$$R = \sum_{n=1}^{N} w_n \left[\log(I) - \log\left(I * G_n\right)\right], \tag{7}$$

where N is the number of scales and w_n stands for the weighting factor. Therefore, the effect of the MSR model is associated with the choice of w_n. Some methods [17, 22] abandoned the usage of the Gaussian function. To estimate the illumination, a variational architecture [54] was adopted with fixed priors instead. However, the complicated computation hindered the practical application of those methods and the fixed priors reduced the robustness to diverse degraded images.

2) Multi-scale Retinex Network: We are inspired to introduce a Multi-scale Retinex Network (MSR-Net) for the purpose of increasing the visibility by estimating and eliminating the illumination. This network can be end-to-end trained from degraded images, rather than relying on designing sophisticated priors [17, 22]. As displayed in Fig. 2, the MSR-Net contains two serial modules, i.e., the Multi-scale Retinex Module (MSRM) and the Dynamic Fusion Module (DFM).

The MSRM leverages the basic principle of the MSR, described in Equation (7), to generate multiple R maps, which are used as the source of the enhanced result. Thus, this module does not require complicated priors. However, the fusion process demonstrated in Equation (7) is implemented using the DFM. The input of this module is the concatenation of the R maps. The output is the final enhancement image, denoted as I_{out}. Through the proposed DFM, we fulfil the adaptive fusion of the R maps and derive the final enhancement result. It should be noted that the DFM uses a different set of weights from those utilized by the dynamic weighting module in the CC-Net even though they share the same architecture.

C. Encoder-Decoder Network

The encoder-decoder network is used by both the dynamic weighting module in the CC-Net and the dynamic fusion module in the MSR-Net. This network is built using the proposed HCAAB.

1) Encoder: As shown in Fig. 3, an *Init Block* is first used to transform the input into a set of initial feature maps at the same resolution as that of the input. In total, the network utilizes four encoder blocks, denoted as Enc_1, Enc_2, Enc_3 and Enc_4, to progressively map the initial feature maps into the latent space.

2) Decoder: Four decoder blocks, denoted as Dec_4, Dec_3, Dec_2 and Dec_1, are symmetrically built. These blocks gradually upsample and decode the latent representation. In addition, skip connection is used to pass the multi-scale feature maps from an encoder block to the corresponding decoder block, which is useful for deriving the finer-grained prediction. An *Out Block* is appended to the last block of the network, i.e., Dec_1. This block is used to transform the output of Dec_1 to the desired prediction. In terms of the dynamic weighting module in the CC-Net and the dynamic fusion module in the MSR-Net, the *Out Block* is implemented as the combination of a convolutional layer and the Sigmoid activation function and a convolutional layer respectively.

D. Hybrid Convolution-Axial Attention Block

Existing UIE methods [27, 28, 29] were normally developed based on CNN blocks. However, it is known that CNNs use limited effective receptive fields [36]. It has been illustrated that a large scope of context is crucial to the UIT task [55]. Recently, the Transformer-based methods [41] achieved the competitive performance in many vision tasks [56, 57, 58]. The key component of Transformers is the self-attention mechanism, which enables the network to exploit a global receptive field. As a result, the Transformer network is able to capture long-range dependencies. Motivated by these studies, we propose a novel Hybrid Convolution-Axial Attention Block (HCAAB), which brings the CNN and the Transformer together, to address the limitations of the CNN-based methods.

1) Hybrid Convolution-Axial Attention Block: As depicted in Fig. 3, the HCAAB consists of two parallel sub-blocks, i.e., a CNN sub-block and a Transformer sub-block. The CNN sub-block comprises two consecutive 3×3 convolutional layers. Each layer is followed by a Batch Normalization (BN) layer and a Rectified Linear Unit (ReLU) activation function. The objective of the CNN sub-block is to extract local characteristics from the image. In contrast, the Transformer sub-block facilitates the utilization of the global context information by exploiting the lightweight axial self-attention mechanism that we deliberately design. Finally, the outputs of both the sub-blocks are fused through an element-wise addition operation.

2) Lightweight Axial Self-attention Mechanism: Given an input $X \in \mathbb{R}^{H\times W\times C}$, the self-attention mechanism of vanilla Transformer [41] is computed as:

$$(Q_i, K_i, V_i) = (X_i W^Q, X_i W^K, X_i W^V), \tag{8}$$

$$\begin{aligned} Y_i &= \mathrm{Attention}(Q_i, K_i, V_i) \\ &= \mathrm{SoftMax}\left(\frac{Q_i\,(K_i)^T}{\sqrt{d}} + B\right) V_i, \end{aligned} \tag{9}$$

where Q, K and V represent the query, key and value feature maps respectively, W^Q, W^K and W^V stand for the learnable projection weights in terms of Q, K and V respectively, $i \in \{1, 2, ..., HW\}$ indicates the location in the feature maps, B is the learnable position embedding and $\mathrm{SoftMax}$ is the softmax activation function. The diagram of the self-attention mechanism is also shown in Fig. 4 (a). It can be observed that the self-attention mechanism has the quadratic complexity with regard to the resolution of the input. In this case, the computational cost will be heavy when the resolution of the input is high.

Considering the recent efforts made on improving the efficiency of Transformers [56, 59], we are inspired to introduce a lightweight axial self-attention mechanism, for purposes of reducing the demand on both the computational and memory resources. Fig. 4 (b) presents the diagram of the lightweight axial self-attention mechanism. To be specific, the input is first downsampled to a relatively small resolution, for example, 16×16 pixels. Then the lightweight axial self-attention is performed along the horizontal and vertical directions instead of operating on all pixel positions. To be specific, we apply the average pooling operation along the horizontal and vertical directions separately, after the query, key and value feature maps have been derived according to Equation (8). The results are two sets of query, key and value feature maps, denoted as $\{Q_i^h, k_i^h, V_i^h\}$ and $\{Q_j^v, k_j^v, V_j^v\}$ ($i \in \{1, 2, ..., W\}$, $j \in \{1, 2, ..., H\}$), respectively.

