Multicolor Attribute Information Fusion for Real-World Underwater Image Enhancement

Xu Liu, *Student Member, IEEE*, Yang Zhao, *Member, IEEE*, Yuan Chen, *Member, IEEE*, Kaichen Chi, Xingguang Li, and Wei Jia, *Member, IEEE*

***Abstract*—In the challenging underwater imaging environment, underwater images often suffer from mixed degradations. Current methods are often insufficient to comprehensively address the issues of light, color, and contrast. To tackle this, we propose a novel text–image joint learning approach called multicolor attribute information fusion (MCAIF), which consists of four enhancement branches for brightness, color, contrast, and fusion, respectively. Initially, we use paired text prompts to supervise the learning of branches in a contrastive manner. To fully use the interdependence of four branches, an implicit interaction module is proposed to facilitate color attribute information collaborative communication. In view of the significant color degradation in underwater images, neural operators are presented to enhance the input image through deep filtering, which includes not only the raw channels, but also the assisted channels by utilizing a learning-based white balance (WB) model. Extensive experiments are conducted on various datasets, including the T200 test set that we collected to evaluate the overall performance of UIE algorithms. Experimental results across datasets demonstrate that our approach outperforms the state-of-the-art methods, producing visually pleasing results that align markedly better with human perception.**

***Index Terms*—Color attribute information fusion, implicit interaction module, neural operator, real-world underwater image enhancement (UIE), text–image joint learning.**

I. Introduction

THE underwater imaging environment is intricate, often resulting in degraded images with low illumination, color distortion, and low contrast [1], [2], [3], [4], [5], [6], [7], [8]. The primary cause of this degradation is the scattering and absorption of light within water, where the image quality degrades depending on the wavelength of light due to these physical phenomena. For the development and protection of marine resources [9], underwater archeology [10], ocean engineering [11], and wearable systems [12], underwater image enhancement (UIE) [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25] becomes an emerging task for both research and practical application.

There exist two main approaches for UIE, that is, traditional methods [26], [27], [28], [29], [30], [31], [32], [33] and learning based methods [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45]. Traditional methods rely on physical and mathematical models to adjust pixel values. Although traditional algorithms can effectively handle specific types of degradation, they usually cause unnatural artifacts such as overexposure, color deviation, and contrast reduction. With the advancement of deep learning, deep learning-based image processing [46], [47], [48], [49], [50] and computer vision algorithms [51], [52], [53] have achieved remarkable progress in diverse fields. In contrast, learning-based methods learn the mappings from degraded images to high-quality images in an end-to-end manner. These learning-based methods are more robust than traditional algorithms, yet the data-driven learning process of mixed degradation disregards the traditional prior knowledge. Compared to images captured on land, underwater images suffer from complex and unknown degradations such as hazing, color distortion, blur, contrast reduction, muddy and low illumination. However, for real-world underwater images, it is difficult to determine the degradation type, and most images contain more than one kind of degradation. By summarizing, we can easily find that most of the degradations can be attributed to problems in brightness, color, and contrast, as in Fig. 1(a). To combine the benefits of both conventional prior and learning-based methods, Li et al. [34] proposed a learning-driven fusion network (Water-Net). It employs three separate sub-networks to perform white balance (WB), histogram equalization (HE), and gamma correction (GC), respectively, followed by fusion via a deep network to generate enhanced underwater images. Specifically, WB is dedicated to color enhancement (CE), HE for contrast stretching (CS), and GC for brightness improvement; this design logic of Water-Net further validates the representativeness of the three degradation categories in addressing real-world underwater image issues. Motivated by Water-Net, this article further proposes a multibranch text–image joint learning model to

Received 1 July 2025; revised 8 November 2025; accepted 27 November 2025. Date of publication 10 December 2025; date of current version 8 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62272142, Grant 62476077, Grant 62262056, and Grant 62277014; and in part by the Fundamental Research Funds for the Central Universities under Grant PA2025GDGP0028. *(Corresponding authors: Yang Zhao; Yuan Chen.)*

Xu Liu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: dalong.xu.liu@ieee.org).

Yang Zhao is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China, and also with Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: yzhao@hfut.edu.cn).

Yuan Chen is with the School of Internet, Anhui University, Hefei 230039, China (e-mail: ychen@ahu.edu.cn).

Kaichen Chi is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: chikaichen@mail.nwpu.edu.cn).

Xingguang Li is with Shenzhen Polytechnical University, Shenzhen 518055, China (e-mail: lxguang@szpu.edu.cn).

Wei Jia is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China, and also with the Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Beijing 100816, China (e-mail: jiawei@hfut.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3642296

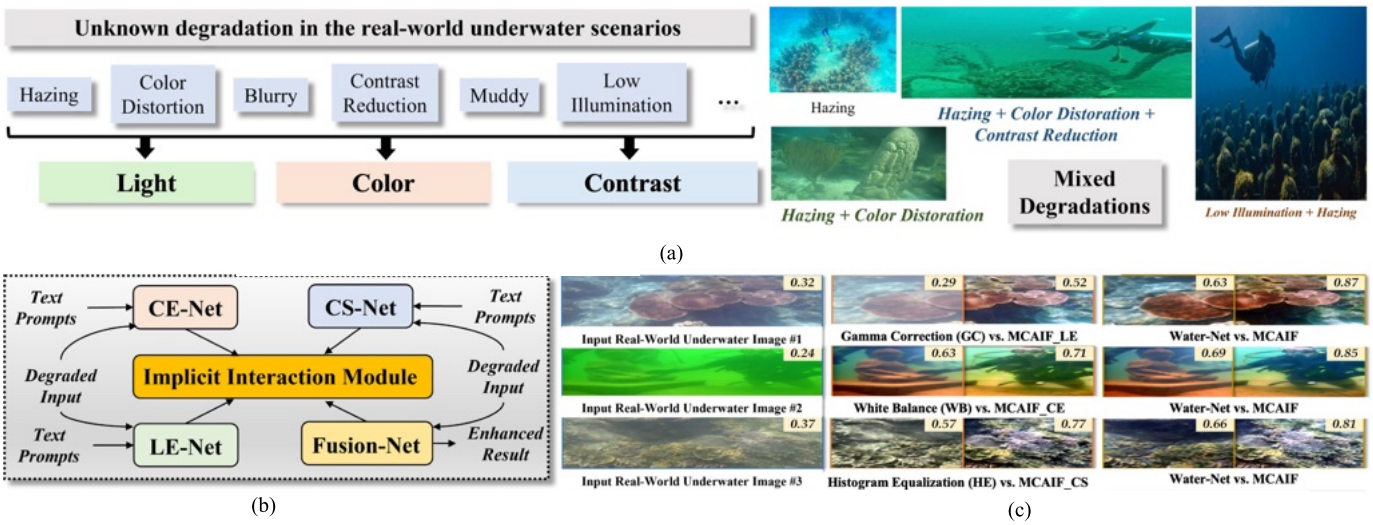

Fig. 1. Underwater images are often subject to a variety of mixed degradations in real-world scenarios. We randomly selected 500 real-world underwater images from the UIEB [34], T200 dataset, and the Internet. The presence of color shift in the input images was determined based on the gray-world assumption [54], where the means of the R, G, and B channels are ideally equal. The normalcy of image contrast was judged by the standard deviation of the luminance channel, and the normalcy of image brightness was evaluated based on the mean of the luminance channel. Finally, the experimental results show that 357 images exhibit color shift (accounting for 89.25%), 184 images have abnormally low contrast (accounting for 36.80%), and 113 images have abnormal brightness (accounting for 22.6%). Based on these observations, the degradation usually can be decomposed into three aspects of enhancement processing, that is, lightness adjustment, color adjustment, and contrast enhancement. Consequently, this article develops the MCAIF, which employs text prompts to assist three basic enhancement networks and presents an implicit interaction module to implicitly process and fuse information interactions. In (c), the top right corner represents the average visual perception scores of the viewers. (a) Summary of the unknown degradation in the real-world underwater scenarios. (b) Overview of our proposed MCAIF. (c) Visual comparison of the MCAIF and classical methods.

make full use of these traditional priors, Fig. 1(b). Visual comparisons between our method and classical UIE methods are illustrated in Fig. 1(c).

Our approach employs light, color, and contrast adjustment branches to support the main enhancement stream (the fusion network), enabling adaptation to a wide spectrum of unknown degradations. Although Water-Net delivers competitive results by embedding traditional methods, it merely learns linear fusion weights. Consequently, its performance ceiling remains tied to the handcrafted primitives. Driven by the success of large-scale vision–language models, text prompts have proven effective for various low-level tasks [55], [56], [57]. Exploiting this insight, we translate conventional algorithms (WB, GC, and HE as utilized in Water-Net) into explicit textual descriptions and inject them, together with a pre-trained text–image encoder, as additional multimodal cues. Each subnetwork thereby specializes in a distinct objective: lightening, color balancing, CS, and fusion. Empirically, we observe an implicit coupling among the four domains: brightening typically entrains improved color and contrast, which in turn reinforce overall restoration. To internalize this reciprocity, we propose an implicit interaction module that enhances cross-communication through the representation of latent features across the four branches. Thanks to these novel architectural and module designs, the multicolor attribute information fusion (MCAIF) exhibits stronger adaptability to unknown degradations than Water-Net. The network dynamically adjusts to these degradations, yielding results that are more comprehensive, robust, and stable. The contributions of this article are summarized as follows.

A. Implicit Interaction Module

We propose an implicit interaction module that enables collaborative communication of color attributes across the four interdependent domains. To our knowledge, this is the first application of implicit neural representation (INR) to color attribute features. Coordinates and corresponding features are fed into the proposed Gabor representation and decoder to yield refined features.

B. Text–Image Joint Learning

We present a novel learning strategy that exploits cross-modal text–image information by establishing text-based semantic quality requirements for the UIE model. Drawing on prior UIE knowledge, text prompts are applied to the three domains of light, color, and contrast.

C. Neural Operators

Chromaticity correction is the greatest challenge in UIE. We propose a neural operator that generates assisted RGB channels via learning-based WB and fuses raw and assisted channels through deep filtering, thereby enhancing brightness, color, and contrast to reproduce high-quality results.

