Underwater Image Co-Enhancement With Correlation Feature Matching and Joint Learning

Qi Qi, *Student Member, IEEE*, Yongchang Zhang, Fei Tian, Q. M. Jonathan Wu, *Senior Member, IEEE*, Kunqian Li, *Member, IEEE*, Xin Luan, and Dalei Song

***Abstract*— In underwater scenes, degraded underwater images caused by wavelength-dependent light absorption and scattering present huge challenges to vision tasks. Underwater image enhancement has attracted much attention due to the significance of vision-based applications in marine engineering and underwater robotics. Numerous underwater image enhancement algorithms have been proposed in the last few years. However, almost all existing approaches focus only on the enhancement of independent images. Considering that images photographed in the same underwater scene usually share similar degradation, related images can provide rich complementary information for each other's enhancement. In this paper, we propose an Underwater Image Co-enhancement Network (UICoE-Net) based on an encoder-decoder Siamese architecture. For joint learning, we introduced correlation feature matching units into the multiple layers of our Siamese encoder-decoder structure in order to communicate the mutual correlation of the two branches. Extensive experiments using the Underwater Image Enhancement Benchmark (UIEB), Underwater Image Co-enhancement Dataset (UICoD) collected from an underwater video dataset with ground-truth reference and Stereo Quantitative Underwater Image Dataset (SQUID) dataset demonstrate the effectiveness of our method.**

***Index Terms*— Underwater image enhancement, underwater image co-enhancement, deep learning, convolutional neural network, Siamese structure, correlation feature matching.**

I. Introduction

COMPARED with mature observation systems on land, underwater observation systems are rather incomplete. With the rapid development of Autonomous Underwater Vehicles (AUV) and Remote Operated Vehicles (ROV), underwater observation has recently entered a new phase. Vision plays an important role in underwater observation [1] and operations [2]. However, due to the degradation induced by wavelength-dependent absorption and scattering, the quality of underwater images is usually unsatisfactory. Improving the visual quality of underwater images is vital to the success of subsequent senior vision tasks.

For the past few decades, underwater image enhancement (UIE) has drawn a lot of attention [3], [4]. Numerous underwater image enhancement algorithms designed from different perspectives have been proposed. Rapidly developed deep learning techniques further promote data-driven approaches [5]. Lots of deep models trained with synthetic [6], [7] and real-world [4] underwater images have been proposed in the community [4], [7]–[11]. However, as far as we know, nearly all underwater image enhancement algorithms focus only on the enhancement of independent underwater images [3]. The information contained in each single image is relatively limited [12], and ambiguity due to limited information may even degrade enhancement. A group of underwater images which are photographed in related underwater scenes usually share similar degradation, and can provide rich complementary information for each other. Inspired by the success of co-processing strategy in diverse computer vision applications [12]–[16], mining the shared common information for underwater image co-enhancement is a promising strategy for improving enhancement performance. The main contribution of this paper is summarized as follows,

1) We first introduce co-processing and joint learning strategy into underwater image enhancement tasks; a strategy which has been demonstrated to effectively improve the enhancement performance of group-wise images captured in related scenes.
2) We propose Underwater Image Co-Enhancement Network (UICoE-Net) with Siamese encoder-decoder structure, in which correlation feature matching units are incorporated into multiple intermediate layers to communicate the mutual correlation of the two branches.
3) We establish an underwater image co-enhancement dataset with ground-truth reference from underwater videos and reorganize the recently proposed Underwater Image Enhancement Benchmark (UIEB) dataset as a group-wise dataset. Then, with comprehensive

Manuscript received February 17, 2021; accepted April 9, 2021. Date of publication April 20, 2021; date of current version March 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61906177, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019BF034, in part by the Significant Applied Technology Innovation Projects for Agriculture of Shandong Province under Grant SD2019NJ020, and in part by the Fundamental Research Funds for the Central Universities under Grant 201813022 and Grant 201964013. This article was recommended by Associate Editor Y. Zhou. *(Corresponding author: Kunqian Li.)*

Qi Qi, Yongchang Zhang, Fei Tian, and Xin Luan are with the College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China (e-mail: qiqi2013@stu.ouc.edu.cn; zhangyongchang@stu.ouc.edu.cn; tianfei@stu.ouc.edu.cn; dr.luanxin@yahoo.com).

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Kunqian Li and Dalei Song are with the College of Engineering, Ocean University of China, Qingdao 266100, China (e-mail: likunqian@ouc.edu.cn; songdalei@ouc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3074197.

Digital Object Identifier 10.1109/TCSVT.2021.3074197

experiments, we verify the effectiveness of the proposed UICoE-Net on these challenging datasets.

The rest of the paper is organized as follows: In Section II, we give a detailed review of the existent deep learning-based approaches in underwater image enhancement and the co-processing strategy in computer vision application. Then, in Section III, we give the definition of our underwater image co-enhancement task and introduce the proposed UICoE-Net in detail. To verify the performance of the proposed method, in Section IV, we conducted comprehensive experiments on two group-wise underwater image co-enhancement datasets with ground-truth reference and an underwater image restoration dataset. Finally, in Section V, we conclude our paper and offer suggestions for further development.

II. Related Works

A. Underwater Image Enhancement With Deep Learning

Underwater image enhancement has attracted widespread attention as it is indispensable to the improvement of the visual quality of recorded images [4], [17], [18]. In recent years, deep learning techniques like Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [19] have delivered excellent performance in many high-level computer vision tasks. Naturally, these techniques have also begun to play important roles in the field of underwater image enhancement. However, quality degradation in underwater imaging, including low contrast, blurred details, and color deviations, makes it difficult to obtain corresponding clear images. Therefore, a variety of deep learning-based enhancement networks adapted to the underwater environment have been proposed.

To make full use of different features of the inputs, various multi-branch-based CNNs have been designed. UIE-Net [8], composed of two fully convolutional neural networks to learn color correction and haze removal tasks, represents pioneering work in using a CNN to enhance underwater images. In [20], Cao *et al.* proposed UIR-Net which is consists of two neural network structures, the background light network (BL-Net) and the transmission map network (TM-Net), to estimate background light and scene depth for the restoration of underwater images. Hou *et al.* [7] proposed an underwater residual CNN (URCNN) model by modifying a VGG neural network [21] to learn the transmission map. To train the proposed URCNN, they synthesized an underwater image dataset consisting of 1000 images from the NYU Depth dataset [22] with realistic depths of object and corresponding clean images. Similarly, by combining an underwater imaging physical model with optical properties from underwater scenes, Li *et al.* [6] synthesized an underwater image degradation dataset and used it to train the proposed light-weight network, i.e. UWCNN. The UWCNN includes three densely-connected building blocks, and each basic building block consists of three densely-connected convolutional layers. In addition, Li *et al.* [4] constructed a large-scale, real-world underwater image enhancement benchmark dataset (i.e., UIEB) which contains 950 real underwater images. Then, they employed a gated fusion network architecture called WaterNet to learn three confidence maps, which are used as the coefficient to combine the three input branches into an enhanced result.

Due to the great success of GAN on image generation, GAN-based underwater image enhancement approaches were also widely explored. WaterGAN [9] is the earliest GAN-based attempt for underwater image enhancement, which uses in-air images with corresponding depth information to generate a synthetic image for specific underwater scenarios. Then, these synthetic images are used to train a two-stage underwater image restoration network. Then, inspired by the cycle-consistent adversarial network (CycleGAN) [23], UGAN [24] was propose to use CycleGAN framework [23] to generate synthetic paired training data, while UWGAN [10] and MCycleGAN [25] were proposed to recover underwater images without paired training data by following the learning strategy of CycleGAN. Liu *et al.* [11] proposed a generic multilevel feature fusion-based conditional GAN (MLFcGAN) to extract features in multiple scales. Considering the running efficiency, Islam *et al.* [26] presented a real-time underwater enhancement method based on conditional GANs called FUnIE-GAN.

In summary, current deep-learning-based methods consisting of complex network structures and well-designed loss functions have promoted the progress of underwater image enhancement. As data-driven approaches, large-scale high-quality training datasets are crucial for performance improvement. However, it is very difficult to obtain very large real underwater image dataset with high-quality references. Moreover, almost all underwater image enhancement methods focus only on the enhancement of single images, which places limits on available information and causes ambiguity which may even degrade the enhancement. Making full use of the limited high-quality training dataset by exploiting the cooperative constraints implied in scene-related images is valuable and promising.

B. Co-Processing in Computer Vision Applications

Co-processing strategies have been explored in many computer vision applications. Co-processing was first proposed in an image co-segmentation task for paired images [27], in which a pair of images are simultaneously segmented by exploiting common foreground constraints. From then on, numerous image co-segmentation approaches from different perspectives have been proposed to deal with different kinds of co-segmentation tasks [12], [28], [29]. Recently, with the rapid development and great success of deep learning techniques in computer vision, deep learning has also been introduced into image co-segmentation [13], [30], [31]. Li *et al.* [13] designed a CNN-based Siamese encoder-decoder architecture for image pair co-segmentation, in which cooperative constraints are introduced with a mutual correlation layer performed on high-level semantic features. Chen *et al.* [30] further proposed an attention mechanism in the bottleneck layer of the deep neural network for the selection of semantically-related features. Li *et al.* [31] designed a co-attention recurrent neural network, whose group representation, generated by co-attention between images fused with multi-scale fine-resolution features,

greatly facilitates the co-segmentation of group-wise common targets.

As a natural extension of image co-segmentation tasks, the video co-segmentation problem also receives a lot of attention [14], [32], [33]. For saliency detection, which aims at simulating human vision's ability to highlight only the salient regions in the image field [34], more and more researchers have found that co-saliency detection for group-wise images helps to mine the most noteworthy and meaningful information from the whole image group [15], [35]–[37]. Studies focused on image pairs [35], multiple images [36], [37] and joint co-processing tasks [15] are widely conducted. Besides pixel-level tasks, existing research shows that object-level tasks, such as object co-localization [16], [38]–[40], also benefit from co-processing strategy.

Given the above, co-processing strategy has been demonstrated to overcome the limitations of using solely independent images in achieving information richness and clarity. The mined cooperative constraint can be used to remove ambiguity and highlight common patterns which carry more valuable information. In low-level computer vision tasks, such as underwater image enhancement, in which visual materials are usually closely related in a specific scene, co-processing strategy can be useful; however, it has not yet been explored. In this paper, we propose an underwater image co-enhancement algorithm, which further demonstrates the effectiveness of co-processing strategy in such low-level tasks.