Furthermore, the attention mechanism can be conducted as:

$$Y_i = \mathrm{SoftMax}\left(\frac{\widetilde{Q_i^h}\widetilde{K_i^h}}{\sqrt{d}}\right) V_i^h + \mathrm{SoftMax}\left(\frac{\widetilde{Q_i^v}\widetilde{K_i^v}}{\sqrt{d}}\right) V_i^v, \tag{10}$$

where $\widetilde{Q_i^h}$, $\widetilde{K_i^h}$, $\widetilde{Q_i^v}$ and $\widetilde{K_i^v}$ are the horizontal and vertical query and key feature maps with the axial position embedding [59] and d denotes the number of channels of the feature maps. Compared with the self-attention mechanism [41], the proposed lightweight axial self-attention mechanism significantly reduces the computational and memory complexity by decomposing the self-attention computation into the horizontal and vertical attentions while effectively capturing the global context. Finally, the resultant feature maps are upsampled using the bilinear interpolation operation to the same resolution as that of the output of the CNN sub-block.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3338611

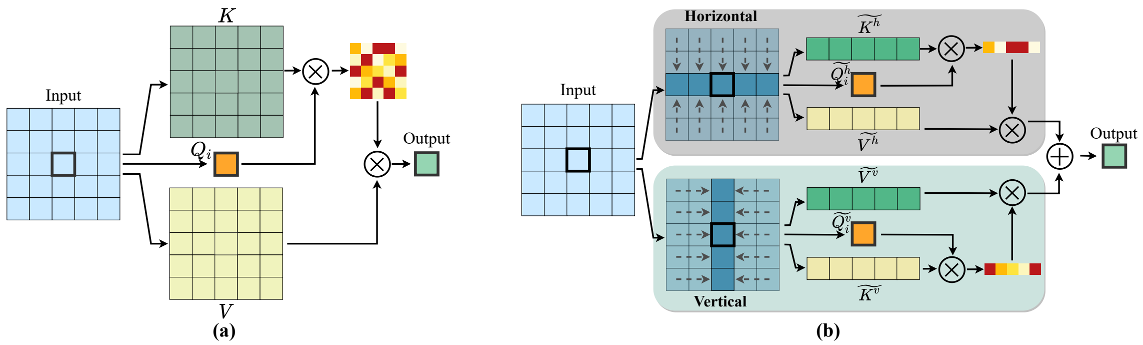

Fig. 4. The self-attention mechanism [41] (a) and the proposed lightweight axial self-attention mechanism (b). The former conducts the computation across all pixel positions, which results in the high memory and computational demands. In contrast, the latter squeezes the number of features and restricts the computation only in both the horizontal and vertical directions, which greatly reduces the memory and computational demands.

IV. Experiments and Results

To validate the effectiveness of the proposed CCMSR-Net, we conducted a series of experiments. In this section, we will report the experimental setup and results in detail. Also, we carried out an extensive ablation study, which will be introduced in this section.

A. Experimental Setup

We will first introduce the implementation details of the proposed method. Then we describe the data sets, baselines and performance measures utilized in our experiments.

1) Implementation Details: We used Pytorch to implement our method. All the experiments were conducted on a GeForce RTX 3080 Graphics Processing Unit (GPU). We trained the proposed CCMSR-Net using the AdamW [60] optimizer. The learning rate was set to 0.0001. Training images were resized to the resolution of 256×256 pixels before they were fed into the network. The mini-batch size was set to 12. We trained the CCMSR-Net for 300 epochs. The MSE loss function was used. Both the color-corrected image I_{CC} and the final enhanced image I_{Out} were supervised by the ground-truth (reference) image I_{GT}. In this case, a combined loss function was built as:

$$L = \lambda \cdot MSE(I_{CC}, I_{GT}) + MSE(I_{Out}, I_{GT}), \tag{11}$$

where the weighting factor λ was set to 0.8.

2) Data Sets: The experiments were carried out on three real-world UIE data sets, including *UIEB* [27], *RUIE* [8] and *SUIM-E* [9], and a synthetic data set, i.e., *EUVP* [61].

With regard to the real-world data sets, we followed the practice that Qi and Dong [50] performed. To be specific, the utilization of these data sets was described as follows. The UIEB [27] data set was split into the training, validation and testing sets, which contained 720, 80 and 80 images respectively. The testing set was denoted as Test-U80. In addition, 60 more challenging images, which did not have reference images, contained in the UIEB data set, were comprised of a second testing set, denoted as Test-C60. For the RUIE [8] data set, the Underwater Color Cast subset (UCCS), which contained 300 images, was used as a third testing set, namely, Test-UCCS. The official testing set of the SUIM-E [9] data set was employed as the fourth testing set, which contained 110 images. This testing set was referred to as Test-S110.

In terms of the synthetic data set, we used 2,185 images contained in the Underwater Scenes subset of the EUVP [61] data set. Among these images, 2,000 images were utilized for training and the rest were used for testing. This testing set was denoted as Test-Scenes.

3) Baselines: We compared the proposed method with 11 existing UIE methods, including two restoration-based methods, four enhancement-based methods, three supervised deep learning methods and two unsupervised learning methods. Regarding the restoration-based methods, the UDCP [10] and GDCP [11] were used. The enhancement-based methods included the MSR [20] method, the Underwater-Retinex [17] approach, the MMLE [15] method and the HLRP [22] method. For the supervised deep learning methods, we utilized WaterNet [27], UColor [28] and a recently proposed physics-aware method, namely, PA-UIENet [50]. Besides, the unsupervised learning methods consisted of the zero-shot [34] and USUIR[35] approaches.

4) Performance Measures: To evaluate the performance of different methods, we utilized three full-reference metrics, including PSNR, SSIM [39] and LPIPS [62], and two non-reference metrics, i.e., UIQM [63] and UCIQE [64]. In terms of these metrics except LPIPS, the higher value indicates the better visual quality. Besides, the number of parameters, Floating Point Operations (FLOPs) and inference latency were calculated as computational performance measures, given that a GeForce RTX 3080 GPU was utilized.