D. Real-World Comprehensive Test Set

Compared with existing real-world underwater image test datasets, such as C60 [34], T40 [58], etc., there are fewer types of degradation. Furthermore, we also collect 200 underwater degraded samples, namely the Tough 200 (T200) test set, in addition to the common benchmark for the overall

performance analysis. Comprehensive subjective and objective experiments demonstrate that the proposed method effectively enhances underwater images, obtaining better results and subjective performance over other state-of-the-art algorithms.

II. Related Work

A. Underwater Image Enhancement

The main purpose of UIE is to improve underwater visual clarity. In general, UIE algorithms can be divided into three categories: prior-based, physical-model-based, and learning-based. Traditional approaches include prior-based and physical-model-based approaches. In prior-based methods, pixel values are directly adjusted. Ancuti et al. [32] proposed a novel WB model, which incorporates red and blue channel priors and grayscale world assumptions. Zhang et al. [27] performed color correction based on histogram priors and combined it with contrast enhancement to emphasize details. Liu et al. [59] presented a rank-one matrix (rank-one prior) for describing the color imaging environment. Although these traditional methods work well when dealing with specific degraded images, they are less robust when dealing with a large number of uncertainly degraded underwater images; the enhancement effect is unstable, and some images still have color casts or artifacts.

In recent years, learning-based image enhancement algorithms have become mainstream [34], [35], [36], [37], [38], [39], [40], [60], [61], [62], [63]. Learning-based UIE methods can be mainly divided into two types: convolutional neural network (CNN)-based [34], [60], [61] and generative adversarial network (GAN)-based [63], [64], [65] methods. Learning-based methods learn the mapping from degraded images to clear images end-to-end, without the need for prior or physical models. Peng et al. [65] presented a novel U-shape transformer GAN that is utilized to address the UIE task, in which a channel-wise and spatially-based attention mechanism based on a transformer enables the removal of color artifacts and casting. Zhou et al. [61] introduced a hybrid contrastive learning model for underwater imaging, which addresses the prevalent problem of suboptimal image quality. Although most learning-based models do not fully improve severely degraded underwater images, they often fail to optimize brightness, color, and contrast at the same time.

B. Application of Text Prompts

Cross-modal learning has gained increasing attention in recent years, particularly with the advent of large models. To learn how text and images match, the CLIP model [66] employs contrastive learning. It has recently been demonstrated that the CLIP model can be used for the learning of images and videos in several studies. Using unsupervised learning, Lee et al. [67] proposed CLIPtone, an approach for adjusting text-based image tones to accommodate natural language descriptions. To achieve generalizable denoising, Cheng et al. [68] developed an asymmetric encoder–decoder denoising network incorporating features from the frozen CLIP ResNet encoder. Liu et al. [69] observed that current enhanced underwater images can be divided into cool and warm tone results. Hence, they use CLIP to generate personalized UIE results for different tone preferences. Inspired by these works, we design a text–image joint learning framework with the pre-trained CLIP model to achieve different enhancing styles.

C. Implicit Neural Representation

Recently, INR has been widely adopted as a means of modeling images. According to the LIIF algorithm [70], the RGB values of pixels are determined based on the HR coordinates using MLPs. Song et al. [71] proposed a parameter-free upsampling module that uses sinusoidal positional encoding based on 2-D Fourier series to achieve lightweight image continuous super-resolution. In NeRCo [55], controllable fitting capabilities provided by INR are used to enhance low-light images. In addition to normalizing lightness degradation, it also eliminates natural noise without any additional operations. Hence, leveraging INR, we propose an implicit interaction module for color and texture co-communication, based on four domains (light, color, contrast, and fusion domains).

III. Methodology

As shown in Fig. 2, the proposed MCAIF enhances the input underwater degraded image into four different results: lightening (LE), CE, CS, and fusion results through the four branches. In order to achieve a more refined chromaticity and better suit the pre-trained CLIP model, we propose a novel WB method for color correction of raw channels. To generate four different neural operators, we use one encoder (with shared weights in four branches) and four decoders. Through the implicit interaction module, the mutual information can be communicated in four branches, which are also responsible for generating the neural operators (named as LENO, CENO, CSNO, and FUSNO).

A. Generation of Basic Channels

While CLIP has been trained on a large number of land images, underwater images are prone to color distortions, such as color deviation. Hence, raw RGB channels cannot be processed effectively using a simple operator, that is, the bilateral filter. Consequently, we propose a learning-based WB to generate assisted RGB channels. According to [32], the green channel is relatively well preserved underwater; the red and blue channels should be compensated before WB. As shown in Fig. 3, we first use the convolution layer Conv and sigmoid function Sig to compute the compensation indexes α_{Rc} and α_{Bc} as follows:

$$\begin{aligned}\alpha_{\mathrm{Rc}} &= \mathrm{Sig}\left(\mathrm{Conv}\left(\mathbf{I}_{\mathrm{R}}(x), \mathbf{I}_{\mathrm{G}}(x)\right)\right)\\ \alpha_{\mathrm{Bc}} &= \mathrm{Sig}\left(\mathrm{Conv}\left(\mathbf{I}_{\mathrm{B}}(x), \mathbf{I}_{\mathrm{G}}(x)\right)\right).\end{aligned} \tag{1}$$

In the same way as in [32], we use the following equations to obtain the compensated red $\mathbf{I}_{\mathrm{Rc}}(x)$ and blue $\mathbf{I}_{\mathrm{Bc}}(x)$ channels:

$$\begin{aligned}\mathbf{I}_{\mathrm{Rc}}(x) &= \mathbf{I}_{\mathrm{R}}(x) + \mathbf{M}_{\mathrm{R}}(x)\\ \mathbf{M}_{\mathrm{R}}(x) &= \alpha_{\mathrm{Rc}} \cdot \left(\overline{\mathbf{I}_{\mathrm{G}}} - \overline{\mathbf{I}_{\mathrm{R}}}\right) \cdot \left(1 - \mathbf{I}_{\mathrm{R}}(x)\right) \cdot \mathbf{I}_{\mathrm{G}}(x)\end{aligned} \tag{2}$$

$$\begin{aligned}\mathbf{I}_{\mathrm{Bc}}(x) &= \mathbf{I}_{\mathrm{B}}(x) + \mathbf{M}_{\mathrm{B}}(x)\\ \mathbf{M}_{\mathrm{B}}(x) &= \alpha_{\mathrm{Bc}} \cdot \left(\overline{\mathbf{I}_{\mathrm{G}}} - \overline{\mathbf{I}_{\mathrm{B}}}\right) \cdot \left(1 - \mathbf{I}_{\mathrm{B}}(x)\right) \cdot \mathbf{I}_{\mathrm{G}}(x)\end{aligned} \tag{3}$$

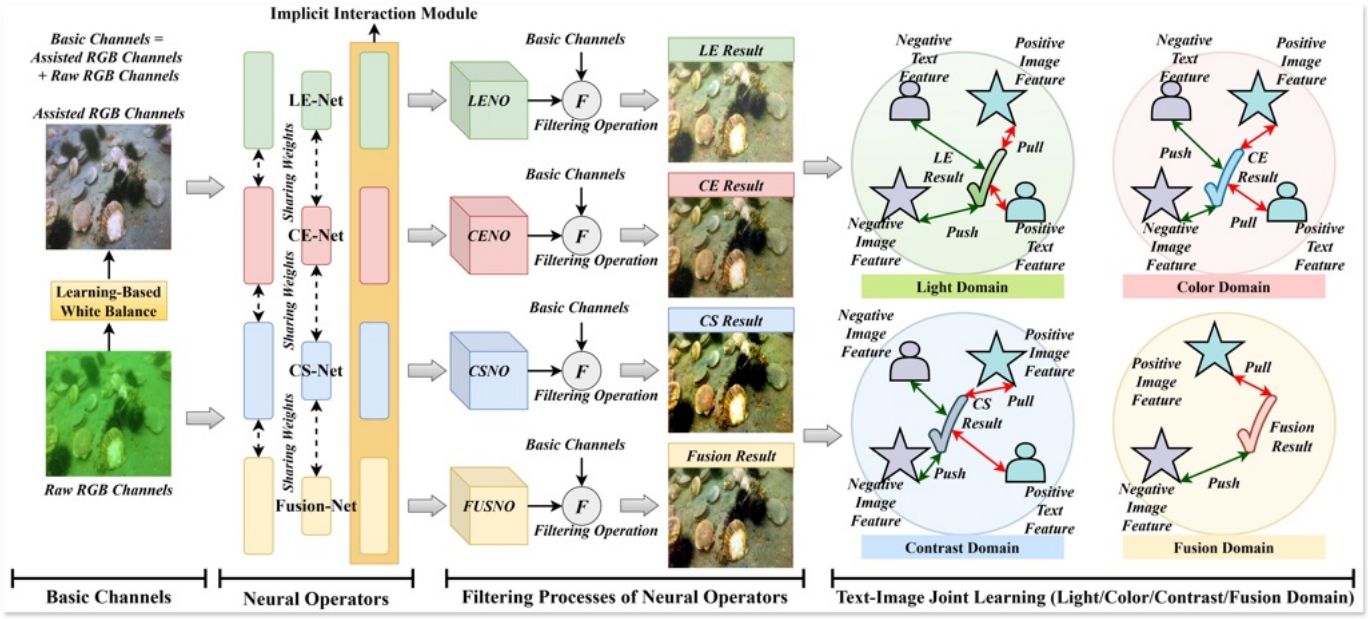

Fig. 2. Details of the MCAIF, which consists of four parts. *Generation of basic channels:* since underwater images are more color-degraded than land images, we first perform a learnable WB on the underwater images and then concatenate them with the degraded input to obtain basic channels. *Generation of neural operators and filtering processes of neural operators:* we propose the neural operators for enhancing the underwater image. Moreover, the three enhanced networks are interconnected; we also present an implicit interaction module to facilitate network learning by serving as a bridge between them. *Text–image joint learning:* the text prompts include negative and positive prompts that can generate the images through the pre-trained models. We design a multimodal joint learning manner to supervise four sub-networks (LE/CE/CS/Fusion Nets). Here, LE-Net refers to the lightening network, CE-Net to the color-enhancement network, CS-Net to the contrast-stretching network, and Fusion-Net to the fusion network.