III. Approach

A. Cooperative Enhancement Strategy

As considerable attention has been paid to underwater vision, more and more underwater image enhancement algorithms have been proposed in the last few years. However, almost all these algorithms focus only on the enhancement of independent underwater images. Due to the fact that it is practically impossible to simultaneously acquire a real-world underwater image and its clear version as reference, the number of high-quality paired training images (real-world underwater image datasets with clear references, such as the UIEB [4]) for the development of data-driven approaches is quite limited. Compared with increasing the budget for sophisticated underwater image acquisition equipment for both raw image collecting and reference image calibration, making full use of the limited high-quality underwater training images that are available is far more promising, although challenging.

Generally speaking, underwater images which are photographed in similar underwater scenes usually contain associated regions which share similar appearances, backgrounds, and objects. Such correlations can provide rich complementary information for each other and improve learning performance for visual quality enhancement. In this section, we introduce a joint learning and co-enhancement strategy to an underwater image enhancement task to improve the enhancement performance for group-wise underwater images captured in related scenes. For convenience of description and understanding, the symbols and their meanings appearing in the rest of the paper are summarized in Table I.

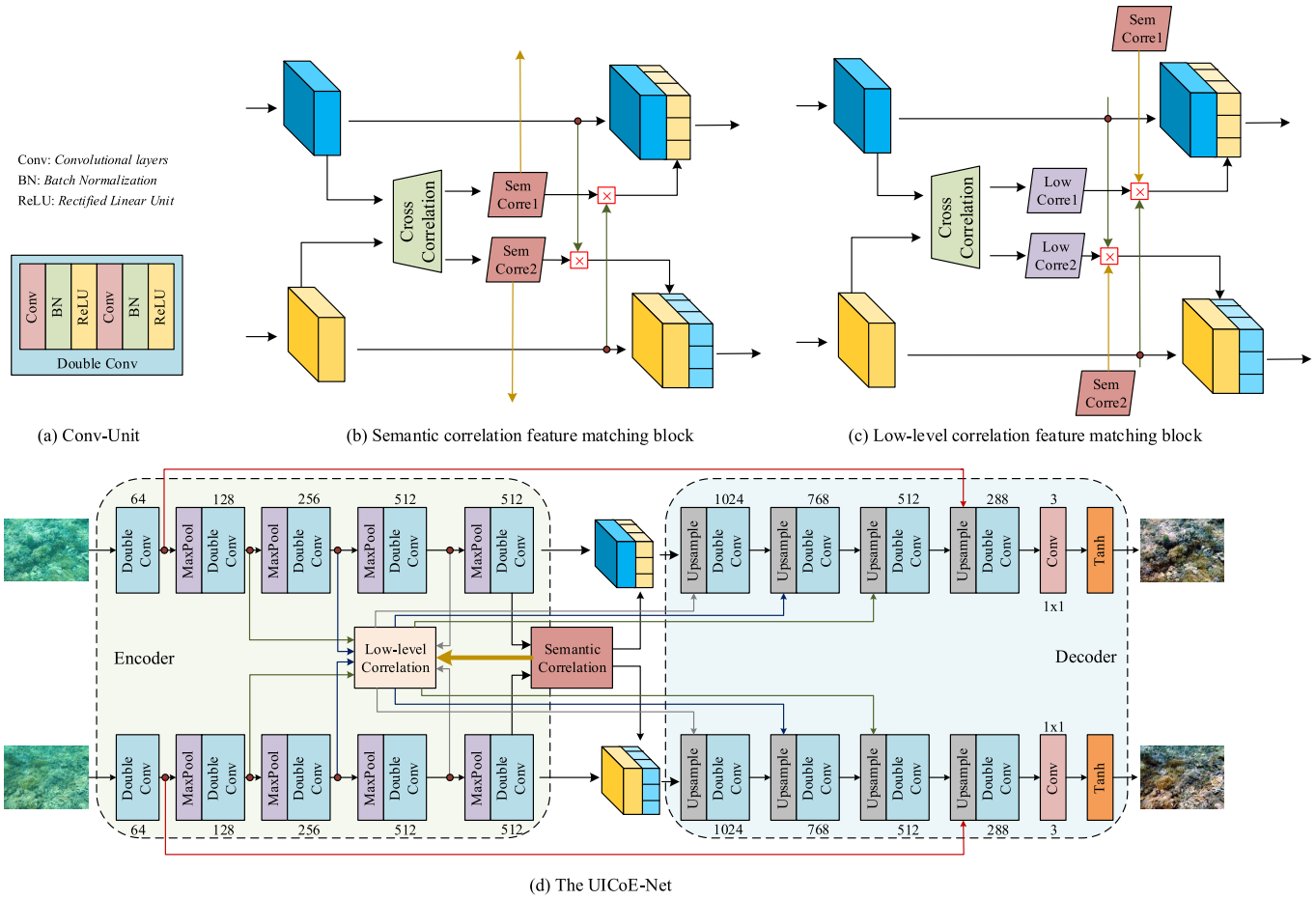
TABLE I
Symbolic Notations and Their Corresponding Meanings

Symbols	Meanings
$\mathcal{I}$	image set
I_A and I_B	paired input underwater images
f_A and f_B	feature maps of I_A and I_B
f_A^{sem} and f_B^{sem}	semantic feature maps of image I_A and I_B
f_A^{low} and f_B^{low}	low-level feature maps of image I_A and I_B
w_{sem} and h_{sem}	width and height of the semantic feature map
w_{low} and h_{low}	width and height of the low-level feature map
C_{AB}^{sem} and C_{AB}^{low}	semantic and low-level correlation maps
k	the channel index of correlation map
$\boldsymbol{\xi}_{(i,j)}$	correlation coefficient vector of C_{AB}^{sem} in location (i, j)
$\boldsymbol{\eta}_{(i,j)}$	correlation coefficient vector of C_{AB}^{low} in location (i, j)
$\lambda_{(i,j)}$	the maximal correlation coefficient in $\boldsymbol{\xi}_{(i,j)}$
$\mu_{(i,j)}$	the maximal correlation coefficient in $\boldsymbol{\eta}_{(i,j)}$
β^{sem}	the matched collaborative feature in f_B^{sem}
β^{low}	the matched collaborative feature in f_B^{low}
F_A and F_B	image-reconstructing features of I_A and I_B

An underwater image set $\mathcal{I}$ which consists of images captured in similar scenes is treated as a whole group for joint learning and co-enhancement. Image enhancement can be viewed as a feature-reorganizing procedure within an encoder-decoder network to realize both color mapping and noise filtering. Specifically, for a pair of underwater images $I_A, I_B \in \mathcal{I}$, which have scene-related clues, the corresponding image-reconstructing features of the related information should be consistent as well. Therefore, we propose to design a deep neural network to simultaneously reorganize raw image features by jointly considering their common information. In order to communicate the mutual correlation of two images, feature-matching units are introduced to reorganize and concatenate matched cooperative features according to their multi-level consistency. The reorganized features of one image can serve as the supplementary information of the other one. By jointly learning, the regions which share similar characteristics will receive consistent transformation enhancement. Finally, the decoder will reconstruct such concatenated features for co-enhancement results.

B. Underwater Image Co-Enhancement Network

To achieve the above-stated cooperative enhancement strategy, we proposed an underwater image co-enhancement network with Siamese structure and correlation feature matching modules. The network is end-to-end trainable for the underwater image co-enhancement task. Figure 1 illustrates the overall structure of UICoE-Net, which consists of a Siamese encoder, multi-level correlation feature-matching modules and a Siamese decoder. Given two underwater images as input, UICoE-Net first reorganizes them into multi-level image features through the Siamese encoder with shared weights. Then, the correlation feature-matching module communicates the mutual correlation of the two feature branches



(a) Conv-Unit (b) Semantic correlation feature matching block (c) Low-level correlation feature matching block

(d) The UICoE-Net

Fig. 1. Architecture of UICoE-Net. UICoE-Net is a CNN-based network with Siamese encoder-decoder structures, in which the correlation feature-matching units (i.e. (b) the semantic correlation feature-matching block and (c) the low-level correlation feature-matching block) are incorporated into multi-level layers of encoder-decoder Siamese structure to communicate mutual correlation of two branches. (a) represents the convolution unit which is used in the encoder and decoder of (d). (b) The semantic correlation-feature matching block performs high-level feature aggregation based on the semantic correlation coefficient of features from the Siamese branches. (c) Low-level correlation feature-matching blocks reorganize and concatenate low-level collaborative features for the Siamese branches according to both low-level feature-correlation coefficients and semantic feature-correlation coefficients.

by reorganizing and concatenating the cooperative features from one branch to another, according to their correlation coefficient. Finally, the Siamese decoder reconstructs the reorganized image features for cooperative enhancement. The following details each part of the architecture and training losses.

1) Siamese Encoder: The first part of the proposed UICoE-Net is a Siamese encoder consisting of two identical feature-extraction CNNs with shared parameters. These two branches are designed using a stack of double Conv-BN-ReLU units (Conv-Units) as shown in Figure 1(a) in order to extract multi-level convolutional image features. We keep the convolutional layers with 3×3 kernel size and set the padding as 1. Before each Conv-Unit module, a max pooling layer (MaxPool in Figure 1) with 2×2 kernel size is used for down-sampling and larger feature-receptive fields. In total, our Siamese encoder contains five Conv-Units with MaxPool layers (except for the first Conv-Unit of each branch). As the raw images pass through the Siamese encoder, a feature map is produced after each Conv-Unit, and, finally, a set of Siamese features $f_{(A,l)}$ and $f_{(B,l)}$, where $l \in \{1, 2, \ldots, 5\}$ is the index of the outputs of different Conv-Units from the low levels to the high level. The size of the input image of the Siamese encoder is unrestricted, such as 256×256 or 512×512. The final output of the encoder is two 512-channel feature maps representing the semantic features of the paired input images.