B. Experimental Results

In this subsection, we report the results derived in terms of the full-reference and non-reference metrics respectively. A qualitative analysis is also provided.

1) Full-Reference Metrics: We compared the proposed CCMSR-Net with 11 baselines on three testing sets, including Test-U80, Test-S110 and Test-Scenes. Since these testing sets contained the ground-truth data, three full-reference metrics were used to measure the performance of these methods. The results are reported in Table I. It can be seen that our CCMSR-Net achieved the best performance with regard to the PSNR, SSIM and LPIPS metrics across all the three testing sets.

TABLE I
COMPARISON OF DIFFERENT METHODS ON THE TEST-U80, TEST-S110 AND TEST-SCENES TESTING SETS IN TERMS OF THREE FULL-REFERENCE METRICS AND TWO NON-REFERENCE METRICS. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN THE **RED BOLD** AND *Blue Italic* FONTS RESPECTIVELY. THIS CONTINUES IN THE FOLLOWING TABLES.

Method	Test-U80					Test-S110					Test-Scenes				
	PSNR↑	SSIM↑	LPIPS↓	UIQM↑	UCIQE↑	PSNR↑	SSIM↑	LPIPS↓	UIQM↑	UCIQE↑	PSNR↑	SSIM↑	LPIPS↓	UIQM↑	UCIQE↑
UDCP [10]	9.37	33.62	0.51	2.69	*0.0079*	10.04	34.28	0.48	1.23	0.0047	13.00	39.71	0.44	1.68	0.0031
GDCP [11]	12.46	67.41	0.46	2.17	**0.0085**	14.67	71.29	0.37	2.12	*0.0059*	13.90	63.31	0.42	1.89	*0.0035*
MSR [20]	16.60	76.66	0.33	2.97	0.0076	14.11	73.98	0.32	1.91	0.0056	12.27	65.34	0.45	*3.92*	0.0031
Water-Retinex [17]	18.12	77.89	0.31	**4.55**	0.0069	16.14	74.40	0.35	**3.41**	0.0053	15.68	68.38	0.44	**4.33**	0.0033
MMLE [15]	18.46	76.42	0.32	*3.98*	0.0070	17.23	76.88	0.29	2.65	0.0055	14.60	61.03	0.32	3.44	0.0033
HLRP [22]	13.36	21.80	0.44	3.48	0.0077	12.52	29.20	0.45	*2.89*	*0.0059*	11.91	17.89	0.50	3.86	0.0034
WaterNet [27]	17.02	70.27	0.54	2.11	0.0067	18.80	75.19	0.46	1.47	0.0050	25.50	83.93	0.23	3.69	0.0028
UColor [28]	20.93	85.19	0.25	3.44	0.0070	20.23	84.23	*0.21*	2.30	0.0058	24.11	81.21	0.29	3.64	0.0027
PA-UIENet [50]	*22.82*	*88.87*	**0.16**	3.60	0.0072	*20.46*	*87.22*	**0.13**	2.35	0.0056	*25.79*	*85.96*	*0.17*	3.65	0.0028
Zero-Shot [34]	13.99	58.72	0.38	3.94	0.0078	13.42	53.62	0.37	1.44	**0.0060**	12.07	41.83	0.44	1.76	**0.0037**
USUIR [35]	16.72	73.56	*0.23*	3.53	0.0073	18.54	78.44	0.25	1.09	0.0054	19.26	74.42	0.34	2.39	0.0032
CCMSR-Net (Ours)	**23.14**	**89.75**	**0.16**	3.57	0.0076	**22.43**	**89.60**	**0.13**	2.33	0.0058	**26.26**	**87.00**	**0.14**	3.70	0.0027

TABLE II
COMPARISON OF DIFFERENT METHODS ON THE TEST-C60 AND TEST-UCCS TESTING SETS WITH REGARD TO TWO NON-REFERENCE METRICS.

Method	Test-C60		Test-UCCS	
	UIQM↑	UCIQE↑	UIQM↑	UCIQE↑
UDCP [10]	0.56	0.0106	0.03	0.0012
GDCP [11]	1.49	0.0093	1.24	0.0018
MSR [20]	1.88	*0.0108*	2.98	*0.0020*
Water-Retinex [17]	**2.81**	0.0091	**4.35**	*0.0020*
MMLE [15]	*2.76*	0.0102	*4.05*	0.0019
HLRP [22]	1.47	0.0108	3.89	**0.0021**
WaterNet [27]	1.28	0.0096	1.25	0.0013
UColor [28]	2.08	0.0102	3.22	0.0016
PA-UIENet [50]	2.02	0.0098	3.24	0.0016
Zero-Shot [34]	1.75	**0.0109**	2.99	**0.0021**
USUIR [35]	0.94	0.0094	0.99	0.0014
CCMSR-Net (Ours)	2.07	0.0105	3.53	0.0015

TABLE III
COMPARISON OF DIFFERENT DEEP UIE METHODS IN TERMS OF MODEL SIZE, COMPUTATIONAL COMPLEXITY AND INFERENCE LATENCY.

Method	#Params (M)	FLOPs (G)	Latency (s)
WaterNet [27]	1.09	71.42	0.0064
UColor [28]	148.04	1402.18	0.0155
PA-UIENet [50]	28.44	49.59	0.0116
Zero-Shot [34]	0.38	3.14	0.0052
USUIR [35]	0.23	14.81	0.0021
CCMSR-Net (Ours)	21.13	43.60	0.0297

Although the WaterNet [27] fused the results of multiple UIE methods while the UColor [28] incorporated the depth information estimated using GDCP [11], they were still inferior to the proposed method. Compared with traditional Retinex-based methods such as Water-Retinex [17], MSR [20] and HLRP [22], the CCMSR-Net outperformed them with large margins. In particular, the supervised deep learning methods, including WaterNet [27], UColor [28] and PA-UIENet [50], usually performed better than the unsupervised deep learning approaches, i.e., Zero-Shot [34] and USUIR [35].