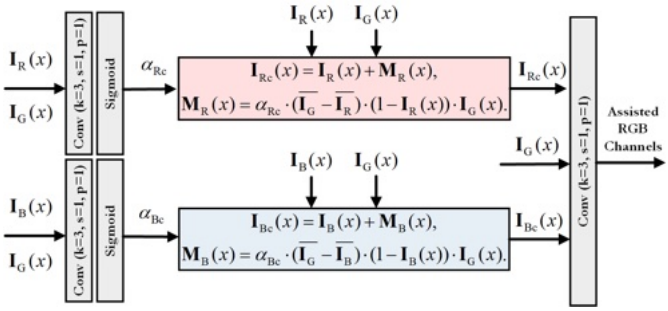

Fig. 3. Indication of the learning-based WB. k, s, and p represent the kernel size, stride, and padding number, respectively.

where $\overline{\mathbf{I}_m}$ is the mean pixel value of this channel. Now, the preparation of learning-based WB is done; as the final step, we use the last convolution layer Conv to obtain the assisted channels $\mathbf{A}_{\mathrm{R}}(x), \mathbf{A}_{\mathrm{G}}(x), and \mathbf{A}_{\mathrm{B}}(x)$:

$$\begin{aligned} \mathbf{A}_{\mathrm{RGB}}(x) &= \mathrm{Conv}\left(\mathbf{I}_{\mathrm{Rc}}(x), \mathbf{I}_{\mathrm{G}}(x), \mathbf{I}_{\mathrm{Bc}}(x)\right) \\ \mathbf{B}_{\mathrm{RGB}}(x) &= \mathrm{Cat}\left(\mathbf{A}_{\mathrm{RGB}}(x), \mathbf{R}_{\mathrm{RGB}}(x)\right) \end{aligned} \tag{4}$$

where $\mathbf{B}_{\mathrm{RGB}}(x)$ indicates the basic channels, which are generated by the concatenation Cat of the assisted RGB channels $\mathbf{A}_{\mathrm{RGB}}(x)$ and raw RGB channels $\mathbf{R}_{\mathrm{RGB}}(x)$ (i.e., $\mathbf{I}_{\mathrm{R}}(x), \mathbf{I}_{\mathrm{G}}(x), and \mathbf{I}_{\mathrm{B}}(x)$).

B. Generation of Neural Operators

The neural operators for the different stylized enhancement results are generated using an encoder (sharing weights between four branches) and four decoders. Taking into account the inter-relationships of four domains (light, color, contrast, and fusion domains), the implicit interaction module is proposed to facilitate interactivity of color attribute information during decoding, as shown in Fig. 4.

1) Implicit Interaction Module: By setting coordinates, the INR can be used to represent the relationship between latent features. Direct neural representation methods [70], however, usually fail to include high-frequency information. Lee and Jin [72] use the Fourier representation, which can learn the frequencies and corresponding Fourier coefficients. Given four features: $\mathbf{F}_{\mathrm{LE}}(x, y)$, $\mathbf{F}_{\mathrm{CE}}(x, y)$, $\mathbf{F}_{\mathrm{CS}}(x, y)$, and $\mathbf{F}_{\mathrm{FUS}}(x, y)$ and the coordinate $\mathbf{C}(x, y)$, we first convert them into the Fourier representation $\mathbf{R}_{\mathrm{Four}}(x, y)$:

$$\begin{aligned} n &= \mathrm{Cat}_*\left(\mathbf{F}_{\mathrm{LE}}, \mathbf{F}_{\mathrm{CE}}, \mathbf{F}_{\mathrm{CS}}, \mathbf{F}_{\mathrm{FUS}}, \mathbf{C}\right) \\ k &= \mathrm{Conv3}D_*(n), \alpha(x, y) = \mathrm{Conv3}D(k) \\ \omega(x, y) &= \mathrm{Conv3}D(k), \phi(x, y) = \mathrm{Conv3}D(k) \\ \mathbf{R}_{\mathrm{Four}}(x, y) &= \alpha \cdot F\begin{bmatrix} \cos(\pi(\omega r + \phi)) \\ \sin(\pi(\omega r + \phi)) \end{bmatrix} \end{aligned} \tag{5}$$

where $r(x, y)$ is the joint feature, $\alpha(x, y)$ represents the coefficient, and $\omega(x, y)$ and $\phi(x, y)$ denote the frequency and phase, respectively. We use the 3-D convolution layer Conv3D_* with $k = (3, 3, 3)$ for latent feature unfolding. Then, we utilize the 3-D convolution layer Conv3D with $k = (1, 1, 1)$ to generate these components. The operation Cat_* is defined as follows:

$$\begin{aligned} n = \mathrm{Cat}_m\big(&\mathrm{Cat}\left(\mathbf{F}_{\mathrm{LE}}, \mathbf{C}\right), \mathrm{Cat}\left(\mathbf{F}_{\mathrm{CE}}, \mathbf{C}\right), \\ &\mathrm{Cat}\left(\mathbf{F}_{\mathrm{CS}}, \mathbf{C}\right), \mathrm{Cat}\left(\mathbf{F}_{\mathrm{FUS}}, \mathbf{C}\right)\big) \end{aligned} \tag{6}$$

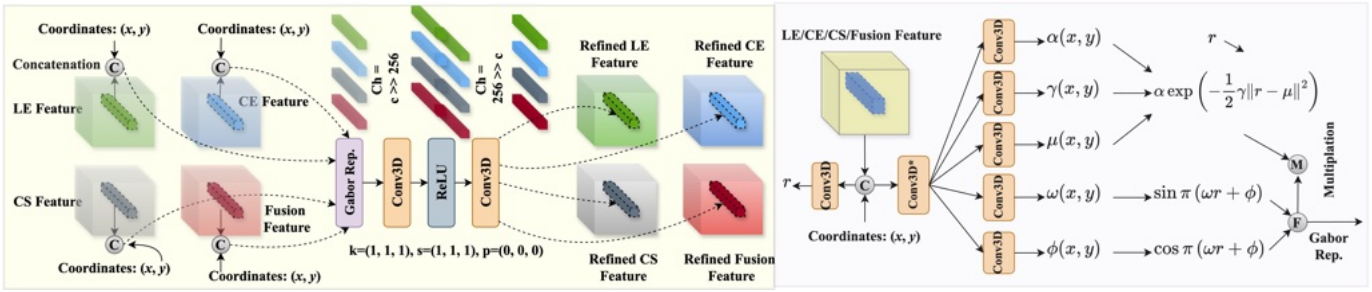

Fig. 4. Demonstration of the implicit interaction module. The block on the right is the Gabor representation. k, s, and p represent the kernel size, stride, and padding number of the 3-D convolution layer, respectively. Ch denotes the number of channels in the hidden layers.

where $n \in \mathbb{R}^{h\times w\times c\times 4}$, Cat represents concatenation in the channel dimension, and Cat_m is concatenation in the modal dimension. F denotes the Conv3D layer for fusion.

While the Fourier-based method supports frequency representation, it has limited spatial capabilities. Motivated by MFN [73], we design a Gabor-based INR representation, capturing both frequency and spatial locality. We incorporate the scale component $\gamma(x, y)$ and the mean component $\mu(x, y)$ into the Fourier representation $\mathbf{R}_{\text{Four}}(x, y)$ to generate the Gabor representation $\mathbf{R}_{\text{Gab}}(x, y)$:

$$\gamma(x,y) = \text{Conv3}D(k), \mu(x,y) = \text{Conv3}D(k)$$
$$\mathbf{R}_{\text{Gab}}(x,y) = \alpha \cdot \exp\left(-\frac{1}{2}\gamma\|r-\mu\|^2\right) \cdot F\begin{bmatrix}\cos(\pi(\omega r+\phi))\\ \sin(\pi(\omega r+\phi))\end{bmatrix}. \tag{7}$$

In Fig. 4, the decoder, which consists of two Conv3D layers and an intermediate ReLU activation, uses the above representation to calculate mutual information. Owing to the Gabor-based INR, four different branches can be interactively learned to improve joint performance.

C. Filtering Processes of Neural Operators

Inspired by deep bilateral filtering [74], different neural operators are designed to filter the underwater degraded image into corresponding results in each branch. The lightness enhancement LENO, CE CENO, contrast enhancement CSNO, and fusion FUSNO are denoted as $\mathbf{O}_{\text{LE}}$, $\mathbf{O}_{\text{CE}}$, $\mathbf{O}_{\text{CS}}$, and $\mathbf{O}_{\text{FUS}}$, respectively. For input raw channels and white-balanced assisted channels, the filtered results of neural operators are calculated as follows:

$$\begin{aligned}\mathbf{V}_{\text{raw}} &= F(\mathbf{R}_{\text{RGB}}(x), \mathbf{O}_{\text{raw}})\\ &= \text{Cat}\left(\sum_{\eta_1\in 1,2,3}\mathbf{R}_{\text{RGB}}\mathbf{O}_{\text{raw}}^{\eta_1} + \mathbf{O}_{\text{raw}}^{4},\right.\\ &\qquad \left.\sum_{\eta_2\in 5,6,7}\mathbf{R}_{\text{RGB}}\mathbf{O}_{\text{raw}}^{\eta_2} + \mathbf{O}_{\text{raw}}^{8}, \sum_{\eta_3\in 9,10,11}\mathbf{R}_{\text{RGB}}\mathbf{O}_{\text{raw}}^{\eta_3} + \mathbf{O}_{\text{raw}}^{12}\right)\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbf{V}_{\text{as}} &= F(\mathbf{A}_{\text{RGB}}(x), \mathbf{O}_{\text{as}})\\ &= \text{Cat}\left(\sum_{\eta_1\in 1,2,3}\mathbf{A}_{\text{RGB}}\mathbf{O}_{\text{as}}^{\eta_1} + \mathbf{O}_{\text{as}}^{4},\right.\\ &\qquad \left.\sum_{\eta_2\in 5,6,7}\mathbf{A}_{\text{RGB}}\mathbf{O}_{\text{as}}^{\eta_2} + \mathbf{O}_{\text{as}}^{8}, \sum_{\eta_3\in 9,10,11}\mathbf{A}_{\text{RGB}}\mathbf{O}_{\text{as}}^{\eta_3} + \mathbf{O}_{\text{as}}^{12}\right)\end{aligned}$$
$$\mathbf{V}_{\text{out}} = \text{Conv}(\mathbf{V}_{\text{raw}} + \mathbf{V}_{\text{as}}) \tag{9}$$

where $\mathbf{O}_{\text{raw}}, \mathbf{O}_{\text{as}} \in \mathbb{R}^{h\times w\times c\times 12}$ denote the raw-component neural operator and assisted-component neural operator of $\mathbf{O}_{\text{LE}}$, $\mathbf{O}_{\text{CE}}$, $\mathbf{O}_{\text{CS}}$, and $\mathbf{O}_{\text{FUS}}$. η_1, η_2, and η_3 represent channels, and F represents the filtering operation. $\mathbf{V}_{\text{raw}}$ and $\mathbf{V}_{\text{as}}$ are results filtered by the operators $\mathbf{O}_{\text{raw}}$ and $\mathbf{O}_{\text{as}}$, respectively. Cat represents concatenation in the channel dimension. Then, we add the convolution layer Conv for fusing the results $\mathbf{V}_{\text{raw}}$ and $\mathbf{V}_{\text{as}}$. Considering that the color distribution of underwater images is uneven, we use full-size neural operators rather than bilinear upsampling of low-resolution operators as in [74].