2) Correlation Feature Matching Module: The correlation feature matching modules play significant roles in UICoE-Net. For ease of understanding, $f_{(A,5)}$ and $f_{(B,5)}$, the outputs of the last Conv-Units of the Siamese encoders are denoted as f_A^{sem} and f_B^{sem} to represent the high-level semantic features of the input images. When $l \in \{1, 2, 3, 4\}$, $f_{(A,l)}$ and $f_{(B,l)}$, which represent the low-level image features, are denoted as $f_{(A,l)}^{low}$ and $f_{(B,l)}^{low}$, respectively. For semantic and low-level features, the related regions indicate the areas in the original images which capture similar scenes. Obviously, the enhancement of these similar scenes should also be consistent. The hope is to strengthen the features of similar regions by communicating and combining the related features of two branches, to improve the final enhancements. Moreover, for those semantic-related regions, subtle variations existing in low-level features not only provide more diverse information for generalized learning, but also more complementary features for the following robust enhancement.

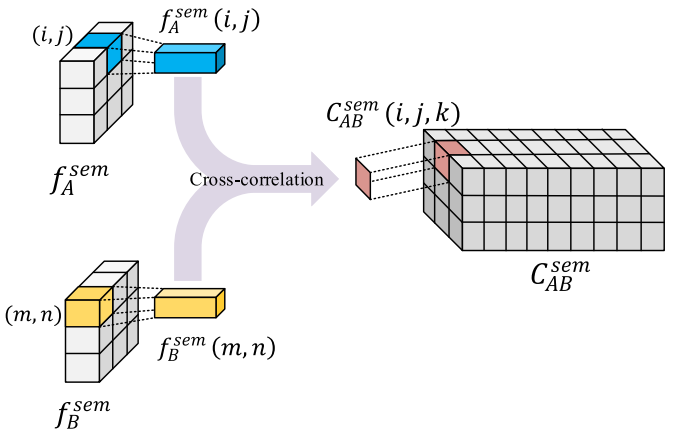


Fig. 2. The diagram for calculating the correlation coefficient map C_{AB} between two input feature maps. Here, we take correlation coefficient map C_{AB}^{sem} for f_A^{sem} and f_B^{sem} as example.

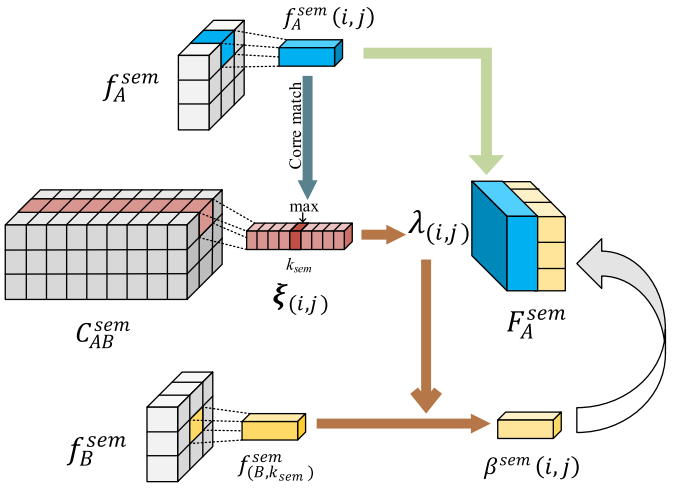


Fig. 3. Diagram of reorganizing and concatenating collaborative features F according to the correlation coefficient map. Here, we take the semantic concatenated feature map F_A^{sem} as the description example.

To communicate the mutual correlation of the two Siamese branches, we introduce the semantic correlation feature-matching block shown in Figure 1(b) and low-level correlation feature-matching block shown in Figure 1(c) to reorganize and concatenate cooperative features from one branch to another, according to their correlation coefficient. Inspired by FlowNet [41], in which the correlation layer is used to estimate the optical flow by matching the feature points between frames, we propose a cross-correlation module to compute the correlation between each pair of locations on the feature maps.

The cross-correlation calculation of f_A^{sem} and f_B^{sem} exemplifies the cross-correlation module's performance of an element-wise comparison between two feature maps. As shown in Figure 2, given a location (i, j) in f_A^{sem} and a location (m, n) in f_B^{sem}, the correlation between feature vectors $f_A^{sem}(i, j)$ and $f_B^{sem}(m, n)$ is defined as

$$C_{AB}^{sem}(i, j, k) = \left\langle f_A^{sem}(i, j), f_B^{sem}(m, n) \right\rangle, \tag{1}$$

where $\langle a, b\rangle$ is the dot product for the given feature vectors a and b, $k = (m-1)w_{sem} + n$ presents the channel index of the correlation map C_{AB}^{sem} and has one-to-one correspondence with each location (m, n), w_{sem} is the width of the feature maps. The output of the correlation module is a correlation map C_{AB}^{sem} with the size of $w_{sem} \times h_{sem} \times D$, where $D = w_{sem} \times h_{sem}$ is the number of channels. The correlation map C_{BA}^{sem} between f_B^{sem} and f_A^{sem} can be computed by the same method. Then, for cooperative enhancement, the related features are combined according to their correlation coefficients. As shown in Figure 3, we denote D-dimensional vector $\boldsymbol{\xi}_{(i,j)}$ as the correlation coefficient vector of location (i, j) in f_A^{sem} comparing it with f_B^{sem}, which is extracted from C_{AB}^{sem} at location (i, j). If the k-th element of $\boldsymbol{\xi}_{(i,j)}$ is large, it indicates that the vector denoted as $f_{(B,k)}^{sem}$ in f_B^{sem} has high correlation with feature $f_A^s(i, j)$. Accordingly, for feature $f_A^{sem}(i, j)$ from I_A, its matched collaborative feature for cooperative enhancement in f_B^{sem} can be represented as a weighted feature as follows

$$\beta_{(i,j)}^{sem} = \lambda_{(i,j)} f_{(B,k_{sem})}^{sem}, \tag{2}$$

where $\lambda_{(i,j)} = \boldsymbol{\xi}_{(i,j)}(k_{sem})$ is the maximal element of $\boldsymbol{\xi}_{(i,j)}$ whose index is k_{sem}, and $f_{(B,k_{sem})}^{sem}$ is the corresponding matched feature in f_B^{sem}. Then, we concatenate the collaborative feature after $f_A^{sem}(i, j)$ as Figure 3 presented.

Low-level features are processed similarly. When matching the collaborative feature for $f_A^{low}(i, j)$, their high-level semantic correlation coefficient in the corresponding location is also considered. Given a location (i, j) in f_A^{low} and a location (m, n) in f_B^{low}, the correlation between low-level features $f_A^{low}(i, j)$ and $f_B^{low}(m, n)$ is defined as

$$\begin{aligned} C_{AB}^{low}(i, j, k) = &\left\langle f_A^{sem}(\hat{i}, \hat{j}), f_B^{sem}(\hat{m}, \hat{n}) \right\rangle \\ &\times \left\langle f_A^{low}(i, j), f_B^{low}(m, n) \right\rangle, \end{aligned} \tag{3}$$

where $(\hat{i}, \hat{j})$ and $(\hat{m}, \hat{n})$ are the corresponding locations of (i, j) and (m, n) in the semantic feature maps, respectively. $\boldsymbol{\eta}_{(i,j)}$ is the D-dimensional correlation coefficient vector at location (i, j) of correlation map C_{AB}^{low} and $D = w_{low} \times h_{low}$. And notably, for both computation efficiency and effective matching, the correlation calculation is only performed on the location pairs whose semantic correlations are matched. Otherwise, their correlation will be directly set to 0 without calculation. Then, the matched collaborative feature of the $f_A^{low}(i, j)$ in f_B^{low} can be acquired as follows

$$\beta_{(i,j)}^{low} = \mu_{(i,j)} f_{(B,k_{low})}^{low}, \tag{4}$$

where $\mu_{(i,j)} = \boldsymbol{\eta}_{(i,j)}(k_{low})$ is the maximal element of $\boldsymbol{\eta}_{(i,j)}$. In particular, $f_{A,1}^{low}$ and $f_{B,1}^{low}$, generated by the first Conv-Unit which does not contain a MaxPool layer, retain the original input size. Calculating the correlation between them requires tremendous graphic processing unit (GPU) resources. Thus, only the low-level correlation block is applied to $f_{(A,l)}^{low}$ and $f_{(B,l)}^{low}$, $l \in \{2, 3, 4\}$. With the above formulas, the features can be reorganized by concatenating f_A and its most-correlated feature vectors from f_B. Conversely, the most-correlated vector for f_B can be collected by changing positions in the formulas. In this way, the concatenated feature maps F_A and F_B generated by co-processing f_A and f_B contain more complementary information provided by each other, which

can serve for the reconstruction of clear underwater images. For clarity, we mark the semantic and low-level concatenated feature maps of I_A (I_B) as F_A^{sem} (F_B^{sem}) and F_A^{low} (F_B^{low}), respectively.

3) Siamese Decoder: The Siamese decoder is the third part of UICoE-Net, which cooperatively reconstructs clear underwater images. By semantic correlation, f_A^{sem} and f_B^{sem} are concatenated with their corresponding maximum response of f_B^{sem} and f_A^{sem}, respectively. Then, the generated F_A^{sem} and F_B^{sem} will serve as the input of the Siamese decoder. The decoder is also designed as a Siamese structure with shared parameters. The first four blocks in the decoder have a similar structure, which is composed of an up-sampling layer followed by the Conv-Unit. After four up-samplings by combining the low-level concatenated features $F_{(A,l)}^{low}$ and $F_{(B,l)}^{low}$, $l \in \{1, 2, 3, 4\}$, we get the tensor, which is the same width and height as the input raw images. Then, a single convolution layer with 1×1 kernel size is set to squeeze the channel number of the tensor from 288 to 3. Before output, the tanh function is applied as the activation function for the results of the previous convolution layer.

4) Network Loss: Because blurring edges results in large errors, inspired by [6], the ℓ_2 loss is used to train our network, as it can well preserve the sharpness of edges and details. Then, the full training loss $\mathcal{L}_{AB}$ for paired image I_A and I_B can be estimated by

$$\mathcal{L}_{AB} = \frac{1}{N_A}\sum_{i}^{N_A}(E_A(i) - R_A(i))^2 + \frac{1}{N_B}\sum_{i}^{N_B}(E_B(i) - R_B(i))^2, \quad (5)$$

where E_A and E_B are the enhancements for I_A and I_B with UICoE-Net, R_A and R_B are their references for training, N_A and N_B are the pixel numbers of I_A and I_B, respectively.