2) Non-reference Metrics: Considering both the Test-C60 and Test-UCCS testing sets were provided without the ground-truth data, two non-reference metrics were utilized for assessing the performance of our method and the 11 baselines on these data sets. The model trained using the *UIEB* [27] data set was used. The results are presented in Table II. Also, the non-reference metrics were computed for the Test-U80, Test-S110 and Test-Scenes data sets. The results are reported in Table I. As can be seen, the performance of our method was slightly inferior to that of its best counterpart. However, this was also the case for other deep learning methods, such as WaterNet [27] and USUIR [35]. In the literature, it has been pointed out that both the UIQM [63] and UCIQE [64] metrics cannot exactly reflect the perception quality of enhanced images [27, 28] and may prefer the results of traditional methods to those of deep learning methods [29]. These findings may explain the relatively low UIQM and UCIQE values produced by our method.

3) Qualitative Analysis: Given the five testing sets, the results derived using the 11 baselines and our method are shown in Figs. 5, 6, 7, 8 and 9 respectively. As can be observed, traditional methods normally did not produce satisfying results. For example, the MSR [20] method usually produced an image with the high brightness but the color was not realistic. Although the other traditional Retinex-based methods [17, 22] increased the contrast, the pale color was produced. The deep learning methods also encountered difficulties in generating promising results. The WaterNet [27] tended to introduce the additional color cast while the UColor [28] method performed poorly on the degraded images captured in the turbid water (see Fig. 9). The PA-UIENet [50] could not completely remove the color cast in degraded images. In contrast, our CCMSR-Net improved different degraded images and produced images with the natural and vivid color. Even if the UIQM [63] or UCIQE [64] value produced by our method was lower than that generated by some baselines, the images that our method enhanced still manifested the satisfying visual quality.

4) Complexity Analysis: As reported in Table III, our method has the moderate space complexity and time complexity. Regarding the inference latency, our method took around 0.0297 seconds to enhance a single image, which suggests a speed of 33 frames per second (FPS). It is also noteworthy that both the UColor [28] and WaterNet [27] use preceding steps to generate the supplementary information. In contrast, our method performs the UIE task in an end-to-end manner.

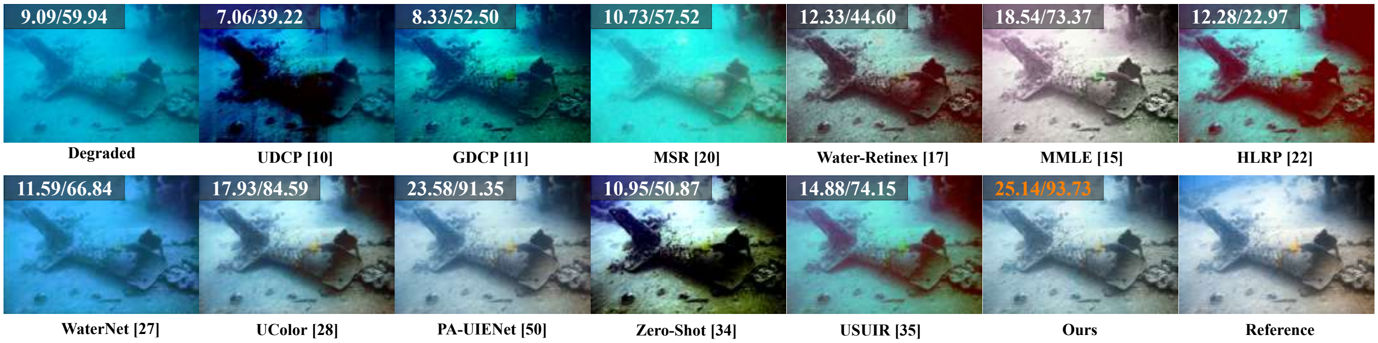

Fig. 5. The results produced by 11 baselines and our method in terms of one degraded images in the Test-U80 testing set. The PSNR and SSIM [39] values are shown at the top-left corner of the degraded image or an enhanced image, which are computed between this image and the reference image.

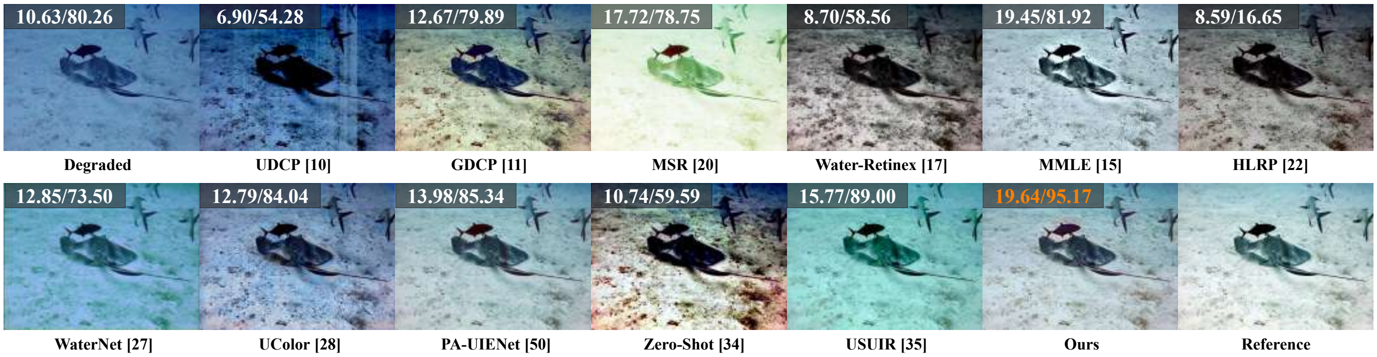

Fig. 6. The results produced by 11 baselines and our method in terms of one degraded images in the Test-S110 testing set. The PSNR and SSIM [39] values are shown at the top-left corner of the degraded image or an enhanced image, which are computed between this image and the reference image.