D. Text–Image Joint Learning

First of all, this article classifies degradations into three categories: light, color, and contrast, as shown in Fig. 2. In order to better learn these different domains, we select the positive prompts and negative prompts in each domain. For example, the positive and negative text prompts can be set as "bright photo." and "dark photo."

Given a degraded underwater image $\mathbf{I}$, the positive text prompts of three domains are $\mathbf{P}_{\text{bri}}$, $\mathbf{P}_{\text{col}}$, and $\mathbf{P}_{\text{con}}$, and the negative text prompts of three domains are $\mathbf{N}_{\text{bri}}$, $\mathbf{N}_{\text{col}}$, and $\mathbf{N}_{\text{con}}$. We can obtain the features from different generated style results ($\mathbf{R}_{\text{bri}}$, $\mathbf{R}_{\text{col}}$, and $\mathbf{R}_{\text{con}}$) and different text prompts ($\mathbf{P}_{\text{bri}}$, $\mathbf{P}_{\text{col}}$, $\mathbf{P}_{\text{con}}$; $\mathbf{N}_{\text{bri}}$, $\mathbf{N}_{\text{col}}$, and $\mathbf{N}_{\text{con}}$) through the text/image encoders of pre-trained CLIP. The image encoder E_{img} is used to generate image features:

$$\mathcal{F}^{\text{img}}_{\text{bri/col/con}} = E_{\text{img}}\left(\mathbf{R}_{\text{bri/col/con}}\right) \tag{10}$$

where $\mathcal{F}^{\text{img}}_{\text{bri}}$, $\mathcal{F}^{\text{img}}_{\text{col}}$, and $\mathcal{F}^{\text{img}}_{\text{con}}$ represent the light, color, and contrast style features, respectively. Similarly, text encoder E_{tex} is applied to obtain text features:

$$\mathcal{F}^{\text{tex}}_{\text{bri}-p/\text{col}-p/\text{con}-p} = E_{\text{tex}}\left(\mathbf{P}_{\text{bri}-p/\text{col}-p/\text{con}-p}\right) \tag{11}$$
$$\mathcal{F}^{\text{tex}}_{\text{bri}-n/\text{col}-n/\text{con}-n} = E_{\text{tex}}\left(\mathbf{N}_{\text{bri}-n/\text{col}-n/\text{con}-n}\right) \tag{12}$$

where $\mathcal{F}^{\text{tex}}_{\text{bri}-p}$, $\mathcal{F}^{\text{tex}}_{\text{col}-p}$, and $\mathcal{F}^{\text{tex}}_{\text{con}-p}$ are positive text features of the light, color, and contrast domains, and $\mathcal{F}^{\text{tex}}_{\text{bri}-n}$, $\mathcal{F}^{\text{tex}}_{\text{col}-n}$, and $\mathcal{F}^{\text{tex}}_{\text{con}-n}$ are negative text features of different domains.

According to [55], we use the CLIP-RN50 baseline as the image and text encoder. Additionally, we select the texts

TABLE I

EVALUATION OF $L1$ LOSS $\|.\|_1$ VERSUS CROSS-ENTROPY LOSS $\|.\|_e$ ON THE T200 TEST SET IN TERMS OF UCIQE [75], UIQM [76], NIQE [77], AND PS METRICS. W/ $L1$ LOSS AND W/ CROSS-ENTROPY LOSS DENOTE (5) WITH THESE TWO LOSSES. PS DENOTES THE PERCEPTUAL SCORE, WHICH IS OBTAINED THROUGH VISUAL EVALUATIONS CONDUCTED BY VOLUNTEERS AND RANGES FROM 0 TO 1. BOLD FONTS INDICATE THE BEST RESULT

Methods	L1 Loss $\|.\|_1$ vs. Cross Entropy Loss $\|.\|_e$			
	UCIQE [75]↑	UIQM [76]↑	NIQE [77]↓	PS↑
w $\|.\|_1$	0.6249	1.4237	4.3742	0.78
w $\|.\|_e$	**0.6317**	**1.4424**	**4.2835**	**0.83**

namely "bright photo.," "colorful photo.," and "high contrast photo." as the positive text prompts, and "dark photo.," "dull photo.," and "low contrast photo." as the negative text prompts, referring to CLIP-IQA [56]. In the research of CLIP-IQA, it conducts an extensive analysis of CLIP's sensitivity to low-level variations such as subtle color shifts or contrast changes. The results demonstrate that these concise prompts do yield meaningful alignment with low-level visual attributes. As illustrated in Fig. 2, to characterize the domain distance, we design the following text–image feature distance measurement:

$$\kappa_{\text{bri / col / con}} = S\left(\text{COS}\left(\mathcal{F}^{\text{tex}}_{\text{bri}-p/\text{col}-p/\text{con}-p}, \mathcal{F}^{\text{img}}_{\text{bri}-p/\text{col}-p/\text{con}-p}\right)\right) \tag{13}$$

$$\beta_{\text{bri / col / con}} = S\left(\text{COS}\left(\mathcal{F}^{\text{tex}}_{\text{bri}-n/\text{col}-n/\text{con}-n}, \mathcal{F}^{\text{img}}_{\text{bri}-n/\text{col}-n/\text{con}-n}\right)\right) \tag{14}$$

where COS computes the cosine distance between features and S represents the softmax function to obtain the score. Therefore, $\kappa_{(.)}$ and $\beta_{(.)}$ represent the positive and negative scores, respectively. To obtain a manner similar to contrastive learning, we set the target positive score to 1 and the target negative score to 0. MM-UIE [69] uses the $L1$ loss $\|.\|_1$ for text–image feature alignment. But we consider that the cross-entropy loss is more suitable for the (0/1) classification problem.

As shown in Table I, the cross-entropy loss achieves better performance than the $L1$ loss. Specifically, significant improvements are observed in multiple metrics, including UCIQE, UIQM, NIQE, and PS. These results demonstrate that it can substantially enhance the image color quality and visual effects. Hence, we use cross-entropy loss $\|.\|_e$ for supervision:

$$L^{\text{tex-img}}_{\text{bri / col / con}} = \left(\left\|\kappa_{\text{bri / col / con}}, 1\right\|_e + \left\|\beta_{\text{bri / col / con}}, 0\right\|_e\right). \tag{15}$$

As for all four domains (light, color, contrast, and fusion), we use the contrastive perceptual loss to control the global content of stylized enhanced and final fusion results:

$$L^{p}_{\text{bri / col / con / fus}} = \sum_{i=1}^{n} w_i \frac{\left\|\phi_i\left(\mathbf{R}_{\text{bri/col/con/fus}}\right) - \phi_i\left(\mathbf{C}\right)\right\|_1}{\left\|\phi_i\left(\mathbf{R}_{\text{bri/col/con/fus}}\right) - \phi_i\left(\mathbf{I}\right)\right\|_1} \tag{16}$$

where $\mathbf{C}$ and $\mathbf{I}$ represent the reference image and degraded image in the training dataset, and $\phi_i, I = 1, 2, \ldots, n$ denotes hidden features of the ith layer of the pre-trained VGG-19 network. We select the 1st, 3rd, 5th, 9th, and 13th layers for computing according to [61]. Then we combine the text–image loss and contrastive perceptual loss by the weight η_a for each stylized result: $L_{\text{bri/col/con}} = L^{p}_{\text{bri/col/con}} + \eta_a \cdot L^{\text{tex-img}}_{\text{bri/col/con}}$.

The generation of the comprehensive result $\mathbf{R}_{\text{fus}}$ is also under the direct image–image supervision of the reference image $\mathbf{C}$. Therefore, we add the $L1$ loss L^{1}_{fus}:

$$L^{1}_{\text{fus}} = \|\mathbf{R}_{\text{fus}} - \mathbf{C}\|_1. \tag{17}$$

Then, we use η_b to combine these two loss functions of comprehensive results: $L_{\text{fus}} = L^{1}_{\text{fus}} + \eta_b \cdot L^{p}_{\text{fus}}$. Additionally, we also use the contrastive perceptual loss to ensure that the results of learning-based WB are consistent with the ground truth. The text loss functions allow the stylized results to pull in features of positive text prompts and push away features from negative text prompts. When it comes to image loss functions, the stylized results will pull in the features of the reference images and push away the features of the inputs.

IV. EXPERIMENTS

First, this section introduces the datasets, metrics, compared methods, and settings we employed in the experiment. In Sections IV-A and IV-B, a series of qualitative and quantitative assessments are conducted to compare existing state-of-the-art approaches. In Section IV-C, we conduct various ablation studies in order to analyze each component of the proposed method.

*Datasets:*The MCAIF is trained using the UIEB dataset [34], where 800 images are used for training and the remaining 90 images (U90 test set) are employed for the ablation study evaluation. We use the color-check7 test set [32] for the color card restoration. Moreover, the Tough 200 (T200) test set, as well as the C60 test set [34], has been selected for further evaluation. Note that the T200 test set is collected by us to fully evaluate the MCAIF in real-world scenarios. In addition, we also collect some underwater videos to verify the effect of our proposed method in processing multiframe images.

In Fig. 5, a portion of the underwater images in T200 test set is captured by the DJI Osmo Action 4[1] equipped with a waterproof case, and the other part is selected from the Internet. T200 can be organized into three categories: landscapes, animals, and people. As for the landscapes, we chose natural and man-made scenes such as oceans, rivers, lakes, etc. In terms of animals, a variety of marine and river creatures were photographed, including fish, turtles, shrimps, etc. To simulate real underwater working conditions, we captured and collected images of single- and multiperson underwater working scenes.