IV. Experiments and Analysis

In this section, we first evaluate the proposed method by comparing it with existing methods on two underwater image enhancement benchmark datasets with reference: the UIEB dataset [4] and UICoD—a new underwater image co-enhancement dataset established by ourselves. We show the performance of the proposed method by comparing it with other non-deep- and deep learning-based methods on the two datasets. We conduct comprehensive ablation study to demonstrate the effect of each component in the network. Besides, we also compare our method with underwater image restoration approaches on the stereo quantitative underwater image dataset (SQUID) [42], which has color charts for quantitative comparison. Finally, necessary discussion is given in the last subsection.

A. Underwater Image Enhancement Benchmark Datasets With Reference and Experimental Implementation Details

As described above, UICoE-Net mainly focuses on group-wise underwater images which are photographed in similar underwater scenes. To verify the performance of our network more effectively, in Section IV-A.1, we reorganized the recently proposed UIEB dataset as a group-wise dataset; a new underwater image co-enhancement dataset with ground truth reference collected from underwater videos will be given in Section IV-A.2. Then, in Section IV-A.3, we will introduce the implementation details of the proposed UICoE-Net.

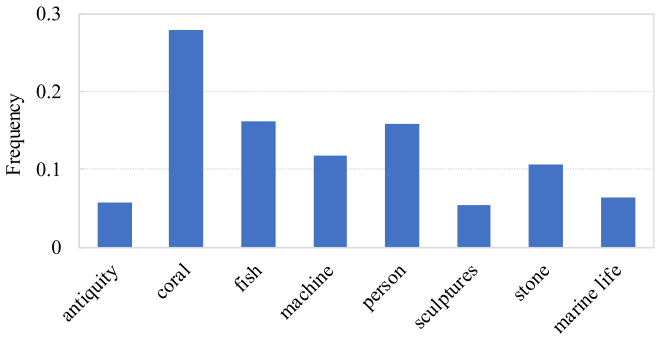


Fig. 4. The scene/object categories and their proportions in the classified UIEB dataset.

1) Classified UIEB Dataset: UIEB dataset offers 890 high-resolution real-world underwater images which are photographed under natural light, artificial light, or a mixture of lighting conditions. Their corresponding reference images are manually selected by volunteers from results generated by 12 typical enhancement methods which match human visual quality perception better. These underwater images contain quite large image resolutions and span diverse scene/main object categories including coral (e.g., fringing reefs and barrier reefs), marine life (e.g., turtles and sharks), etc. Li *et al.* [4] provided a statistic of scene/main object categories, where the images of UIEB dataset are divided into seven main object categories and an "others" class. However, the images are mixed together in UIEB dataset. As shown in Figure 4, we reorganized and classified them into eight categories. For different categories, the training, verification, and test images were randomly allocated at a ratio of 3:1:1.

2) UICoD Collected From Underwater Videos: Inspired by [4], to facilitate the study of the underwater video enhancement problem we created an underwater video dataset called UVE-38K with ground-truth reference, which was recently partially released.[1] The paper and whole dataset will be publicly available soon. There are 50 video clips, capturing fish (e.g., cheilinus, cuttlefish, mobula rays), coral, turtles, and diverse underwater scenes. In UVE-38K dataset, to get the enhancement reference for the videos, we follow [4] to invite 10 volunteers with professional background to vote for the best video enhancement from the enhancement candidate pool, which is generated with 12 typical underwater image enhancement approaches. Then, with the top vote getter as the intermediate references, we further perform inter-frame color transfer to reduce unpleasant temporal inconsistency. With this video dataset, we further established an underwater image co-enhancement dataset (UICoD) with ground-truth reference. Video clips were randomly divided into training,

[1] https://github.com/TrentQiQ/UVE-38K

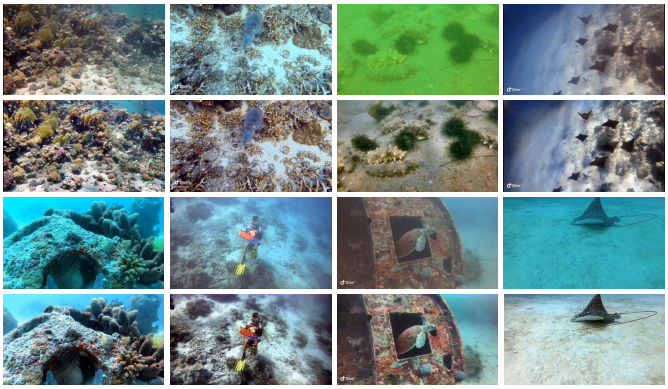


Fig. 5. Sample images in the newly established UICoD benchmark. Rows 1 and 3: raw underwater images taken in diverse underwater scenes; rows 2 and 4: the corresponding references.

verification, and test at a ratio of 3:1:1. Since the number of frames contained in different video clips is not fixed, adaptive thresholds were set to extract frames at different intervals, so as to balance the number of frames collected from different videos. Finally, frames collected from the same video form an image group. UICoD includes 1,272 images from 46 different video sequences. Figure 5 shows some sample images in our co-enhancement dataset.

3) Implementation Details: We implemented the proposed UICoE-Net using the PyTorch platform. During the training, a batch-mode learning strategy with a batch size of four was applied. The filter weights of each layer were initialized with Kaiming initialization. We used the Adam optimizer with a 1e-5 weight decay for our network optimization, and learning rate was set to 1e-4. The learning rate remained fixed for the entire training procedure. During the training, both the input and output resolutions were 256×256. The parameters updated with ℓ_2 loss as formulated in Section III-B.4, which characterizes the deviation between ground truth and the prediction. The experiments were conducted on a PC with an NVIDIA RTX Titan GPU, a 2.3GHz Intel Xeon processor, 128GB RAM and the Ubuntu 18.04 operation system.

B. Enhancement Comparison and Analysis

To demonstrate the advantages achieved by the proposed UICoE-Net, we compared it against several typical underwater image enhancement methods, including CLAHE [43], the fusion-based method [44], the Retinex-based method [45], the histogram prior method [46], UDCP [47], CBF [48], Two-Step [49], ULAP [50], UWCNN [6], MLFcGAN [11], FUnIEGAN [26], U-Net [51] and WaterNet [4], using the classified UIEB and UICoD datasets. All the above-mentioned methods were tested with the author-provided projects or publicly available implementations. For UWCNN, MLFcGAN, and FUnIEGAN, we used their pre-trained models trained with their own paired and unpaired training images. UWCNN adopted a training set with 5000 samples. For the two GAN-based approaches, MLFcGAN used 6,000 paired images for training while FUnIE-GAN used 11,000 paired images generated with UGAN and 7,500 unpaired images. For U-Net and WaterNet, we used the same training data as for our UICoE-Net.

1) Evaluation on Classified UIEB Dataset: In this section, we evaluate the proposed method using the classified UIEB dataset. The comparison, shown in Figure 6, verifies the qualitative superiority of the proposed UICoE-Net. The presented raw images are photographed in multiple typical underwater environments, which involve obvious hazy and color-distorted degradation. Compared with several prior and contemporary methods, our method achieves clearer enhancement with more balanced color and reasonable stretched contrast. More importantly, compared with the competing results our enhancements are much closer to the given references, which represent the visual quality pleasing to human beings. Some enhancements of the competing non-deep-learning methods such as CLAHE, the fusion-based method and UDCP retained much of their green overlay, indicating that these methods cannot excel with remit color cast. Among the traditional non-leaning-based approaches, the Retinex-based method, the histogram prior method, CBF and Two-step method achieved relatively better visual performance. However, some enhancements of the Retinex-based method loss natural true colors, such as the last two images. Histogram prior method sometimes causes severe color cast and excessive contrast and brightness in some scenes, such as the first image, the images of a shoal of fish or a diver. CBF and Two-step method effectively removes blue and green distortion while sometimes introduces reddish component, such as the fish and diver images. Because of the various underwater environments of the test images, UDCP can hardly construct a single degradation model to adapt to such diverse scenes. Through careful observation, we found that MLFcGAN usually produced over-enhancement and over-saturation, and that the results of this method in varied underwater images were unstable. As to the enhancements of FUnIEGAN, there usually remained residual greenish distortion and a hazy appearance. In contrast, the U-Net-based enhancement model and the proposed UICoE-Net, which were trained on the non-synthetic paired underwater images, showed obvious improvement in visual clarity and color appearance. UICoE-Net demonstrated better color appearance than the U-Net-based enhancement model. The enhancements achieved by WaterNet and UWCNN were darker than the results of other methods, demonstrating low visual contrast.

Conversely, as a totally data-driven approach, UICoE-Net achieves quite good result according to its consistency to the given reference. In most cases, our method not only improved the visibility of the underwater images, but also achieved aesthetically pleasing visual quality with genuine natural colors. We also notice that the visual quality of enhancement for the first image in Fig. 6 is not satisfactory enough, which is largely due to the references used for learning are not strictly veritable ground-truth. Part of the poor references for training have also led to those unsatisfactory enhancements in the test stage. While, with the proposed co-enhancement and joint learning framework, the high consistency between the enhancement and the current reference suggests that the UICoE-Net can produce enhancement with better visual quality by using more