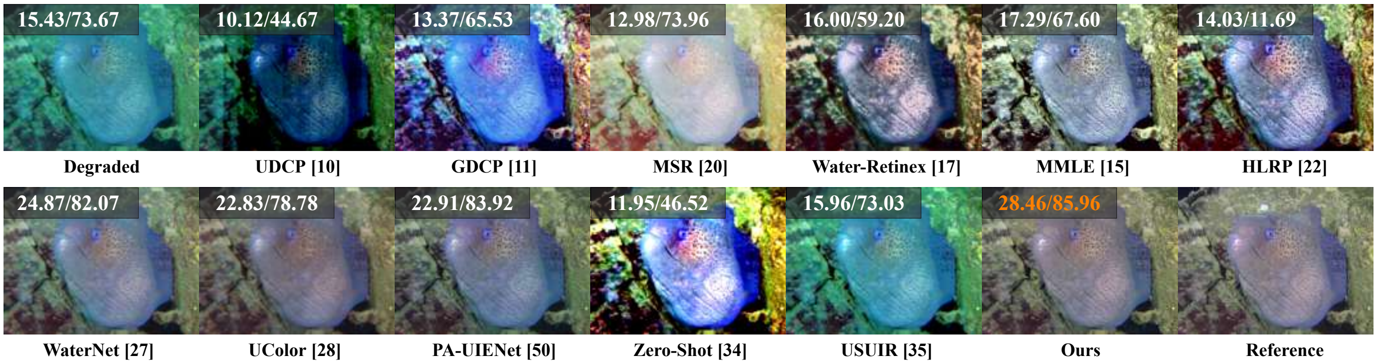

Fig. 7. The results produced by 11 baselines and our method in terms of one degraded images in the Test-Scenes testing set. The PSNR and SSIM [39] values are shown at the top-left corner of the degraded image or an enhanced image, which are computed between this image and the reference image.

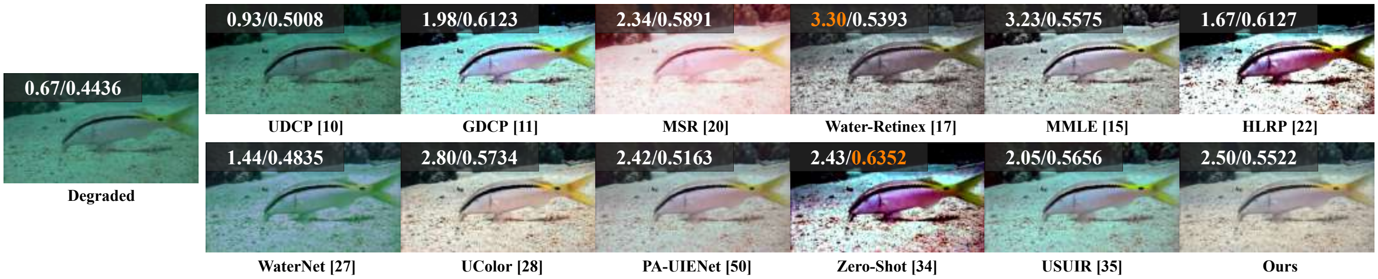

Fig. 8. The results produced by 11 baselines and our method in terms of one degraded images in the Test-C60 testing set. Both the UIQM [63] and UCIQE [64] values are shown at the top-left corner of the degraded image and enhanced images.

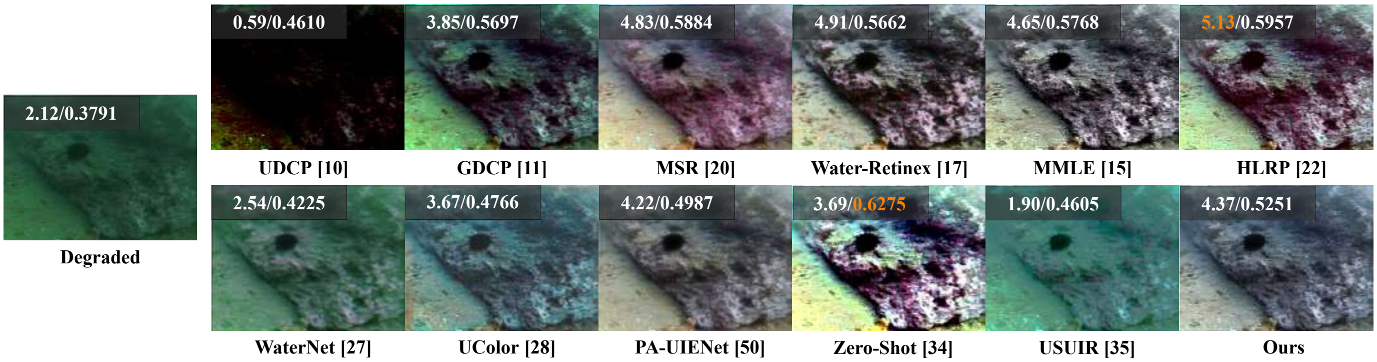

Fig. 9. The results produced by 11 baselines and our method in terms of one degraded images in the Test-UCCS testing set. Both the UIQM [63] and UCIQE [64] values are shown at the top-left corner of the degraded image and enhanced images.

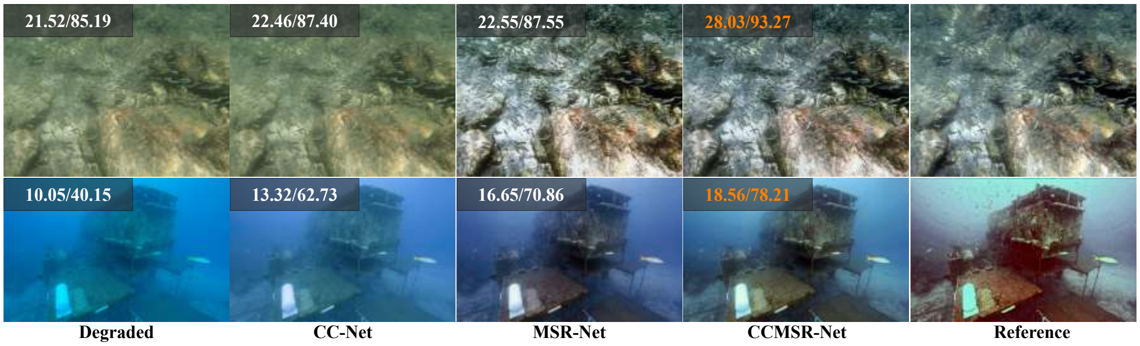

Fig. 10. Two sets of results obtained from two real-world degraded underwater images in the *UIEB* [27] data set respectively. Within each set, the degraded image, three enhanced images produced by the CC-Net, MSR-Net and CCMSR-Net, and the associated reference image are displayed from left to right. The PSNR and SSIM [39] values, computed between a degraded or enhanced image and the reference image, are shown at the top-left corner of the image.