*Evaluation Metrics:*For real-world underwater scenarios, we analyze five no-reference assessments, that is, PI [80], MA [81], NIQE [77], UIQM [76], and UCIQE [75]. These indicators focus on the quality of the image, the contrast, and the color level of the image. UIQM and UCIQE are two unique indicators for underwater images, which can be used to evaluate comprehensively. For color card restoration, we use the CIEDE2000 [82] metric. Moreover, because the above metrics are not accurate in some cases, we also conducted a survey following [34], [60], [65], and [83], the results of which are stated as the perception score. Twenty volunteers

[1]DJI Osmo Action 4: https://www.dji.com/cn/osmo-action-4

Fig. 5. Demonstration of our Tough 200 (T200) test set, which has three parts: landscapes, animals, and people.

with visual-esthetics training rated the perceptual quality of each enhanced image on a 0-to-1 scale (higher averages denote higher quality) using five criteria: color fidelity, brightness naturalness, contrast improvement, content consistency, and overall visual appeal. Method identities are concealed, and all versions of the same image are presented simultaneously in random order without time limits to ensure thorough observation and comparison.

*Comparison Methods:*In this article, we choose 10 state-of-the-art comparison methods including five traditional methods: ROP+ [59], MMLE [27], ERH [30], CBF [32], and OCM [78], and five learning-based methods: TUDA [62], U-shaped [65], Ucolor [60], Water-Net [34], and HCLR-Net [61]. Given the development of novel network architectures, we have also incorporated the CLIP-UIE [79] and MM-UIE [69] algorithms.

*Implementation Details:*The proposed method is trained with the AdamW optimizer for 500 epochs. The initial learning rate is set to 0.0001, and then halved after every 50 epochs. The batch size is set to 16. η_a and η_b are set to 2 and 0.1, respectively. All training images have been resized to 224 $\times$ 224 patches, and then normalized to the range $[0, 1]$. The experiments are implemented with the PyTorch platform on two NVIDIA GeForce RTX 3090 GPUs. The MCAIF uses the U-shaped architecture [74] with encoder, latent, and decoder blocks. Referring to [71], we normalize all the coordinates to the range $[-1, 1]$. Table II provides a summary of the architectural configurations and training settings used to achieve the reported results.

TABLE II
SUMMARY OF THE IMPLEMENTATION DETAILS

Parameter	MCAIF
Gabor representation (channels)	Same as the input tensor channels.
Hidden layers of implicit interaction module (channels)	256
$Conv$	$k = (3, 3),\ s = (1, 1), p = (1, 1)$
$Conv3D_*$	$k = (3, 3, 3),\ s = (1, 1, 1), p = (1, 1, 1)$
$Conv3D$	$k = (1, 1, 1),\ s = (1, 1, 1), p = (0, 0, 0)$
Text-image loss weight for stylized results η_a	2
Contrastive perceptual loss weight for stylized results	1
L1 loss weight for fusion results	1
Contrastive perceptual loss weight for fusion results η_b	0.1
Contrastive perceptual loss weight for learning-based white balance results	1
Initial learning rate	0.0001
Decay rate	0.5
Decay milestones	Every 50 epochs.

A. Qualitative Evaluation

The subjective results of these methods on color card, T200, and C60 test sets are illustrated in Figs. 6–8. In this section, we also provide some detail comparisons of the T200 test set and several real-world videos, as shown in Fig. 9. Based on these figures, we can make the following observations for summary. Traditional algorithms are capable of improving color degradation, resulting in improved results without obvious greenish or bluish tones. In spite of this, these processed images are often characterized by monotonous colors and severe noise in the details. The colors MCAIF restored are more vivid than those of other algorithms. There has been a significant improvement in brightness with the ROP+ and OCM methods. However, ROP+ tends to overexpose, whereas OCM exhibits color distortion. The results of MMLE do not display vibrant colors. Also, there are uneven haze effects in the results of ERH and CBF. When it comes to learning-based methods, most of them show better color and well-processed details than traditional methods. However, it should be noted that their results still contain some limitations. According to the results of ERH, CBF, U-shaped, Ucolor, and Water-Net, there is still an uneven amount of haze. Moreover, the colors of our MCAIF are more accurate than those of TUDA and HCLR-Net, and the contrast between our method and the traditional method is greater. The MCAIF algorithm presented in this article offers significant advantages in both overall and detailed color, as well as better visual effects than state-of-the-art methods.

Fig. 6 demonstrates the visual results of different color cards. To begin with, traditional algorithms are capable of improving color degradation, resulting in improved results without obvious greenish or bluish tones. In spite of this, these processed images are often characterized by monotonous

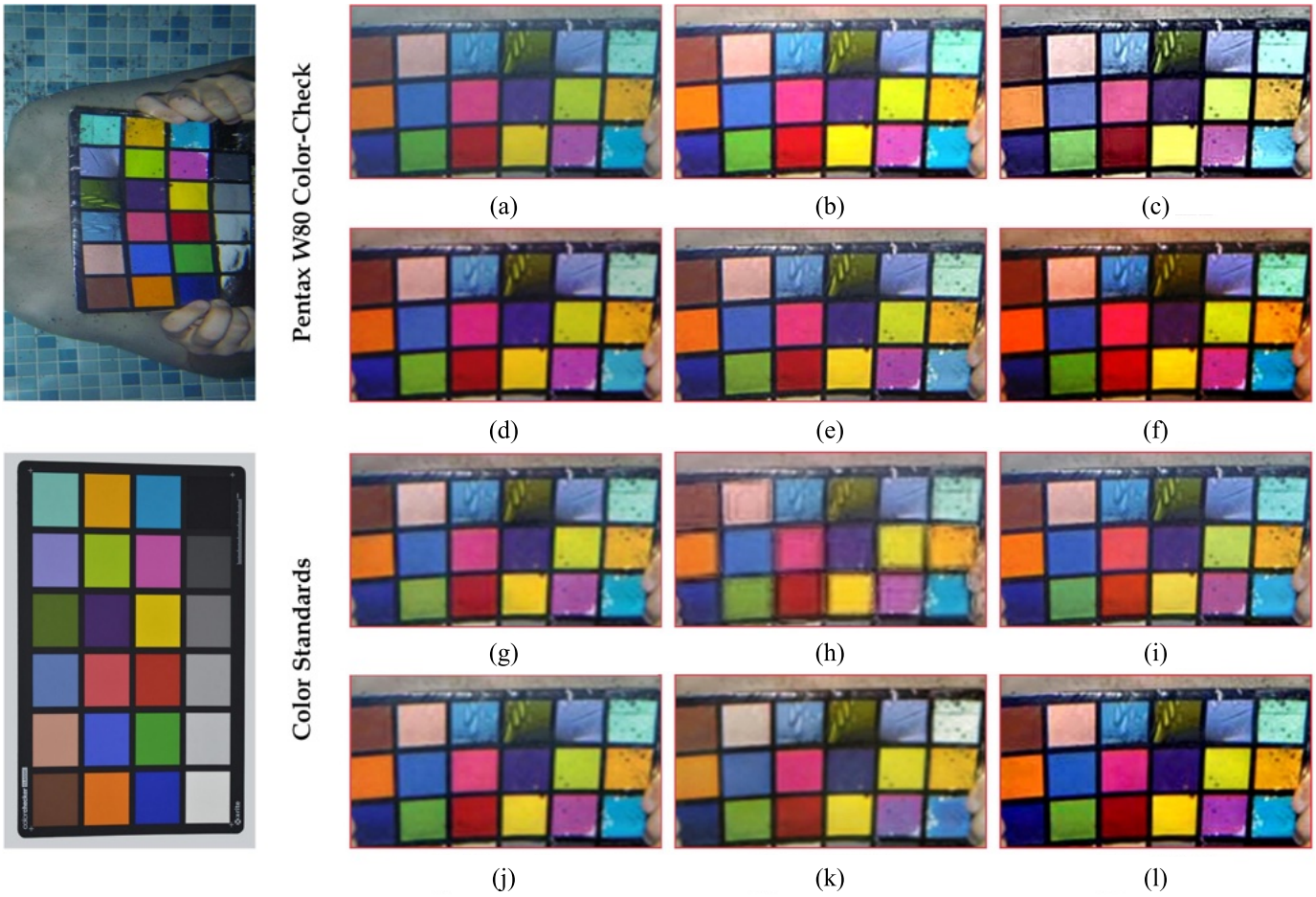

Fig. 6. Visual comparison of the color card test set. (a) Input, (b) ROP+ [59], (c) MMLE [27], (d) ERH [30], (e) CBF [32], (f) OCM [78], (g) TUDA [62], (h) U-shaped [65], (i) Ucolor [60], (j) Water-Net [34], (k) HCLR-Net [61], and (l) MCAIF.

Fig. 7. Visual comparison on the T200 test set. (a) Input, (b) TUDA [62], (c) U-shaped [65], (d) Ucolor [60], (e) Water-Net [34], (f) HCLR-Net [61], and (g) MCAIF.

colors and severe noise in the details. On the color-check7 benchmark, MCAIF reconstructs the majority of color patches more accurately than competing approaches, yielding distinctly more vivid hues. Additional visual evidence is supplied in Figs. 7 and 8 for the T200 dataset and in Fig. 9 for several real-world underwater videos. Specifically, Fig. 7 demonstrates that U-shaped and Ucolor produce severely under-exposed results, TUDA suffers from desaturated colors and weak contrast, whereas HCLR-Net elevates saturation yet reduces brightness. Fig. 8 further reveals that MCAIF removes haze more effectively in Images 1 and 2, preserves finer facial details in Image 3, and restores more vibrant, detailed colors in Image 4. Finally, Fig. 9 verifies that the proposed method achieves multiframe enhancement with consistent temporal behavior.