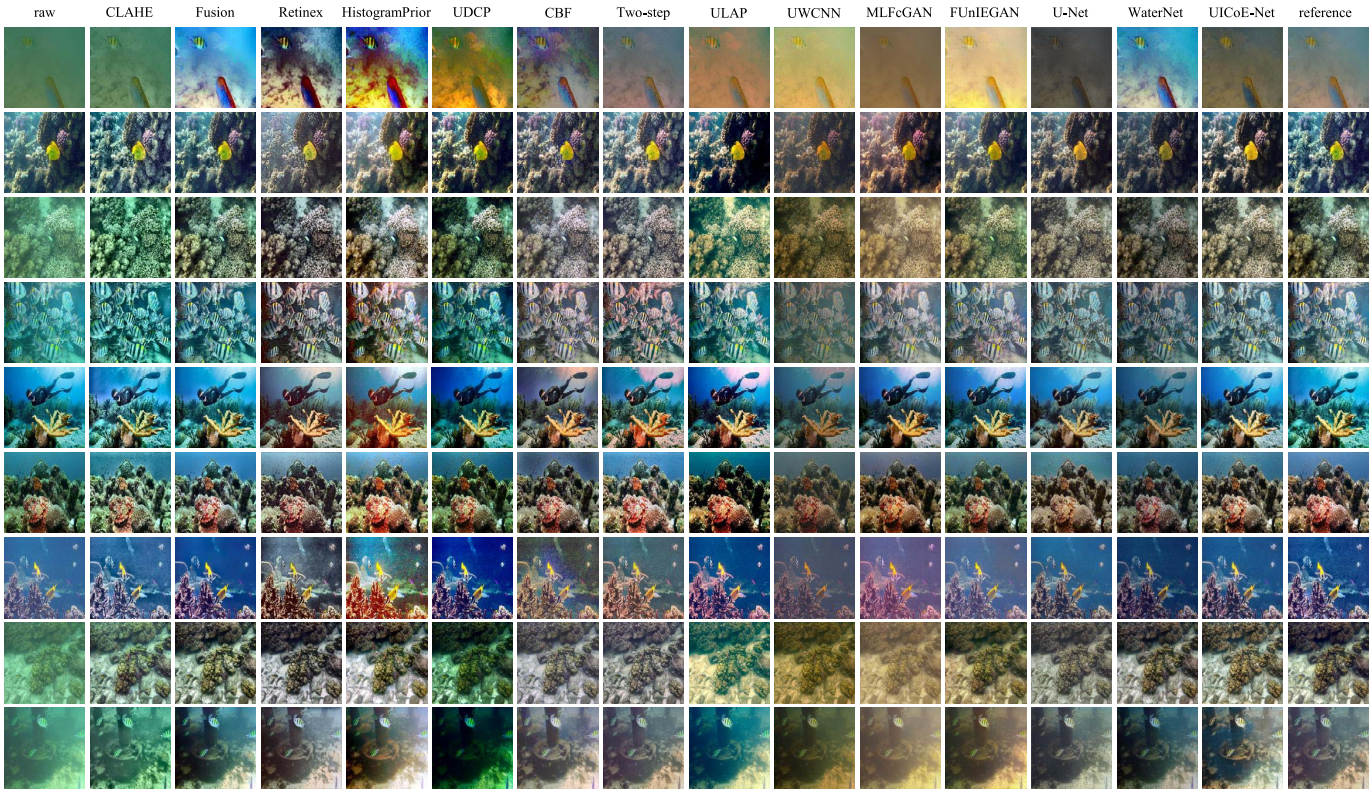


Fig. 6. Subjective comparisons on underwater images from classified UIEB test set. From left to right are raw underwater images, and the results of CLAHE [43], the fusion-based method [44], Retinex [45], HistogramPrior [46], UDCP [47], CBF [48], Two-Step [49], ULAP [50], UWCNN [6], MLFcGAN [11], FUnIEGAN [26], U-Net [51], WaterNet [4], the proposed UICoE-Net, and reference images.

TABLE II

FULL-REFERENCE IMAGE QUALITY ASSESSMENT IN TERMS OF MSE, PSNR, AND SSIM ON CLASSIFY UIEB TEST SET. TRADITIONAL METHODS, DEEP-LEARNING-BASED METHODS WITH PRE-TRAINED MODELS, AND DEEP-LEARNING-BASED METHODS TRAINED WITH PAIRED REFERENCE IMAGES ARE SEPARATED WITH LINES

Method	SSIM↑	PSNR↑	MSE ($\times 10^3$) ↓
CLAHE [43]	0.7957	16.8448	1.5619
Fusion [44]	0.8649	20.1696	0.8541
Retinex [45]	0.7999	17.4437	1.4200
HistogramPrior [46]	0.7910	16.9184	1.5922
UDCP [47]	0.6521	13.7478	3.3368
CBF [48]	0.8374	17.9034	1.2740
Two-Step [49]	0.8667	20.5817	1.0373
ULAP [50]	0.7584	16.8836	1.8617
UWCNN [6]	0.6801	14.1921	2.7882
MLFcGAN [11]	0.6407	16.5759	1.7473
FUnIEGAN [26]	0.7945	18.8872	1.0637
U-Net [51]	0.8578	20.2393	0.7607
WaterNet [4]	0.8532	19.5732	0.9744
UICoE-Net	**0.8944**	**21.7456**	**0.4711**

high-quality references for training. More related discussion can be found in Section IV-E.

The numerical comparison on UIEB dataset is given in Table II. Performance was measured by three different metrics: mean square error (MSE), peak signal-to-noise ratio (PSNR), and the structural similarity index metric (SSIM) [52]. For MSE and PSNR metrics, lower MSE (higher PSNR) denotes that the result is closer to ground truth in terms of image content. Higher SSIM scores mean the result is closer to ground truth in terms of image structure and texture. The presented results are average scores. The values in bold represent the best results. Noticeably, among all underwater image enhancement methods tested, our method achieved the best performance across all metrics. Although both our network and U-Net use encoder-decoder-like architecture, our network introduces a co-processing enhancement scheme by reorganizing and aggregating features at different levels to further increase the richness of reconstruction features and improve the consistency of visual quality enhancement. As to the PSNR response, our method (21.7456) showed about 1.2 improvement over the second-best method (20.5817), and the value of MSE was only half of WaterNet's [4]. Similarly, our SSIM score was also much higher than those of the compared methods.

2) Evaluation on UICoD: To further verify the effectiveness and robustness of our method, we conducted experiments using UICoD. Visual comparisons with other methods are presented in Figure 7. The fusion-based method, Retinex-based method, histogram prior method, UDCP, CBF, Two-step and ULAP usually introduced obvious artificial color or color deviations. For example, for the underwater image in the first row, the result of the fusion-based method was introduced more blue in the background; while the Retinex-based method was left with more yellow than other methods, as well as

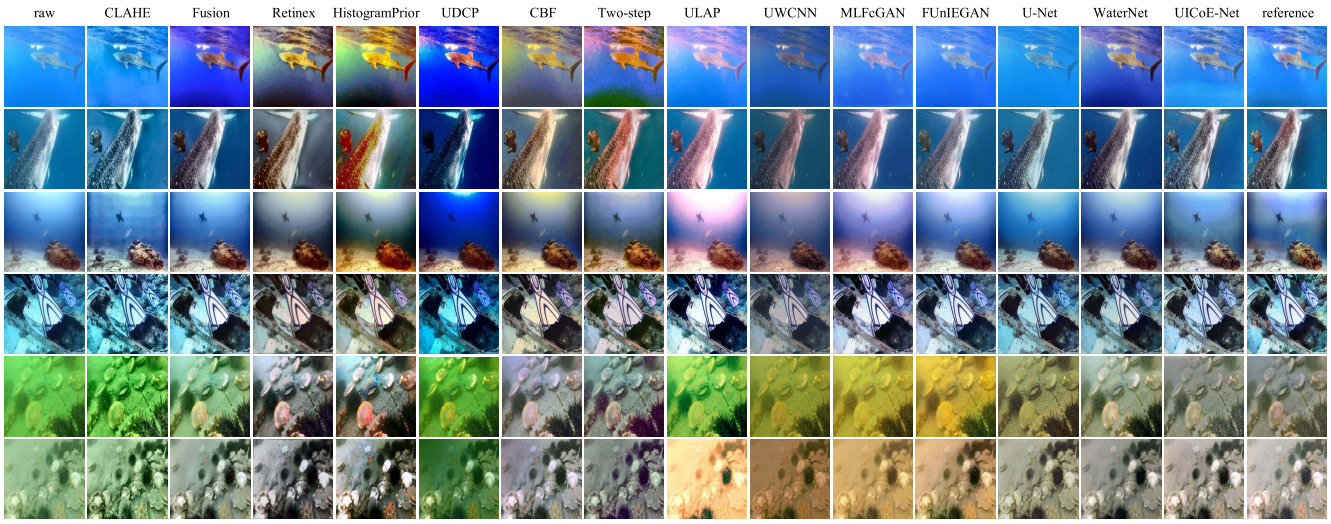


Fig. 7. Subjective comparisons of underwater images from UICoD test set. From left to right are raw underwater images, and the results of CLAHE [43], fusion-based method [44], Retinex [45], HistogramPrior [46], UDCP [47], CBF [48], Two-Step [49], ULAP [50], UWCNN [6], MLFcGAN [11], FUnIEGAN [26], U-Net [51], WaterNet [4], the proposed UICoE-Net, and reference images.

the given reference. CLAHE and UDCP stretched the visual contrast and improved the visual clarity; however, CLAHE had little effect on the color correction and UDCP even made the color distortion heavier. Conversely, WaterNet corrected the color appearance, but introduced unpleasantly low visual contrast. By observing the enhancement results of the GAN-based methods [11], [26], it can be noted that their enhancement results show excessive yellow and red, which led to serious color cast in the complex scenes. For varying underwater scenes, UWCNN first synthesized underwater image degradation datasets into an underwater imaging physical model with the optical properties of the underwater scenes. Since the prior knowledge can hardly cover diverse underwater scenes, the results UWCNN achieved with our test set cannot be compared to its effect on the synthetic data set. The proposed UICoE-Net showed promising results on the test images, which are quite close to the references, and barely introduce artificial colors.

The quality of the recovered images using the UICoD test set has been quantified. As shown in Table III, among all underwater image enhancement methods we tested, our method achieved the best performance across all metrics. Compared with others, the UDCP obtained lower scores in terms of SSIM and PSNR, which is consistent with its performance on the qualitative analysis shown above; further demonstrating that the traditional physical model-based methods generally cannot adapt to varied underwater environments. Similarly, the UWCNN, which is based on the underwater scene prior, was subject to the same concern. Since UWCNN contains multiple models matching diverse water types, the best performing model was selected to run the test. Overall, compared to the testing on UIEB, the performance of most enhancement methods on the UICoE test set was slightly decreased, which to some extent demonstrates that the UICoD is a more challenging dataset. However, the performance of our method not only did not decrease, but increased slightly, further verifying that our method achieves better enhancement effects on images captured in similar scenes.