C. Ablation Study

To investigate the effect of different components of the CCMSR-Net, we conducted a series of ablation experiments. For simplicity, only the Test-U80 testing set was used.

1) Effect of the Subnetwork: We investigated the effect of the subnetwork on the performance of the proposed CCMSR-Net by removing the MSR-Net or CC-Net separately. In this case, only the CC-Net or MSR-Net was remained. As reported in Table IV, either the CC-Net or MSR-Net produced the lower values than those generated by the CCMSR-Net (also see Fig. 10) which comprises both the subnetworks, in terms of the PSNR, SSIM, LPIPS and UCIQE metrics. In particular, a significant performance drop was observed in the PSNR (23.14db to 17.23db), SSIM (89.75 to 78.65) and LPIPS (0.16 to 0.31) values when only the CC-Net was utilized. This finding should be attributed to the fact that the CC-Net was only designed for color correction while it could not remove the illumination.

2) Effect of the Color Correction Method: For the purpose of examining the effect of the color correction method on the performance of our CCMSR-Net, we removed the CC-Net or the dynamic weighting module. In this case, we derived two variants of the CCMSR-Net, denoted as "w/o CC-Net" and "w/o dynamic weighting". The results produced by the two variants and the CCMSR-Net are reported in Table V. As can be seen, the best performance was achieved by the CCMSR-Net together with the entire CC-Net with regard to the PSNR, SSIM, LPIPS and UCIQE metrics. In particular, the PSNR, SSIM and LPIPS values derived using our network without the CC-Net were only 22.81db, 87.88 and 0.18 respectively. Given that the dynamic weighting module was removed and only the static weighting module was used for color correction, the SSIM and LPIPS values reached to 89.34 and 0.17 respectively. The results were further improved when the dynamic weighting module was added into the network.

TABLE IV
COMPARISON BETWEEN EACH OF THE TWO SUBNETWORKS AND THE PROPOSED CCMSR-NET.

	CC-Net	MSR-Net	CCMSR-Net (Ours)
PSNR↑	17.23	*22.81*	**23.14**
SSIM↑	78.65	*87.88*	**89.75**
LPIPS↓	0.31	*0.18*	**0.16**
UIQM↑	2.06	**3.63**	*3.57*
UCIQE↑	0.0067	*0.0075*	**0.0076**

To further assess the impact of the three color correction methods on color distribution equilibrium, we plot the histograms of the red, green and blue channels of a degraded image and the enhanced images produced by the three methods in Fig. 11. As can be observed, the remaining network tended to retain the residual color distortion after the entire CC-Net had been removed. In the case that the static weighting module was utilized, the enhanced image showed the more balanced color distributions. However, this image suffered from the reduced color vibrancy due to the limited adaptability of the

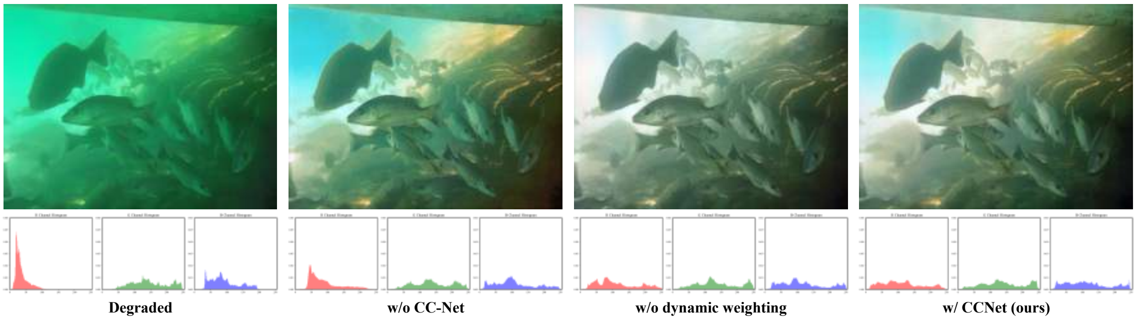

Fig. 11. The impact of color correction methods on balancing the distributions of different colors. Here, a degraded image contained in the *UIEB* [27] data set and three associated enhanced images obtained using two variants of the CCMSR-Net and the CCMSR-Net itself are shown in the upper row in turn. The histograms of the red, green and blue channels of each image are displayed below this image.

TABLE V
THE EFFECT OF THE COLOR CORRECTION METHOD TO THE PERFORMANCE OF THE CCMSR-NET.

	w/o CC-Net	w/o dynamic weighting	w/ CC-Net (Ours)
PSNR↑	*22.81*	22.64	**23.14**
SSIM↑	87.88	*89.34*	**89.75**
LPIPS↓	0.18	*0.17*	**0.16**
UIQM↑	**3.63**	*3.62*	3.57
UCIQE↑	*0.0075*	0.0074	**0.0076**

TABLE VI
THE EFFECT OF THE BASIC BLOCK THAT THE CC-NET AND MSR-NET UTILIZE.

CC-Net	MSR-Net	PSNR↑	SSIM↑	LPIPS↓	UIQM↑	UCIQE↑
CNN	CNN	21.90	88.34	0.19	3.56	0.0070
HCAAB	CNN	21.75	88.65	*0.18*	*3.61*	0.0070
CNN	HCAAB	*22.57*	*89.26*	*0.18*	**3.64**	*0.0073*
HCAAB	HCAAB	**23.14**	**89.75**	**0.16**	3.57	**0.0076**

static weighting module. In contrast, our method achieved the comprehensive color distortion elimination and produced the vivid result with balanced color distributions, benefiting from the entire CC-Net.