B. Quantitative Evaluation

In addition to qualitative analysis, we also conducted some quantitative experiments. For the color card restoration experiment in Table III, it is evident that the algorithm presented in this article has certain advantages in the CIEDE2000 scores

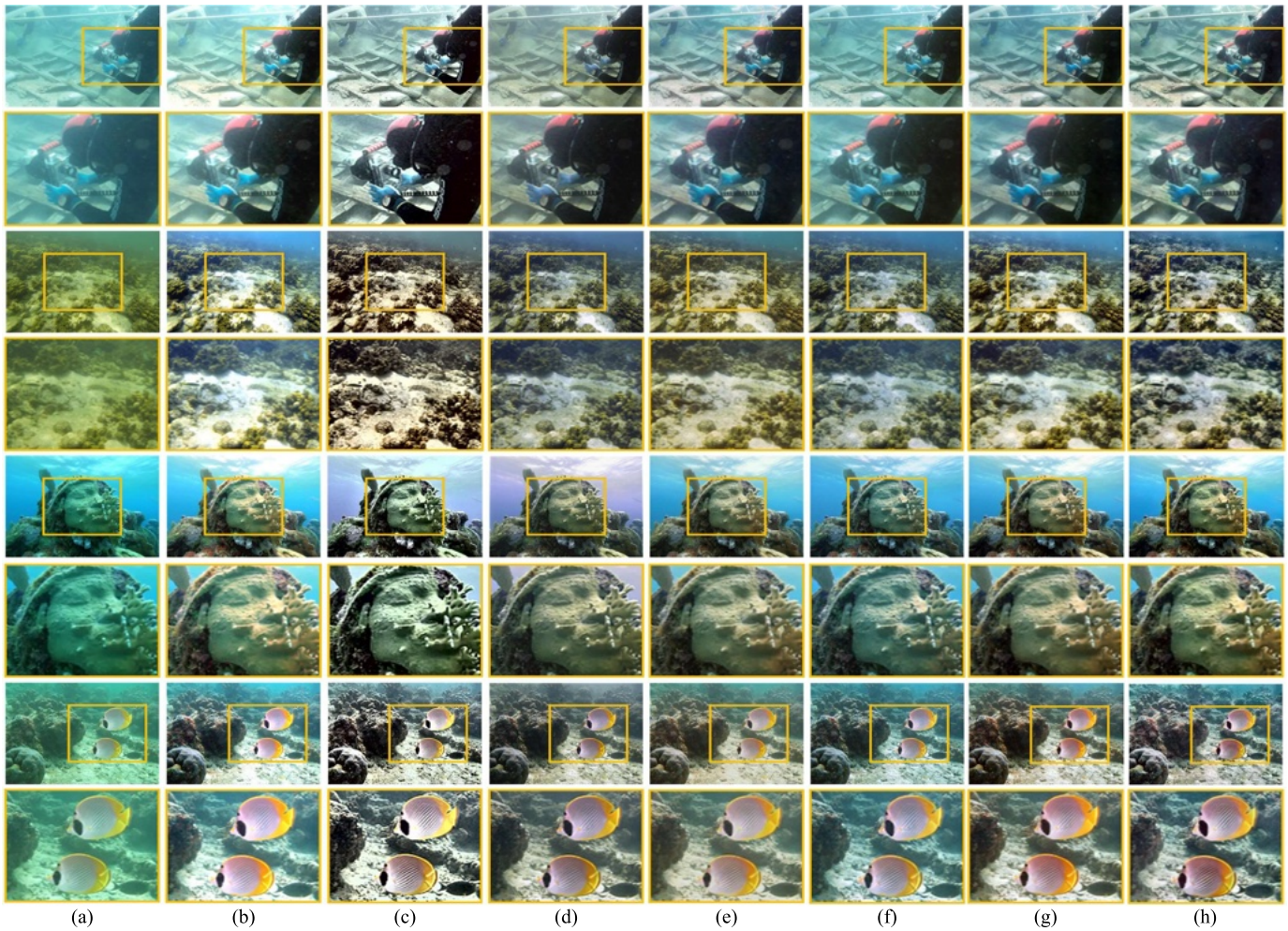

Fig. 8. Detail comparisons on the T200 test set. (a) Input, (b) ROP+ [59], (c) MMLE [27], (d) CBF [32], (e) Ucolor [60], (f) Water-Net [34], (g) HCLR-Net [61], and (h) MCAIF.

TABLE III

EVALUATIONS OF DIFFERENT METHODS ON THE COLOR-CHECKER7 TEST SET IN TERMS OF CIEDE2000 METRICS. RED AND BLUE BOLD FONTS INDICATE THE BEST TWO RESULTS

Method	Color-Check7							
	Canon D10	Fuji Z33	Olympus T6000	Olympus T8000	Panasonic TS1	Pentax W60	Pentax W80	Average
OCM [78]	17.23	24.73	15.48	16.31	12.40	14.70	15.84	16.67
MMLE [27]	14.91	17.88	18.68	13.02	11.86	16.01	15.30	15.38
ROP+ [59]	16.79	13.13	17.56	12.09	11.99	18.13	14.59	14.90
Water-Net [34]	12.05	**10.80**	13.37	10.33	10.03	12.36	10.76	11.38
ERH [30]	10.10	11.31	10.73	12.16	10.48	10.87	12.61	11.18
TUDA [62]	11.94	12.20	**10.14**	10.10	**9.28**	12.39	12.07	11.16
Ucolor [60]	**9.10**	12.76	12.34	**8.94**	12.18	9.92	10.04	10.76
HCLR-Net [61]	9.88	12.76	11.87	9.59	10.96	9.99	10.09	10.74
CBF [32]	10.98	11.35	10.95	9.51	9.47	11.54	10.83	10.66
U-Shaped [65]	**8.99**	11.37	10.71	10.29	13.26	**9.52**	**9.58**	**10.53**
MCAIF	9.35	**10.76**	**10.29**	**9.16**	**9.43**	**8.97**	**9.85**	**9.69**

of multiple color cards, and achieves the optimal average score. This indicates that the MCAIF is capable of adjusting to multiple camera parameters and of restoring color cards taken by multiple cameras. For no-reference tests, we conduct experiments on T200 and C60 test sets, and use five metrics (PI, MA, NIQE, UIQM, and UCIQE). As shown in Tables IV and V, our proposed method still has significant advantages over excellent algorithms in terms of five metrics (PI, MA, NIQE, UIQM, and UCIQE). We also conduct a perception score experiment using volunteers on the T200 test set, and the results are demonstrated in Table VI. It can be found that the perception score (PS) values of our method have

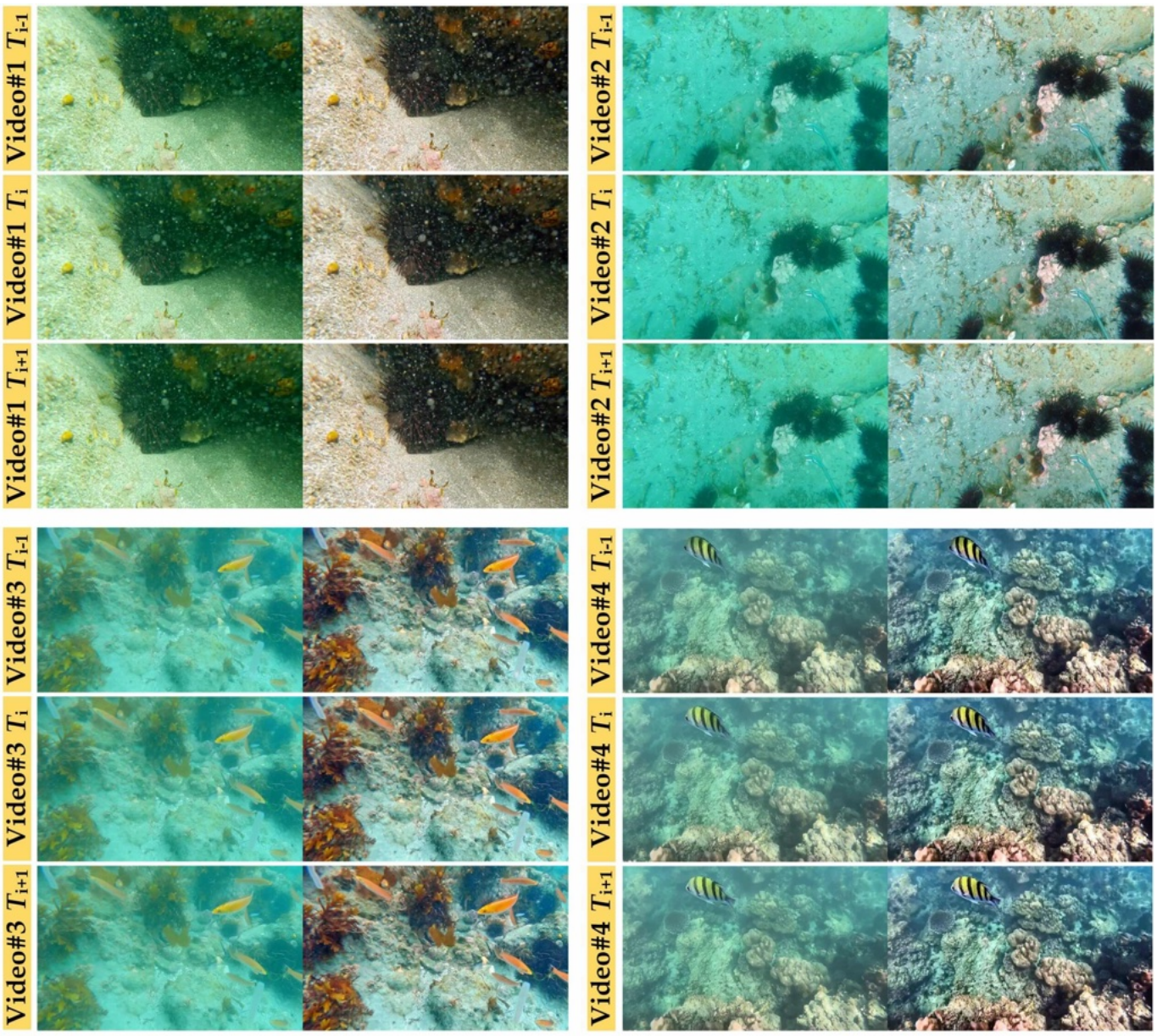

Fig. 9. Visual results of the real-world underwater videos with T_{i-1}, T_i, and T_{i+1} frames.