TABLE III

FULL-REFERENCE IMAGE QUALITY ASSESSMENT IN TERMS OF MSE, PSNR, AND SSIM ON UICOD TEST SET. TRADITIONAL METHODS, DEEP-LEARNING-BASED METHODS WITH PRE-TRAINED MODELS AND DEEP-LEARNING-BASED METHODS TRAINED WITH PAIRED REFERENCE IMAGES ARE SEPARATED WITH LINES

Method	SSIM↑	PSNR↑	MSE ($\times 10^3$) ↓
CLAHE [43]	0.7515	16.7253	1.6143
Fusion [44]	0.7915	18.7363	1.3235
Retinex [45]	0.7143	15.5938	2.2678
HistogramPrior [46]	0.7231	15.6859	2.2629
UDCP [47]	0.5740	12.3389	4.4362
CBF [48]	0.7862	17.1231	1.8320
Two-Step [49]	0.7953	17.8514	1.4649
ULAP [50]	0.7873	16.5439	1.7492
UWCNN [6]	0.6674	14.2166	2.7788
MLFcGAN [11]	0.7066	16.7418	1.5977
FUnIEGAN [26]	0.7702	17.8983	1.4846
U-Net [51]	0.8379	20.6821	0.7779
WaterNet [4]	0.7742	19.3246	0.9747
UICoE-Net	**0.9124**	**23.1643**	**0.3756**

The evaluation conducted on UIEB and UICoD uses full-reference metrics, where the manually-selected enhanced images serve as the clear references. Due to the limited number of candidates for dataset establishment, references for some tough cases are not quite satisfactory. As Figure 8 presented, UICoE-Net sometimes even produces better enhancements than the given references. For the results shown on the first and second row, the removal for greenish distortion of our enhancements are more thorough. While for the image on the third row, our enhancements recover more details and keep consistency colors for the regions with same semantics, such as the undersea ground.

TABLE IV
QUANTITATIVE RESULTS FOR THE CLASSIFIED UIEB AND THE UICOD TEST SET

Metric	DataSet	UICoE-Net-w/o-LC	UICoE-Net-w/o-SC	UICoE-Net-w/o-S	UICoE-Net
SSIM	UIEB	0.8803	0.8837	0.8578	0.8944
	UICoD	0.8830	0.8877	0.8379	0.9124
PSNR	UIEB	20.3718	20.4537	20.2393	21.7456
	UICoD	21.1751	20.6574	20.6821	23.1643
MSE($\times 10^3$)	UIEB	0.6701	0.6304	0.7607	0.4711
	UICoD	0.6019	0.6062	0.7779	0.3756

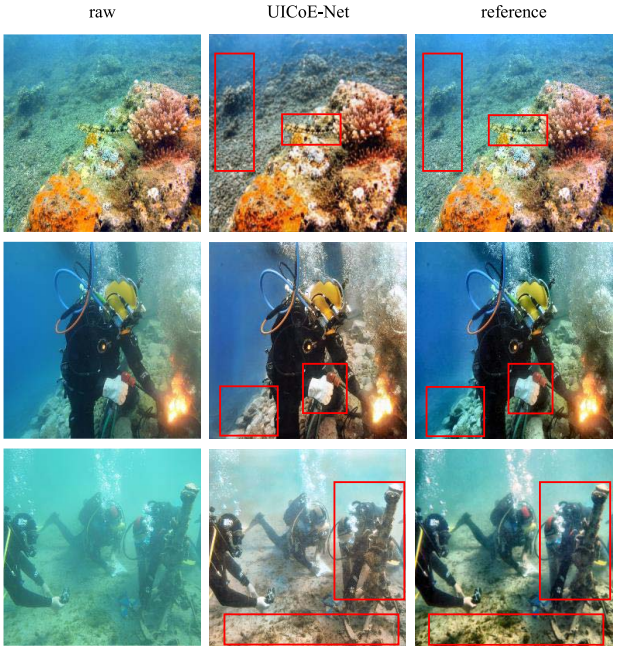


Fig. 8. Examples of the enhancements produced by UICoE-Net, which have better visual quality than the given references. The red rectangles highlight some obvious visual quality improvements.

C. Ablation Study

To demonstrate the effect of each component in our network, we carried out an ablation study involving the following experiments: (i) UICoE-Net without low-level correlation (UICoE-Net-w/o-LC), (ii) UICoE-Net without semantic correlation (UICoE-Net-w/o-SC), (iii) UICoE-Net without Siamese architecture (UICoE-Net-w/o-S). The quantitative evaluations were performed on the classified UIEB and UICoD datasets. The average scores in terms of SSIM, PSNR and MSE are reported in Table IV.

The proposed joint learning strategy for co-enhancement is built on the Siamese architecture, which provides the paths for correlation feature matching and communication. Compared with the single-branch learning strategy (i.e., UICoE-Net-w/o-S), our complete UICoE-Net with Siamese architecture improved SSIM, PSNR, and MSE scores on UIEB from 0.8578, 20.2393 and 0.7607 to 0.8944, 21.7456 and 0.4711, respectively. It also demonstrated that the co-enhancement strategy helps to more effectively mine consistent visual reconstruction information from scene-related underwater images for better enhancement.

Moreover, compared with the performance of UICoE-Net-w/o-LC, the SSIM score of UICoE-Net increased from 0.8803 to 0.8944, the PSNR score increased from 20.3718 to 21.7456, and the MSE score decreased from 670.1 to 471.1 on UIEB, respectively. The performance scores in terms of SSIM, PSNR and MSE on UICoD show the same trend. Similarly, compared with UICoE-Net-w/o-SC, UICoE-Net also showed obvious improvement on both UIEB and UICoD according to all three performance metrics. The above-mentioned quantitative comparisons demonstrate that both semantic and low-level correlation feature matching contribute to the high performance of the proposed UICoE-Net. This improvement mainly benefits from the additional enhancement constraints brought by the paired training data through the correlation feature matching modules. Specially, the low-level correlation modules help to better perceive local detailed corrections from the references of its paired images; and the semantic correlation modules provides assisted global appearance constraints and guidance for low-level feature matching.

In Figure 9, we show several examples of the enhancement from the ablation study. Compared with its ablated versions, the proposed UICoE-Net produces more visually pleasing enhancements, which are also more consistent with the given reference. In contrast, due to the lack of high-level constraint, UICoE-Net-w/o-SC usually only focuses on the local enhancement while sometimes fails to learn the global appearance consistency with the reference. Without low-level constraints, UICoE-Net-w/o-LC sometimes receives worse recovery for local details, while UICoE-Net-w/o-S usually produces under-enhancement, which shows as low global contrast and unclear local details.

To further explore the contribution of different low-level features, we further evaluated enhancement performance by communicating different combinations of low-level features. The evaluation was conducted on the UIEB dataset and the comparison results are given in Figure 10. Compared with leaving different low-level features out, using all the three levels led to the best performance, which further demonstrates that the correlation feature matching module also benefits from the multi-level feature matching strategy. It is also notable that communicating higher low-level features usually achieves better performance than with lower level features. While, by comparing the results with $l = \{3, 4\}$ and $l = \{2, 3, 4\}$, we find that lower low-level features also contribute a lot to the global performance gains of UICoE-Net.

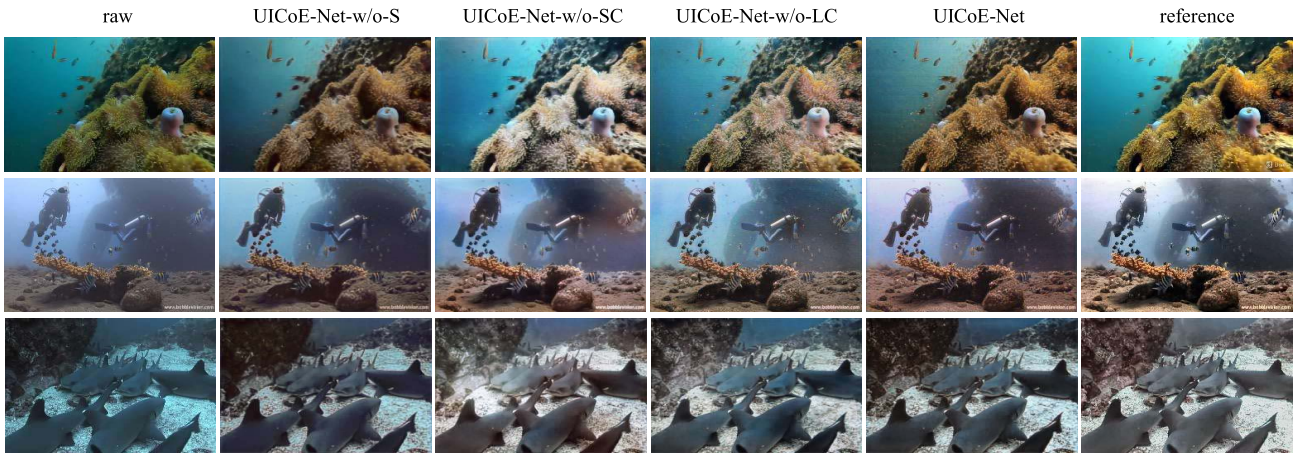


Fig. 9. Examples of the ablation study. From left to right, raw images, enhancements with UICoE-Net by removing Siamese architecture, semantic correlation and low-level correlation, enhancements with complete UICoE-Net and reference images are presented, respectively.

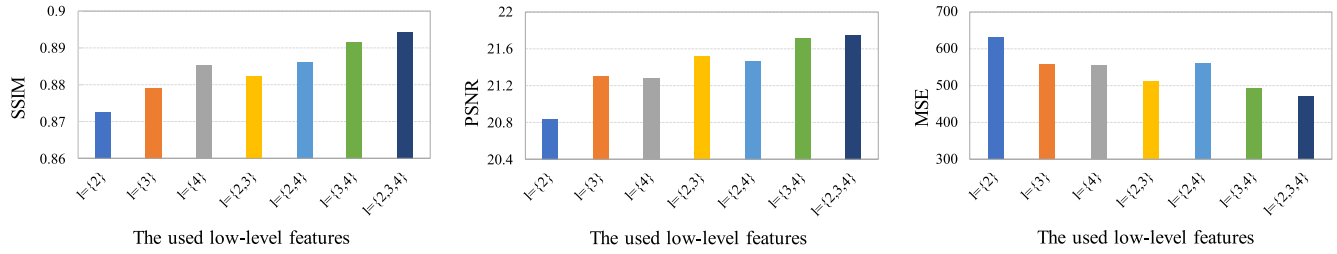


Fig. 10. Enhancement of performance by communicating different low-level features. The evaluation was conducted on the test set of the UIEB dataset. l is the level index of low-level features as mentioned in Section III-B.