3) Effect of the Basic Block: To investigate the effect of the basic block on which the CCMSR-Net was built, the CNN sub-block in the proposed HCAAB was used to build the CC-Net and/or the MSR-Net instead of the entire HCAAB. In total, we derived four variants of the CCMSR-Net based on different combinations of the blocks. As shown in Table VI, the three variants, which used the HCAAB in one or two subnetworks, performed better than that built only on top of the CNN sub-block. In terms of the PSNR, SSIM, LPIPS and UCIQE metrics, the best result was derived using our CCMSR-Net. In this case, the effectiveness of the HCAAB was indicated. It is noteworthy that, however, the variant which only used the CNN sub-block still outperformed all the baselines except the PA-UIENet [50] (see Table I), regarding the PSNR, SSIM and LPIPS metrics. This finding also implies the effectiveness of our progressive enhancement network.

V. APPLYING ENHANCED IMAGES TO VISION TASKS

To investigate the usefulness of the proposed method to downstream vision tasks, we applied the enhanced images produced by our method to four different tasks, including object detection, edge detection, key-point matching and salient object detection. Correspondingly, the YOLOv5 [65], Canny [66], Scale Invariant Feature Transform (SIFT) [67] and U^{2}Net

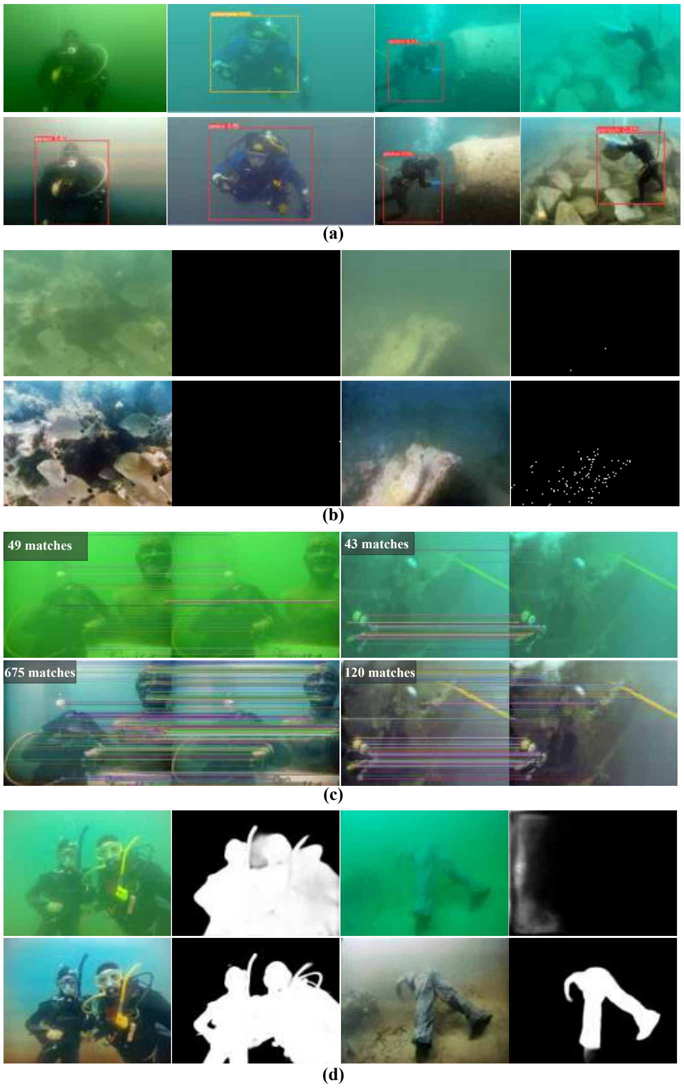

Fig. 12. Illustration of the usefulness of the enhanced images produced by our method for different vision tasks, including object detection (a), edge detection (b), key-point matching (c) and salient object detection (d), performed using YOLOv5 [65], Canny [66], SIFT [67] and U^{2}Net [68], respectively. Within each group, the upper row shows the results derived using degraded images while the lower row displays the results obtained using the enhanced images.

[68] algorithms were utilized for the four tasks respectively. The results obtained in these tasks are shown in Fig. 12.

It can be seen from Fig. 12(a) that the performance of YOLOv5 [65] was boosted by applying the enhanced images in which the visibility of objects was increased. In particular, not only missing or wrongly identified objects were correctly detected, but also the confidence of the objects detected was increased, compared with the results derived using degraded images. As shown in Fig. 12(b), our method significantly increased the performance of the edge detection task. This result should be due to the ability of our method to enhance the fuzzy details in degraded images. According to the results displayed in Fig. 12(c), the use of the enhanced images also improved the accuracy of key-point matching. As shown in Fig. 12(d), U^{2}Net [68] produced the more precise and fine-grained salient object detection results using the enhanced images produced by our method than those derived using degraded images. These results imply the usefulness of the proposed method, which is able to boost the performance of downstream vision tasks by enhancing the quality of degraded underwater images.

VI. Conclusion

In this paper, we proposed a new deep Color-Corrected Multi-scale Retinex Network, referred to as CCMSR-Net. This network comprised a Color Correction subnetwork (CC-Net) and a Multi-scale Retinex subnetwork (MSR-Net). The former was used to reduce the color distortion while the latter was designed for illumination estimation and elimination. Since the CC-Net could be end-to-end trained, it was more robust to diverse underwater images than traditional color correction algorithms. On the other hand, the MSR-Net did not rely on the priors that traditional Retinex-based methods used because it could also be end-to-end trained from degraded images. We designed a Hybrid Convolution-Axial Attention Block (HCAAB), which combined convolutions and a lightweight axial self-attention mechanism. This block was used to build the CC-Net and the MSR-Net. In contrast to CNN-based UIE methods, our CCMSR-Net was able to capture both local characteristics and the global context. The CCMSR-Net performed better than, or at least comparably to, the 11 baselines tested in this study across five testing sets. We believe that this performance should be due to the effective combination of the color correction and illumination elimination operations, conducted using the CC-Net and the MSR-Net respectively, which benefited from the joint exploitation of CNNs and Transformers. Besides, the results produced by applying the enhanced images, which were derived using our method, to four vision tasks suggested that our method can be used as a pre-processing algorithm for downstream vision tasks, e.g., object detection and key-point matching.