TABLE IV

EVALUATIONS OF DIFFERENT METHODS ON THE T200 TEST SET IN TERMS OF NO-REFERENCE (PI, MA, NIQE, UCIQE, AND UIQM) METRICS. RED AND BLUE BOLD FONTS INDICATE THE BEST TWO RESULTS

Method	No-Reference				
	PI↓	Ma↑	NIQE↓	UIQM↑	UCIQE↑
TUDA [62]	**2.9985**	8.3885	**4.3855**	1.3968	0.5973
U-Shaped [65]	3.5329	7.6624	4.7283	1.3357	0.5765
Ucolor [60]	3.4752	8.0678	5.0182	1.3388	0.5814
Water-Net [34]	3.6894	**8.3975**	5.7764	1.4209	0.6125
HCLR-Net [61]	3.0501	8.3565	4.4567	1.4346	**0.6241**
CLIP-UIE [79]	3.0877	8.3936	4.4890	1.4321	0.6201
MM-UIE-Normal [69]	3.0972	8.3433	4.4228	**1.4402**	0.6207
MCAIF	**2.9234**	**8.4367**	**4.2835**	**1.4424**	**0.6317**

certain advantages over other algorithms. This confirms that MCAIF can produce results that are more consistent with human perception. And the FLOPs of our method is the lowest, which indicates its great performance.

TABLE V

EVALUATIONS OF DIFFERENT METHODS ON THE C60 TEST SET IN TERMS OF NO-REFERENCE (PI, MA, NIQE, UCIQE, AND UIQM) METRICS. RED AND BLUE BOLD FONTS INDICATE THE BEST TWO RESULTS

Method	No-Reference				
	PI↓	Ma↑	NIQE↓	UIQM↑	UCIQE↑
TUDA [62]	**3.9516**	7.4201	**5.7435**	**1.2847**	0.5701
U-Shaped [65]	4.4887	6.8359	5.8134	1.2069	0.5384
Ucolor [60]	4.7888	6.9711	6.5488	1.1853	0.5386
Water-Net [34]	4.3408	**7.6011**	6.2818	1.2778	**0.5828**
HCLR-Net [61]	4.3046	7.3349	5.9441	1.2521	0.5803
CLIP-UIE [79]	3.9709	7.3625	5.9507	1.2838	0.5805
MM-UIE-Normal [69]	3.9913	7.3435	5.9059	1.2622	0.5793
MCAIF	**3.9221**	**7.6723**	**5.5166**	**1.2852**	**0.5872**

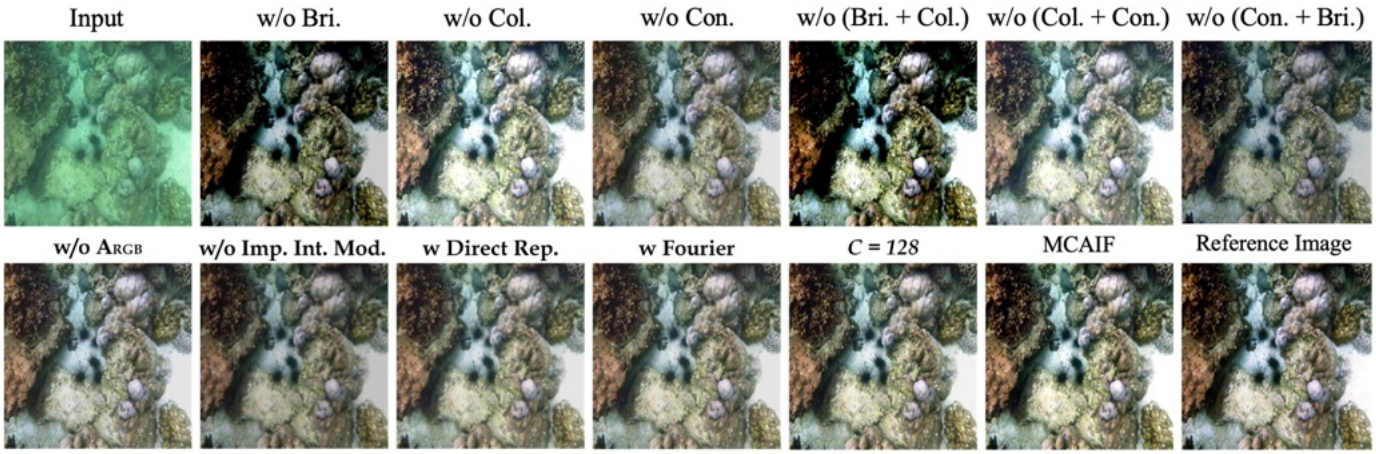

Fig. 10. Visual comparison of the ablation study on the U90 test set [60]; w/o means without. Bri., Col., and Con. represent the LE-Net, CE-Net, and CS-Net, respectively. $\mathbf{A}_{\mathrm{RGB}}$ denotes the assisted RGB channel. Imp. Int. Mod. represents the implicit interaction module. Direct Rep. is the direct neural representation, similar to the format in [70]. $C = 128$ fixes the channel count of the implicit interaction module (hidden layers) to 128.

TABLE VI

PERFORMANCE COMPARISON OF DIFFERENT METHODS. PS DENOTES THE PERCEPTUAL SCORE, WHICH RANGES FROM 0 TO 1 AND IS TESTED ON THE T200 TEST SET. RED, BLUE, AND BOLD FONTS INDICATE THE BEST THREE RESULTS. #P DENOTES THE NUMBER OF PARAMETERS

Method	Performance		
	PS↑	FLOPs (G)↓	#P (M)↓
TUDA [62]	0.56	**174.40**	31.36
U-Shaped [65]	0.49	419.48	65.60
Ucolor [60]	0.72	2805.34	148.77
Water-Net [34]	0.58	1937.00	**1.09**
HCLR-Net [61]	**0.75**	401.97	**4.87**
MCAIF	**0.83**	**116.95**	44.16

TABLE VII

EVALUATIONS OF ABLATION STUDY ON THE U90 TEST SET INDICATED BY THE UCIQE METRIC. BOLD FONT INDICATES THE BEST RESULT

Method	Ablation Study
	UCIQE↑
w/o Bri.	0.6234
w/o Col.	0.6199
w/o Con.	0.6224
w/o (Bri. + Col.)	0.5977
w/o (Col. + Con.)	0.6086
w/o (Con. + Bri.)	0.6076
w/o $\mathbf{A}_{\mathrm{RGB}}$	0.6141
w/o Imp. Int. Mod.	0.6185
w Direct Rep.	0.6199
w Fourier	0.6283
$C = 128$	0.6402
MCAIF	**0.6478**

C. Ablation Study

An ablation experiment is designed to demonstrate the efficacy of each component. Fig. 10 and Table VII show visual and evaluation results of the ablation study; w/o means without. Bri., Col., and Con. represent the LE-Net, CE-Net, and CS-Net, respectively. $\mathbf{A}_{\mathrm{RGB}}$ denotes the assisted RGB channel. Imp. Int. Mod. represents the implicit interaction module. Direct Rep. is the direct neural representation, similar to the format in [70]. $C = 128$ fixes the channel count of the implicit interaction module (hidden layers) to 128.

First of all, we can observe that the light domain will increase the brightness, the color domain will make the colors more vivid, and the contrast domain will enhance the contrast to a great extent. The lack of branches will impact the final result. For example, if the color branch is not present, the contrast and brightness will be oversaturated, and the final result will be severely overexposed. Without the contrast

branch, the final result will exhibit low contrast and fogging, and the overall image will be dark in the absence of the brightness branch. Without $\mathbf{A}_{\mathrm{RGB}}$, it is evident that it cannot accurately restore the color details of some scenes due to the complex color degradation of underwater images. Adopting the Gabor representation and setting the hidden layer channel dimension to 128 yields the best performance.

V. Conclusion

In this article, we propose a novel text-heuristic UIE method, which can generate different style clear underwater images with four branches, that is, light, color, contrast, and fusion. A text–image joint learning manner is designed for these branches based on the pre-trained CLIP model and corresponding positive and negative text prompts. In consideration of the inter-correlation of these four domains, an implicit interaction module is proposed to fully use the mutual information of different branches. In addition, the neural operators are presented for better color and content transfer. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods in both subjective and objective scores. In future work, we will embed the pre-trained multimodal model into a video enhancement backbone, using it as a bridge between single-frame and multiframe CE networks to lift existing underwater image enhancers to the video domain.

Acknowledgment

The computation is completed on the HPC Platform of Hefei University of Technology.

References

[1] S. Raveendran, M. D. Patil, and G. K. Birajdar, "Underwater image enhancement: A comprehensive review, recent trends, challenges and applications," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 5413–5467, Oct. 2021.

[2] S. P. González-Sabbagh and A. Robles-Kelly, "A survey on underwater computer vision," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–39, Dec. 2023.

[3] F. Zhang, S. You, Y. Li, and Y. Fu, "Atlantis: Enabling underwater depth estimation with stable diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11852–11861.

[4] Y. Tang, C. Zhu, R. Wan, C. Xu, and B. Shi, "Neural underwater scene representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11780–11789.

[5] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *Int. J. Comput. Vis.*, vol. 133, no. 11, pp. 1–19, Nov. 2025.

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

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

[8] S. Liu, H. Fan, S. Lin, Q. Wang, N. Ding, and Y. Tang, "Adaptive learning attention network for underwater image enhancement," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5326–5333, Apr. 2022.

[9] C. M. Duarte et al., "Rebuilding marine life," *Nature*, vol. 580, no. 7801, pp. 39–51, 2020.

[10] H. Renkewitz, S. Matz, S. Thomas, J. Schwendner, and J. Albiez, "Evaluation of a high-end laserscanner for underwater archeology," in *Proc. OCEANS: San Diego Porto*, Sep. 2021, pp. 1–5.

[11] B.-H. Jun and H. Shim, "A dexterous crabster robot explores the seafloor," *XRDS, Crossroads, ACM Mag. Students*, vol. 20, no. 3, pp. 38–45, Mar. 2014.

[12] H. Xia et al., "Perspective on wearable systems for human underwater perceptual enhancement," *IEEE Trans. Cybern.*, vol. 55, no. 2, pp. 698–711, Feb. 2025.

[13] H. Xia, B. Bao, F. Liao, J. Chen, B. Wang, and Z. Li, "A patch-based method for underwater image enhancement with denoising diffusion models," *IEEE Trans. Cybern.*, vol. 55, no. 1, pp. 269–281, Jan. 2025.