D. Experiments on Underwater Image Restoration Dataset

To set a challenging and high-quality benchmark for color restoration approaches, Berman *et al.* [42] established SQUID dataset with color charts for quantitative comparison. SQUID contains 57 stereo pairs photographed in four different sites with different depth ranges. To facilitate the quantitative comparison with typical underwater image restoration approaches, we test our method on the same 8 test images whose quantitative evaluations have been provided by the authors of the SQUID benchmark. Considering that the proposed method is learning based, we fine-tune our model on UIEB dataset with 30 right camera images from SQUID dataset. The 30 images used for fine-tuning are randomly selected from the extra 49 images (except for the 8 test images) and the rest of them are served as the validation set. Since the SQUID dataset does not provide the ground truth restoration images, we use restorations of other algorithms to serve as the training references. As reported in [42] and [54] performs the best on SQUID dataset. However, since there is no publicly available implementation for it, as a compromise scheme, we generated the references for 30 training images by transferring [55] the color of restorations[2] generated by [54] to the raw images.

Besides [54], we also compared our approach with [42], [48], [53] and the baseline of our method—UICoE-Net-w/o-S. As the restorations presented in Figure 11, though UICoE-Net is fine-tuned by taking 30 recovered images of [54] as training data, its results on the test images even show more merits than the results of [54]. For example, UICoE-Net recovered finer details (see the results of RGT_3008) and removed local "overexposure" (see the results of RGT_4491). This mainly benefits from the correlation matching scheme, which helps to better recover the matched regions with the constraint of communicated features; and the U-Net-like encoder-decoder structure, which introduces global conceptions into the enhancements and helps to reduce inappropriate local corrections. Compared with the results of [42], [48], and [53] the results of UICoE-Net show better global recovery since it has learned the good restoration ability from [54].

To give quantitative comparison, we follow [42] to calculate the color correction error according to the standard color chart. As presented in Figures 12 and 13, the average reproduction angular error $\overline{\psi}$ on each image is reported, which shows the deviation between the corrected colors and the given standard colors. We can notice that UICoE-Net outperforms all the competitors on the gray-scale patches, including [54]. For the other colors, the reproduction angular errors are relatively large for all the approaches, which shows that recovering the colors for such severely degraded cases is still very challenging. Specifically, [48] performs the best of all the tested approaches, whose average $\overline{\psi}$ on all

[2]http://csms.haifa.ac.il/profiles/tTreibitz/datasets/ambient_forwardlooking/index.html

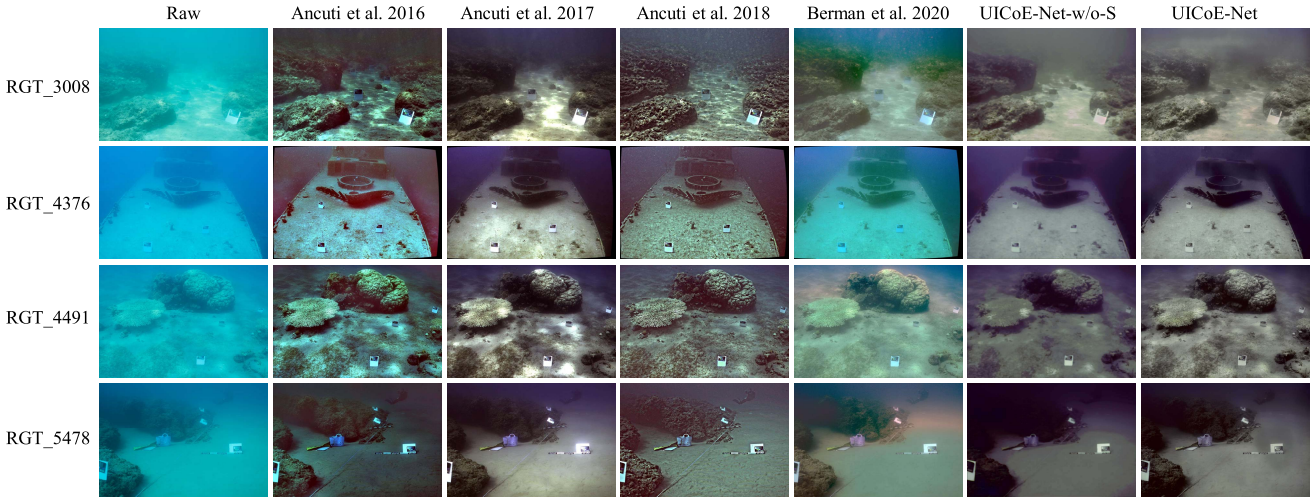


Fig. 11. Underwater image restoration examples for images from SQUID. From left to right, restoration results of Ancuti *et al.* 2016 [53], Ancuti *et al.* 2017 [54], Ancuti *et al.* 2018 [48], Berman *et al.* 2020 [42], UICoE-Net-w/o-S and the proposed UICoE-Net are presented, respectively.

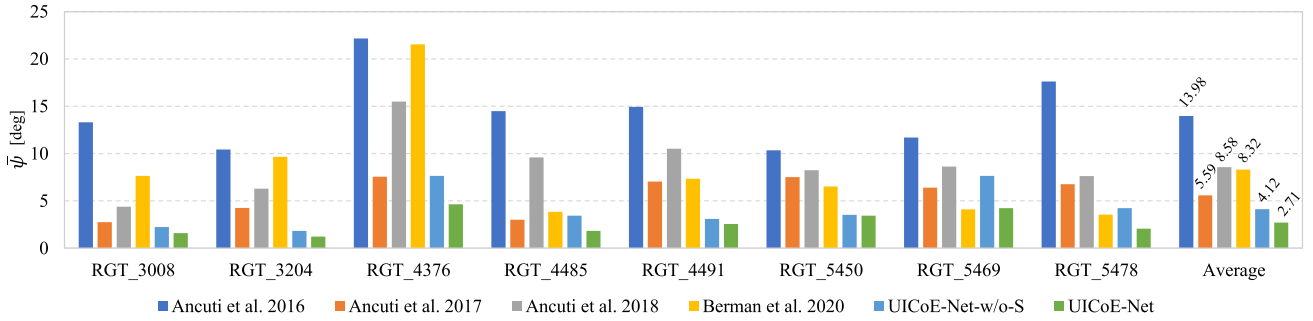


Fig. 12. The average reproduction angular error $\overline{\psi}$ between the gray-scale patches and the pure gray colors. For each test image, the sub-bargraph reports the average performance of each comparison method on all the color charts. The average performance of each approach is reported on the right.

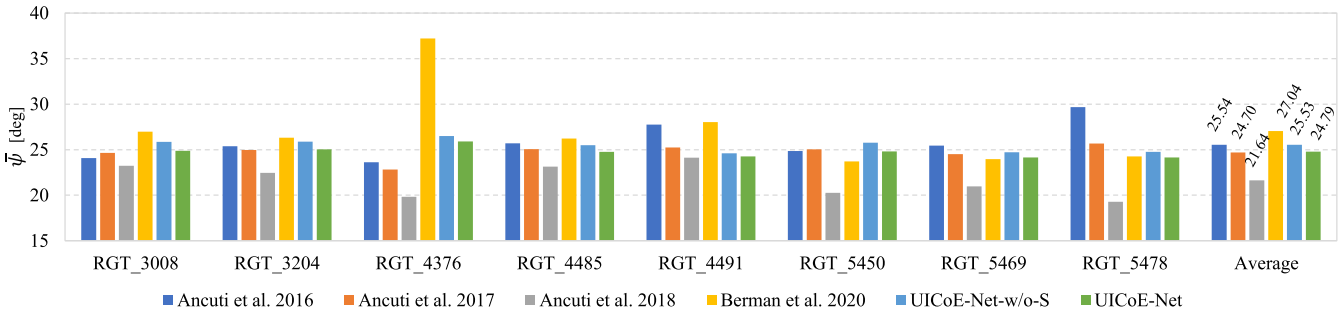


Fig. 13. The average reproduction angular error $\overline{\psi}$ between the color patches and the standard colors in color chart. For each test image, the sub-bargraph reports the average performance of each comparison method on all the color charts. The average performance of each approach is reported on the right.

the test images is 21.64. Since the 30 training images do not provide correct color corrections for learning, the performance of UICoE-Net is relatively lower than its own performance on the gray patches and close to that of [54]. The above experimental results indicate once again that our enhancement network with Siamese structures for correlation matching can better learn the correction relations implied in the given training data, whose quality has become the vital factor for the performance of learning-based underwater image restoration.

E. Discussion

1) Running Time Analysis: In this paper, we introduce a cooperative enhancement scheme which has been demonstrated to effectively improve the visual quality of group-wise underwater images captured in related scenes. Compared with the basic encoder-decoder network, correlation feature matching units are the main calculation modules in the proposed UICoE-Net, which are used to communicate mutual correlation of the two branches. To give a comprehensive running time analysis, we decomposed the time cost of the

TABLE V
TIME COST PER IMAGE IN TESTING STAGE. UICOE-NET AND ITS SIMPLIFIED VERSIONS ARE TESTED ON IMAGES WITH TWO DIFFERENT SIZES (128 × 128 AND 256 × 256)

Seconds per image	Encoding	Correlation feature matching		Decoding	Average
		Semantic	Low-level		
UICoE-Net-w/o-LC (128×128)	0.002	0.013	—	0.003	0.018
UICoE-Net-w/o-LC (256×256)	0.003	0.053	—	0.004	0.060
UICoE-Net-w/o-SC (128×128)	0.002	—	0.668	0.028	0.698
UICoE-Net-w/o-SC (256×256)	0.003	—	2.788	0.131	2.922
UICoE-Net-w/o-S (128×128)	0.001	—	—	0.001	0.002
UICoE-Net-w/o-S (256×256)	0.002	—	—	0.002	0.004
UICoE-Net (128×128)	0.002	0.010	0.334	0.021	0.367
UICoE-Net (256×256)	0.003	0.038	1.289	0.070	1.400

model-testing stage according to different calculation modules, summarized in Table V. We tested UICoE-Net and its simplified versions on images with two different sizes, i.e., 128×128 and 256×256.