References

[1] G. Han, M. Wang, H. Zhu, and C. Lin, "Uiegan: Adversarial learning-based photorealistic image enhancement for intelligent underwater environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[2] J. Zhou, B. Li, D. Zhang, J. Yuan, W. Zhang, Z. Cai, and J. Shi, "Ugif-net: An efficient fully guided information flow network for underwater image enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[3] R. Chen, Z. Cai, and W. Cao, "Mffn: An underwater sensing scene image enhancement method based on multiscale feature fusion network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[4] A. S. M. Shihavuddin, N. R. Gracias, R. García, A. C. R. Gleason, and B. Gintert, "Image-based coral reef classification and thematic mapping," *Remote. Sens.*, vol. 5, pp. 1809–1841, 2013.

[5] C. Fernández-Isla, P. J. Navarro, and P. M. Alcover, "Automated visual inspection of ship hull surfaces using the wavelet transform," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–12, 2013.

[6] L. Paull, S. Saeedi, M. Seto, and H. Li, "Auv navigation and localization: A review," *IEEE Journal of Oceanic Engineering*, vol. 39, no. 1, pp. 131–149, 2014.

[7] Y. Y. Schechner and N. Karpel, "Recovery of underwater visibility and structure by polarization analysis," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 3, pp. 570–587, 2005.

[8] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[9] 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 Transactions on Image Processing*, vol. 31, pp. 6816–6830, 2022.

[10] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 825–830.

[11] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2856–2868, 2018.

[12] D. Berman, T. Treibitz, and S. Avidan, "Diving into haze-lines: Color restoration of underwater images," in *Proc. British Machine Vision Conference (BMVC)*, vol. 1, no. 2, 2017.

[13] K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*. USA: Academic Press Professional, Inc., 1994, p. 474–485.

[14] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 81–88.

[15] 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 Transactions on Image Processing*, vol. 31, pp. 3997–4010, 2022.

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

[17] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 4572–4576.

[18] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[19] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE transactions on image processing*, vol. 6, no. 3, pp. 451–462, 1997.

[20] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proceedings of 3rd IEEE international conference on image processing*, vol. 3. IEEE, 1996, pp. 1003–1006.

[21] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image processing*, vol. 6, no. 7, pp. 965–976, 1997.

[22] P. Zhuang, J. Wu, F. Porikli, and C. Li, "Underwater image enhancement with hyper-laplacian reflectance priors," *IEEE Transactions on Image Processing*, vol. 31, pp. 5442–5455, 2022.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[25] T. Ren, H. Xu, G. Jiang, M. Yu, X. Zhang, B. Wang, and T. Luo, "Reinforced swin-convs transformer for simultaneous underwater sensing scene image enhancement and super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[26] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep

underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020.

[27] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.

[28] 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 Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021.

[29] X. Cai, N. Jiang, W. Chen, J.-H. Hu, and T. Zhao, "Cure-net: A cascaded deep network for underwater image enhancement," *IEEE Journal of Oceanic Engineering*, 2023.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[31] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7159–7165.

[32] Y. Guo, H. Li, and P. Zhuang, "Underwater image enhancement using a multiscale dense generative adversarial network," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 3, pp. 862–870, 2019.

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

[34] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 205–16 215.

[35] Z. Fu, H. Lin, Y. Yang, S. Chai, L. Sun, Y. Huang, and X. Ding, "Unsupervised underwater image restoration: From a homology perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 643–651.

[36] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *NIPS*, 2016.

[37] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[38] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5896–5905.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[40] 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, 2017.

[41] 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.

[42] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[43] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756–1769, 2011.

[44] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4952–4956.

[45] R. Hummel, "Image enhancement by histogram transformation," *Unknown*, 1975.

[46] J. Yuan, Z. Cai, and W. Cao, "Tebcf: Real-world underwater image texture enhancement model based on blurriness and color fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[47] H. Wang, S. Sun, and P. Ren, "Underwater color disparities: Cues for enhancing underwater images toward natural color consistencies," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[48] N. Jiang, W. Chen, Y. Lin, T. Zhao, and C.-W. Lin, "Underwater image enhancement with lightweight cascaded network," *IEEE Transactions on Multimedia*, vol. 24, pp. 4301–4313, 2021.

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

[50] H. Qi and X. Dong, "Physics-aware semi-supervised underwater image enhancement," 2023.

[51] H. Wang, S. Sun, X. Bai, J. Wang, and P. Ren, "A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes," *IEEE Journal of Oceanic Engineering*, vol. 48, no. 2, pp. 443–461, 2023.

[52] H. Wang, S. Sun, and P. Ren, "Meta underwater camera: A smart protocol for underwater image enhancement," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 462–481, 2023.

[53] H. Lu, Y. Li, L. Zhang, and S. Serikawa, "Contrast enhancement for images in turbid water," *JOSA A*, vol. 32, no. 5, pp. 886–893, 2015.

[54] R. Kimmel, M. Elad, D. Shaked, R. Keshet, and I. Sobel, "A variational framework for retinex," *International Journal of computer vision*, vol. 52, pp. 7–23, 2003.

[55] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *Oceans 2010 Mts/IEEE Seattle*. IEEE, 2010, pp. 1–8.

[56] 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," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021.

[57] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[58] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[59] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," *ArXiv*, vol. abs/2301.13156, 2023.

[60] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv*, vol. abs/1711.05101, 2017.

[61] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[63] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.

[64] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[65] "Yolo v5," https://github.com/ultralytics/yolov5, 2020.

[66] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[67] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[68] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.