[14] L. Shen et al., "U^{2}PNet: An unsupervised underwater image-restoration network using polarization," *IEEE Trans. Cybern.*, vol. 54, no. 9, pp. 5164–5177, Sep. 2024.

[15] F. Zhou, S. Zhang, Y. Huang, P. Zhu, and Y. Zhang, "SCN: A novel underwater images enhancement method based on single channel network model," *IEEE J. Ocean. Eng.*, vol. 50, no. 2, pp. 758–775, Apr. 2025.

[16] Y. Qing, L. Shen, Z. Fang, and Y. Wang, "HG2former: HSV-gamma guided transformers for efficient underwater image enhancement," *IEEE J. Ocean. Eng.*, vol. 50, no. 2, pp. 866–878, Apr. 2025.

[17] X. Zhou, M. Peng, Q. Jiang, R. Cong, J. Wang, and Y. Chen, "CA-Net: Cascaded adaptive network for underwater image enhancement," *IEEE J. Ocean. Eng.*, vol. 50, no. 2, pp. 879–897, Apr. 2025.

[18] X. Ding, Y. Sui, and J. Zhang, "Vector quantized underwater image enhancement with transformers," *IEEE J. Ocean. Eng.*, vol. 50, no. 1, pp. 136–149, Jan. 2025.

[19] G. Han, S. Yu, H. Zhu, and Y. Zhu, "UMCTN: Real-world underwater image enhancement based on transformer with multikernel convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4207715.

[20] L. Zhou, H. Tan, T. Liu, Y. Zhu, J. Liang, and Y. Tao, "WDFN: Wavelet-based decomposition-fusion network for low-light underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4207917.

[21] S. Li et al., "Realistic simulation of underwater scene for image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5620314.

[22] H. Zhang, H. Xu, X. Yu, X. Zhang, X. Gao, and C. Wu, "CDF-UIE: Leveraging cross-domain fusion for underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4203715.

[23] X. Liu, Y. Jiang, Y. Wang, T. Liu, and J. Wang, "MDA-Net: A multidistribution aware network for underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5603713.

[24] H. Wang, K. Köser, and P. Ren, "Large foundation model empowered discriminative underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5609317.

[25] Y. Zhang, J. Yuan, and Z. Cai, "DCGF: Diffusion-color-guided framework for underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4201012.

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

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

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

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

[30] H. Song, L. Chang, Z. Chen, and P. Ren, "Enhancement-registration-homogenization (ERH): A comprehensive underwater visual reconstruction paradigm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6953–6967, Oct. 2022.

[31] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2822–2837, Aug. 2021.

[32] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2018.

[33] J. Zhou, S. Wang, Z. Lin, Q. Jiang, and F. Sohel, "A pixel distribution remapping and multi-prior retinex variational model for underwater image enhancement," *IEEE Trans. Multimedia*, vol. 26, pp. 7838–7849, 2024.

[34] C. Li et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2020.

[35] Y. Wang, Y. Cao, J. Zhang, F. Wu, and Z.-J. Zha, "Leveraging deep statistics for underwater image enhancement," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 3s, pp. 1–20, Oct. 2021.

[36] P. Mu, H. Xu, Z. Liu, Z. Wang, S. Chan, and C. Bai, "A generalized physical-knowledge-guided dynamic model for underwater image enhancement," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7111–7120.

[37] Y. Tang, H. Kawasaki, and T. Iwaguchi, "Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5419–5427.

[38] P. Mu, H. Qian, and C. Bai, "Structure-inferred bi-level model for underwater image enhancement," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2286–2295.

[39] Z. Zhang, Z. Jiang, J. Liu, X. Fan, and R. Liu, "WaterFlow: Heuristic normalizing flow for underwater image enhancement and beyond," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7314–7323.

[40] P. Sharma, I. Bisht, and A. Sur, "Wavelength-based attributed deep neural network for underwater image restoration," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–23, Jan. 2023.

[41] C. Zhao, W. Cai, C. Dong, and C. Hu, "Wavelet-based Fourier information interaction with frequency diffusion adjustment for underwater image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 8281–8291.

[42] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18145–18155.

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

[44] Q. Jiang, Y. Kang, Z. Wang, W. Ren, and C. Li, "Perception-driven deep underwater image enhancement without paired supervision," *IEEE Trans. Multimedia*, vol. 26, pp. 4884–4897, 2024.

[45] Z. Wang, L. Shen, Y. Yu, and Y. Hui, "UIERL: Internal-external representation learning network for underwater image enhancement," *IEEE Trans. Multimedia*, vol. 26, pp. 9252–9267, 2024.

[46] Q. Li, Y. Yuan, and Q. Wang, "Multiscale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5523110.

[47] K. Chi, J. Li, W. Jing, Q. Li, and Q. Wang, "Neural implicit Fourier transform for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5628110.

[48] K. Chi, S. Guo, J. Chu, Q. Li, and Q. Wang, "RSMamba: Biologically plausible retinex-based mamba for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5606310.

[49] K. Chi, J. Li, J. Chu, Q. Li, and Q. Wang, "A diffusion model with physically plausible gradient for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5639311.

[50] K. Chi, W. Jing, J. Li, Q. Li, and Q. Wang, "Cross-modal spherical aggregation for weakly supervised remote sensing shadow removal," *IEEE Trans. Multimedia*, early access, Oct. 6, 2025, doi: 10.1109/TMM.2025.3618537.

[51] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multibranch mutual-guiding learning for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5605710.

[52] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5643510.

[53] J. Zhou, Z. He, D. Zhang, S. Liu, X. Fu, and X. Li, "Spatial residual for underwater object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 6, pp. 4996–5013, Jun. 2025.

[54] J. Cepeda-Negrete and R. E. Sanchez-Yanez, "Gray-world assumption on perceptual color spaces," in *Image and Video Technology* (Lecture Notes in Computer Science), R. Klette, M. Rivera, and S. Satoh, Eds., Berlin, Germany: Springer, 2014, pp. 493–504.

[55] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12918–12927.

[56] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, 2023, pp. 2555–2563.

[57] Y. Wei et al., "Visual and large multimodal models promote image restoration and enhancement: Research progress," *J. Image Graph.*, vol. 30, no. 5, pp. 1197–1219, 2025.

[58] X. Liu, S. Lin, K. Chi, Z. Tao, and Y. Zhao, "Boths: Super lightweight network-enabled underwater image enhancement," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[59] J. Liu, R. W. Liu, J. Sun, and T. Zeng, "Rank-one prior: Real-time scene recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8845–8860, Jul. 2023.

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

[61] J. Zhou et al., "HCLR-Net: Hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement," *Int. J. Comput. Vis.*, vol. 132, no. 10, pp. 4132–4156, Oct. 2024, doi: 10.1007/s11263-024-01987-y.

[62] Z. Wang, L. Shen, M. Xu, M. Yu, K. Wang, and Y. Lin, "Domain adaptation for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 32, pp. 1442–1457, 2023.

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

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

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

[66] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[67] H. Lee, K. Kang, J. Ok, and S. Cho, "CLIPtone: Unsupervised learning for text-based image tone adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2942–2951.

[68] J. Cheng, D. Liang, and S. Tan, "Transfer CLIP for generalizable image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25974–25984.

[69] X. Liu, Y. Zhao, K. Chi, Z. Zhang, Y. Chen, and W. Jia, "Toward individual tone preference in underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4212211.

[70] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8624–8634.

[71] G. Song, Q. Sun, L. Zhang, R. Su, J. Shi, and Y. He, "OPE-SR: Orthogonal position encoding for designing a parameter-free upsampling module in arbitrary-scale image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10009–10020.

[72] J. Lee and K. H. Jin, "Local texture estimator for implicit representation function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1919–1928.

[73] R. Fathony, A. K. Sahu, D. Willmott, and J. Z. Kolter, "Multiplicative filter networks," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[74] Z. Zheng et al., "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16180–16189.

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

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

[77] 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, Mar. 2013.

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

[79] S. Liu, K. Li, Y. Ding, and Q. Qi, "Underwater image enhancement by diffusion model with customized CLIP-classifier," *Pattern Recognit.*, vol. 171, Mar. 2026, Art. no. 112232.

[80] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The PIRM challenge on perceptual super resolution," in 2018, *arXiv:1809.07517*.

[81] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.

[82] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Res. Appl.*, vol. 26, no. 5, pp. 340–350, 2001.

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

Xu Liu (Student Member, IEEE) received the B.E. degree in electronic and information engineering from Liaoning Technical University, Huludao, China, in 2022, with a minor in data science and big data technology (Tencent Premier Class) jointly offered by Tencent Cloud Computing (Beijing) Company Ltd., Beijing, China, Shanghai Motong Huakai Education Technology Company Ltd., Shanghai, China, and his home university, and the M.E. degree in information and communication engineering from Hefei University of Technology, Hefei, China, in 2025.

In the same year he was honored as one of Liaoning Technical University's Top Ten Student Role Models in Science and Technology. His research interests include deep learning and image processing.

Mr. Liu was awarded an Incentive Fund that year by Prof. Tat-Seng Chua (Fellow of the Singapore National Academy of Science).

Yang Zhao (Member, IEEE) received the B.E. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively.

From September 2013 to October 2015, he was a Post-Doctoral Fellow with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Beijing, China. He is currently a Full Professor with the School of Computer and Information, Hefei University of Technology, Hefei. His research interests include image processing and pattern recognition.

Yuan Chen (Member, IEEE) received the B.E. degree in information security and the Ph.D. degree in computer science from Hefei University of Technology, Hefei, China, in 2017 and 2022, respectively.

She is currently an Assistant Professor with the School of Internet, Anhui University, Hefei. Her research interests include computer vision and image processing.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

His research interests include image processing and deep learning.

Xingguang Li received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively.

He is currently a Faculty Member with the School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen, China. His research interests include biometrics, visual perception, and large time-series models.

Wei Jia (Member, IEEE) received the B.Sc. degree in informatics from Central China Normal University, Wuhan, China, in 1998, the M.Sc. degree in computer science from Hefei University of Technology, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence systems from the University of Science and Technology of China, Hefei, in 2008.

He has been a Research Assistant and an Associate Professor at Hefei Institutes of Physical Science, Chinese Academy of Science, from 2008 to 2016. He is currently a Full Professor at the School of Computer and Information, Hefei University of Technology. His research interests include computer vision, biometrics, pattern recognition, image processing, and machine learning.