The time cost of the testing stage shown in Table V shows that our UICoE-Net can process an image with a size of 128×128 within 0.4s. For 256×256 images, enhancing each image costs 1.4s on average, with the low-level correlation feature matching accounting for more than 90% of total time used. In contrast, the testing speed of UICoE-Net-w/o-LC, which only keeps the semantic correlation feature matching in the proposed joint learning framework, can reach about 55 FPS (16FPS) on 128×128 (256×256) images. From the ablation study, it is apparent that UICoE-Net-w/o-LC achieves a quite obvious performance improvement over the traditional non-cooperative enhancement strategy. Therefore, when applied to high-efficiency-required applications, UICoE-Net-w/o-LC would be an ideal alternative. For high-quality-required enhancement tasks for which efficiency is less important, the complete UICoE-Net is a much better choice.

2) Future Works: As mentioned in Section IV-B, the references of UIEB and UICoD used for learning are not strictly veritable ground-truth, but manually selected by volunteers from candidate enhancement pool, a small number of reference images are not visually good enough when the candidate pool itself does not include satisfactory enhancements. Those poor references for training will inevitably introduce improper learning constraints and lead to unsatisfactory enhancements in the test stage. Thanks to the proposed cooperative learning framework, which helps to better learn the mapping relation implied in the raw and reference images, UICoE-Net is able to produce better enhancement by introducing more high-quality references for training. In future, we plan to establish bigger benchmark and introduce extra correction strategy in dataset reference selection, so as to further improve the visual performance of data-driven approaches and facilitate the research of UIE community.

Besides, to break through the efficiency bottleneck, we intend to explore more efficient correlation communication strategies; for example, by using the attention models to highlight correlation regions, like the strategies introduced in [30] and [56]. We also would like to extend the cooperative strategy to underwater video enhancement task, which is an under-explored area and also of high research value, since frames in the same video share a large amount of consistency information. As discussed in [3], inter-frame consistency in underwater video enhancement tasks needs to be addressed. Introducing cooperative constraints into the enhancement model for adjacent or content-related frames could be a promising solution.

V. CONCLUSION

In this paper, we proposed the first underwater image co-enhancement network, called UICoE-Net, to learn inter-image cooperative information for better enhancement performance. The proposed correlation feature matching units are incorporated into multiple layers of the Siamese encoder-decoder structures to communicate the mutual correlation of two branches. We conducted experiments on two underwater image enhancement datasets and a challenging underwater image restoration benchmark to verify the effectiveness and generalization of our network. Ablation studies further demonstrated the effectiveness of the proposed low-level correlation modules, semantic correlation modules, and Siamese architecture used in our network. In future, we intend to explore more efficient correlation-calculating strategies and introduce the cooperative strategy into underwater video enhancement tasks.

REFERENCES

[1] R. Gibson, R. Atkinson, and J. Gordon, "A review of underwater stereo-image measurement for marine biology and ecology applications," *Oceanogr. Mar. Biol., Annu. Rev.*, vol. 47, pp. 257–292, Apr. 2016.
[2] T. Łuczyński, P. Łuczyński, L. Pehle, M. Wirsum, and A. Birk, "Model based design of a stereo vision system for intelligent deep-sea operations," *Measurement*, vol. 144, pp. 298–310, Oct. 2019.
[3] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du, and K. Hu, "An in-depth survey of underwater image enhancement and restoration," *IEEE Access*, vol. 7, pp. 123638–123657, 2019.
[4] C. Li *et al.*, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2020.
[5] S. Anwar and C. Li, "Diving deeper into underwater image enhancement: A survey," 2019, *arXiv:1907.07863*. [Online]. Available: http://arxiv.org/abs/1907.07863
[6] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107038.
[7] M. Hou, R. Liu, X. Fan, and Z. Luo, "Joint residual learning for underwater image enhancement," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4043–4047.

[8] Y. Wang, J. Zhang, Y. Cao, and Z. Wang, "A deep CNN method for underwater image enhancement," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1382–1386.

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

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

[11] X. Liu, Z. Gao, and B. M. Chen, "MLFcGAN: Multilevel feature fusion-based conditional GAN for underwater image color correction," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1488–1492, Sep. 2020.

[12] K. Li, J. Zhang, and W. Tao, "Unsupervised co-segmentation for indefinite number of common foreground objects," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1898–1909, Apr. 2016.

[13] W. Li, O. H. Jafari, and C. Rother, "Deep object co-segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 638–653.

[14] L. Wang, G. Hua, R. Sukthankar, J. Xue, Z. Niu, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2074–2088, Oct. 2017.

[15] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCO3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8838–8847.

[16] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognit.*, vol. 88, pp. 113–126, Apr. 2019.

[17] G. Hou, X. Zhao, Z. Pan, H. Yang, L. Tan, and J. Li, "Benchmarking underwater image enhancement and restoration, and beyond," *IEEE Access*, vol. 8, pp. 122078–122091, 2020.

[18] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4861–4875, Dec. 2020.

[19] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[20] K. Cao, Y.-T. Peng, and P. C. Cosman, "Underwater image restoration using deep networks to estimate background light and scene depth," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, Apr. 2018, pp. 1–4.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

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

[25] J. Lu, N. Li, S. Zhang, Z. Yu, H. Zheng, and B. Zheng, "Multi-scale adversarial network for underwater image restoration," *Opt. Laser Technol.*, vol. 110, pp. 105–113, Feb. 2019.

[26] 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.

[27] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching–incorporating a global constraint into MRFs," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 993–1000.

[28] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 12–27, Jan. 2016.

[29] K. Li, S. Qi, H. Yang, L. Zhou, and D. Song, "Extensible image object co-segmentation with sparse cooperative relations," *Inf. Sci.*, vol. 521, pp. 422–434, Jun. 2020.

[30] H. Chen, Y. Huang, and H. Nakayama, "Semantic aware attention based deep object co-segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 435–450.

[31] B. Li, Z. Sun, Q. Li, Y. Wu, and H. Anqi, "Group-wise deep object co-segmentation with co-attention recurrent neural network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8518–8527.

[32] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.

[33] J. Zhang, K. Li, and W. Tao, "Multivideo object cosegmentation for irrelevant frames involved videos," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 785–789, Jun. 2016.

[34] K. Li, D. Shi, Y. Zhang, Q. J. Wu, X. Luan, and D. Song, "CascNet: No-reference saliency quality assessment with cascaded applicability sorting and comparing network," *Neurocomputing*, vol. 425, pp. 231–242, 2021.

[35] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[36] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[37] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[38] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1464–1471.

[39] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-Wolfe algorithm," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 253–268.

[40] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 19–34.

[41] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[42] 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.*, early access, Mar. 2, 2020, doi: 10.1109/TPAMI.2020.2977624.

[43] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Morgan Kaufmann, 1994, pp. 474–485.

[44] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 81–88.

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

[46] 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.

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

[48] 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.

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

[50] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Advances in Multimedia Information Processing—PCM*. Cham, Switzerland: Springer, 2018, pp. 678–688.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

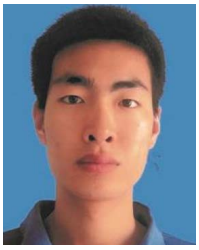
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[53] C. Ancuti, C. O. Ancuti, C. De Vleeschouwer, R. Garcia, and A. C. Bovik, "Multi-scale underwater descattering," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 4202–4207.

[54] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, L. Neumann, and R. Garcia, "Color transfer for underwater dehazing and depth estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 695–699.

[55] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 4, pp. 34–41, Jul. 2001.
[56] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

Qi Qi (Student Member, IEEE) received the B.S. and M.S. degrees from the University of Jinan, Jinan, China, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Ocean University of China, Qingdao, China. His research interests include image and video enhancement, underwater computer vision, image processing, and visual recognition.



Yongchang Zhang received the B.S. degree from Qingdao University, Qingdao, China, in 2018. He is currently pursuing the master's degree with the College of Information Science and Engineering, Ocean University of China, Qingdao. His research interests include underwater computer vision, underwater image/video enhancement, and deep learning.

Fei Tian received the B.S. degree from Shanghai Normal University, Shanghai, China, in 2018. She is currently pursuing the master's degree with the College of Information Science and Engineering, Ocean University of China, Qingdao, China. Her research interests include underwater image processing, visual recognition, and machine learning.

Q. M. Jonathan Wu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. From 1995 to 2005, he was affiliated with the National Research Council of Canada, Ottawa, ON, Canada, where he became a Senior Research Officer and a Group Leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published over 300 peer-reviewed articles in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks. He is a fellow of the Canadian Academy of Engineering. He holds the Tier 1 Canada Research Chair of Automotive Sensors and Information Systems. He has served on technical program committees and international advisory committees for many prestigious conferences. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON CYBERNETICS, *Cognitive Computation, and the Neurocomputing.*

Kunqian Li (Member, IEEE) received the B.S. degree from the China University of Petroleum (UPC), Qingdao, China, in 2012, and the Ph.D. degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2018. He is currently a Lecturer with the College of Engineering, Ocean University of China, Qingdao. He has published over 20 peer-reviewed articles in computer vision and image processing. His research interests include image processing, visual recognition, and underwater computer vision. He serves as a Reviewer for many journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON MULTIMEDIA.

Xin Luan received the B.S. and M.S. degrees from the School of Computer Science and Technology, Harbin Engineering University. She has been a Lecturer, an Associate Professor, a Professor, and a Doctoral Tutor with the College of Information Science and Engineering, Ocean University of China, Qingdao, China. She is currently an Extramural Doctoral Tutor with the Ocean University of China. She is mainly engaged in research on ocean observation technology and artificial intelligence.

Dalei Song received the Ph.D. degree from Harbin Industrial University, Harbin, China, in 1999. From 1999 to 2001, he was a Senior Engineer with Lucent Technologies. He is currently a full Professor with the Department of Automation and Measurement, College of Engineering, Ocean University of China, Qingdao, China. His research interests include machine intelligent perception, ocean observation technology, robot control technology, and computer vision.